Soft Biometric Analysis: Multi-Person and Real-Time Pedestrian Attribute Recognition in Crowded Urban Environments

Ehsan Yaghoubi

Thesis for obtaining the doctorate degree in
Computer Engineering
(3º cycle of study)

Supervisor: Prof. Dr. Hugo Pedro Proença

November 2021

Jury of the Doctoral Exam

The doctoral exam took place on November 11, 2021, at 14:00, with a jury consisting of the following members:

Doctor Joaquim Mateus Paulo Serra, Vice-Rector of the University of Beira Interior as the president of the jury;

Doctor Ruben Vera Rodriguez, associate professor at the Autonomous University of Madrid, Spain.

Doctor Joao Manuel Ribeiro da Silva Tavares, full professor at the Faculty of Engineering at the University of Porto;

Doctor Ana Luisa Nobre Fred, associate professor at the Higher Technical Institute of the University of Lisbon;

Doctor Luis Filipe Barbosa de Almeida Alexandre, full professor at the University of Beira Interior;

Doctor Joao Carlos Raposo Neves, assistant professor at the University of Beira Interior.

Doctor Hugo Pedro Martins Carrico Proenca, full professor at the University of Beira Interior;

This thesis was prepared at the University of Beria Interior, IT - Instituto de Telecomunicações, Soft Computing and Image Analysis Laboratory (SOCIA Lab), Covilhã Delegation, and was submitted to the University of Beira Interior for defense in a public examination session.

This thesis was supported in part by the FCT/MEC through National Funds and Co-Funded by the FEDER-PT2020 Partnership Agreement under Project UIDB/50008/2019, Project UIDB/50008/2020, Project POCI-01-0247-FEDER-033395 and in part by operation Centro-01-0145-FEDER-000019 - C4 - Centro de Competências em Cloud Computing, co-funded by the European Regional Development Fund (ERDF) through the Programa Operacional Regional do Centro (Centro 2020), in the scope of the Sistema de Apoio à Investigação Científica e Tecnológica - Programas Integrados de IC&DT. This work was also supported by "IT: Instituto de Telecomunicações" and "TOMI: City's Best Friend" under Project UID/EEA/50008/2019.

List of Publications

Publications: Articles included in the main body of the thesis resulting from this doctoral research program

1. **Yaghoubi, E.**, Khezeli, F., Borza, D., Kumar, S.V., Neves, J. and Proença, H., 2020. Human Attribute Recognition—A Comprehensive Survey. Applied Sciences, 10(16), p.5608.
2. **Yaghoubi, E.**, Kumar, A. and Proença, H., 2021. SSS-PR: A short survey of surveys in person re-identification. Pattern Recognition Letters, 143, pp.50-57.
3. **Yaghoubi, E.**, Alirezazadeh, P., Assunção, E., Neves, J.C. and Proença, H., 2019, September. Region-Based CNNs for Pedestrian Gender Recognition in Visual Surveillance Environments. In 2019 International Conference of the Biometrics Special Interest Group (BIOSIG) (pp. 1-5). IEEE.
4. **Yaghoubi, E.**, Borza, D., Neves, J., Kumar, A. and Proença, H., 2020. An attention-based deep learning model for multiple pedestrian attributes recognition. Image and Vision Computing, 102, p.103981.
5. **Yaghoubi, E.**, Borza, D., Kumar, S.A. and Proença, H., 2021. Person re-identification: Implicitly defining the receptive fields of deep learning classification frameworks. Pattern Recognition Letters, 145, pp.23-29.
6. **Yaghoubi, E.**, Borza, D., Degardin, B. and Proença, H., 2021. You look so different! Haven't I seen you a long time ago?. Image and Vision Computing, 115, p.104288.

Collaborative Publications: Other publications resulting from this doctoral research program not included in the body of the thesis

1. Alirezazadeh, P., **Yaghoubi, E.**, Assunção, E., Neves, J.C. and Proença, H., 2019, September. Pose Switch-based Convolutional Neural Network for Clothing Analysis in Visual Surveillance Environment. In 2019 International Conference of the Biometrics Special Interest Group (BIOSIG) (pp. 1-5). IEEE.
2. Proença, H., **Yaghoubi, E.** and Alirezazadeh, P., 2020. A Quadruplet Loss for Enforcing Semantically Coherent Embeddings in Multi-Output Classification Problems. IEEE Transactions on Information Forensics and Security, 16, pp.800-811.
3. Borza, D., **Yaghoubi, E.**, Neves, J. and Proença, H., All-in-one "HairNet": A Deep Neural Model for Joint Hair Segmentation and Characterization. In 2020 IEEE International Joint Conference on Biometrics (IJCB) (pp. 1-10). IEEE.
4. Kumar, S.A., **Yaghoubi, E.**, Das, A., Harish, B.S. and Proença, H., 2020. The P-DESTRE: A Fully Annotated Dataset for Pedestrian Detection, Tracking, and Short/Long-Term Re-Identification From Aerial Devices. IEEE Transactions on Information Forensics and Security, 16, pp.1696-1708.

Acknowledgments

First of all, I would like to express my sincere gratitude to my supervisor, Prof. Hugo Proença, for his consistent support and encouragement throughout my PhD, without which it would be very difficult for me to conclude my PhD course. Also, I would like to thank Prof. Rúben Vera-Rodriguez, who gave me the opportunity to work with him as my internship supervisor.

It wasn't easy to go through all challenges of getting a PhD in abroad without the support of my wonderful wife, Zeinab. I would like to thank her not only for the unconditional love she gives me but also for her patience in many short and long periods we had to live far from each other. Life is short but beautiful and valuable. Throughout these three years, we couldn't visit our families, and I would like to thank them, especially our mothers, for enduring the great pain and suffering caused by our far distance.

Last but not least, I would like to thank Dr. Diana Borza, Dr. Aruna Kumar, Dr. João Neves, and Eng. Farhad Khezeli, who collaborated in my researches and gave me great comments. I would also like to thank my helpful friends Vasco Lopes, Miguel Fernandes, Nuno Pereira, and Bruno Carneiro da Silva who helped me during the first months of my PhD to come up with many difficulties. Also, I would like to thank my great friends Dr. Hamzeh Mohammadi, Mostafa Razavi, Bruno Degardin, António Gaspar, Leonice Souza Pereira, Eduardo Assunção, João Brito, and Tiago Roxo with whom I shared precious memories and great moments.

Abstract

Traditionally, recognition systems were only based on human hard biometrics. However, the ubiquitous CCTV cameras have raised the desire to analyze human biometrics from far distances, without people attendance in the acquisition process. High-resolution face close-shots are rarely available at far distances such that face-based systems cannot provide reliable results in surveillance applications. Human soft biometrics such as body and clothing attributes are believed to be more effective in analyzing human data collected by security cameras.

This thesis contributes to the human soft biometric analysis in uncontrolled environments and mainly focuses on two tasks: Pedestrian Attribute Recognition (PAR) and person re-identification (re-id). We first review the literature of both tasks and highlight the history of advancements, recent developments, and the existing benchmarks. PAR and person re-id difficulties are due to significant distances between intra-class samples, which originate from variations in several factors such as body pose, illumination, background, occlusion, and data resolution. Recent state-of-the-art approaches present end-to-end models that can extract discriminative and comprehensive feature representations from people. The correlation between different regions of the body and dealing with limited learning data is also the objective of many recent works. Moreover, class imbalance and correlation between human attributes are specific challenges associated with the PAR problem.

We collect a large surveillance dataset to train a novel gender recognition model suitable for uncontrolled environments. We propose a deep residual network that extracts several pose-wise patches from samples and obtains a comprehensive feature representation. In the next step, we develop a model for multiple attribute recognition at once. Considering the correlation between human semantic attributes and class imbalance, we respectively use a multi-task model and a weighted loss function. We also propose a multiplication layer on top of the backbone features extraction layers to exclude the background features from the final representation of samples and draw the attention of the model to the foreground area.

We address the problem of person re-id by implicitly defining the receptive fields of deep learning classification frameworks. The receptive fields of deep learning models determine the most significant regions of the input data for providing correct decisions. Therefore, we synthesize a set of learning data in which the destructive regions (e.g., background) in each pair of instances are interchanged. A segmentation module determines destructive and useful regions in each sample, and the label of synthesized instances are inherited from the sample that shared the useful regions in the synthesized image. The synthesized learning data are then used in the learning phase and help the model rapidly learn that the identity and background regions are not correlated. Meanwhile, the proposed solution could be seen as a data augmentation approach that fully preserves the label information and is compatible with other data augmentation techniques.

When re-id methods are learned in scenarios where the target person appears with

identical garments in the gallery, the visual appearance of clothes is given the most importance in the final feature representation. Cloth-based representations are not reliable in the long-term re-id settings as people may change their clothes. Therefore, developing solutions that ignore clothing cues and focus on identity-relevant features are in demand. We transform the original data such that the identity-relevant information of people (e.g., face and body shape) are removed, while the identity-unrelated cues (i.e., color and texture of clothes) remain unchanged. A learned model on the synthesized dataset predicts the identity-unrelated cues (short-term features). Therefore, we train a second model coupled with the first model and learns the embeddings of the original data such that the similarity between the embeddings of the original and synthesized data is minimized. This way, the second model predicts based on the identity-related (long-term) representation of people.

To evaluate the performance of the proposed models, we use PAR and person re-id datasets, namely BIODI, PETA, RAP, Market-1501, MSMT-V2, PRCC, LTCC, and MIT and compared our experimental results with state-of-the-art methods in the field.

In conclusion, the data collected from surveillance cameras have low resolution, such that the extraction of hard biometric features is not possible, and face-based approaches produce poor results. In contrast, soft biometrics are robust to variations in data quality. So, we propose approaches both for PAR and person re-id to learn discriminative features from each instance and evaluate our proposed solutions on several publicly available benchmarks.

Keywords

Pedestrian Attribute Recognition, Person Re-Identification, Multi-task learning, Human Soft-Biometric Analysis, Attention Mechanism, Multi-Person Soft Biometric Estimation, Face and Body Attribute Recognition, Clothing Attribute Recognition, Visual Surveillance Data Analysis, Cloth-Changing Person Re-Identification.

Contents

List of Figures

List of Tables

Acronyms

APiS Attributed Pedestrians in Surveillance

ACN Attributes Convolutional Net

BBs Bounding Boxes

BCE Binary Cross-Entropy

BIODI Biometria e Deteção de Incidentes

BN Batch Normalization

CAA Clothing Attribute Analysis

CAD Clothing Attributes Dataset

CAMs Class Activation Maps

CBCL Center for Biological and Computational Learning

CCTV Closed-Circuit TeleVision

CNN Convolutional Neural Network

CRF Conditional Random Field

CRP Caltech Roadside Pedestrians

CVPR Computer Vision and Pattern Recognition

CSD Color Structure Descriptor

CTD Clothing Tightness Dataset

DNN Decompositional Neural Network

DukeMTMC Duke Multi-Target, Multi-Camera

DPM Deformable Part Model

FAA Facial Attribute Analysis

FAR Full-body Attribute Recognition

FCN Fully Connected Network

FCL Fully Connected Layer

GAN Generative Adversarial Network

GCN Graph Convolutional Network

GRID underGround Re-IDentification

HAR Human Attribute Recognition

HAT Human ATtributes

He-Reid Heterogeneous re-id

HD High Definition

Ho-Reid Homogeneous re-id

HOG Histogram of Oriented Gradients

ICCV International Conference on Computer Vision

KITTI Karlsruhe Institute of Technology and Toyota Technological Institute

re-id re-identification

LSTM Long Short Term Memory

MAP Maximum A Posterioris

MCSH Major Colour Spectrum Histogram

mAP mean Average Precision

MLCNN Multi-Label Convolutional Neural Network

MSCR Maximally Stable Colour Regions

OPC Office of the Privacy Commissioner of Canada

P-DESTRE Pedestrian Detection, Tracking, Re-Identification and Search

PARSe27k Pedestrian Attribute Recognition in Sequences

PETA PEdesTrian Attribute

PET Privacy-Enhancing Technologies

PAR Pedestrian Attribute Recognition

PASCAL-VOC PASCAL Visual Object Classes

PGRN Pedestrian Gender Recognition Network

PSN Pose-Sensitive Network

RAP Richly Annotated Pedestrian

RCB Residual Convolutional Block

ResNet Residual Networks

RHSP Recurrent Highly-Structured Patches

RNN Recurrent Neural Networks

RoI Regions of Interest

RPN Region Proposal Network

SE-Net Squeeze-and-Excitation Networks

SGD Stochastic Gradient Descent

SIFT Scale-Invariant Feature Transform

SoBiR Soft Biometric Retrieval

SPR Spatial Pyramid Representation

SPPE Single Person Pose Estimator

SSD Single Shot Detector

STN Spatial Transformer Network

SVM Support Vector Machine

UAV Unmanned Aerial Vehicle

UDA Unsupervised Domain Adaptation

VAE Variational Auto-Encoders

VGG Visual Geometry Group

YOLO You Only Look Once

Chapter 1

Introduction

In recent decades, the growing demand for video surveillance systems in public places such as metro stations, malls, and streets has been emerging new research tracks for monitoring people and the environments themselves [1].
In general, the automatic analysis of video surveillance is a practical approach that enhances the quality of public services. For example, suppose that parents lost their child in an amusement park and they only provide the security officers with some photos and some traits related to their child to find her/him. In such situations, even if some CCTVs have recorded the children's activity, the manual inspection of the recorded content may take too much time; whereas a drone equipped with a camera and a re-identification (re-id) framework may fly over the most probable areas and automatically find the locations of the similar children to the query child [2]. Another important application of video surveillance analysis lies in the domain of security and forensic measurements, such that upon an accident, the authorities inspect the available recorded data to investigate the situation [3]. Traditionally, huge amounts of collected data were reviewed by human operators that were time-consuming and accompanied by human errors caused by tiredness, hurry, and biased opinion. Recently, computer-based analysis of visual surveillance data has significantly helped human operators expedite the inspection process –e.g., by highlighting the suspicious parts of the recorded data [4].
Analyzing video surveillance data has many different perspectives and components such as scene understanding [5], human interactions [6] and behavior understanding [7], action and activity recognition [8] and prediction [9], human emotions detection [10], person re-id [11], Human Attribute Recognition (HAR) [12], and privacy concerns [13]. Among these fields of study, we focus on soft biometric analysis in the wild, narrowed explicitly to the problems of person re-id and Pedestrian Attribute Recognition (PAR) from data collected at for distances. Although the other fields are related to video surveillance analysis and are active research areas, scholars consider them different tasks that demand different benchmarks and approaches. For instance, human behavior understanding techniques usually deal with body skeleton data over several consecutive frames and could be implemented successfully without accessing any RGB videos, but only skeleton information.

1.1 Challenges and Motivations

As mentioned previously, the field of soft-biometric analyses includes a wide range of problems, and in this thesis, our focus is on the problems of person re-id and PAR. Person re-id is the task of recognizing the visual data of a query identity and retrieving most

similar identities that have been captured in different situations, e.g., various physical places or different occasions; and, Human Attribute Recognition (HAR) (also known as PAR) aims to estimate the soft biometric attributes associated to people.

Fig. 1.1 shows some general challenges in the visual analysis of CCTV data: the presence of more than one person in one shot, high range variation in illumination, significant misalignment in shots, low-resolution data, the existence of wide background area in one shot. When the intensity of these challenges goes beyond some extent, even humans cannot provide reliable responses. In general, the face area is the most informative region that could revile the persons' identity. However, usually, the satisfactory data are not available either because the camera captures the back of the person or there are large camera distances from the subject, which causes blurred face shots such that the state-of-the-art face-based re-id systems cannot provide reliable results. Furthermore, the illumination of the captured data from a subject in a shadowed area has high variations with images captured from the same person under the sunshine. The variations in brightness change the observations in clothing and skin colors and consequently introduce some challenges to each stage of the system, from annotation and learning to estimation processes [14].

Usually, the input data of the image-based HAR and person re-id systems are one full-body close shot (bounding box) of the person such that the shots include as little as possible background region. The bounding box shots are extracted from a full frame containing a wide area, including several persons. Person detectors [15] are the primary tools used for extracting full-body close shots. However, the performance of the person detectors is not perfect [16]; therefore, we should expect a percentage of error (misalignment) in bounding box extraction, which causes misaligned shots, in which either some body parts of the person are missed, or extra regions of the background area are included in the extracted bounding box. The challenge of misalignment impacts the person re-id systems more than HAR systems since we usually perform a cross-matching between the image of the query person and the gallery images to re-identify people. Hence, missing body parts or extra areas of background reduces the matching confidence. Whereas when we want to perform attribute recognition, misalignment may degrade the quality of the final representation only because of the presence(/absence) of the destructive(/useful) features. Thus, the HAR systems are intrinsically more robust to misalignment. The presence of more than one person in each shot is another challenge since usually image-based person re-id and HAR models provide one feature representation for each available shot. Therefore, when one shot contains more than one person, the features extracted from other persons are entangled with the features of the target person, and consequently, the quality of the final representation of the target person is degraded and affects the model performance [17].

The number of captured images from each subject is limited such that this number may be less than few images in some existing person re-id and HAR datasets. Therefore, in some fractions of the dataset, each subject may appear in an environment with a unique background. This situation causes some difficulties since the background features will be entangled with person features in the final representation, mainly because the model

Figure 1.1: General challenges in person re-id and HAR frameworks. From left to right, each image shows a challenge: presence of more than one person in one shot, illumination variations, missing body parts, low resolution data, wide background area in one shot. Samples are from the RAP, PETA, and Market1501 datasets.

cannot automatically distinguish between the body-associated features and background features.

One possibility to perform an accurate person re-id is to use people's hard biometrics features. Hard biometrics, also known as biometrics, are some features that are uniquely associated with only one person, e.g., iris. However, acquiring biometric information demands an attentive collaboration of people because it cannot be performed from far distances. Also, variations in noise (e.g., illumination) and data resolution highly degrade the performance of biometric systems. Therefore, person re-id cannot be successful using hard biometrics information mainly because the data collected by CCTVs have poor resolution such that the required information (e.g., iris) are not available.

Unlike hard biometrics, soft biometrics are human understandable features that help to distinguish one person from another. Traits such as hairstyle, gender, body figure, height, hair color, and clothing style are examples of human characteristics that people use to distinguish one person from another. Soft biometric attributes are prone to be altered and counterfeit easily and are not appropriate to be used solely in secured verification systems, e.g., accessing bank account; however, they are robust to some extent of noise caused by low-resolution data and illumination variation. More importantly, soft biometrics could be captured without people's collaboration and could speed up the search process for the query person in the verification systems. For example, if the query person is confirmed to be a male, the search process for this person in the galley data could be done only among males, improving the system overall performance when identifying/verifying the user [18].

Soft biometric characteristics are also known as semantic features, improving the quality of the final feature representation of the subjects. Convolutional Neural Networks are believed to be successful in obtaining representative feature maps from data; however, person re-id is a challenging task such that the final representations obtained by holistic CNNs are insufficient for an accurate re-id task. In general, human operators decision is based on matching characteristics originated from soft biometric attributes, whereas, computer-based person re-id systems exploit low-level and mid-level features such as

textures, colors, and spatial structures. Therefore, successful estimation of people's soft biometrics can mimic human ability and provide a different and valuable source of information. Further, this information can be fused with CNN-based features and represent richer final representations of input data [19].

1.2 Objectives

Generally, the objective of this research is to study HAR and person re-id problems based on image data collected by video surveillance cameras in uncontrolled environments. Our specific objectives follow the goals of two practical projects that support this thesis: *BIODI: Biometria e Deteção de Incidentes* [1] and *CLOUD-S-POLIS: Cloudification of Autonomous Security Agents for Urban Environments* [2].
In the scope of the BIODI project, the industrial partner, TOMI WORLD company [3], aimed to set up some urban information panels all over Portugal, in which the soft biometrics of people were required. It is believed that the quantity and quality of learning data can directly affect the performance of the PAR models. However, the data in the existing PAR datasets have low variability and, more importantly, do not match the data that our proposed PAR model needs to work with later (the inference phase). Therefore, the existing domain gap between our data and the existing datasets lead us to collect a new dataset for our industrial needs. Our first objective was to collect and annotate a massive dataset from Portugal and Brazil in different parts of the day, weather, illumination, and environments. The annotation was performed for several full-body soft biometrics such as gender, height, weight, ethnicity, hair color, hairstyle, upper body clothes, lower body clothes, carrying objects, action, wearing glasses, and hats. The next objective was to develop a PAR model to be learned on the collected data and compared its performance with the existing cutting-edged PAR frameworks.
In the scope of the CLOUD-S-POLIS project, we aimed to study the existing state-of-the-art person re-id techniques and design a deep learning framework for person re-id based on surveillance data. The proposed solutions are then evaluated and compared with the state-of-the-art techniques.

1.3 Contributions

The main contributions of this thesis are as follows.

- We provide a comprehensive review of the HAR datasets and methods. We categorize the HAR benchmarks into four groups: full-body, face, fashion, and synthetic datasets, and discuss the critical points to provide an insight regarding the future data collection and annotation tasks. We also propose a challenge-based

[1] https://www.it.pt/Projects/Index/4558
[2] http://wordpress.ubi.pt/c4/cloud-applications/
[3] https://tomiworld.com/pt/meet-tomi/

taxonomy for PAR approaches and categorize the existing methods in five clusters: localization, limited data, attribute relation, occlusion, and class imbalance.

- We perform a short survey of surveys to and propose a multi-dimensional taxonomy to categorize various person re-id studies: deep based versus hand-craft based approaches, types of learning based on the amount of supervision, close and open-world identification settings, strategies of learning, data modality, the data type of queries, and contextual versus non-contextual approaches. We also discuss privacy and security concerns caused by processing people's visual data collected by CCTVs.

- We present a pose-sensitive region-based framework for pedestrian gender recognition from full-body images of people collected from surveillance cameras from far distances in the wild. The proposed framework takes advantage of human detection and tracking algorithms to capture the bounding boxes of persons. Then, we use an off-the-shelf body skeleton detector to infer the rough body pose (front, back, side) of the person and extract several regions of interest (raw, head, convhull of body). Finally, considering each pair of the considered body pose and the extracted regions of interest, we feed the data to nine specialized CNNs and consider the output of the most confident CNN as the final output, which means that the model decides based on an optimum perspective.

- We propose an attention-based multi-task PAR model to predict multiple attributes of pedestrians at once. To provide the attention of the body region and filter the destructive background feasters, we present a multiplication layer situated on top of the convolutional layers and multiplies a binary mask with the feature maps. In addition, to implicitly consider the correlation between persons' attributes, we integrate a multi-branch classifier into the model. This helps to relativize the importance of each group of attributes using a weighted loss function.

- We address the short-term person re-id task. We present an image-processing technique integrated into the learning process of deep learning architectures as a data augmentation process. The proposed technique implicitly defines the receptive fields of CNNs by providing a set of synthesized data for the training phase. In practice, we use a segmentation algorithm to obtain the background region and the body area of subjects and then interchange these segments with other samples in the learning set. As a result, the model learns from the synthesized data that the background region is changeable and identity labels are only correlated to the body area. The proposed solution has some benefits: it is compatible and integrable with the existing data augmentation techniques, it fully preserves the label information of the original data, it is a parameter-learning-free technique.

- We study the problem of long-term person re-id setting in which the query subjects may appear with different clothing styles in the gallery set. The proposed solution takes advantage of an image transformation step that facilitates the extraction of identity-unrelated features of persons, including the background area and cloth

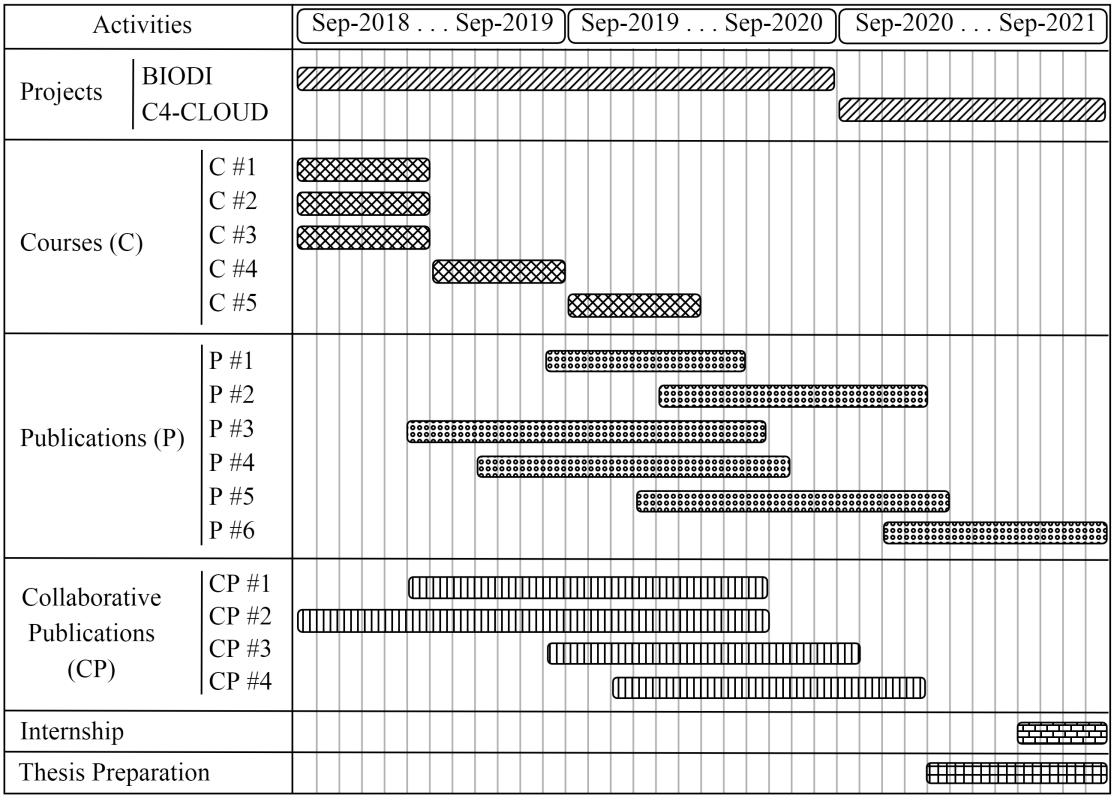


Figure 1.2: Gantt chart: the research progress path including passed courses, industrial research projects, publications, internship period and thesis preparation time line.

textures. Next, we employ a simple CNN equipped with a cosine similarity loss function to only focus on the identity-related features by learning some embeddings that are dissimilar to the previously obtained identity-unrelated features. The main idea of this strategy is to enhance the quality of the final feature representations of people learned by the CNNs in the learning phase; so, the image transformation process and the step performed for extraction of the identity-unrelated features are skipped during the inference phase.

1.4 Research Progress Path

In Fig. 1.2, we illustrate the progress path of this research thesis in a Gantt Chart, including the passed courses, accomplished industrial projects, published papers, timeline of the internship and thesis preparation..

The industrial research projects, namely *BIODI: Biometria e Deteção de Incidentes* and *CLOUD-S-POLIS: Cloudification of Autonomous Security Agents for Urban Environments* provided the financial supports for conducting this PhD research for 2 and 1 years, respectively. Regarding the BIODI project, first, we collected and annotated a comprehensive full-body biometric dataset, and in the remaining time, we implemented two solutions for pedestrian attribute recognition from low-resolution images in the wild. During the CLOUD project period, we focused on the task of person re-id in both short-

term and long-term scenarios and proposed competitive frameworks compared to the existing state of the art methods.

The third cycle of study (PhD) in the University of Beira Interior is a course and research based degree, in which the student first passes several courses prior to entering into the research activities. To accomplish this PhD thesis, totally 5 courses were passed: $C\#1$) *Advanced Topics in Computer Engineering*, $C\#2$) *Neural Networks*, $C\#3$) *Thesis and Seminar Project*, $C\#4$) *Biometric Systems*, $C\#5$) *Cloud Computing Architecture Topics* (see Fig. 1.2).

The objective of the *Advanced Topics in Computer Engineering* course was to provide the attendee with the scientific skills and knowledge of research methodologies. Another aim of this course was to prepare the student for conducting a survey study on the state of the arts of a selected topic. The *Neural Networks* course was taken in the same semester so that the combined knowledge acquired from both previous courses resulted in commencing two survey publications. In the next semester, the course of *Thesis and Seminar Project* was attended to gain the knowledge of preparing a research proposal for the remainder of the PhD.

At the beginning of the second year, the courses *Biometric Systems* and *Cloud Computing Architecture Topics* were participated to improve the general knowledge of recent computer vision techniques in biometrics and cloud-based platforms such as google colab [4]. Specifically, the objective of the *Biometric Systems* course was to provide the attendee with deep insight about the knowledge behind the cutting edge commercial biometric products such as Microsoft Azure [5], Face ++ [6] and Aura Vision [7].

The contribution of this thesis in the biometric field of study is 6 first-authored articles and 4 collaborative publications. The primary publications include 2 survey articles (published in the Applied Sciences and Pattern Recognition Letters journals) and 4 technical papers, from which 2 were published in the Image and Vision Computing and Pattern Recognition Letters journals, one was presented in the BIOSIG-2019 conference in Germany, and one has been recently submitted to a journal media. The contributions of these publications were described in detail in section 1.3. In addition, the timeline of the collaborative publications have been illustrated in Fig. 1.2, and the body of these papers have been presented in attachments 9. The collaborative publications were in line with the objectives of the BIODI and CLOUD projects, in which we first collected and annotated two pedestrian datasets respectively using standstill panels in the urbane environments and using a drone with a mobile camera. Then, developed different solutions to compete with the existing state of the art methods in the field.

The internship period was accomplished in collaboration with Prof. Ruben Vera-Rodriguez, associate professor at the Universidad Autonoma de Madrid. As a result of this collaboration, one paper idea is under process, which is about studying the effect of synthesized data (using human 3D models) on enhancing the generalization ability of

[4] https://colab.research.google.com/

[5] https://azure.microsoft.com/

[6] https://www.faceplusplus.com/

[7] https://auravision.ai/

CNNs for person re-id and pedestrian attribute recognition tasks.

1.5 Thesis Structure

As discussed previously, soft biometric analysis with a focus on attribute recognition and person re-id are long-lasting research topics, mainly because of the continuing demand for monitoring the public environments and social behaviors. Over the last decade, deep convolutional neural networks caused remarkable improvements in the area of pedestrian attribute recognition and person re-id and showed that the performance of the modern surveillance systems even could reach human recognition ability.

In chapter 2, we review the literature of PAR methods and data. We discuss five main existing challenges: localization, limited data, attribute relation, occlusion, and class imbalance. Generally, an optimum localization-based PAR model recognizes attributes based on their expected location; for instance, people's hair color and hairstyle are detected from the head and shoulder area. The challenge of limited data refers to the fact that the existing learning datasets annotated with human attributes are finite, limiting the generalization ability of the model. Attribute relation is another factor required to be considered since the occurrence probability of some attributes is correlated together. For example, the probability of having a beard is very low for a person detected as female. Occlusion is another challenge that requires attention because, in uncontrolled environments, others or objects may block some body parts of the subject person. All the visual attributes may not appear in everybody (e.g., wearing a hat), resulting in the fact that human attribute datasets become very imbalanced in some classes. The challenges mentioned above are repeatedly addressed to different extents in PAR literature; therefore, in chapter 2, we propose a challenged-based taxonomy to categorize them.

We survey several person re-id surveys in chapter 3. Based on several recent surveys, we suggest that the existing re-id strategies can be categorized from five points of view such as scalability, pre-processing and augmentation, model architecture design, post-processing strategies, and robustness to noise. Works with a focus on scalability try to propose efficient techniques to improve the speed and accuracy of person re-id frameworks and perform on-board processing. For example, hashing and transfer learning are two hot topics that lie in the area of scalability-based techniques. Pre-processing and augmentation approaches improve the quality (e.g., generating occluded body parts) or quantity (e.g., synthesizing new samples) of learning data. Some state-of-the-art person re-id frameworks come up with novel deep architectures or processing blocks to improve the final representation of the person by extracting the most useful local and global features. In general, person re-id models receive an image of the target person and deliver a list of images of persons that have the most similarity with the target. Approaches that attempt to re-order the detection list are known as post-processing strategies or re-ranking techniques. Last but not least, person re-id frameworks have to manage the noises resulted from inaccurate bounding boxes of persons, occluded body parts, and wrong

annotations, which is investigated as the last perspective to the literature of the person re-id field. In short, in chapter 3, we address the person re-id problem from five perspectives and elaborate on each of them to highlight the recent advances in the field.

In chapter 4, we propose a pose-sensitive region-based gender classification framework. Considering the assumption that regional features and body pose can improve the quality of the final representation of people, we suggest a framework that provides several classification scores based on the subject's pose and some Regions of Interest (RoI)s. Our experimental results on three datasets such as BIODI, PETA, and MIT show that the aggregation of these classification scores contributes to solid improvements in gender recognition accuracy from full-body images in the wild.

In chapter 5, we propose a model that estimates multiple attributes of people at the same time. To this end, we implement a multi-task framework to consider the semantic correlation between pedestrian attributes and suggest an element-wise multiplication layer to remove destructive features, i.e., background area. Additionally, we present a weighted-sum loss function to manage the importance of each task (groups of attributes) in the course of model training. Finally, we train and test the proposed framework on two well-known PAR datasets (i.e., PETA and RAP) and compare the performance with several state-of-the-art methods.

In chapter 6, we propose a data augmentation technique for person re-id frameworks that help to define the receptive fields of the CNN implicitly. Furthermore, considering the harmful effects of background features on the performance of person re-id models, we present a pre-processing image approach that increases the quantity of learning data such that the person re-id model interprets that the identity and cluttered descriptions are not correlated. The presented model is evaluated on several person re-id datasets: RAP, Market1501, and MSMT17-V2.

In chapter 7 we address the problem of person re-id with an assumption that the same people may appear with different clothing styles. First, we propose to extract the ID-unrelated features of each person by synthesizing an image from each instance in the learning set. Then, we employ a model to learn the long-term representation of persons from the original samples, such that the loss function of the model imposes the embeddings to be dissimilar to the previously extracted ID-unrelated embeddings. This way, the person re-id model learns the ID-related features of people and ignores the background and clothes information. To evaluate the suggested approach, we use two long-term person re-id datasets namely PRCC, and LTCC. Finally, we compare our experimental results with several current methods to evaluate the effectiveness of the proposed framework.

Finally, in chapter 8, we present the conclusions, including discussions on the proposed solutions, memorization of our contributions, and highlights of several future research directions. We discuss that the performance of state-of-the-art methods has exponentially improved over the recent years. However, some scenarios have not been studied profoundly and require more attention to fill the gap between the studies conducted in the laboratories and the industry demands.

Bibliography

[1] N. Dilshad, J. Hwang, J. Song, and N. Sung, "Applications and challenges in video surveillance via drone: A brief survey," in *2020 International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, 2020, pp. 728–732. 1

[2] A. Grigorev, S. Liu, Z. Tian, J. Xiong, S. Rho, and J. Feng, "Delving deeper in drone-based person re-id by employing deep decision forest and attributes fusion," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 1s, pp. 1–15, 2020. 1

[3] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *arXiv preprint arXiv:1610.02984*, 2016. 1

[4] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *IMAGE VISION COMPUT*, vol. 32, no. 4, pp. 270–286, 2014. 1

[5] J. M. Grant and P. J. Flynn, "Crowd scene understanding from video: a survey," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 13, no. 2, pp. 1–23, 2017. 1

[6] N. Khalid, M. Gochoo, A. Jalal, and K. Kim, "Modeling two-person segmentation and locomotion for stereoscopic action identification: A sustainable video surveillance system," *Sustainability*, vol. 13, no. 2, p. 970, 2021. 1

[7] A. B. Mabrouk and E. Zagrouba, "Abnormal behavior recognition for intelligent video surveillance systems: A review," *Expert Systems with Applications*, vol. 91, pp. 480–491, 2018. 1

[8] D. R. Beddiar, B. Nini, M. Sabokrou, and A. Hadid, "Vision-based human activity recognition: a survey," *Multimedia Tools and Applications*, vol. 79, no. 41, pp. 30 509–30 555, 2020. 1

[9] Q. Ke, M. Fritz, and B. Schiele, "Time-conditioned action anticipation in one shot," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9925–9934. 1

[10] J. Arunnehru and M. K. Geetha, "Automatic human emotion recognition in surveillance video," in *Intelligent Techniques in Signal Processing for Multimedia Security*. Springer International Publishing, Oct. 2016, vol. 660, pp. 321–342. [Online]. Available: https://doi.org/10.1007/978-3-319-44790-2_15 1

[11] E. Yaghoubi, A. Kumar, and H. Proença, "Sss-pr: A short survey of surveys in person re-identification," *Pattern Recognit. Lett.*, vol. 143, pp. 50–57, 2021. 1

[12] E. Yaghoubi, D. Borza, J. Neves, A. Kumar, and H. Proença, "An attention-based deep learning model for multiple pedestrian attributes recognition," *IMAGE*

VISION COMPUT., pp. 1–25, 2020. [Online]. Available: https://doi.org/10.1016/j.imavis.2020.103981 1

[13] E. Bentafat, M. M. Rathore, and S. Bakiras, "A practical system for privacy-preserving video surveillance," in *International Conference on Applied Cryptography and Network Security*. Springer, 2020, pp. 21–39. 1

[14] X. Wang, S. Zheng, R. Yang, B. Luo, and J. Tang, "Pedestrian attribute recognition: A survey," *arXiv preprint arXiv:1901.07474*, 2019. 2

[15] A. Brunetti, D. Buongiorno, G. F. Trotta, and V. Bevilacqua, "Computer vision and deep learning techniques for pedestrian detection and tracking: A survey," *Neurocomputing*, vol. 300, pp. 17–33, 2018. 2

[16] Y. Liu, H. Yang, and Q. Zhao, "Hierarchical feature aggregation from body parts for misalignment robust person re-identification," *Applied Sciences*, vol. 9, no. 11, p. 2255, 2019. 2

[17] E. Yaghoubi, F. Khezeli, D. Borza, S. Kumar, J. Neves, and H. Proença, "Human attribute recognition—a comprehensive survey," *Applied Sciences*, vol. 10, no. 16, p. 5608, 2020. 2

[18] F. Becerra-Riera, A. Morales-González, and H. Méndez-Vázquez, "A survey on facial soft biometrics for video surveillance and forensic applications," *Artificial Intelligence Review*, vol. 52, no. 2, pp. 1155–1187, 2019. 3

[19] B. Hassan, E. Izquierdo, and T. Piatrik, "Soft biometrics: a survey," *Multimedia Tools and Applications*, Mar. 2021. [Online]. Available: https://doi.org/10.1007/s11042-021-10622-8 4

Chapter 2

Human Attribute Recognition: A Comprehensive Survey

Abstract. Over the last decade, the field of HAR has dramatically changed, mainly due to the improvements brought by deep learning solutions. This survey reviews the progress obtained in HAR, considering the transition from the traditional hand-crafted to deep-learning approaches. The most relevant works on the field are analyzed concerning the advances proposed to address the HAR's typical challenges. Furthermore, we outline the applications and typical evaluation metrics used in the HAR context and provide a comprehensive review of the publicly available datasets for the development and evaluation of novel HAR approaches.

2.1 Introduction

Over recent years, the increasing amount of multimedia data available in the Internet or supplied by Closed-Circuit TeleVision (CCTV) devices deployed in public/private environments has been raising the requirements for solutions able to automatically analyse human appearance, features and behavior. Hence, HAR has been attracting increasing attentions in the computer vision/pattern recognition domains, mainly due to its potential usability for a wide range of applications (e.g., crowd analysis [1], person search [2; 3], detection [4], tracking [5], and re-identification [6]). HAR aims at describing and understanding the subjects' traits (such as their hair color, clothing style [7], gender [8], etc.) either from full-body or facial data [9]. Generally, there are four main sub-categories in this area of study:

- Facial Attribute Analysis (FAA). Facial attribute analysis aims at estimating the facial attributes or manipulating the desired attributes. The former is usually carried out by extracting a comprehensive feature representation of the face image, followed by a classifier to predict the face attributes. On the other hand, in manipulation works, face images are modified (e.g., glasses are removed or added) using generative models.
- Full-body Attribute Recognition (FAR). Full-body attribute recognition regards the task of inferring the soft-biometric labels of the subject, including clothing style, head-region attributes, recurring actions (talking to the phone) and role (cleaning lady, policeman), regardless of the location or body position (eating in a restaurant).
- PAR. As an emerging research sub-field of HAR, PAR focuses on the full-body human data that have been exclusively collected from video surveillance cameras or panels, where persons are captured while walking, standing, or running.

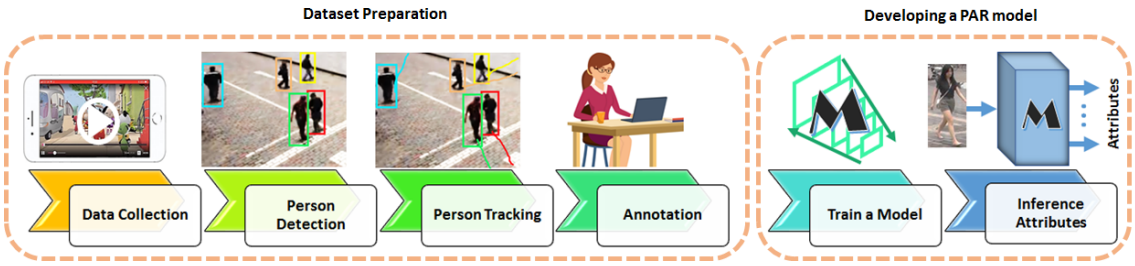


Figure 2.1: Typical pipeline to develop a HAR model.

- Clothing Attribute Analysis (CAA). Another sub-field of human attribute analysis that is exclusively focused on clothing style and type. It comprises several sub-categories such as in-shop retrieval, costumer-to-shop retrieval, fashion landmark detection, fashion analysis, and cloth attribute recognition, each of which requires specific solutions to handle the challenges in the field. Among these sub-categories, cloth attribute recognition is similar to pedestrian and full-body attribute recognition and studies the clothing types (e.g., texture, category, shape, style).

The typical pipeline of the HAR systems is given in Figure 2.1, which indicates the requirement of a dataset preparation prior to designing a model. As shown in Figure 2.1, preparing a dataset for this problem typically comprises four steps:

1. Capturing raw data, which can be accomplished using mobile cameras (e.g., drone) or stationary cameras (e.g., CCTV). Also, the raw data might even be collected from images/videos publicly available (e.g., *Youtube*, or similar sources).

2. In most supervised training approaches, HAR models consider one person at a time (instead of analyzing a full-frame with multiple persons). Therefore, detecting the bounding boxes of each subject is essential and can be done by state-of-the-art object detection solutions (i.e., Mask R-CNN [10], You Only Look Once (YOLO) [11], Single Shot Detector (SSD) [12], etc.)

3. If the raw data is in video format, spatio-temporal information should be kept. in such cases, the accurate tracking of each object (subject) in the scene can significantly ease the annotation process.

4. Finally, in order to label the data with semantic attributes, all the bounding boxes of each individual are displaced to human annotators. based on human perception, the desired labels (e.g., 'gender' or 'age') are then associated to each instance of the dataset.

Regarding the data-type and available annotations, there are many possibilities for designing HAR models. Early researches were based on crafted feature extractors. Typically, the linear Support Vector Machine (SVM) was used with different descriptors (such as ensemble of localized features, local binary patterns, color histograms, histogram of oriented gradients) to estimate the human attributes. However, as the correlation between human attributes were ignored in traditional methods, one single model was

not suitable for estimating several attributes. For instance, descriptors suitable for gender recognition could not be effective enough to recognize the hairstyle. Therefore, conventional methods mostly focused on obtaining independent feature extractors for each attribute. After the advent of Convolutional Neural Network (CNN)s and using it as a holistic feature extractor, a growing number of methods focused on models that can estimate multiple attributes at once. Earlier deep-based methods used shallow networks (e.g., 8-layer AlexNet [13]), while later models moved towards deeper architectures (e.g., Residual Networks (ResNet)) [14].

The difficulties in HAR originates mainly due to the high-variability in human appearance particularly in intra-class samples. Nevertheless, the following factors have been identified as the basis for the development of robust HAR systems:

- learn in an end-to-end manner and yield multiple attributes at once;
- extract a discriminative and comprehensive feature representation from the input data;
- leverage the intrinsic correlations between attributes;
- consider the location of each attribute in a weakly supervised manner;
- are robust to primary challenges such as low-resolution data, pose variation, occlusion, illumination variation, and cluttered background;
- handle the classes imbalance;
- manage the limited-data problem effectively.

Despite the relevant advances and many research articles published, HAR can be considered still in its early stages. For the community to come up with original solutions, it is necessary to be aware of the history of advancements, state-of-the-art performance, and the existing datasets related to this field. Therefore, in this study, we discuss a collection of HAR related works, starting from the traditional one to the most recent proposals, and explain their possible advantages/drawbacks. We further analyze the performance of recent studies. Moreover, although we identified more than 15 publicly available HAR datasets, to the best of our knowledge, we do not have a clear discussion on the aspects that one should observe while collecting a HAR dataset. Thus, after taxonomizing the datasets and describing their main features and data collection setups, we discuss the critical issues of the data preparation step.

Regarding the previously published surveys that addressed similar topics, we particularly mention Zheng et al. [15], where the facial attribute manipulation and estimation methods have been reviewed. However, to date, there is no solid survey on the recent advances in other sub-categories of human attribute analysis. As the essence of full-body, pedestrian, and cloth attribute recognition methods are similar to each other; in this paper, we cover all of them with a particular focus on the pedestrian attribute recognition methods. Meanwhile, Reference [16] is the only work similar to our survey that is about pedestrian attribute recognition. Several points distinguish our work from Reference [16]:

- The recent literature on HAR has been mostly focused on addressing some particular challenges of this problem (such as class imbalance, attribute localization, etc.) rather devising a general HAR system. Therefore, instead of providing a methodological categorization of the literature as in Reference [16], our survey proposes a challenge-based taxonomy, discussing the state-of-the-art solutions and the rationale behind them;
- Contrary to Reference [16], we analyze the motivation of each work and the intuitive reason for its superior performance;
- The datasets main features, statistics and types of annotation are compared and discussed in detail;
- Beside the motivations, we discuss HAR applications, divided into three main categories: security, commercial, and related research directions.

Motivation and Applications

Human attribute recognition methods extract semantic features that describe human-understandable characteristics of the individuals in a scene, either from images or video sequences, ranging from demographic information (gender, age, race/ethnicity), appearance attributes (body weight, face shape, hairstyle and color etc.), emotional state, to the motivation and attention of people (head pose, gaze direction). As they provide vital information about humans, such systems have already been integrated into numerous real-world applications, and are entwined with many technologies across the globe.

Indisputably, HAR is one of the most important steps in any visual surveillance system. Biometric identifiers are extracted to identify and distinguish between the individuals. Based on the biometric traits, humans are uniquely identified, either based on their facial appearance [17–19], iris patterns [20] or on behavioral traits (gait) [21; 22]. With the increase of surveillance cameras worldwide, the research focus has shifted from hard biometric (e.g., iris recognition and palm print) to soft biometric identifiers. The latter describe human characteristics, taxonomized into a humanly understandable manner, but are not sufficient to uniquely differentiate between individuals. Instead, they are descriptors used by humans to categorize their peers into several classes.

On a top level, HAR applications can be divided into three main categories: *security and safety*, *research* directions, and *commercial applications*.

Yielding high-level semantic information, HAR could provide auxiliary information for different computer vision tasks, such as person re-identification ([23; 24]), human action recognition [25], scene understanding, advanced driving assistance systems, and event detection ([26]).

Another fertile field where HAR could be applied is in human drone surveillance. Drones or Unmanned Aerial Vehicle (UAV), although initially designed for military applications, are rapidly extending to various other application domains, due to their reduced size, swiftness, and ability to navigate through remote and dangerous environments.

Researchers in multiple fields have started to use UAVs drones in their research work, and, as a result, the Scopus database has shown an increase in the papers related to UAVs, from 11 (4.7×10^6 of total papers) papers published in 2009 to 851 (270.0×10^6 of total articles) published in 2018 [27]. In terms of human surveillance, drones have been successfully used in various scenarios, ranging from rescue operations and victim identification, people counting and crowd detection, to police activities. All these applications require information about human attributes.

Nowadays, researchers in universities and major car industries work together to design and build the self-driving cars of the future. HAR methods have important implications in such systems as well. Although numerous papers addressed the problem of pedestrian detection, pedestrian attribute recognition is one of the keys to future improvements. Cues about the pedestrians' body and head orientation provide insights about their intent, and thus avoiding collisions. The pedestrians' age is another aspect that should be analyzed by advanced driving assistance systems to decrease vehicle speed when children are on the sidewalk. Finally, other works suggest that even pedestrians' accessories could be used to avoid collisions: starting from the statistical evidence that collisions between pedestrians and vehicles are more frequent on rainy days, in Reference [28] authors suggest that detecting whether a pedestrian has on open umbrella could reduce traffic incidences.

As mentioned above, the applications of biometric cues are not limited to surveillance systems. Such traits have necessary implications also in commercial applications (logins, medical records management) and government applications (ID cards, border, and passport control) [29]. Also, a recent trend is to have advertisement displays in malls and stores equipped with cameras and HAR systems to extract socio-demographic attributes of the audience and present appropriate and targeted ads based on the audience's gender, generation or age.

Of course, this application list is not exhaustive, and numerous other practical uses of HAR can be envisioned, as this task has implications in all fields interested in and requiring (detailed) human description.

In the remainder of this paper, we first describe the HAR preliminaries—dataset preparation, and the general difference between the earliest and most recent model approaches. In Section 2.3, we survey the HAR techniques from their main challenge point-of-view, in order to increase the reader's creativity in introducing novel ideas for solving the task of HAR. Further, in Sections 2.4 and 2.5, we detail the existing PAR, FAR, and CAA datasets and commonly used evaluation metrics for HAR models. In Section 2.6, we discuss the advantages and disadvantages of the above-presented methods and compare their performance over the well-known HAR datasets.

2.2 Human Attribute Recognition Preliminaries

To recognize the human full-body attributes, it is necessary to follow a two-step pipeline, as depicted in Figure 2.1. In the remainder of this section, each of these steps is described

in detail.

2.2.1 Data Preparation

Developing a HAR model requires relevant annotated data, such that each person is manually labeled based on its semantic attributes. As discussed in Section 2.4, there are different types of data sources such as fashion, aerial, and synthetic datasets, which could be collected from the Internet resources (e.g., Flickr) or through static or mobile cameras in indoor/outdoor locations. HAR models are often developed to recognize human attributes from person bounding boxes (instead of analyzing an entire frame comprising multiple persons). That is why, after the data collection step, it is required to pre-process the data and extract the bounding box of each person. Earlier methods use human annotators to specify the person locations in each image, and then assign soft biometric labels to each of person bounding boxed, while recent approaches take advantage of the CNN-based person detectors (e.g., Reference [10])—or trackers [30], if the data is collected as videos—to provide the human annotators with person bounding boxes for more labeling processes. We refer the interested reader to Reference [31] for more information on person detection and tracking methods.

2.2.2 HAR Model Development

In this part, we discuss the main problem in HAR and highlight the differences between the earlier methods and the most recent deep learning-based approaches.
In machine learning, classification is most often seen as a supervised learning task, in which a model learns from the labeled input data to predict the appeared classes in the unseen data. For example, given many person images with gender labels (‘male’ or ‘female’), we develop an algorithm to find the relationship between images and labels, based on which we predict the labels of the new images. Fisher’s linear discriminant [32], support vector machine [33], decision trees [34; 35], and neural networks [36; 37] are examples of classification algorithms. As the input data is large or suspected to have redundant measures, before analyzing it for classification, the image is transformed into a reduced set of features. This transformation can be performed using neural networks [38] or different feature descriptors [39]—such as Major Colour Spectrum Histogram (MCSH) [40], Color Structure Descriptor (CSD) [41; 42], Scale-Invariant Feature Transform (SIFT) [43; 44], Maximally Stable Colour Regions (MSCR) [45; 46], Recurrent Highly-Structured Patches (RHSP), and Histogram of Histogram of Oriented Gradients (HOG) [47–49]. Image descriptors are not generalized to all the computer vision problems and may be suitable only for specific data type—for example, color descriptors are only suitable for color images. Therefore, models based on feature descriptors are often called hand-crafted methods, in which we should define and apply proper feature descriptors to extract a comprehensive and distinct set of features from each input image. This process may require more feature engineering, such as dimensionality reduction, feature selection, and fusion. Later, based on the extracted

features, multiple classifiers are learned, such that each one is specialized in predicting specific attributes of the given input image. As the reader may have noticed, these steps are offline (the result of each step should be saved on the disk as the input of the next step). On the contrary, deep neural networks are capable of modeling the complex non-linear relationships between the input image and labels, such that the feature extraction and classifier learning are performed simultaneously. Deep neural networks are implemented as multi-level (large to small feature-map dimensions) layers, in which different processing filters are convoluted with the output of the previous layer. In the first levels of the model, low-level features (e.g., edges) are extracted, while mid-layers and last-layers extract the mid-level features (e.g., texture) and high-level features (e.g., expressiveness of the data), respectively. To learn the classification, several fully connected layers are added on top of the convolutional layers (known as a backbone) to map the last feature map to a feature vector with several neurons equal to the number of class labels (attributes).

Several major advantages of deep learning approaches moved the main research trend towards the deep neural network methods. First, CNNs are end-to-end (i.e., both the feature extraction and classification layers are trained simultaneously). Second, the deep neural networks' high generalization ability has provided the possibility of transferring the knowledge of other similar fields to scenarios with limited data. As an example, applying the weights of a model that has been trained on a large dataset (e.g., ImageNet [50]) not only has shown positive effects on the accuracy of the model but also has decreased the convergence time and over-fitting problem [51–53]. Thirdly, CNNs could be designed to handle multiple tasks and labels in a unified model [54; 55].

To fully understand the discussion on the state of the arts in HAR, we encourage the newcomer readers to read about different architectures of deep neural networks and their components in References [56; 57]. Meanwhile, common evaluation metrics are explained in Section 2.5.

2.3 Discussion of Sources

As depicted in Figure 2.2, we identified five major challenges frequently addressed by the literature on HAR—localization, limited learning data, attribute relation, body-part occlusion, and data class imbalance.

HAR datasets only provide the labels for a bounding box of person, but the locations related to each attribute are not annotated. Finding which features are related to which parts of the body is not a trivial task (mainly because body posture is always changing), and not fulfilling it may cause an error in prediction. For example, recognizing the 'wearing sunglasses' attribute in a full-body image of a person without considering the eyeglasses' location may lead to omitting the sunglasses feature information due to extensive pooling layers and a small region of the eyeglasses, compared to the whole image. This challenge is known as localization (Section 2.3.1), as in which we attempt to extract features of different spatial locations of the image to be certain no information is lost, and we can

extract distant features from the input data.

Earlier methods used to work with limited data as the mathematical calculations were computationally expensive, and increasing the amount of data could not justify the exponential computational cost and the amount of improvement in the accuracy. After the deep learning breakthrough, more data proved to be effective in the generalization ability of the models. However, collecting and annotating very large datasets is prohibitively expensive. This issue is known as limited data challenge, which has been the subject of many studies in the deep neural network fields of study, including deep-based HAR, addressed in Section 2.3.2.

In the context of HAR, dozens of attributes are often analyzed together. As humans, we know that some of these attributes are highly correlated, and knowing one can improve the recognition probability of the other attributes. For example, for a person wearing a 'tie,' it is less likely to wear a 'Pyjama' and more likely to wear a 'shirt' and 'suit'. Studies that address the relationship between attributes as their main contribution are categorized in the 'attribute relation' taxonomy and discussed in Section 2.3.3.

Body parts occlusion is another challenge when dealing with HAR data that has not yet been addressed by many studies. The challenge in occluded body parts is not only about the missing information of the body parts, but also the presence of some misleading features of other persons or objects. Further, because in HAR, some attributes are related to specific regions, considering the occluded parts before the prediction is important. For example, for a person with an occluded lower body, yielding predictions about the attributes located in the lower body region is questionable. In Section 2.3.4, we discuss the methods and ideas that have particularly addressed the occlusion in HAR data.

Another critical challenge in HAR is the imbalanced number of samples in each class of data. Naturally, an observer sees fewer persons wearing long coats, while there are many persons in the community that appear with a pair of jeans. That is why the HAR datasets are intrinsically imbalanced and cause the model to be biased/over-fitted on some classes of data. Many studies address this challenge in HAR data, which have been discussed in Section 2.3.5.

Among the major challenges in HAR, considering attribute correlation and extracting fine-grained features from local regions of the given data have attracted the most attention, such that recent works [58; 59] attempt to develop some models that could address both challenges at the same time. Data class imbalance is another contribution of many HAR methods which is often handled by applying weighted loss functions to increase the importance of the minority samples and decrease the effect of the samples from classes with many samples. To deal with limited data challenges, scholars frequently apply the existing holistic transfer learning and augmentation techniques in computer vision and pattern recognition. In this section, we discuss the significant contributions of the literature works in alleviating the main challenges in HAR.

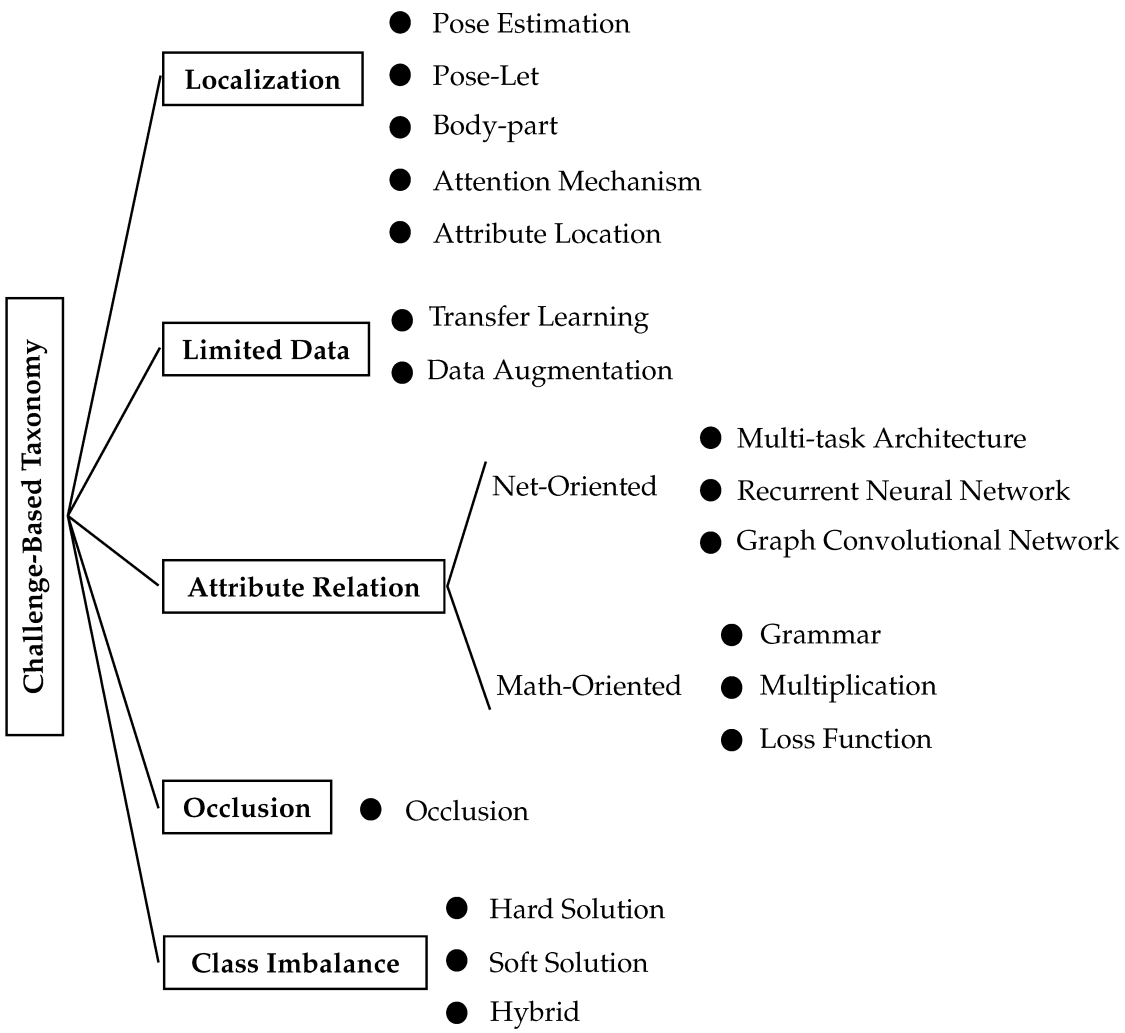


Figure 2.2: The proposed taxonomy for main challenges in HAR.

2.3.1 Localization Methods

Analyzing human full-body images only yields the global features; therefore, to extract distinct features from each identity, analyzing local regions of the image becomes important [60]. To capture the human fine-grained features, typical methods divide the person's image into several strides or patches and aggregate all the decisions on parts to yield the final decision. The intuition behind these method is that, decomposition of human-body and comparing it with others is intuitively similar to localizing the semantic body-parts and then describing them. In the following, we survey 5 types of localization approaches—(1) attribute location-based methods that consider the spatial location of each attribute in the image (e.g., glasses features are located in the head area, while shoes features are in the lower part of the image); (2) attention mechanism-based techniques that attempts to automatically find on the most important locations of the image based on the ground truth labels; (3) body part-based models, in which the model first locates the body parts (i.e., head, torso, hands, and legs) and then extract the related features from each body parts and aggregate them; (4) pose-let-based techniques that extracts the features from many random locations of the image and aggregate them; (5) pose

estimation-based methods that use the coordination of the body skeleton/joints to extract the local features.

2.3.1.1 Pose Estimation-Based Methods

Considering the effect of the body-pose variation of the feature representation, [61] proposes to learn multiple attribute classifiers so that each of them is suitable for a specific body-pose. Therefore, authors use the Inception architecture [62] as the backbone feature extractor, followed by three branches to capture the specific features of the front, back, and side views of the individuals. Simultaneously, a view-sensitive module analyzes the extracted features from the backbone to refine each branch's scores. The final results are the concatenation of all the scores. Ablation studies on the PEdesTrian Attribute (PETA) dataset show that a plain Inception model achieves an 84.4 $F1$-score, while for the model with a pose-sensitive module, this metric increases to 85.5.

Reference [63] is another research that takes advantage of pose estimation for improving the performance of pedestrian attribute recognition. In this work, Li et al. suggested a two-stream model whose results are fused, allowing the model to benefit from both regular global and pose-sensitive features. Given an input image, the first stream extracts the regular global features. The pose-sensitive branch comprises three steps—(1) coarse pose estimator (body-joint coordinates predictor) by applying the approach proposed in Reference[64], (2) region localization that uses the body-pose information to spatially transform the desired region, originally proposed in References [65], (3) fusion layer that concatenates the features of each region. In the first step, pose coordinates are extracted to be shared with the second module, in which body parts are localized by using spatial transformer networks [65]. A specific classifier is then trained for each region. Finally, the extracted features from both streams are concatenated to return a comprehensive feature representation of the given input data.

2.3.1.2 Pose-Let-Based Methods

The main idea of pose-let based methods is to provide a bag-of-features from the input data using different patching technique. As earlier methods lacked accurate body part detectors, overlapping patches of the input images were used to extract local features.

Reference [66] is one of the first techniques in this group that uses Spatial Pyramid Representation (SPR) [67] to divide the images into grids. Unlike a standard bag-of-features method that extracts the features from a uniform patching distribution, they suggest a recursive splitting technique, in which each grid has a parameter that is jointly learned with the weight vector. Intuitively, the spatial grids are varying for each class, which leads to better feature extraction.

In Reference [68], hundreds of pose-lets are detected from the input data; a classifier is trained for each pose-let and semantic attribute. Then, another classifier aggregates the body-part information, with emphasis on the pose-lets taken from usual viewpoints that have discriminative features. A third classifier is then used to consider the relationship

between the attributes. This way, by using the obtained feature representation, the body pose and viewpoint are implicitly decomposed.

Noticing the importance of accurate body-part detection when dealing with clothing appearance variations, Reference [69] proposes to learn a comprehensive dictionary that considers various appearance part types (e.g., representing the lower-body in different appearances from bare legs to long skirts). To this end, all the input images are divided into static overlapping cells, each of which is represented by a feature descriptor. Then, as a result of feature clustering into K clusters, they represent k types of appearance parts.

In Reference [70], the authors targeted the human attributes and action recognition from still images. To this end, supposing that the available human bounding boxes are located in the center of the image, the model learns the scale and positions of a series of image partitions. Later, the model predicts the labels based on the reconstructed image from the learned partitions.

To address the large variation in articulation, angle, and body-pose [71] proposes a CNN-based features extractor, in which each pose-let is fed to an independent CNN. Then, a linear SVM classifier learns to distinguish the human attributes based on the aggregation between the full-body and pose-let features.

References [72; 73] showed that not only CNNs can yield a high-quality feature representation from the input, but also they are better at classification than SVM classifiers. In this context, Zhu et al. propose to predict multiple attributes at-once, by implicit regard to the attribute dependencies. Therefore, the authors divide the image into 15 static patches and analyze each one with a separate CNN. To consider the relationship between attributes and patches, they connect the output of some specific CNNs to the relevant static patches. For example, the upper splits of the images are connected to the head and shoulder's attributes.

Reference [74] claims that in previous pose-let works, the location information of the attributes is ignored. For example, to recognize whether a person wears a hat or not, knowing that this feature is related to the upper regions of the image can guide the model to extract more relevant features. To implement this idea, the authors used an Inception [62] structure, in which the features of three different levels (low, middle, and high levels) are fed to three identical modules. These modules extract different patches from the whole and part of the input feature maps. The aggregation of the three branches yields the final feature representation. By following this architectural design, the model implicitly learns the regions related to each attribute in a weakly supervised method. Surprisingly, the baseline (the same implementation without the proposed module) achieves better results on the PETA dataset (84.9 vs. 83.4 of $F1$), while on Richly Annotated Pedestrian (RAP) dataset, the results of the model equipped with their proposed module ($F168.6$) is better with a margin of 2.

Reference [75] receives the full frames and uses the scene features (i.e., hierarchical contexts) to help the model learn the attributes of the targeted person. For example, in a sports scene, it is expected that people have sporty style clothing. Using Fast R-CNN [76], the bounding box of each individual is detected, and several pose-let are extracted. After

feeding the input frame and its Gaussian pyramids into several convolutional layers, four fully connected branches are added to the top of the network to yield four scores (from human bounding box, pose-lets, nearest neighbors of the selected parts, and full-frame) for a final concatenation.

2.3.1.3 Part-Based Methods

Extracting discriminative fine-grained features often requires first to localize patches of the relevant regions in the input data. Unlike pose-let-based methods that detect the patches from the entire image, part-based methods aim to learn based on accurate body parts (i.e., head, torso, arms, and legs). Optimal part-based models are (1) pose sensitive (i.e., for similar poses, shows strong activations); (2) extendable to all samples; (3) discriminative on extracting features. CNNs can handle all these factors to some extend, and [77] empirical experiments confirm that for deeper networks, accurate body-parts are less significant.

As one of the first part-based works, inspired by a part detector (i.e., deformable part model [78], which captures viewpoint and pose variations), Zhang et al. [79] propose two descriptors that learn based on the part annotations. Their main objective is to localize the semantic parts and obtain a normalized pose representation. To this end, the first descriptor is fed by correlated body parts, while for the second descriptor, the input body splits have no semantic correlation. Intuitively, the first descriptor is based on the inherent semantics of the input image, and the second descriptor learns the cross-component correspondences between the body parts.

Later, in this context, Reference [77] proposes a model composed of a CNN-based body-part detector, including an SVM classifier (trained on the full-body and body parts, that is, head, torso, and legs) to predict the human attributes and action. Given an input image, a Gaussian pyramid is obtained, each level is fed to several convolutional layers to produce pyramids of feature maps. The convolution of each feature-level with each body-part produces scores correspond to that body-part. Therefore, the final output is a pyramid of part model scores suitable for learning an SVM classifier. The experiments indicate that using body-part analysis and making the network deeper improve the results.

As earlier part-based methods used separate feature extractors and classifiers, the parts could not be optimized for recognizing the semantic attributes. Moreover, the detectors, at that time, were inaccurate in detection. Therefore, Reference [80] proposed an end-to-end model, in which the body partitions are generated based on the skeleton information. As authors augment a large skeleton estimation dataset (MPII [81]) for human skeleton information (which is less prone to error for annotation in comparison with bounding box annotations for body parts), their body detector is more accurate in detecting the relevant partitions, leading to better performance.

To encode both global and fine-grained features and implicitly relate them to the specific attributes, References [82] proposes to add several branches on top of a ResNet50 network, such that each branch explores particular regions of the input data and learns an exclusive classifier. Meanwhile, before the classifier stage, all branches share a layer,

which passes the 6 static regions of features to the attribute classifiers. For example, the head attribute classifier is fed only with the two upper strips of the feature maps. Experimental results on the Market-1501 dataset [24] show that applying a layer that feeds regional features to the related classifiers can improve the mA from 85.0 to 86.2. Further, repeating the experiments while adding a branch to the architecture of the model for predicting the person ID (as an extra-label) improves the mA result from 84.9 to 86.1. These experiments show that simultaneous ID prediction without any purpose could slightly diminish the accuracy.

2.3.1.4 Attention Based Methods

By focusing on the most relevant regions of the input data, human beings recognize the objects and their attributes without the background's interference. For example, when recognizing the head-accessories attributes of an individual, special attention is given to the facial region. Therefore, many HAR methods have attempted to implement an attention module to be inserted at multiple levels of CNN. Attention heat maps (also called localization score map [83; 84] or class activation map [85]) are colorful localization score maps that make the model interpretable and are usually faded over the original image to show the model's ability to focus on the relevant regions.

In order to eliminate the need for body-part detection and prior correspondence among the patches, Reference [86] proposed to refine the Class Activation Map network [85], in which the relevant regions of the image to each attribute are highlighted. The model comprises a CNN feature extraction backbone with several branches on its top, which yield the scores for all the attributes and their regional heat maps. The fitness of the attention heat maps is measured using an exponential loss function, while the score of the attributes is derived from a classification loss function. The evaluation of the model is performed using two different convolutional backbones (i.e., Visual Geometry Group (VGG) [87] and AlexNet [13] models), and the result for the deeper network (VGG16) is better than the other one.

To extract more distinctive global and local features, Liu et al. [88] propose an attention module that fuses several feature layers of the relevant regions and yields attention maps. To take full advantage of the attention mechanism, they apply the attention module to different model levels. Obtaining the attentive feature maps from various layers of the network means that the model has captured multiple levels of the input sample's visual patterns so that the attention maps from higher blocks can cover more extensive regions, and the lower blocks focus on smaller regions of the input data.

Considering the problem of cloth classification and landmark detection, Reference [89] proposes an attentive fashion grammar network, in which both the symmetry of the cloths and effect of body motion is captured. To enhance the clothing classification, authors suggest to (1) develop supervised attention using the ground truth landmarks to learn the functional parts of the clothes and (2) use a bottom-up, top-down network [90], in which a successive down and up-sampling are performed on the attention maps to learn the global attention. The evaluation results of their model for clothing attribute prediction

improved the counterpart methods by a large margin (30% to 60% top-5 accuracy on the DeepFashoin-C dataset [91]).

With a view to select the discriminative regions of the input data, Reference [92] proposes a model considering three aspects: (1) Using the parsing technique [93], they split features of each body-part and help the model learns the location-oriented features by pixel-to-pixel supervision. (2) Multiple attention maps are assigned to each label due to empowering the features from the relevant regions to that label and suppressing the other features. Different from the previous step, the supervision in this module is performed on the image-level. (3) Another module learns the relevant regions for all the attributes and learns from a global perspective. The quantitative results on several datasets show that the full version of the model improves the plain model's performance slightly (e.g., for the RAP dataset, the F1 metric improves from 79.15 to 79.98).

Reference [94] is another research that has focused on localizing the human attributes engaging multi-level attention mechanisms in full-frame images. First, supervised coarse learning is performed on the target person, in which the extracted features of each residual block is multiplied by the ground truth mask. Then, inspired by Reference [95], to further boost the attribute-based localization, an attention module uses the labels to refine the aggregated features of multiple levels of the model.

To alleviate the complex background and occlusion challenges in HAR, Reference [96] introduces a coarse attention layer that uses the multiplication between the output of the CNN backbone and ground truth human masks. Further, to guide the model to consider the semantic relationships among the attributes, authors use a multi-task architecture with a weighted loss function. This way, the CNN learns to find the relevant regions to the attributes in the foreground regions. Their ablation studies show that considering the correlation between attributes (multi-task learning) is more effective than coarse attention on the foreground region, although both improve the model performance.

2.3.1.5 Attribute Based Methods

Noticing the effectiveness of the additional information (e.g., pose, body-part and viewpoint) in the global feature representation, Reference [97] introduces a method that improves the localization ability of the model by locating the attributes' regions in the images. The model comprises two branches, one of them extracts the global features and provides the Class Activation Maps (CAMs) [98] (attention heat-maps), and the other one uses [99] to produce some RoI for extracting the local features. To localize each attribute, the authors consider regions with high overlap between the CAMs and RoI as the attribute location. Finally, the local and global features are aggregated using an element-wise sum. Their ablation studies on the RAP dataset show that for the model without localization $F1$ metric is about 77%, while the full-version model improves the results to about 80%.

As a weakly supervised method, Reference [100] aims to learn the regions in the input data related to the specific attributes. Thereby, the input image is fed into a Batch Normalization (BN)-Inception model [101], and the features from three levels of the model (low, mid, and high) are concatenated together to be ready for three separate

localization process. The localization module is built from a Squeeze-and-Excitation Networks (SE-Net) [102] (that considers the channel relationships) proceeded with a Spatial Transformer Network (STN) (that performs conditional transformations on the feature maps) [65]. The training is weakly supervised because instead of using the ground truth coordinates of the attribute region, the STN is treated as a differentiable RoI pooling layer that is learned without box annotations. The $F1$ metric on the RAP dataset for BN-Inception plain model is around 78.2 while this number fro the full version of the model is 80.2.
Considering that both the local and global features are important for making a prediction, most of the literature's localization-based methods have introduced modular techniques. Therefore, the proposed module could be used in multiple levels of the model (from the first convolutional layers to the final classification layers) to capture both the low-level and high-level features. Intuitively, the implicit location of each attribute is learned in a weakly supervised manner.

2.3.2 Limited Data

Although deep neural networks are powerful in the attribute recognition task, an insufficient amount of data causes an early overfitting problem and hinders them from extracting a generalized feature representation from the input data. Meanwhile, the deeper the networks are, the more data are required to learn a wide range of layer weight parameters. Data augmentation and transfer learning are two primary solutions that address the challenge of limited data in computer vision tasks. In the context of HAR, there are few researches that have studied the effectiveness of these methods that are discussed in the following.
(A) Data Augmentatio. In this context, Bekele et al. [103] studied the effectiveness of 3 basic data augmentation techniques on their proposed solution and observed that the $F1$ score is improved from 85.7 to 86.4 for an experiment on the PETA dataset. Further, [104] discussed that ResNet could take advantage of the skipped connections to avoid overfitting. Their experimental results on the PETA dataset confirm the superiority of ResNet without augmentation over the SVM-based and plain CNN models.
(B) Transfer Learning. In clothing attribute recognition, some works may deal with two domains (types of images): (1) in shop images that are high-quality in specific poses; (2) in-the-wild images that vary in the pose, illumination, and resolution. To address the problem of limited labeled data, we can transfer the knowledge of one domain to the other domain. In this context, inspired by curriculum learning, Dong et al. [105] suggest a two-step framework for curriculum transfer of knowledge from shop clothing images to in-the-wild *similar* clothing images. To this end, they train a multi-task network with easy samples (in-shop) and copy its weights to a triplet-branch curriculum transfer network. At first, these branches have identical weights; however, in the second training stage (with harder examples), the feature similarity values between the target and the positive branches become larger than between the target and negative branches. The ablation studies confirm the effectiveness of the authors' idea and show that the mean average

(mA) improved from 51.4 to 58.8 for plain multi-task and proposed model, respectively, on the Cross-Domain clothing dataset [106]. Moreover, this work indicates that curriculum learning versus end-to-end learning achieves better results, with 62.3 and 64.4 of mA, respectively.

2.3.3 Attributes Relationship

Both the spatial and semantic relationships among the attributes affect the performance of the PAR models. For example, hairstyle and footwear are correlated, while related to different regions (i.e., spatial distributions) of the input data. Regarding the semantic relationship, pedestrian attributes may either conflict with each other or are mutually confirming. For instance, wearing jeans and a skirt is an unexpected outfit, while wearing a T-shirt and sports shoes may co-appear with high probability. Therefore, taking these intuitive interpretations into account could be considered as a refinement step that improves the prediction-list of the attributes [107]. Furthermore, considering the contextual relation between various regions improve the performance of the PAR models. To consider the correlation among the attributes there are several possibilities such as using multi-task architecture [96], multi-label classification with weighted loss function [108], Recurrent Neural Networks (RNN) [109], Graph Convolutional Network (GCN) [110]. We have classified them into two main groups:

- Network-Oriented methods that take advantage of the various implementation of convolutional layers/blocks to discover the relation between attributes,
- math-oriented methods that may or may not extract the features using CNNs, but perform some mathematical operations on the features to modify them regarding the existing intrinsic correlations among the attributes.

In the following, we discuss the literature of both categories.

2.3.3.1 Network-Oriented Attribute Correlation Consideration

(A) Multi-task Learning. In [55], Lu et al. discuss that the intuition-based design of multi-task models is not an optimal solution for sharing the relevant information over multiple tasks, and they propose to gradually widen the structure of the model (add new branches) using an iterative algorithm. Consequently, in the final architecture, correlated tasks share most of the convolutional blocks together, while uncorrelated tasks will use different branches. Evaluation of the model on the fashion dataset [91] shows that by widening the network to 32 branches, the accuracy of the model cannot compete with other counterparts; however, the speed increases (from 34 ms to 10 ms) and the number of parameters decreases from 134 million to 10.5 million.
In a multi-task attribute recognition problem, each task may have a different convergence rate. To alleviate this problem and jointly learn multiple tasks, Reference [111] proposes a weighted loss function that updates the weights for each task in the course of learning.

The experimental evaluation on the Market-1501 dataset [24] shows an improvement in accuracy from 86.8% to 88.5%.

In [112; 113], the authors study the multi-task nature of PAR and attempt to build an optimal grouping of the correlated tasks, based on which they share the knowledge between tasks. The intuition is that, similar to the human brain, the model should learn more manageable tasks first and then uses them for solving more complex tasks. The authors claim that learning correlated tasks needs less effort, while uncorrelated tasks require specific feature representations. Therefore, they apply a curriculum learning schedule to transfer the knowledge of the easier tasks (strongly correlated) to the harder ones (weakly correlated). The baseline results show that learning the tasks individually yields 71.0% accuracy on the Soft Biometric Retrieval (SoBiR) dataset [114], while this number for learning multiple tasks at once is 71.3% and for a curriculum-based multi-task model is 74.2%.

Considering HAR as a multi-task problem, Reference [54] proposes to improve the model architecture in terms of feature sharing between tasks. Authors claim that by learning a linear combination of features, the inter-dependency of the channels is ignored, and the model cannot exchange spatial information. Therefore, after each convolutional block in the model, they insert a shared module between tasks to share the information. This module considers three aspects: (1) fusing the features of each two tasks together, (2) generating attention maps regarding the location of the attributes [115], and (3) keeping the effect of the original features of each task. Ablation studies over this module's positioning indicate that adding it at the end of the convolutional blocks yields the best results. However, the performance is approximately stable when different branches of the module (one at a time) are ablated.

(B) RNN. In [116], authors discuss that person re-id focuses on the global features, while attribute recognition relies on local aspects of individuals. Therefore, Liu et al. [116] propose a network consisted of three parts that work together to learn the person's attributes and re-identification (re-id). Further, to capture the contextual spatial relationships and focus to the location of each attribute, they use the RNN-CNN backbone feature extractor followed by an attention model.

To mine the relation of attributes, Reference [117] uses a model based on Long Short Term Memory (LSTM). Intuitively, using several successive stages of LSTM preserves the necessary information along the pipeline and forgets the uncorrelated features. In this work, the authors first detect three-body pose-lets based on the skeleton information. They consider the full-body as another pose-let followed by several fully connected layers to produce several groups of features (for each attribute, one group of features). Each group of features is passed to an LSTM block, followed by a fully-connected layer. Finally, the concatenation of all features is considered as the final feature representation of the input image. Considering that LSTM blocks are successively connected to each other, they carry the useful information of previous groups of features to the next LSTM. The ablation study in this work shows that the plain Inception-v3 on PETA dataset attains 85.7 of $F1$ metric, and adding LSTM blocks on top of the baseline improves its performance to 86.0,

while the full version of the model that processes the body-parts achieves to $F1$ 86.5.

Regarding the functionality of RNN in contextual combinations in the sequenced data, Reference [118] introduces two different methods to localize the semantic attributes and capture their correlations implicitly. In the first method, the input image's extracted features are divided into several groups; then, each group of features is given to an LSTM layer followed by a regular convolution block and a fully connected layer, while all the LSTM layers are connected together successively. In the second method, all the extracted features from the backbone are multiplied (spatial point-wise multiplication) by the last convolution block's output to provide global attention. The experiments show that dividing the features into groups from global to local features yields better results than random selection.

Inspired by image-captioning methods, Reference [119] introduced a Neural PAR that converts attributes recognition to the image-captioning task. To this end, they generated sentence vectors to describe each pedestrian image using a random combination of attribute-words. However, there are two major disruptions in designing an image-caption architecture for attribute classification: (1) variable length of sentences (attribute-words) for different pedestrians and (2) finding relevance between attributes vectors and spatial space. To address these challenges, the authors used RNNs units and lookup-table, respectively.

To deal with low-resolution images, Wang et al. [109] formulated the PAR task as a sequential prediction problem, in which a two-step model is used to encode and decode the attributes for discovering both the context of intra-individual attributes and the inter-attribute relation. To this end, Wang et al. took advantage of LSTMs in both encode and decode steps for different purposes, such that in the encoding step the context of the intra-person attributes is learned, while in the decoding step, LSTMs is utilized to learn the inter-attributes correlation and predict the attributes as a sequence prediction problem.

(C) GCN. In Reference [110], Li et al. introduce a sequential-based model that relies on two graph convolutional networks, in which the semantic attributes are used as the nodes of the first graph, and patches of the input image are used as the nodes of the second graph. To discover the correlation between regions and semantic attributes, they embedded the output of the first graph as the extra inputs into the second graph and vise versa (the output of the second graph is embedded as the extra inputs into the first graph). To avoid a closed loop in the architecture, they defined two separate feed-forward branches, such that the first branch receives the image patches and presents the spatial context representation of them. This representation is then mapped into the semantic space to produce the features that capture the similarity between regions. The second branch input is semantic attributes that are processed using a graph network and mapped into spatial graphs to capture the semantic-aware features. The output of both branches is fused to let and end-to-end learning. The ablation studies show that in comparison with a plain ResNet50 network, the $F1$ results could improve by margins of 3.5 and 1.3 for the PETA and RAP datasets, respectively.

Inspired by Reference [110], in Reference [107], Li et al. present a GCN-based model

to yield the human parsing alongside the human attributes. Therefore, a graph is built upon the image features so that each group of features corresponds to one node of the graph. Afterward, to capture the relationships among the groups of attributes, a graph convolution is performed. Finally, for each node, a classifier is learned to predict the attributes. To produce the human parsing results, they apply a residual block that uses both the original features and the output of the graph convolution in the previous branch. Based on the ablation study, a plain ResNet50 on the PETA dataset achieves a $F1$ score of 85.0, while a model based on body parts yields a $F1$ score of 84.4, and this number for the model equipped with the above-mentioned idea is 87.9.

Tan et al. [120] observed the close relationship between some of the human attributes and claimed that in multi-task architectures, the final loss function layer is the critical point of learning, which may not have sufficient influence for obtaining a comprehensive representation for explaining the attribute correlations. Moreover, the limitation in receptive fields of CNNs [121] hinders the model's ability to effectively learn the contextual relations in the data. Therefore, to capture the structural connections among attributes and contextual information, the authors use two GCN [122]. However, as image data is not originally structured as graphs, they use the extracted attribute-specific features (each feature corresponds to one attribute) from a ResNet backbone to obtain the first graph. For the second graph, clusters of regions (pixels) in the input image are considered as the network nodes. The clusters are learned using the share ResNet backbone—with the previous graph). Finally, the outputs of both graph-based branches are averaged. As LSTM also considers the relationship between parts, authors have replaced their proposed GCNs with LSTMs in the model and observed a slight drop in the model's performance. The ablation strides on three pedestrian datasets show that the $F1$ metric performance of a vanilla model improves with a margin of 2.

Reference [123] recognized the clothing style by mixing extracted features from the body parts. They applied a graph-based model with Conditional Random Field (CRF)s to explore the correlation between clothes attributes. Specifically, using the weighted sum of body-part features, they trained an SVM for each of the attributes and used CRFs to learn the relationships between attributes. By training the CRFs with output probability scores from SVM classifiers, the attributes' relationship is explored. Although using CRFs was successful in this work, there are yet some disadvantages: (a) due to extensive computational cost, CRFs is not an appropriate solution when a broad set of attributes are considered, and (b) CRFs cannot capture the spatial relation between attributes [110] (c) models can not simultaneously optimize classifiers and CRFs [110], so it is not useful in an end-to-end model.

2.3.3.2 Math-Oriented Attribute Correlation Consideration

(A) Grammar. In [124], Park et al. addressed the need for an interpretable model that can jointly yield the body-pose information (body joints coordinates) and human semantic attributes. To this end, authors implemented an and-or grammar model, in which they integrated three types of grammars—(1) simple grammars that break down the full-body

into smaller nodes; (2) dependency grammar that indicates which nodes (body parts) are connected to each other and models the geometric articulations; (3) attribute grammar that assigns the attributes to each node. The ablation studies for attribute prediction showed that the performance is better if the best pose estimation for each attribute is used for predicting the corresponding attribute score.

(B) Multiplication. In [125], authors discussed that a plain CNN could not handle human multi-attribute classifications effectively, as for each image, several labels have been entangled. To address this challenge, Han et al. [125] proposed to use a ResNet50 backbone followed by multiple branches to predict the occurrence probability of each attribute. Further, to improve the results, they provided a matrix from ground truth labels to obtain the conditional probability of each label (semantic attribute) given another attribute. The multiplication of this matrix by the previously obtained probability provides the models with a priori knowledge about the correlation of attributes. The ablation study indicated that the baseline (plain ResNet50) on the PETA dataset achieves 85.8 of $F1$ metric, while this number for a simple multi-branch model and full-version model is 86.6 and 87.6, respectively.

In order to mitigate the correlation between the visual appearance and the semantic attributes, Reference [126] uses a fusion attention mechanism and provides a balanced-weight between the image-guided and attribute-guided features. First, attributes are embedded in a latent space with the same dimension of the image features. Next, a nonlinear function is applied to the image features to obtain its feature distribution. Then, the image-guided features are obtained via an element-wise multiplication between the feature distribution of the image and the embedded attribute features. To obtain the attribute-guided features, they embed the attributes to a new latent space; next, the results of the element-wise multiplication between image features and embedded attribute features are considered as the input of a nonlinear function, for which its output provides attribute-guided features. Meanwhile, to consider the class imbalance, authors use the focal loss function to train the model. The ablation study shows that the $F1$ metric performance of the baseline on the PETA dataset is 85.6, which improves to 85.9 when the model is equipped with the above-mentioned idea.

In Reference [127], authors propose a multi-task architecture, in which each attribute corresponds to one separate task. However, to consider the relationship between attributes, both the input image and category information are projected into another space, where the latent factors are disentangled. By applying the element-wise multiplication between the feature representation of the image and its class information, the authors define a discriminant function. When using it, a logistic regression model can learn all the attributes simultaneously. To show the efficiency of the methods, authors evaluate their proposed approach in several attribute datasets of animals, objects, and birds.

(C) Loss Function. Li et al. [128] discussed the attribute relationships and introduced two models to demonstrate the effectiveness of their idea. Considering HAR as a binary classification problem, the authors proposed a plain multi-label CNN that predicts all the

attributes at-once. They also equipped the previous model with a weighted-loss function (cross-entropy), in which each attribute classifier has a specific weight to update the network weights for the next epoch. The experimental results on the PETA dataset with 35 attributes indicated that weighted cross-entropy loss function could improve the accuracy prediction in 28 attributes and increase the mA by 1.3 percent.

2.3.4 Occlusion

In HAR, occlusion is a primary challenge, in which parts of the useful information of the input data may be covered with other subjects/objects [129]. As this situation is likely to occur in real-world scenarios, it is necessary to be handled. In the context of person re-id, Reference [130] claims that inferring the occluded body parts could improve the results, and in the HAR context, Reference [131] suggests that using sequences of pedestrian images somehow alleviates the occlusion problem.

Considering the low-resolution images and partial occlusion of the pedestrian's body, Reference [132] proposed to manipulate the dataset with occurring frequent partial occlusions and degraded the resolution of the data. Then, the authors trained a model to rebuild the images with high resolution and do not suffer from occlusion. This way, the reconstruction model will help to manipulate the original dataset before training a classification model. As rebuild is performed with a GAN, the generated images are different from the original annotated dataset and somehow lost part of the annotations, which degrade the overall performance of the system compared to when one uses the original dataset for training. However, the ablation study in this paper shows that if two identical classification networks are separately trained on corrupted and generated data, the performance of the model that learns from the reconstructed data is better with a high margin.

To tackle the problem of occlusion, Reference [133] proposes to use a sequence of frames for recognizing human attributes. First, they extract the frame-level spatial features using a shared ResNet-50 backbone feature extractor [134]. The extracted features are then processed in two separate paths, one of them learns the body pose and motion, and the other branch learns the semantic attributes. Finally, each attribute's classifier uses an attention module that generates an attention vector showing the importance of each frame for attribute recognition.

To address the challenge of partial occlusion, References [129; 131] adopted video datasets for attributes recognition as often occlusions are a temporary situation. Reference [129] divided each video clip to several pieces and extracted a random frame from each piece to create a new video clip with a few frame length. The final recognition confidence of each attribute is obtained by aggregating the recognition probability on the selected frames.

2.3.5 Classes Imbalance

The existence of large differences between the number of samples for each attribute (class) is known as data class imbalance. Generally, in multi-class classification problems, the

ideal scenario would be to use the same amount of data for each class, in order to preserve the learning importance of all the classes at the same level. However, the classes in HAR datasets are naturally imbalanced since the number of samples of some attributes (e.g., wearing skirts) are lower than others (e.g., wearing jeans). Large class imbalance causes over-fitting in classes with limited data, while classes with large number of samples need more training epochs to converge. To address this challenge, some methods attempt to balance the number of samples in each class as a pre-processing step [135–137], which are called *hard solutions*. Hard solutions are classified into three groups—(1) up-sampling the minority classes, (2) down-sampling the large classes, and (3) generating new samples. On the other hand, *soft solutions* are interested in handling the data class imbalance by introducing new training methods [138] or novel loss functions, in which the importance of each class is weighted based on the frequencies of the data [139–141]. Furthermore, the combination of both solutions has been the subject of some studies [142].

2.3.5.1 *Hard* Solutions

The earlier hard solutions are focused either on interpolation between the samples [135; 143], or clustering the dataset and oversampling by cluster-based methods [144]. The primary way of up-sampling in deep learning is to augment the existing samples –as discussed in Section 2.3.2. However, excessive up-sampling may lead to over-fitting when the classes are highly imbalanced. Therefore, some works down-sample the majority classes [145]. Random down-sampling may be an easy choice, but Reference [146] proposes to use the boundaries among the classes to remove redundant samples. However, loss of information is an inevitable part of down-sampling, as some samples are removed, which may carry useful information.

To address these problems, Fukui et al. [28] designed a multi-task CNN, in which classes (attributes) with fewer samples are given more importance in the learning phase. The batch of samples in conventional learning methods are selected randomly; therefore, the rare examples are less likely to be in the mini-batch. Meanwhile, data augmentation cannot be sufficient for balancing the dataset as ordinary data augmentation techniques generate new samples regardless of their rarity. Therefore, Fukui et al. [28] defines a rarity rate for each sample in the dataset and perform the augmentation for rare samples. Later, from the created mini-batches, those with appropriate sample balance are selected for training the model. The experimental results on a dataset with four attributes show a slight improvement in the average recognition rate, though the superiority is not consistent for all the attributes.

2.3.5.2 Soft Solutions

As previously mentioned, soft solutions focus on boosting the learning methods' performance, rather than merely increasing/decreasing the number of samples. Designing loss functions is a popular approach for guiding the model to take full advantage of the minority samples. For instance, Reference [126] proposes the combination of focal

loss [147] and cross-entropy loss functions to introduce a focal cross-entropy loss function (see Section 2.3.3.2 for the analytical review over [126]).
Considering the success of curriculum learning [148] in other fields of studies, in Reference [138], the author addressed the challenge of imbalance-distributed data in HAR by batch-based adjustment of data sampling strategy and loss weights. It was argued that providing balanced distribution from a highly imbalanced dataset (using sampling strategies) for the whole learning process may cause the model to disregard the samples with most variations (i.e., classes with majority samples) and only emphasizes on the minority class. Moreover, the weighted terms in loss functions play an essential role in the learning process. Therefore, both the classification loss (often cross-entropy) and metric learning loss (which aims to learn feature embedding for distinguishing between samples) should be handled based on their importance. To consider these aspects, authors defined two schedules, one for adjusting the sampling strategy by re-ordering the data from imbalanced to balanced and easy to hard; and the other curriculum schedule handles the loss importance between classification and distance metric learning. The ablation study in this work showed that the sampling scheduler could increase the results of a baseline model form 81.17 to 86.58, and adding loss scheduler to it could improve the results to 89.05.
To handle the class imbalance problem, Reference [149] modifies the focal loss function [147] and apply it for an attention-based model to focus on the hard samples. The main idea is to add a scaling factor to the binary cross-entropy loss function to down-weight the effect of easy samples with high confidence. Therefore, the hard misclassified samples of each attribute (class) add larger values to the loss function and become more critical. Considering the usual weakness of attention mechanism that does not consider the location of an attribute, the authors modified the attention masks in multiple levels of the model using attribute confidence weighting. Their ablation studies on the WIDER dataset [75] with ResNet-101 backbone feature extractor [134] showed the plain model achieves mA 83.7 and applying the weighted focal loss function improve the results to 84.4 while adding the multi-scale attention increased it to 85.9.

2.3.5.3 Hybrid Solutions

Hybrid approaches use the combination of the above-mentioned techniques. Performing data augmentation over the minority classes and applying a weighted loss function or a curriculum learning strategy are examples of hybrid solutions for handling the class data imbalance. In Reference [142], the authors discuss that learning from an unbalanced dataset leads to biased classification, with higher classification accuracy over the majority classes and lower performance over the minority classes. To address this issue, Chawla et al. [142] proposed an algorithm that focuses on difficult samples (misclassified). To implement this strategy, the authors took advantage of Reference [143], which generates new synthetic instances in each training iteration from the minority classes. Consequently, the weights for the minority samples (false negatives) are increased, which improves the model's performance.

2.3.6 Part-Based And Attribute Correlation-Based Methods

"Whether considering a group of attributes together improve the results of an attribute recognition model or not?" is the question that Reference [150] tries to answer by addressing the correlation between attributes using a CRF strategy. Concerning the calculated probability distribution over each attribute, all the Maximum A Posterioris (MAP) are estimated, and then, the model searches for the most probable mixture in the input image. To also consider the location of each attribute, authors extract the part patches based on the bounding box around the full-body, as in fashion datasets pose variations are not significant. A comparison between several simple baselines shows that the CRF-based method ($0.516F1$ score) works slightly better than a localization-based CNN ($0.512F1$ score) on the Chictopia dataset [151], while a global-based CNN $F1$ performance is 0.464.

2.4 Datasets

As opposed to other surveys, instead of merely enumerating the datasets, in this manuscript, we discuss the advantages and drawbacks of each dataset, with emphasis on data collection methods/software. Finally, we discuss the intrinsically imbalanced nature of HAR datasets and other challenges that arise when gathering data.

2.4.1 PAR datasets

- PETA dataset. PETA [152] dataset combines 19,000 pedestrian images gathered from 10 publicly available datasets; therefore the images present large variations in terms of scene, lighting conditions and image resolution. The resolution of the images varies from 17 $\times$ 39 to 169 $\times$ 365 pixels. The dataset provides rich annotations: the images are manually labeled with 61 binary and 4 multi-class attributes. The binary attributes include information about demographics (gender: *Male*, age: *Age16–30, Age31–45,Age46–60, AgeAbove61*), appearance (*long hair*), clothing (*T-shirt, Trousers* etc.) and accessories (*Sunglasses, Hat, Backpack* etc.). The multi-class attributes are related to (eleven basic) color(s) for the upper-body and lower-body clothing, shoe-wear, and hair of the subject. When gathering the dataset, the authors tried to balance the binary attributes; in their convention, a binary class is considered balanced if the maximal and minimal class ratio is less than 20:1. In the final version of the dataset, more than half of the binary attributes (31 attributes) have a balanced distribution.

- RAP dataset. Currently, there are two versions of the RAP dataset. The first version, RAP-v1 v1 [153] was collected from a surveillance camera in shopping malls over a period of three months; next, 17 hours of video footage were manually selected for attribute annotation. In total, the dataset comprises 41,585 annotated human silhouettes. The 72 attributes labeled in this dataset include demographic information (*gender* and *age*), accessories (*backpack, single shoulder*

bag, handbag, plastic bag, paper bag etc.), human appearance (*hair style, hair color, body shape*) and clothing information (*clothes style, clothes color, footware style, footware color* etc.). In addition, the dataset provides annotations about occlusions, viewpoints and body-parts information.

The second version of the RAP dataset [108] is intended as a unifying benchmark for both person retrieval and person attribute recognition in real-world surveillance scenarios. The dataset was captured indoor, in a shopping mall and contains 84,928 images (2589 person identities) from 25 different scenes. High-resolution cameras (1280 $\times$ 720) were used to gather the dataset, and the resolution of human silhouettes varies from 33 $\times$ 81 to 415 $\times$ 583 pixels. The attributes annotated are the same as in RAP v2 (72 attributes, and occlusion, viewpoint, and body-parts information).

- Duke Multi-Target, Multi-Camera (DukeMTMC) dataset. The DukeMTMC dataset [154] was collected in Duke's university campus and contains more than 14 h of video sequences gathered from 8 cameras, positioned such that they capture crowded scenes. The main purpose of this dataset was person re-identification and multi-camera tracking; however, a subset of this dataset was annotated with human attributes. The annotations were provided at the identity level, and they included 23 attributes, regarding the gender (male, female), accessories: wearing hat (yes, no), carrying a backpack (yes, no), carrying a handbag (yes, no), carrying other types of the bag (yes, no), and clothing style: shoe type (boots, other shoes), the color of shoes (dark, bright), length of upper-body clothing (long, short), 8 colors of upper-body clothing (black, white, red, purple, gray, blue, green, brown) and 7 colors of lower-body clothing (black, white, red, gray, blue, green, brown). Due to violation of civil and human rights, as well as privacy issues, since June 2019, Duke University has terminated the DukeMTMC dataset page.

- PA-100K dataset. The PA-100k dataset [88] was developed with the intention to surpass the existing HAR datasets both in quantity and in diversity; the dataset contains more than 100,000 images captured in 598 different scenarios. The dataset was captured by outdoor surveillance cameras; therefore, the images provide large variance in image resolution, lighting conditions, and environment. The dataset is annotated with 26 attributes, including demographic (age, gender), accessories (handbag, phone) and clothing information.

- Market-1501 dataset. Market-1501 attribute [24; 155] dataset is a version of the Market-1501 dataset augmented with the annotation of 27 attributes. Market-1501 was initially intended for cross camera person re-identification, and it was collected outdoor in front of a supermarket using 6 cameras (5 high-resolution cameras and one low resolution). The attributes are provided at the identity level, and in total, there are 1501 annotated identities. In total, the dataset has 32,668 bounding boxes for these 1501 identities. The attributes annotated in *Market-1501 attribute* include demographic information (gender and age), information about accessories

(*wearing hat, carrying backpack, carrying bag, carrying handbag*), appearance (*hair length*) and clothing type and color (*sleeve length, length of lower-body clothing, type of lower-body clothing, 8 color of upper-body clothing, 9 color of lower-body clothing*).

- Pedestrian Detection, Tracking, Re-Identification and Search (P-DESTRE) Dataset. Over the recent years, as their cost has diminished considerably, UAVs applications extended rapidly in various surveillance scenarios. As a response, several UAVs datasets have been collected and made publicly available to the scientific community. Most of them are intended for human detection [156; 157], action recognition [158] or re-identification [159]. To the best of our knowledge, the P-DESTRE [160] dataset is the first benchmark that addresses the problem of HAR from aerial images.

  P-DESTRE dataset [160] was collected in the campuses of two Universities from India and Portugal, using DJI-Phantom-4 drones controlled by human operators. The dataset provides annotations both for person re-identification, as well as for attribute recognition. The identities are consistent across multiple days. The annotations for the attributes include demographic information: *gender, ethnicity* and *age*, appearance information: *height, body volume, hair color, hairstyle, beard, moustache*; accessories information: *glasses, head accessories, body accessories*; *clothing* information and *action* information. In total, the dataset contains over 14 million person bounding boxes, belonging to 261 known identities.

2.4.2 FAR datasets

- Pedestrian Attribute Recognition in Sequences (PARSe27k) dataset. PARSe27k dataset [161] contains over 27,000 pedestrian images, annotated with 10 attributes. The images were captured by a moving camera across a city environment; every 15th video frame was fed to the Deformable Part Model (DPM) pedestrian detector [78] and the resulting bounding boxes were annotated with the 10 attributes based on binary or multinomial propositions. As opposed to other datasets, the authors also included an N/A state (i.e., the labeler cannot decide on that attribute). The attributes from this dataset include gender information (3 categories—*male, female, N/A*), accessories (*Bag on Left Shoulder, Bag on Right Shoulder Bag in Left Hand, Bag in Right Hand, Backpack*; each with three possible states: *yes, no, N/A*), orientation (with *4 + N/A* or *8 + N/A* discretizations) and action attributes: *posture* (*standing, walking, sitting* and *N/A*) and *isPushing* (*yes, no, N/A*). As the images were initially processed by a pedestrian detector, the images of this dataset consist of a fixed-size bounding region of interest, and thus are strongly aligned and contain only a subset of possible human poses.

- Caltech Roadside Pedestrians (CRP) dataset. CRP [162] dataset was captured in real world conditions, from a moving vehicle. The position (bounding-box) of each pedestrian, together with 14 body joints are annotated in each video frame. The CRP dataset comprises 4222 video tracks, with 27,454 pedestrian bounding boxes.

The following attributes are annotated for each pedestrian—age (5 categories: *child*, *teen*, *young adult*, *middle aged* and *senior*), gender (2 categories—*female* and *male*), weight (3 categories: *Under*, *Healthy* and *Over*), and clothing style (4 categories—*casual*, *light athletic*, *workout* and *dressy*). The original, un-cropped videos together with the annotations are publicly available.

- Describing People dataset. Describing People dataset [68] comprises 8035 images from the H3D [163] and the PASCAL Visual Object Classes (PASCAL-VOC) 2010 [164] datasets. The images from this database are aligned, in the sense that for each person, the image is cropped (by leaving some margin) and then scaled so that the distance between the hips and the shoulders is 200 pixels. The dataset features 9 binary (True/False) attributes, as follows: gender (*is male*), appearance (*long hair*), accessories (*glasses*) and several clothing attributes (*has hat, has t-shirt, has shorts, has jeans, long sleeves, long pants*). The dataset was annotated on Amazon Mechanical Turk by five independent labelers; the authors considered a valid label if at least four of the five annotators agreed on its value.

- Human ATtributes (HAT) dataset. HAT [66; 78] contains 9344 images gathered from Flickr website; for this purpose, the authors used more than 320 manually specified queries to retrieve images related to people and then, employed an off-the-shelf person detector to crop the humans in the images. The false positives were manually removed. Next, the images were labeled with 27 binary attributes; these attributes incorporate information about the gender (Female), age (Small baby, Small kid, Teen aged, Young (college), Middle Aged, Elderly), clothing (Wearing tank top, Wearing tee shirt, Wearing casual jacket, Formal men suit, Female long skirt, Female short skirt, Wearing short shorts, Low cut top, Female in swim suit, Female wedding dress, Bermuda/beach shorts), pose (Frontal pose, Side pose, Turned Back), Action (Standing Straight, Sitting, Running/Walking, Crouching/bent, Arms bent/crossed) and occlusions (Upper body). The images have high variations both in image size and in the subject's position.

- WIDER dataset. The WIDER Attribute dataset [75] comprises a subset of 13,789 images selected from the WIDER database [165], by discarding the images full of non-human objects and the images in which the human attributes are indistinguishable; the human bounding boxes from these images are annotated with 14 attributes. The images contain multiple humans under different and complex variations. For each image, the authors selected a maximum of 20 bounding boxes (based on their resolution), so in total, there are more than 57,524 annotated individuals. The attributes follow a ternary taxonomy: positive, negative and unspecified, and include information about age (*Male*), clothing (*Tshirt, longSleeve, Formal, Shorts, Jeans, Long Pants, Skirt), accessories (Sunglasses, Hat, Face Mask, Logo*), appearance (*Long Hair*). In addition, each image is annotated into one of 30 event classes (meeting, picnic, parade, etc.), thus allowing to correlate the human attributes with the context they were perceived in.

- Clothing Attributes Dataset (CAD) dataset. CAD [123] uses images gathered from the website Sartorialist (`https://www.thesartorialist.com/`) and Flikcr website. The authors downloaded several images, mostly of pedestrians, and applied an upper-body detector to detect humans; they ended up with 1856 images. Next, the ground truth was established by labelers from Amazon Mechanical Turk. Each image was annotated by 6 independent individuals, and a label was accepted as ground truth if it has at least 5 agreements. The dataset is annotated with the gender of the wearer, information about the accessories (*Wearing scarf, Collar presence, Placket presence*) and with several attributes regarding the clothing appearance (*clothing pattern, major color, clothing category, neckline shape* etc.).
- Attributed Pedestrians in Surveillance (APiS) dataset. The APiS dataset [166] gathers images from four different sources: the Karlsruhe Institute of Technology and Toyota Technological Institute (KITTI) database [167], Center for Biological and Computational Learning (CBCL) Street Scenes [168] (`http://cbcl.mit.edu/software-datasets/streetscenes/`), INRIA database [48] and some video sequences collected by the authors at a train station; in total APiS comprises 3661 images. The human bounding boxes are detected using an off-the-shelf pedestrian detector, and the results are manually processed by the authors: the false positives and the low-resolution images (smaller than 90 pixels in height and 35 pixels in width) are discarded. Finally, all the images of the dataset are normalized in the sense that the cropped pedestrian images are scaled to 128×48 pixels. These cropped images are annotated with 11 ternary attributes (positive, negative, and ambiguous) and 2 multi-class attributes. These annotations include demographic (gender) and appearance attributes (*long hair*), as well as information about accessories (*back bag, S-S (Single Shoulder) bag, hand carrying*) and clothing (*shirt, T-shirt, long pants, M-S (Medium and Short) pants, long jeans, skirt, upper-body clothing color, lower-body clothing color*). The multi-class attributes are the two attributes related to the clothing color. The annotation process is performed manually and divided into two stages: annotation stage (the independent labeling of each attribute) and validation stage (which exploits the relationship between the attributes to check the annotation; also, in this stage, the controversial attributes are marked as ambiguous).

2.4.3 Fashion Datasets

- DeepFashion Dataset. The DeepFashion dataset [91] was gathered from shopping websites, as well as image search engines (blogs, forums, user-generated content). In the first stage, the authors downloaded 1,320,078 images from shopping websites and 1,273,150 images from Google images. After a data cleaning process, in which duplicate, out-of-scope, and low-quality images were removed, 800,000 clothing images were finally selected to be included in the DeepFashion dataset. The images are annotated solely with clothing information; these annotations are divided into categories (50 labels: dress, blouse, etc.) and attributes (1000 labels: adjectives

Table 2.1: Pedestrian attributes datasets. Symbol † indicates that the dataset has been permanently suspended regarding privacy issues. Titles **1** to **5** stand for demographic, accessories, appearance, clothing and color, respectively and M is the abbreviation for million.

Dataset Type	Dataset	#images	1	2	3	4	5	Setup
Pedestrian	PETA [152]	19,000	✓	✓	✓	✓	✓	10 databases
	RAP v1 [153]	41,585	✓	✓	✓	✓	✓	indoor static camera
	RAP v2 [108]	84,928	✓	✓	✓	✓	✓	indoor static camera
	DukeMTMC †	34,183	✓	✓	✗	✓	✓	outdoor static camera
	PA-100K [88]	100,000	✓	✓	✗	✓	✗	outdoor surveillance
	Market-1501 [24]	1501	✓	✓	✓	✓	✓	outdoor
	P-DESTRE [160]	14M	✓	✓	✓	✓	✗	UAV
Full body	PARSe27k [161]	27,000	✓	✓	✗	✗	✗	outdoor moving camera
	CRP [162]	27,454	✓	✓	✗	✗	✗	moving vehicle
	APiS [166]	3661	✓	✓	✓	✓	✓	3 databases
	HAT [66]	9344	✓	✓	✗	✓	✗	Flickr
	CAD [123]	1856	✓	✓	✗	✓	✓	website crawling
	DP [68]	8035	✓	✓	✗	✓	✗	2 databases
	WIDER [75]	13,789	✓	✓	✓	✓	✗	website crawling
Synthetic	CTD [169]	880	✗	✗	✗	✓	✓	generated data
	CLOTH3D [170]	2.1M	✗	✗	✗	✓	✓	generated data

describing the categories). The categories were annotated by expert labelers, while for the attributes, due to their huge number, the authors resorted to meta-data annotation (provided by Google search engine or by the shopping website). In addition, a set of clothing landmarks, as well as their visibility, are provided for each image.

DeepFashion is split into several benchmarks for different purposes: category and attribute prediction (classification of the categories and the attributes), in-shop clothes retrieval (determine if two images belong to the same clothing item), consumer-to-shop clothes retrieval (matching consumer images to their shop counterparts) and fashion landmark detection.

2.4.4 Synthetic Datasets

Virtual reality systems and synthetic image generation have become prevalent in the last few years, and their results are more and more realistic and of high resolution. Therefore, we also discuss some data sources comprising computer-generated images. It is a well-known fact that the performance of deep learning methods is highly dependent on the amount and distribution of data they were trained on, and synthetic datasets could theoretically be used as an inexhaustible source of diverse and balanced data. In theory, any combination of attributes in any amount could be synthetically generated.

- DeepFashion—Fashion Image Synthesis. The authors of DeepFashion [91] introduce FashionGAN, an adversarial network for generating clothing images on a wearer [171]. FashionGAN is organized into two stages: on a first level, the network generates a semantic segmentation map modeling the wearer's pose. In the second level, a generative model renders an image with precise regions and textures conditioned on this map. In this context, the DeepFashion dataset was extended with 78,979 images (taken for the In-shop Clothes Benchmark), associated with several caption sentences and a segmentation map.

- Clothing Tightness Dataset (CTD). CTD [169] comprises 880 3D human models, under various poses, both static and dynamic, "dressed" with 228 different outfits. The garments in the dataset are grouped under various categories, such as "T/long shirt, short/long/down coat, hooded jacket, pants, and skirt/dress, ranging from ultra-tight to puffy". CTD was gathered in the context of a deep learning method that maps a 3D human scan into a hybrid geometry image. This synthetic dataset has important implications in virtual try-on systems, soft biometrics, and body pose evaluation. The main drawbacks of this dataset are that it cannot capture exaggerated human postures of low 3D human scans.

- Cloth-3D Dataset. Cloth-3D [170] comprises thousands of 3D sequences of animated human silhouettes, "dressed" with different garments. The dataset features a large variation on the garment shape, fabric, size, and tightness, as well as human pose. The main applications of this dataset listed by the authors include—"human pose and action recognition in-depth images, garment motion analysis, filling missing vertices of scanned bodies with additional metadata (e.g., garment segments), support designers and animators tasks, or estimating 3D garment from RGB images".

2.5 Evaluation Metrics

This section reviews the most common metrics used in the evaluation of HAR methods. Considering that HAR is a multi-class classification problem, Accuracy (Acc), Precision ($Prec$), Recall (Rec), and $F1$ score are the most common metrics for measuring the performance of these methods. In general, these metrics can be calculated at two different levels: label-level and sample-level.
The evaluation at label-level considers each attribute independently. As an example, if the gender and height attributes are considered with the labels (male, female) and (short, medium, high), respectively, the label-level evaluation will measure the performance of each attribute-label combination. The metric adopted in most papers for label-level evaluation is the mean accuracy (mA):

$$mA = \frac{1}{2N}\sum_{i=1}^{N}(\frac{TP_i}{P_i} + \frac{TN_i}{N_i}), \quad (2.1)$$

where i refers to each of the N attributes. mA determines the average accuracy between the positive and negative examples of each attribute.
In the sample-level evaluation, the performance is measured for each attribute disregarding the number of labels that it comprises. $Prec$, Rec, Acc, and $F1$ score for the i^{th} attribute are thus given by:

$$Prec_i = \frac{TP_i}{P_i}, \quad Rec_i = \frac{TP_i}{N_i}, \quad Acc_i = \frac{TP_i + TN_i}{P_i + N_i}, \quad F_i = \frac{2 * Prec * Rec}{Prec + Rec}. \quad (2.2)$$

The use of these metrics is very common for providing a comparative analysis of the different attributes. The overall system performance can be either measured by the mean Acc_i over all the attributes or using mA. However, these metrics can diverge significantly, when attributes are highly unbalanced. mA is preferred when authors deliberately want to evaluate the effect of data unbalancing.

2.6 Discussion

2.6.1 Discussion Over HAR Datasets

In recent years, HAR has received much interest from the scientific community, with a relatively large number of datasets developed for this purpose; this is also demonstrated by the number of citations. We performed a query for each HAR related database on the Google Scholar (`scholar.google.com`) search engine, and extracted its corresponding number of citations; the results are graphically presented in Figure 2.3. In the past decade, more than 15 databases related to this research field have been published, and most of them received hundreds of citations.
In Table 2.1, we chose to taxonomize the attributes semantically into demographic attributes (gender, age, ethnicity), appearance attributes (related to the appearance of the subject, such as hairstyle, hair color, weight, etc.), accessory information (which indicate the presence of a certain accessory, such as a hat, handbag, backpack etc.) and clothing attributes (which describe the garments worn by the subjects). In total, we have described 17 datasets, the majority containing over ten thousand images. These datasets can be seen as a continuous effort made by researchers to provide large amounts of varied data required by the latest deep learning neural networks.

1. Attributes definition. The first issues that should be addressed when developing a new dataset for HAR are: (1) *which attributes should be annotated?* and (2) *how many and which classes are required to describe an attribute properly?*. Obviously, both these questions depend on the application domain of the HAR system. Generally, the ultimate goal on a HAR, regardless of the application domain, would be to accurately describe an image in terms of human-understandable semantic labels, for example, “a five-year-old boy, dressed in blue jeans, with a yellow T-shirt carrying a striped backpack”. As for the second question, the answer is straightforward for some attributes, such as gender, but it becomes more

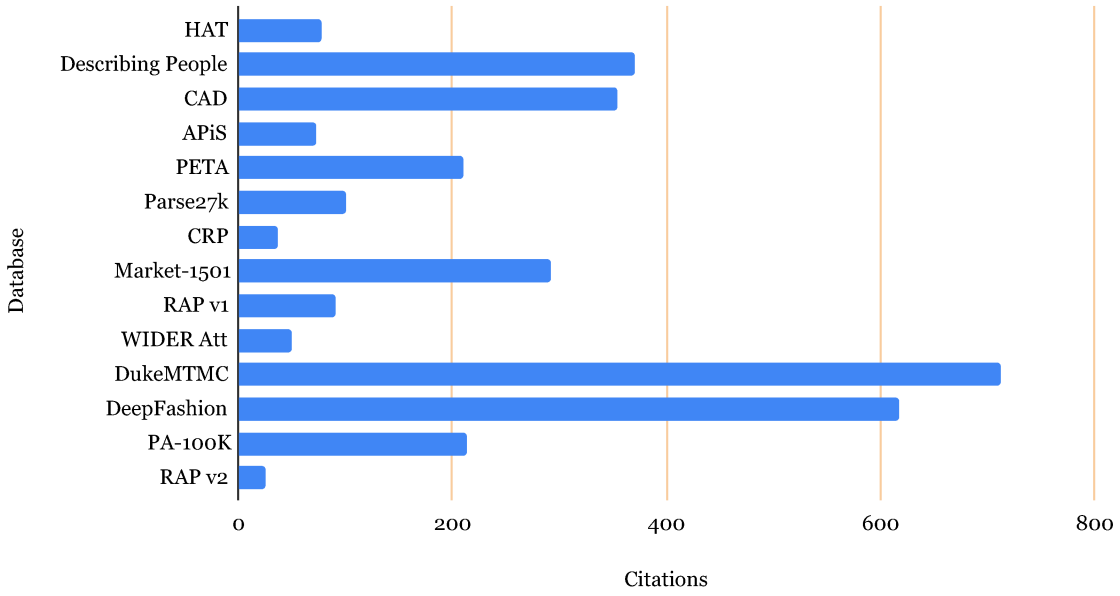


Figure 2.3: Number of citations to HAR datasets. The datasets are arranged in an increasing order by their publication date. The "oldest" dataset being HAT, published in 2009, while the latest is RAP v2, published in 2018.

complex and subjective for other attributes, such as age or clothing information. Let's take for example, the age label; different datasets provided different classes for this information: PETA distinguishes between *AgeLess15, Age16-30, Age31-45, Age46-60, AgeAbove61*, while the CRP dataset adopted a different age classification scheme: *child*, *teen*, *young adult*, *middle aged* and *senior*. Now, if a HAR analyzer is integrated into a surveillance system in a crowded environment, such as Disneyland, and this system should be used to locate a missing child, the age labels from the PETA dataset are not detailed enough, as the "lowest" age class is AgeLess15. Secondly, these differences between the different taxonomies make it difficult to assess the performance of a newly developed algorithm across different datasets.

2. Unbalanced data. An important issue in any dataset is related to unbalanced data. Although some datasets were developed by explicitly striking for balanced classes, some classes are not that frequent (especially those related to clothing information), and fully balanced datasets are not a trivial problem. The problem of imbalance also affects the demographic attributes. In all HAR datasets, the class of young children is poorly represented. To illustrate the problem of unbalanced classes, we selected two of the most prominent HAR related datatsets which are labeled with age information: CRP and PETA. In Figure 2.4, for each of these two datasets, we plot a pie charts to show age distribution of the labeled images.

   Furthermore, as datasets are usually gathered in a single region (city, country, continent), the data tends to be unbalanced in terms of ethnicity. This is an important issue as some studies [172] proved the existence of *the other race effect* –-the tendency to more easily recognize faces from the same ethnicity-– for machine

learning classifier.

3. Data context. Strongly linked to the problem of data unbalance is the context or environment in which the frames were captured. The environment has a great influence on the distribution of the clothing and demographic (age, gender) attributes. In [75] the authors noticed "strong correlations between image event and the frequent human attributes in it". This is quite logical, as one would expect to encounter more casual outfits in a picnic or sporting event, while at ceremonies (wedding, graduation proms), people tend to be more elegant and dressed-up. The same is valid for the demographic attributes: if the frames are captured in the backyard of a kindergarten, one would that most of the subjects to be children. Ideally, a HAR dataset should provide images captured from multiple and variate scenes. Some datasets explicitly annotated the context in which the data was captured [75], while others address this issue by merging images from various datasets [152]. From another point of view, this leads our discussion to how the images from the datasets are presented. Generally speaking, the dataset provides the images either aligned (all the images have the same size and cropped around the human silhouette with a predefined margin; for example, [68]), or make the full video frame/image available and specify the bounding box of each human in the image. We consider that the latter approach is preferable, as it also incorporates context information and allows researches to decide how to handle the input data.

4. Binary attributes. Another question in database annotation is what happens when the attribute to annotate is indistinguishable due to low resolution and degraded images, occlusions, or other ambiguities. The majority of datasets tend to ignore this problem and classify the presence of an attribute or provide a multi-class attribute scheme. However, in a real-world setup, we cannot afford this luxury, as the case of indistinguishable attributes might occur quite frequently. Therefore, some datasets [161; 166] formulate the attribute classification task with N + 1 classes (+1 for the N/A label). This approach is preferable, as it allows taking both views over the data: depending on the application context, one could simply ignore the N/A attributes or, make the classification problem more interesting, integrate the N/A value into the classification framework.

5. Camera configuration. Another aspect that should be taken into account when discussing HAR datasets is the camera setup used to capture the images or video sequences. We can distinguish between fixed-camera and moving-camera setups; obviously, this choice again depends on the application domain into which the HAR system will be integrated. For automotive applications or robotics, one should opt for a moving camera, as the camera movement might influence the visual properties of the human silhouettes. An example of a moving-camera dataset is PARSe27k dataset [161]. For surveillance applications, a static camera setup will suffice. In another way, we could distinguish between indoor or outdoor camera setups; for example, RAP dataset [153] uses an indoor camera, while PARSe27k dataset [161]

comprises outdoor video sequences. Indoor captured datasets, such as [153], although captured in real-world scenarios, do not pose that many challenges as outdoor captured datasets, where the weather and lighting conditions are more volatile. Finally, the last aspect regarding the camera setup is related to the presence of a photographer. If the images are captured by a (professional) photographer some bias is introduced, as a human decides how and when to capture the images, such that it will enhance the appearance of the subject. Some databases, such as CAD [123] or HAT [66; 78] use images downloaded from public websites. However, in these images, the persons are aware of being photographed and perhaps even prepared for this (posing for the image, dressed up nicely for a photo session, etc.). Therefore, even if some datasets contain *in-the-wild* images gathered for a different system, they might still contain important differences from *real-world* images in which the subject is unaware of being photographed, the image is captured automatically, without any human intervention, are the subjects are dressed normally and performing natural dynamic movements.

6. Pose and occlusion labeling. Another nice to have feature for a HAR dataset is the annotation of pose and occlusions. Some databases already provide this information [66; 78; 108; 153]. Amongst other things, these extra labels prove useful in the evaluation of HAR systems, as they allow researchers to diagnose the errors of HAR and examine the influence of various factors.

7. Data partitioning strategies. When dealing with HAR, the datasets partitioning scheme (into the train, validation, and test splits) should be carefully engineered. A common pitfall is to split the frames into the train and validation splits randomly, regardless of the person's identity. This can lead to an unfair assignment of a subject into one of these splits, and inducing bias in the evaluation process. This is even more important, as the current state-of-the-art methods generally rely on deep neural network architectures, which have a black-box behavior in nature, and it is not so straightforward to determine which image features lead to the final classification result.

   Solutions to this problem include extracting each individual (along with its tracklets) from the video sequence or providing the annotations at the identity level. Then, each person could be randomly assigned to one of the dataset splits.

8. Synthetic data. Recently, significant advances have been made in the field of computer graphics and synthetic data generation. For example, in the field of drone surveillance, generated data [173] has proven its efficiency in training accurate machine vision systems. In this section, we have presented some computer-generated datasets which contain human attribute annotations. We consider that synthetically generated data is worth taking into consideration, as theoretically, it can be considered an inexhaustible source of data, which could be able to generate subjects with various attributes, under different poses, in diverse scenarios. However, state-of-the-art generative models rely on deep learning, which is known

to be ”hungry” for data, so data is needed to build a realistic generative model. Therefore, this solution might prove to be just a vicious circle.

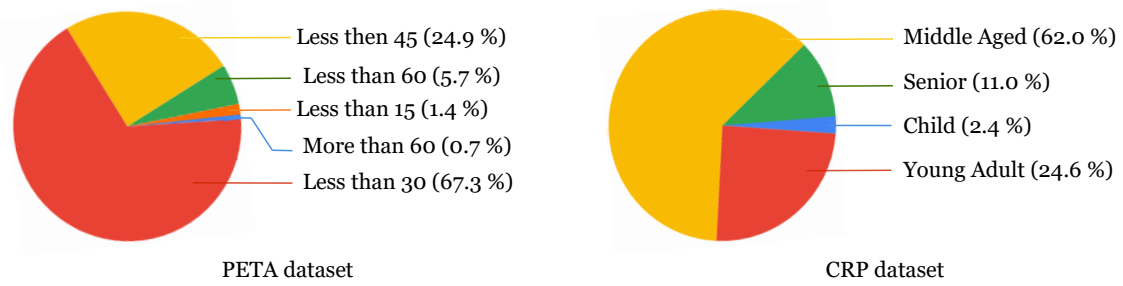
9. Privacy issues. In the past, as traditional video surveillance systems were simple and involved only human monitoring, privacy was not a major concern; however, these days, the pervasiveness of systems equipped with cutting-edge technologies in public places (e.g., shopping malls, private and public buildings, bus and train stations) have aroused new privacy and security concerns. For instance, the Office of the Privacy Commissioner of Canada (OPC) is an organization that helps people report their privacy concerns and enforces the enterprises to manage people's personal data in their business activities based on restricting standards (`https://www.priv.gc.ca/en/report-a-concern/`).

   When gathering a dataset with real-world images, we deal with privacy and human rights violations. Ideally, HAR datasets should contained images captured by real-world surveillance cameras, with the subjects are unaware of being filmed, such that their behavior is as natural as possible. From an ethical perspective, humans should consent before their images are annotated and publicly distributed. However, this is not feasible for all scenarios. For example, the *Brainwash* [174] dataset was gathered inside a private cafe for the purpose of head detection, and comprised 11,917 images. Although this benchmark is not very popular, it is seen in the lists of the popular datasets for commercial and military applications, as it has captured the regular customers without their awareness. The DukeMTMC [152] dataset targets the task of multi-person re-identification from full-body images taken by several cameras. This dataset was collected in a university campus in an outdoor environment and contains over 2 million frames of 2000 students captured by 8 cameras at 1080p. *MS-Celeb-1M* [175] is another large dataset of 10 million faces collected from the Internet.

   However, despite the success of these datasets (if we evaluate success by the number of citations and database downloads), the authors decided to shout-down the datasets due to human rights and privacy violation issues.

   According to Pew Research Center Privacy Panel Survey conducted from 27 January to 16 February 2015, among 461 adults, more than 90 percent agreed that two factors are critical for surveillance systems: (1) *who* can access to their information? (2) *what* information is collected about them? Moreover, it is notable that they consent to share confidential information with someone they trust (93%); however, it is important not to be monitored without permission (88%).

   As people's faces contain sensitive information that could be captured in the wild, authorities have published some standards (`https://gdpr-info.eu/`) to enforce enterprises respect the privacy of their costumers.



PETA dataset CRP dataset

Figure 2.4: Frequency distribution of the labels describing the 'Age' class in the PETA [152] (on the **left**) and CRP [162] (on the **right**) databases.

2.6.2 Critical Discussion and Performance Comparison

As mentioned, the main objective of the localization method is to extract distinct fine-grained features, by careful analyses of different pieces of the input data and aggregating them. Although the extracted localized features create a detailed feature representation of the image, dividing the image to several pieces has several drawbacks:

- the expressiveness of the data is lost (e.g., when processing a jacket only by several parts, some global features that encode the jacket's shape and structure are ignored).
- as the person detector cannot always provide aligned and accurate bounding boxes, rigid partitioning methods are prone to error in body-part captioning, mainly when the input data includes a wide background. Therefore, methods based on stride/grid patching of the image are not robust to misalignment errors in the person bounding boxes, leading to degradation in prediction performance.
- different from gender and age, most human attributes (such as glasses, hat, scarf, shoes, etc.) belong to small regions of the image; therefore, analyzing other parts of the image may add irrelevant features to the final feature representation of the image.
- some attributes are view-dependent and highly changeable due to human body-pose, and ignoring them reduces the model performance; for example, glasses recognition in the side-view images is more laborious than front-view, while it may be impossible in back-view images. Therefore, in some localization methods (e.g., pose-let based techniques), regardless of this fact, features of different parts may be aggregated to perform a prediction on an unavailable attribute.
- some localization methods rely on the body-parsing techniques [176] or body-part detection methods [177] to extract local features. Not only requires training such part detectors rich annotations of data but also errors in body-parsing and body-part detection methods directly affect the performance of the HAR model.

There are several possibilities to address some of these issues, which mostly attempt to guide the learning process using additional information. For instance, as discussed in Section 2.3, some works use novel model structures [72] to capture the relationships

and correlations between the parts of the image, while others try to use prior body-pose coordinates [63] (or develop a view-detector in the main structure [61]) to learn the view-specific attributes. Some methods develop attention modules to find the relevant body parts, while some approaches extract various pose-lets [163] of the image by slicing-window detectors. Using the semantic attributes as a constraint for extracting the relevant regions is another solution to look for localized attributes [100]. Moreover, developing accurate body-part detectors, body-parsing algorithms and introducing datasets with part annotations are some strategies that can help the localization methods.

Limited data is the other main challenge in HAR. The primary solutions for solving the problem of limited data are synthesizing artificial samples or augmenting the original data. One of the popular approaches for increasing the size of the dataset is to use generative models (i.e., Generative Adversarial Network (GAN) [178], Variational Auto-Encoders (VAE) [179], or a combination of both [180]). These models are powerful tools for producing new samples, but are not widely used for extending human full-body datasets for three reasons:

- in opposition to the breakthrough in face generative models [181], full-body generative models are still in early stages and their performance is still unsatisfactory,
- the generated data is unlabelled, while HAR is yet far from the stage to be implemented based on unlabeled data. It worth mentioning that, automatic annotations is an active research area in object detection [182].
- not only takes learning high-quality generative models for human full-body too much time, but it also requires a large amount of high-resolution learning data, which is yet not available.

Therefore, researchers [71; 82; 103; 129; 183–185] mostly either perform transfer learning to capture the useful knowledge of large datasets or resort to the simple yet useful label-persevering augmentation techniques from basic data augmentation (flipping, shifting, scaling, cropping, resizing, rotating, shearing, zooming, etc.) to more sophisticated methods such as random erasing [186] and foreground augmentation [187].

Due to the lack of sufficient data in some data classes (attributes), augmentation methods should be implemented carefully. Suppose that we have very few data from some classes (e.g., ‘age 0–11’, ‘short winter jacket’) and much more data from other classes (e.g., ‘age 25–35’, ‘t-shirt’). A blind data augmentation process would exacerbate the data class imbalance and increase the over-fitting problem in minority classes. Furthermore, some basic augmentations are not label persevering. For example, for a dataset annotated for body weight, scratching the images of a thin person may be interpreted as a medium or fat person, while it may be acceptable for color-based labels. Therefore, visualizing a set of augmented data and careful studying of the annotation data are highly suggested before performing augmentation.

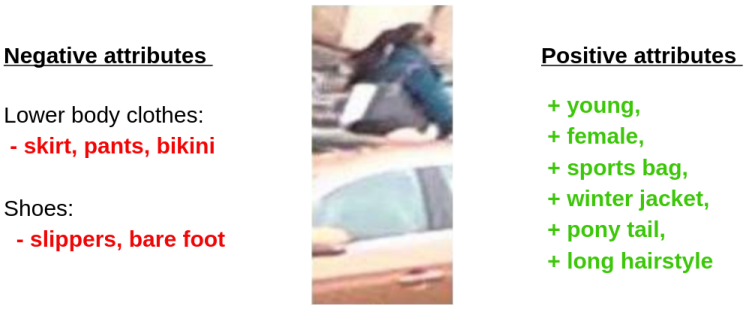


Figure 2.5: As human, not only we describe the available attributes in occluded images but also we can predict the covered attributes in a negative strategy based on the attribute relations.

Using *proper* pre-trained models (transfer learning) not only reduces the training time but also increases the system's performance. To have an effective transfer learning from task A to task B we should consider the following conditions [188–190],:

1. There should be some relationships between the data of task A and task B. For example, applying pre-trained weights of the ImageNet dataset [50] on HAR task is beneficial as both domains are dealing with RGB images of objects, including human data, while transferring the knowledge of medical imagery (e.g., CT/MRI) are not useful and may only impose some heavy parameters to the model.
2. The data in task A is much more than the data in task B as transferring the knowledge of other small datasets cannot guarantee performance improvements.

Generally, there are two useful strategies for applying transfer learning to HAR problems, in which we suggest to discard the classification layers (i.e., fully connected layers that are on top of the model), and use the pre-trained model as a feature extractor (backbone). Then,

- we can freeze the backbone model and adding several classification layers on top of the model for fine-tuning.
- we can add the proper classification layers on top of the model and train all the model layers in several steps: (1) freeze the backbone model and fine-tune the last layers, (2) considering a lower learning rate, we unfreeze high-level feature extractor layers and fine-tune the model, (3) we unfreeze mid-level and low-level layers in other steps and train them with a lower learning rate, as these features are normally common between most tasks with the same data types.

Considering attribute correlations can boost the performance of HAR models. Some works (e.g., multi-task and RNN based models) attempt to extract the semantic relationship between the attributes from the visual data. However, lack of enough data and also the type of annotations in HAR datasets (the region of attributes are not annotated) lead to the poor performance of these models in capturing the correlation of attributes. Even in GCN and CRF based models that are known to be effective in

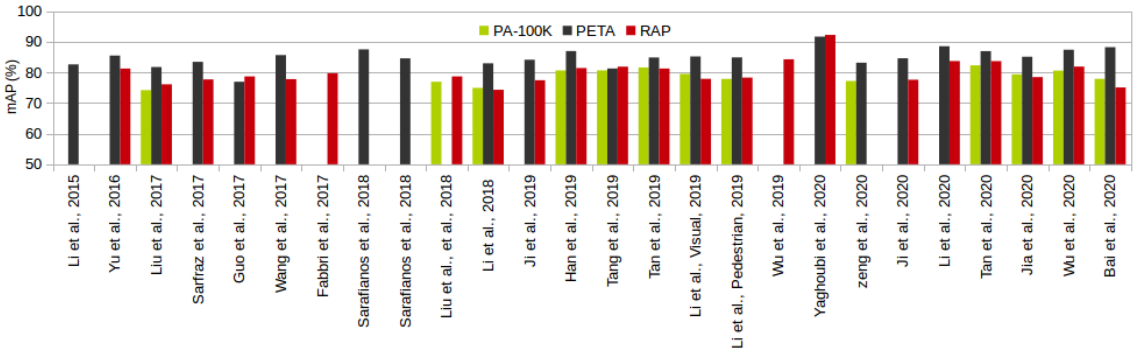


Figure 2.6: State-of-the-art mAP results on three well-known PAR datasets.

capturing the relationship between defined nodes, yet these are no explicit mathematical expressions about several aspects: what is the optimal way to convert the visual data to some nodes, and what is the optimum number of nodes? When fusing the visual features with correlation information, how much should we give importance to the correlation information? How would be the performance of a model if it learns the correlation between attributes from external text data (particularly from the aggregation of several HAR datasets)?

Although occlusion is a primary challenge, yet few studies address it in HAR data. As surveyed in Section 2.3.4, several works have proposed to use video data, which is a rational idea only if more data are available. However, in still images, we know that even if most parts of the body (and even face of a person) is occluded, as human, we still are able to easily decide about many attributes of the person (see Figure 2.5). Another idea that could be considered in HAR data is labeling a/an (occluded) person with certain labels that are not correct. For example, suppose that the input data is a person with legging, even if the model is not certain about the correct lower body clothes, yet it could yield some labels indicating that the lower body cloth is not certainly a dress/skirt. Later, this information could be beneficial when considering the correlation between attributes. Moreover, introducing a HAR dataset composed of different degrees of occlusion could trigger more domain-specific studies. In the context of person re-id, Reference [191] provided an occluded dataset based on DukeMTMC dataset, which is not publicly available anymore (`https://megapixels.cc/duke_mtmc/`).

Last but not least, studies based on class imbalance challenge attempt to promote the importance of the minority classes and (or) decrease the importance of majority classes, by proposing hard and (or) soft solutions. As mentioned earlier, providing more data blindly (collecting or augmenting the existing data) cannot guarantee better performance and may increase the gap between the number of samples in data classes. Therefore, we should provide a trade-off between the down-sampling and up-sampling strategies, while using the proper loss functions to learn more from minority samples. As discussed in Section 2.3.5, these ideas have been developed to some extent; however, other challenges in HAR have been neglected in the final proposal.

Table 2.2 shows the performance of the HAR approaches over the last decade and indicates a consistent improvement of methods over time. In 2016, the $F1$ performance evaluation

of [74] on the RAP and PETA datasets was 66.12 and 84.90, respectively, while these numbers were improved to 79.98 and 86.87 in the year 2019 [92] and to 82.10 and 88.30 in year 2020. Furthermore, according to Table 2.2, it is clear that challenges of attributes localization and attributes correlation have attracted the most attention over the recent years, which indicates that extracting distinctive fine-grained features from relevant locations of the given input images is the most important aspect of HAR models.

Despite the early works that analyzed the human full-body data in different locations and situations, recent works have focused on attribute recognition from surveillance data, which arouses some privacy issues.

Appearing comprehensive evaluation metrics is another noticeable change over the last decade. Due to the intrinsic, large class imbalance in the HAR datasets, mA cannot provide a comprehensive performance evaluation over different methods. Suppose that in a binary classification situation, if 99% of the samples belong to persons with glasses and 1% of samples belong to persons without glasses, the model can recognize all the test samples as persons with glasses and still has 99% of accuracy in recognition. Therefore, for a fair performance comparison with the state of the arts, it is necessary to consider metrics such as $Prec$, Rec, Acc, and $F1$ – which are discussed in Section 2.5.

Table 2.2 also shows that the RAP, PETA, and PA-100K datasets have attracted the most attention in the context of attribute recognition –which excludes person re-id. In Figure 2.6 we illustrate the state-of-the-art results obtained on these datasets for mAP metric. As seen, the PETA dataset sounds easier than other datasets, despite the smaller size and lower quality data compared with the RAP dataset.

Table 2.2: Performance comparison of HAR approaches over the last decade for different benchmarks.

Ref., Year, Cat.	Taxonomy	Dataset	mA	Acc.	prec.	rec.	F1
[66], 2011, FAR	Pose-Let	HAT [66]	53.80	-	-	-	-
[68], 2011, FAR	Pose-Let	[68]	82.90	-	-	-	-
[123], 2012, FAR and CAA	Attribute relation	[123]	-	84.90	-	-	-
		D.Fashion [91]	35.37 (top-5)	-	-	-	-
[79], 2013, FAR	Body-Part	HAT [66]	69.88	-	-	-	-
[69], 2013, FAR	Pose-Let	HAT [66]	59.30	-	-	-	-
[70], 2013, FAR	Pose-Let	HAT [66]	59.70	-	-	-	-
[77], 2015, FAR	Body-Part	DP [68]	83.60	-	-	-	-
[128], 2015, PAR	Loss function	PETA [152]	82.6	-	-	-	-
[77], 2015, FAR	Body-Part	DP [68]	83.60	-	-	-	-
[150], 2015, CAA	Attribute location and relation	Dress [150]	-	84.30	65.20	70.80	67.80

Table 2.2: *Cont.*

Ref., Year, Cat.	Taxonomy	Dataset	mA	Acc.	prec.	rec.	F1
[75], 2016, FAR	Pose-Let	WIDER [75]	92.20	-	-	-	-
[74], 2016, PAR	Pose-Let	RAP [108]	81.25	50.30	57.17	78.39	66.12
		PETA [152]	85.50	76.98	84.07	85.78	84.90
[91], 2016, CAA	Limited data	D.Fashion [91]	54.61 (top-5)	-	-	-	-
[86], 2017, FAR	Attention	WIDER [75]	82.90	-	-	-	-
		Berkeley [68]	92.20	-	-	-	-
[88], 2017, PAR	Attention	RAP [108]	76.12	65.39	77.33	78.79	78.05
		PETA [152]	81.77	76.13	84.92	83.24	84.07
		PA-100K [88]	74.21	72.19	82.97	82.09	82.53
[124], 2018, FAR	Grammar	DP [68]	89.40	-	-	-	-
[61], 2018, PAR and FAR	Pose Estimation	RAP [108]	77.70	67.35	79.51	79.67	79.59
		PETA [152]	83.45	77.73	86.18	84.81	85.49
		WIDER [75]	82.40	-	-	-	-
[86], 2017, PAR	Attention	RAP [108]	78.68	68.00	80.36	79.82	80.09
		PA-100K [88]	76.96	75.55	86.99	83.17	85.04
[109], 2017, PAR	RNN	RAP [108]	77.81	-	78.11	78.98	78.58
		PETA [152]	85.67	-	86.03	85.34	85.42
[104], 2017, PAR	Loss Function - Augmentation	PETA [152]	-	75.43	-	70.83	-
[132], 2017, PAR	Occlusion	RAP [108]	79.73	83.97	76.96	78.72	77.83
[105], 2017, CAA	Transfer Learning	[105]	64.35	-	64.97	75.66	-
[111], 2017, PAR	Multitask	Market [24]	-	88.49	-	-	-
		Duke [152]	-	87.53	-	-	-
[192], 2017, CAA	Multiplication	D.Fashion [91]	30.40 (top-5)	-	-	-	-
[89], 2018, CAA	Attention	D.Fashion [91]	60.95 (top-5)	-	-	-	-
[113], 2018, PAR	Soft-Multitask	SoBiR [114]	74.20	-	-	-	-
		VIPeR [193]	84.00	-	-	-	-
[149], 2018, PAR and FAR	Soft solution	PETA [152]	87.54	-	-	-	-
		WIDER [75]	86.40	-	-	-	-
		PETA [152]	84.59	78.56	86.79	86.12	86.46
[63], 2018, PAR	Pose Estimation	PETA [152]	82.97	78.08	86.86	84.68	85.76
		RAP [108]	74.31	64.57	78.86	75.90	77.35
		PA-100K [88]	74.95	73.08	84.36	82.24	83.29

Table 2.2: *Cont.*

Ref., Year, Cat.	Taxonomy	Dataset	mA	Acc.	prec.	rec.	F1
[97], 2018, PAR	Attribute location	RAP [108]	78.68	68.00	80.36	79.82	80.09
		PA-100K [88]	76.96	75.55	86.99	83.17	85.04
[117], 2018, PAR	RNN	RAP [108]	-	77.81	78.11	78.98	78.58
		PETA [152]	-	85.67	86.03	85.34	85.42
[126], 2019, PAR	Soft solution	RAP [108]	77.44	65.75	79.01	77.45	78.03
		PETA [152]	84.13	78.62	85.73	86.07	85.88
[125], 2019, PAR	Multiplication	PETA [152]	86.97	79.95	87.58	87.73	87.65
		RAP [108]	81.42	68.37	81.04	80.27	80.65
		PA-100K [88]	80.65	78.30	89.49	84.36	86.85
[118], 2019, PAR	RNN	RAP [108]	-	77.81	78.11	78.98	78.58
		PETA [152]	-	86.67	86.03	85.34	85.42
[133], 2019, PAR	Occlusion	Duke [152]	-	89.31	-	-	73.24
		MARS [194]	-	87.01	-	-	72.04
[100], 2019, PAR	Attribute Location	RAP [108]	81.87	68.17	74.71	86.48	80.16
		PETA [152]	86.30	79.52	85.65	88.09	86.85
		PA-100K [88]	80.68	77.08	84.21	88.84	86.46
[92], 2019, PAR	Attention	PA-100K [88]	81.61	78.89	86.83	87.73	87.27
		RAP [108]	81.25	67.91	78.56	81.45	79.98
		PETA [152]	84.88	79.46	87.42	86.33	86.87
		Market [24]	87.88	-	-	-	-
		Duke [152]	87.88	-	-	-	-
[110], 2019, PAR	GCN	RAP [108]	77.91	70.04	82.05	80.64	81.34
		PETA [152]	85.21	81.82	88.43	88.42	88.42
		PA-100K [88]	79.52	80.58	89.40	87.15	88.26
[107], 2019, PAR	GCN	RAP [108]	78.30	69.79	82.13	80.35	81.23
		PETA [152]	84.90	80.95	88.37	87.47	87.91
		PA-100K [88]	77.87	78.49	88.42	86.08	87.24
[94], 2019, PAR and FAR	Attention	RAP [108]	84.28	59.84	66.50	84.13	74.28
		WIDER [75]	88.00	-	-	-	-
[96], 2020, PAR	Attention	RAP [108]	92.23	-	-	-	-
		PETA [152]	91.70	-	-	-	-
[54], 2020, PAR	Multi-task	PA-100K [88]	77.20	78.09	88.46	84.86	86.62
		PETA [152]	83.17	78.78	87.49	85.35	86.41
[195], 2020, PAR	RNN	RAP [108]	77.62	67.17	79.72	78.44	79.07
		PETA [152]	84.62	78.80	85.67	86.42	86.04

Table 2.2: *Cont.*

Ref., Year, Cat.	Taxonomy	Dataset	mA	Acc.	prec.	rec.	F1
[58], 2020, PAR	RNN and attention	RAP [108]	83.72	-	81.85	79.96	80.89
		PETA [152]	88.56	-	88.32	89.62	88.97
[120], 2020, PAR	GCN	RAP [108]	83.69	69.15	79.31	82.40	80.82
		PETA [152]	86.96	80.38	87.81	87.09	87.45
		PA-100K [88]	82.31	79.47	87.45	87.77	87.61
[196], 2020, PAR	Baseline	RAP [108]	78.48	67.17	82.84	76.25	78.94
		PETA [152]	85.11	79.14	86.99	86.33	86.09
		PA-100K [88]	79.38	78.56	89.41	84.78	86.25
[59], 2020, PAR	RNN and attention	PA-100K [88]	80.60	-	88.70	84.90	86.80
		RAP [108]	81.90	-	82.40	81.90	82.10
		PETA [152]	87.40	-	89.20	87.50	88.30
		Market [24]	88.50	-	-	-	-
		Duke [152]	88.80	-	-	-	-
[197], 2020, PAR	Hard solution	PA-100K [88]	77.89	79.71	90.26	85.37	87.75
		RAP [108]	75.09	66.90	84.27	79.16	76.46
		PETA [152]	88.24	79.14	88.79	84.70	86.70
[198], 2020, CAA	—	Fashionista [199]	-	88.91	47.72	44.92	39.42
[200], 2020, PAR	Math-oriented	Market [24]	92.90	78.01	87.41	85.65	86.52
		Duke [152]	91.77	76.68	86.37	84.40	85.37

We observe that the performance of the state of the arts is yet far from the reliable range to be used in forensic affairs and enterprises, and it requires more attention in both introducing novel datasets and proposing robust methods.
Among PAR, FAR, and CAA fields of study, most of the papers have focused on the PAR task. The reason is not apparent, but at least we know that (1) PAR data are often collected from CCTV and surveillance cameras, and analyzing such data is critical for forensic and security objectives, (2) person re-id is a hot topic that mainly works with the same data type and could be highly influenced by powerful PAR methods.

2.7 Conclusions

This survey reviewed the most relevant works published in the context of HAR problem over the last decade. Contrary to the previous published reviews, which provided a methodological categorization of the literature, in this survey we privileged a challenge-based taxonomy, that is, methods were organized based on the challenges of the HAR problem that they were devised to address. According to this type of organization, readers can easily understand the most suitable strategies for addressing each of the

typical challenges of HAR and simultaneously learn which strategies perform better. In addition, we comprehensively reviewed the available HAR datasets, outlining the relative advantages and drawbacks of each one with respect to others, as well as the data collection strategy used. Also, the intrinsically imbalanced nature of the HAR datasets is discussed, as well the most relevant challenges that typically arise when gathering data for this problem.

Bibliography

[1] G. Tripathi, K. Singh, and D. K. Vishwakarma, "Convolutional neural networks for crowd behaviour analysis: a survey," *The Visual Computer*, vol. 35, no. 5, pp. 753–776, 2019. 13

[2] B. Munjal, S. Amin, F. Tombari, and F. Galasso, "Query-guided end-to-end person search," in *Proc. CVPR*, 2019, pp. 811–820. 13

[3] Y. Yan, Q. Zhang, B. Ni, W. Zhang, M. Xu, and X. Yang, "Learning context graph for person search," in *Proc. CVPR*, June 2019, pp. 2158–2167. 13

[4] C. V. Priscilla and S. A. Sheila, "Pedestrian detection-a survey," in *International Conference on Information, Communication and Computing Technology*. Springer, 2019, pp. 349–358. 13

[5] N. Narayan, N. Sankaran, S. Setlur, and V. Govindaraju, "Learning deep features for online person tracking using non-overlapping cameras: A survey," *IMAGE VISION COMPUT*, vol. 89, pp. 222–235, 2019. 13

[6] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *arXiv preprint arXiv:2001.04193*, 2020. 13

[7] J. Xiang, T. Dong, R. Pan, and W. Gao, "Clothing attribute recognition based on rcnn framework using l-softmax loss," *IEEE Access*, vol. 8, pp. 48 299–48 313, 2020. 13

[8] B. H. Guo, M. S. Nixon, and J. N. Carter, "A joint density based rank-score fusion for soft biometric recognition at a distance," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 3457–3462. 13

[9] N. Thom and E. M. Hand, "Facial attribute recognition: A survey," *Computer Vision: A Reference Guide*, pp. 1–13, 2020. 13

[10] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proc. IEEE ICCV*, 2017, pp. 2961–2969. 14, 18

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. CVPR*, 2016, pp. 779–788. 14

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Proc. ECCV*. Springer, 2016, pp. 21–37. 14

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105. 15, 25

[14] E. Bekele and W. Lawson, "The deeper, the better: Analysis of person attributes recognition," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 2019, pp. 1–8. 15

[15] X. Zheng, Y. Guo, H. Huang, Y. Li, and R. He, "A survey of deep facial attribute analysis," *International Journal of Computer Vision*, pp. 1–33, 2020. 15

[16] X. Wang, S. Zheng, R. Yang, B. Luo, and J. Tang, "Pedestrian attribute recognition: A survey," *arXiv preprint arXiv:1901.07474*, 2019. 15, 16

[17] I. Masi, Y. Wu, T. Hassner, and P. Natarajan, "Deep face recognition: A survey," in *2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)*. IEEE, 2018, pp. 471–478. 16

[18] G. B. Huang, H. Lee, and E. Learned-Miller, "Learning hierarchical representations for face verification with convolutional deep belief networks," in *Proc. CVPR*. IEEE, 2012, pp. 2518–2525.

[19] Y. Sun, D. Liang, X. Wang, and X. Tang, "Deepid3: Face recognition with very deep neural networks," *arXiv preprint arXiv:1502.00873*, 2015. 16

[20] M. De Marsico, A. Petrosino, and S. Ricciardi, "Iris recognition through machine learning techniques: A survey," *Pattern Recognit. Lett.*, vol. 82, pp. 106–115, 2016. 16

[21] F. Battistone and A. Petrosino, "Tglstm: A time based graph deep learning approach to gait recognition," *Pattern Recognit. Lett.*, vol. 126, pp. 132–138, 2019. 16

[22] P. Terrier, "Gait recognition via deep learning of the center-of-pressure trajectory," *Applied Sciences*, vol. 10, no. 3, p. 774, 2020. 16

[23] R. Layne, T. M. Hospedales, S. Gong, and Q. Mary, "Person re-identification by attributes." in *Bmvc*, vol. 2, 2012, p. 8. 16

[24] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang, "Improving person re-identification by attribute and identity learning," *Pattern Recognition*, vol. 95, pp. 151–161, 2019. 16, 25, 29, 37, 41, 53, 54, 55

[25] J. Liu, B. Kuipers, and S. Savarese, "Recognizing human actions by attributes," in *CVPR 2011*. IEEE, 2011, pp. 3337–3344. 16

[26] J. Shao, K. Kang, C. Change Loy, and X. Wang, "Deeply learned attributes for crowded scene understanding," in *Proc. CVPR*, 2015, pp. 4657–4666. 16

[27] N. Tsiamis, L. Efthymiou, and K. P. Tsagarakis, "A comparative analysis of the legislation evolution for drone use in oecd countries," *Drones*, vol. 3, no. 4, p. 75, 2019. 17

[28] H. Fukui, T. Yamashita, Y. Yamauchi, H. Fujiyoshi, and H. Murase, "Robust pedestrian attribute recognition for an unbalanced dataset using mini-batch training with rarity rate," in *Proc. IEEE IV*. IEEE, 2016, pp. 322–327. 17, 34

[29] S. Prabhakar, S. Pankanti, and A. K. Jain, "Biometric recognition: Security and privacy concerns," *IEEE security & privacy*, vol. 1, no. 2, pp. 33–42, 2003. 17

[30] Y. Xiu, J. Li, H. Wang, Y. Fang, and C. Lu, "Pose flow: Efficient online pose tracking," *arXiv preprint arXiv:1802.00977*, 2018. 18

[31] J. Neves, F. Narducci, S. Barra, and H. Proença, "Biometric recognition in surveillance scenarios: a survey," *Artif. Intell. Rev.*, vol. 46, no. 4, pp. 515–541, 2016. 18

[32] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936. 18

[33] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995. 18

[34] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE transactions on systems, man, and cybernetics*, vol. 21, no. 3, pp. 660–674, 1991. 18

[35] B. Kamiński, M. Jakubczyk, and P. Szufel, "A framework for sensitivity analysis of decision trees," *Central European journal of operations research*, vol. 26, no. 1, pp. 135–159, 2018. 18

[36] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943. 18

[37] G. P. Zhang, "Neural networks for classification: a survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 30, no. 4, pp. 451–462, 2000. 18

[38] T. Georgiou, Y. Liu, W. Chen, and M. Lew, "A survey of traditional and deep learning-based feature descriptors for high dimensional data in computer vision," *International Journal of Multimedia Information Retrieval*, pp. 1–36, 2019. 18

[39] R. Satta, "Appearance descriptors for person re-identification: a comprehensive review," *arXiv preprint arXiv:1307.5748*, 2013. 18

[40] M. Piccardi and E. D. Cheng, "Track matching over disjoint camera views based on an incremental major color spectrum histogram," in *IEEE Conference on Advanced Video and Signal Based Surveillance, 2005.* IEEE, 2005, pp. 147–152. 18

[41] S.-Y. Chien, W.-K. Chan, D.-C. Cherng, and J.-Y. Chang, "Human object tracking algorithm with human color structure descriptor for video surveillance systems," in *2006 IEEE International Conference on Multimedia and Expo*. IEEE, 2006, pp. 2097–2100. 18

[42] K.-M. Wong, L.-M. Po, and K.-W. Cheung, "Dominant color structure descriptor for image retrieval," in *2007 IEEE International Conference on Image Processing*, vol. 6. IEEE, 2007, pp. VI–365. 18

[43] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004. 18

[44] J. M. Iqbal, J. Lavanya, and S. Arun, "Abnormal human activity recognition using scale invariant feature transform," *International Journal of Current Engineering and Technology*, vol. 5, no. 6, pp. 3748–3751, 2015. 18

[45] P.-E. Forssén, "Maximally stable colour regions for recognition and matching," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8. 18

[46] S. Basovnik, L. Mach, A. Mikulik, and D. Obdrzalek, "Detecting scene elements using maximally stable colour regions," in *Proceedings of the EUROBOT Conference*, 2009. 18

[47] N. He, J. Cao, and L. Song, "Scale space histogram of oriented gradients for human detection," in *2008 International Symposium on Information Science and Engineering*, vol. 2. IEEE, 2008, pp. 167–170. 18

[48] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. CVPR*, vol. 1. IEEE, 2005, pp. 886–893. 40

[49] H. Beiping and Z. Wen, "Fast human detection using motion detection and histogram of oriented gradients." *JCP*, vol. 6, no. 8, pp. 1597–1604, 2011. 18

[50] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009. 19, 50

[51] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016. 19

[52] P. Alirezazadeh, E. Yaghoubi, E. Assunção, J. C. Neves, and H. Proença, "Pose switch-based convolutional neural network for clothing analysis in visual surveillance environment," in *Proc. BIOSIG*. Darmstadt, Germany: IEEE, 2019, pp. 1–5.

[53] E. Yaghoubi, P. Alirezazadeh, E. Assunção, J. C. Neves, and H. Proençaã, "Region-based cnns for pedestrian gender recognition in visual surveillance environments," in *Proc. BIOSIG*. IEEE, 2019, pp. 1–5. 19

[54] H. Zeng, H. Ai, Z. Zhuang, and L. Chen, "Multi-task learning via co-attentive sharing for pedestrian attribute recognition," *arXiv preprint arXiv:2004.03164*, 2020. 19, 29, 54

[55] Y. Lu, A. Kumar, S. Zhai, Y. Cheng, T. Javidi, and R. Feris, "Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification," in *Proc. CVPR*, 2017, pp. 5334–5343. 19, 28

[56] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27–48, 2016. 19

[57] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artif. Intell. Rev.*, pp. 1–62, 2020. 19

[58] Y. Li, H. Xu, M. Bian, and J. Xiao, "Attention based cnn-convlstm for pedestrian attribute recognition," *Sensors*, vol. 20, no. 3, p. 811, 2020. 20, 55

[59] J. Wu, H. Liu, J. Jiang, M. Qi, B. Ren, X. Li, and Y. Wang, "Person attribute recognition by sequence contextual relation learning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020. 20, 55

[60] J. Krause, T. Gebru, J. Deng, L.-J. Li, and L. Fei-Fei, "Learning features and parts for fine-grained recognition," in *2014 22nd International Conference on Pattern Recognition*. IEEE, 2014, pp. 26–33. 21

[61] M. S. Sarfraz, A. Schumann, Y. Wang, and R. Stiefelhagen, "Deep view-sensitive pedestrian attribute inference in an end-to-end model," *arXiv preprint arXiv:1707.06089*, 2017. 22, 49, 53

[62] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. CVPR*, 2015, pp. 1–9. 22, 23

[63] D. Li, X. Chen, Z. Zhang, and K. Huang, "Pose guided deep model for pedestrian attribute recognition in surveillance scenarios," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2018, pp. 1–6. 22, 49, 53

[64] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proc. CVPR*, 2018, pp. 79–88. 22

[65] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc NIPS - Volume 2*, ser. NIPS'15. Cambridge, MA, USA: MIT Press, 2015, p. 2017–2025. 22, 27

[66] G. Sharma and F. Jurie, "Learning discriminative spatial representation for image classification," in *BMVC 2011 - British Machine Vision Conference*, J. Hoey, S. J. McKenna, and E. Trucco, Eds. Dundee, United Kingdom: BMVA Press, 2011, pp. 1–11. [Online]. Available: https://hal.inria.fr/hal-00722820 22, 39, 41, 46, 52

[67] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 2169–2178. 22

[68] L. Bourdev, S. Maji, and J. Malik, "Describing people: A poselet-based approach to attribute classification," in *Proc. ICCV*. IEEE, 2011, pp. 1543–1550. 22, 39, 41, 45, 52, 53

[69] J. Joo, S. Wang, and S.-C. Zhu, "Human attribute recognition by rich appearance dictionary," in *Proc. ICCV*, 2013, pp. 721–728. 23, 52

[70] G. Sharma, F. Jurie, and C. Schmid, "Expanded parts model for human attribute and action recognition in still images," in *Proc. CVPR*, June 2013. 23, 52

[71] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev, "Panda: Pose aligned networks for deep attribute modeling," in *Proc. CVPR*, 2014, pp. 1637–1644. 23, 49

[72] J. Zhu, S. Liao, D. Yi, Z. Lei, and S. Z. Li, "Multi-label cnn based pedestrian attribute learning for soft biometrics," in *2015 International Conference on Biometrics (ICB)*. IEEE, 2015, pp. 535–540. 23, 48

[73] J. Zhu, S. Liao, Z. Lei, and S. Z. Li, "Multi-label convolutional neural network based pedestrian attribute classification," *IMAGE VISION COMPUT*, vol. 58, pp. 224–229, 2017. 23

[74] K. Yu, B. Leng, Z. Zhang, D. Li, and K. Huang, "Weakly-supervised learning of mid-level features for pedestrian attribute recognition and localization," *arXiv preprint arXiv:1611.05603*, 2016. 23, 52, 53

[75] Y. Li, C. Huang, C. C. Loy, and X. Tang, "Human attribute recognition by deep hierarchical contexts," in *Proc. ECCV*. Springer, 2016, pp. 684–700. 23, 35, 39, 41, 45, 53, 54

[76] R. Girshick, "Fast r-cnn," in *Proc. ICCV*, 2015, pp. 1440–1448. 23

[77] G. Gkioxari, R. Girshick, and J. Malik, "Actions and attributes from wholes and parts," in *Proc. ICCV*, 2015, pp. 2470–2478. 24, 52

[78] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE TPAMI*, vol. 32, no. 9, pp. 1627–1645, 2009. 24, 38, 39, 46

[79] N. Zhang, R. Farrell, F. Iandola, and T. Darrell, "Deformable part descriptors for fine-grained recognition and attribute prediction," in *Proc. ICCV*, December 2013. 24, 52

[80] L. Yang, L. Zhu, Y. Wei, S. Liang, and P. Tan, "Attribute recognition from adaptive parts," *arXiv preprint arXiv:1607.01437*, 2016. 24

[81] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *Proc. CVPR*, 2014, pp. 3686–3693. 24

[82] Y. Zhang, X. Gu, J. Tang, K. Cheng, and S. Tan, "Part-based attribute-aware network for person re-identification," *IEEE Access*, vol. 7, pp. 53 585–53 595, 2019. 24, 49

[83] X. Fan, K. Zheng, Y. Lin, and S. Wang, "Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation," in *Proc. CVPR*, 2015, pp. 1347–1355. 25

[84] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free?-weakly-supervised learning with convolutional neural networks," in *Proc. CVPR*, 2015, pp. 685–694. 25

[85] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Advances in neural information processing systems*, 2014, pp. 487–495. 25

[86] H. Guo, X. Fan, and S. Wang, "Human attribute recognition by refining attention heat map," *Pattern Recognit. Lett.*, vol. 94, pp. 38–45, 2017. 25, 53

[87] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. 25

[88] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang, "Hydraplus-net: Attentive deep features for pedestrian analysis," in *Proc IEEE ICCV*, 2017, pp. 350–359. 25, 37, 41, 53, 54, 55

[89] W. Wang, Y. Xu, J. Shen, and S.-C. Zhu, "Attentive fashion grammar network for fashion landmark detection and clothing category classification," in *Proc. CVPR*, June 2018. 25, 53

[90] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *ECCV*. Springer, 2016, pp. 483–499. 25

[91] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proc IEEE CVPR*, 2016, pp. 1096–1104. 26, 28, 40, 42, 52, 53

[92] Z. Tan, Y. Yang, J. Wan, H. Wan, G. Guo, and S. Z. Li, "Attention based pedestrian attribute analysis," *IEEE transactions on image processing*, 2019. 26, 52, 54

[93] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. CVPR*, 2017, pp. 2881–2890. 26

[94] M. Wu, D. Huang, Y. Guo, and Y. Wang, "Distraction-aware feature learning for human attribute recognition via coarse-to-fine attention mechanism," *arXiv preprint arXiv:1911.11351*, 2019. 26, 54

[95] F. Zhu, H. Li, W. Ouyang, N. Yu, and X. Wang, "Learning spatial regularization with image-level supervisions for multi-label image classification," in *Proc. CVPR*, 2017, pp. 5513–5522. 26

[96] E. Yaghoubi, D. Borza, J. Neves, A. Kumar, and H. Proença, "An attention-based deep learning model for multiple pedestrian attributes recognition," *IMAGE VISION COMPUT.*, pp. 1–25, 2020. [Online]. Available: https://doi.org/10.1016/j.imavis.2020.103981 26, 28, 54

[97] P. Liu, X. Liu, J. Yan, and J. Shao, "Localization guided learning for pedestrian attribute recognition," *arXiv preprint arXiv:1808.09102*, 2018. 26, 54

[98] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. CVPR*, 2016, pp. 2921–2929. 26

[99] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. ECCV*. Springer, 2014, pp. 391–405. 26

[100] C. Tang, L. Sheng, Z. Zhang, and X. Hu, "Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization," in *Proc. ICCV*, October 2019, pp. 4997–5006. 26, 49, 54

[101] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015. 26

[102] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, 2018, pp. 7132–7141. 27

[103] E. Bekele, W. E. Lawson, Z. Horne, and S. Khemlani, "Implementing a robust explanatory bias in a person re-identification network," in *Proc. CVPRW*, 2018, pp. 2165–2172. 27, 49

[104] E. Bekele, C. Narber, and W. Lawson, "Multi-attribute residual network (maresnet) for soft-biometrics recognition in surveillance scenarios," in *Proc. FG)*. IEEE, 2017, pp. 386–393. 27, 53

[105] Q. Dong, S. Gong, and X. Zhu, "Multi-task curriculum transfer deep learning of clothing attributes," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2017, pp. 520–529. 27, 53

[106] Q. Chen, J. Huang, R. Feris, L. M. Brown, J. Dong, and S. Yan, "Deep domain adaptation for describing people based on fine-grained clothing attributes," in *Proc. CVPR*, 2015, pp. 5315–5324. 28

[107] Q. Li, X. Zhao, R. He, and K. Huang, "Pedestrian attribute recognition by joint visual-semantic reasoning and knowledge distillation," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, 2019, pp. 833–839. 28, 30, 54

[108] D. Li, Z. Zhang, X. Chen, and K. Huang, "A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios," *IEEE T IMAGE PROCESS*, vol. 28, no. 4, pp. 1575–1590, 2018. 28, 37, 41, 46, 53, 54, 55

[109] J. Wang, X. Zhu, S. Gong, and W. Li, "Attribute recognition by joint recurrent learning of context and correlation," in *Proc. ICCV*, 2017, pp. 531–540. 28, 30, 53

[110] Q. Li, X. Zhao, R. He, and K. Huang, "Visual-semantic graph reasoning for pedestrian attribute recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8634–8641. 28, 30, 31, 54

[111] K. He, Z. Wang, Y. Fu, R. Feng, Y.-G. Jiang, and X. Xue, "Adaptively weighted multi-task deep network for person attribute classification," in *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 2017, pp. 1636–1644. 28, 53

[112] N. Sarafianos, T. Giannakopoulos, C. Nikou, and I. A. Kakadiaris, "Curriculum learning for multi-task classification of visual attributes," in *Proc. ICCVW*, 2017, pp. 2608–2615. 29

[113] ——, "Curriculum learning of visual attribute clusters for multi-task classification," *Pattern Recognition*, vol. 80, pp. 94–108, 2018. 29, 53

[114] D. Martinho-Corbishley, M. S. Nixon, and J. N. Carter, "Soft biometric retrieval to describe and identify surveillance images," in *2016 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*. IEEE, 2016, pp. 1–6. 29, 53

[115] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in *Proc. ECCV*, September 2018. 29

[116] H. Liu, J. Wu, J. Jiang, M. Qi, and B. Ren, "Sequence-based person attribute recognition with joint ctc-attention model," *arXiv preprint arXiv:1811.08115*, 2018. 29

[117] X. Zhao, L. Sang, G. Ding, Y. Guo, and X. Jin, "Grouping attribute recognition for pedestrian with joint recurrent learning." in *IJCAI*, 2018, pp. 3177–3183. 29, 54

[118] X. Zhao, L. Sang, G. Ding, J. Han, N. Di, and C. Yan, "Recurrent attention model for pedestrian attribute recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 9275–9282. 30, 54

[119] Z. Ji, W. Zheng, and Y. Pang, "Deep pedestrian attribute recognition based on lstm," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 151–155. 30

[120] Z. Tan, Y. Yang, J. Wan, G. Guo, and S. Z. Li, "Relation-aware pedestrian attribute recognition with graph convolutional networks." in *Proc. AAAI*, 2020, pp. 12 055–12 062. 31, 55

[121] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Advances in neural information processing systems*, 2016, pp. 4898–4906. 31

[122] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016. 31

[123] H. Chen, A. Gallagher, and B. Girod, "Describing clothing by semantic attributes," in *Proc. ECCV*. Springer, 2012, pp. 609–623. 31, 40, 41, 46, 52

[124] S. Park, B. X. Nie, and S. Zhu, "Attribute and-or grammar for joint parsing of human pose, parts and attributes," *IEEE TPAMI*, vol. 40, no. 7, pp. 1555–1569, 2018. 31, 53

[125] K. Han, Y. Wang, H. Shu, C. Liu, C. Xu, and C. Xu, "Attribute aware pooling for pedestrian attribute recognition," *arXiv preprint arXiv:1907.11837*, 2019. 32, 54

[126] Z. Ji, E. He, H. Wang, and A. Yang, "Image-attribute reciprocally guided attention network for pedestrian attribute recognition," *Pattern Recognit. Lett.*, vol. 120, pp. 89–95, 2019. 32, 34, 35, 54

[127] K. Liang, H. Chang, S. Shan, and X. Chen, "A unified multiplicative framework for attribute learning," in *Proc. ICCV*, December 2015. 32

[128] D. Li, X. Chen, and K. Huang, "Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios," in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*. IEEE, 2015, pp. 111–115. 32, 52

[129] Y. Zhao, X. Shen, Z. Jin, H. Lu, and X.-s. Hua, "Attribute-driven feature disentangling and temporal aggregation for video person re-identification," in *Proc. CVPR*, 2019, pp. 4913–4922. 33, 49

[130] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Vrstc: Occlusion-free video person re-identification," in *Proc. CVPR*, June 2019. 33

[131] J. Xu and H. Yang, "Identification of pedestrian attributes based on video sequence," in *2018 IEEE International Conference on Advanced Manufacturing (ICAM)*. IEEE, 2018, pp. 467–470. 33

[132] M. Fabbri, S. Calderara, and R. Cucchiara, "Generative adversarial models for people attribute recognition in surveillance," in *2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS)*. IEEE, 2017, pp. 1–6. 33, 53

[133] Z. Chen, A. Li, and Y. Wang, "A temporal attentive approach for video-based pedestrian attribute recognition," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 2019, pp. 209–220. 33, 54

[134] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778. 33, 35

[135] B.-H. M. Hui Han, Wen-Yuan Wang, "Borderline-smote: a new over-sampling method in imbalanced data sets learning," *International Conference on Intelligent Computing, Springer*, 2015. 34

[136] E. A. G. Haibo He, "Learning from imbalanced data," *IEEE T. KNOWL. DATA EN.*, 2009.

[137] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," *IEEE International Joint Conference on Neural Networks*, 2008. 34

[138] Y. Wang, W. Gan, J. Yang, W. Wu, and J. Yan, "Dynamic curriculum learning for imbalanced data classification," in *Proc. ICCV*, 2019, pp. 5017–5026. 34, 35

[139] Y. Tang, Y.-Q. Zhang, N. V. Chawla, and S. Krasser, "Svms modeling for highly imbalanced classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 2008. 34

[140] Z.-H. Z. . X.-Y. Liu, "Training cost-sensitive neural networks with methods addressing the class imbalance problem," *IEEE T. KNOWL. DATA EN.*, 2005.

[141] B. Z. . J. L. . N. Abe, "Cost-sensitive learning by cost-proportionate example weighting," *Third IEEE International Conference on Data Mining*, 2003. 34

[142] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "Smoteboost: Improving prediction of the minority class in boosting," *European Conference on Principles of Data Mining and Knowledge Discovery(PKDD)*, 2003. 34, 35

[143] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of artificial intelligence research,*, 2002. 34, 35

[144] T. Jo and N. Japkowicz., "Class imbalances versus small disjuncts," *ACM Sigkdd Explorations Newsletter*, 2004. 34

[145] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Systems with Applications*, vol. 73, pp. 220–239, 2017. 34

[146] e. a. GMiroslav Kubat, Stan Matwin, "Addressing the curse of imbalanced training sets: one-sided selection," *ICML*, 1997. 34

[147] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. ICCV*, 2017, pp. 2980–2988. 35

[148] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 41–48. 35

[149] N. Sarafianos, X. Xu, and I. A. Kakadiaris, "Deep imbalanced attribute classification using visual attention aggregation," in *Proc. ECCV*, 2018, pp. 680–697. 35, 53

[150] K. Yamaguchi, T. Okatani, K. Sudo, K. Murasaki, and Y. Taniguchi, "Mix and match: Joint model for clothing and attribute recognition." in *BMVC*, vol. 1, 2015, p. 4. 36, 52

[151] K. Yamaguchi, T. L. Berg, and L. E. Ortiz, "Chic or social: Visual popularity analysis in online fashion networks," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 773–776. 36

[152] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," in *Proceedings of the 22Nd ACM International Conference on Multimedia*, ser. MM '14. New York, NY, USA: ACM, 2014, pp. 789–792. [Online]. Available: http://doi.acm.org/10.1145/2647868.2654966 36, 41, 45, 47, 48, 52, 53, 54, 55

[153] D. Li, Z. Zhang, X. Chen, H. Ling, and K. Huang, "A richly annotated dataset for pedestrian attribute recognition," *arXiv preprint arXiv:1603.07054*, 2016. 36, 41, 45, 46

[154] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. ECCVW*, 2016. 37

[155] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE ICCV*, 2015, pp. 1116–1124. 37

[156] P. Zhu, L. Wen, X. Bian, H. Ling, and Q. Hu, "Vision meets drones: A challenge," *arXiv preprint arXiv:1804.07437*, 2018. 38

[157] M. Barekatain, M. Martí, H.-F. Shih, S. Murray, K. Nakayama, Y. Matsuo, and H. Prendinger, "Okutama-action: An aerial view video dataset for concurrent human action detection," in *Proc. CVPRW*, 2017, pp. 28–35. 38

[158] A. G. Perera, Y. W. Law, and J. Chahl, "Drone-action: An outdoor recorded drone video dataset for action recognition," *Drones*, vol. 3, no. 4, p. 82, 2019. 38

[159] S. Zhang, Q. Zhang, Y. Yang, X. Wei, P. Wang, B. Jiao, and Y. Zhang, "Person re-identification in aerial imagery," *IEEE Trans. Multimed.*, p. 1–1, 2020. [Online]. Available: http://dx.doi.org/10.1109/TMM.2020.2977528 38

[160] S. Aruna Kumar, E. Yaghoubi, A. Das, B. Harish, and H. Proença, "The p-destre: A fully annotated dataset for pedestrian detection, tracking, re-identification and search from aerial devices," *arXiv*, pp. arXiv–2004, 2020. 38, 41

[161] P. Sudowe, H. Spitzer, and B. Leibe, "Person attribute recognition with a jointly-trained holistic cnn model," in *Proc. ICCVW*, 2015, pp. 87–95. 38, 41, 45

[162] D. Hall and P. Perona, "Fine-grained classification of pedestrians in video: Benchmark and state of the art," in *Proc. CVPR*, 2015, pp. 5482–5491. 38, 41, 48

[163] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3d human pose annotations," in *Proc. ICCV*. IEEE, 2009, pp. 1365–1372. 39, 49

[164] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010. 39

[165] Y. Xiong, K. Zhu, D. Lin, and X. Tang, "Recognize complex events from static images by fusing deep channels," in *Proc. CVPR*, 2015, pp. 1600–1609. 39

[166] J. Zhu, S. Liao, Z. Lei, D. Yi, and S. Li, "Pedestrian attribute classification in surveillance: Database and evaluation," in *Proc. ICCVW*, 2013, pp. 331–338. 40, 41, 45

[167] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013. 40

[168] S. M. Bileschi and L. Wolf, "Cbcl streetscenes," Center for Biological and Computational Learning (CBCL) at MIT, Tech. Rep., 2006. [Online]. Available: http://cbcl.mit.edu/software-datasets/streetscenes/ 40

[169] X. Chen, A. Pang, Y. Zhu, Y. Li, X. Luo, G. Zhang, P. Wang, Y. Zhang, S. Li, and J. Yu, "Towards 3d human shape recovery under clothing," *CoRR*, vol. abs/1904.02601, 2019. [Online]. Available: http://arxiv.org/abs/1904.02601 41, 42

[170] H. Bertiche, M. Madadi, and S. Escalera, "Cloth3d: Clothed 3d humans," *arXiv preprint arXiv:1912.02792*, 2019. 41, 42

[171] S. Zhu, S. Fidler, R. Urtasun, D. Lin, and C. C. Loy, "Be your own prada: Fashion synthesis with structural coherence," in *Proc. ICCV*, October 2017. 42

[172] P. J. Phillips, F. Jiang, A. Narvekar, J. Ayyad, and A. J. O'Toole, "An other-race effect for face recognition algorithms," *ACM Trans. Appl. Percept.*, vol. 8, no. 2, pp. 1–11, 2011. 44

[173] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "Airsim: High-fidelity visual and physical simulation for autonomous vehicles," in *Field and service robotics*. Springer, 2018, pp. 621–635. 46

[174] R. Stewart, M. Andriluka, and A. Y. Ng, "End-to-end people detection in crowded scenes," *Proc. CVPR*, pp. 2325–2333, 2016. 47

[175] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *Proc. ECCV*. Springer, 2016, pp. 87–102. 47

[176] T. Wang and H. Wang, "Graph-boosted attentive network for semantic body parsing," in *Proc. ICANN*. Springer, 2019, pp. 267–280. 48

[177] S. Li, H. Yu, and R. Hu, "Attributes-aided part detection and refinement for person re-identification," *Pattern Recognition*, vol. 97, p. 107016, 2020. 48

[178] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *NIPS*, 2014. 49

[179] B. Kim, S. Shin, and H. Jung, "Variational autoencoder-based multiple image captioning using a caption attention map," *Applied Sciences*, vol. 9, no. 13, p. 2699, 2019. 49

[180] W. Xu, S. Keshmiri, and G. Wang, "Adversarially approximated autoencoder for image generation and manipulation," *IEEE Trans. Multimed.*, vol. 21, no. 9, pp. 2387–2396, 2019. 49

[181] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. CVPR*, 2019, pp. 4401–4410. 49

[182] H. Jiang, R. Wang, Y. Li, H. Liu, S. Shan, and X. Chen, "Attribute annotation on large-scale image database by active knowledge transfer," *IMAGE VISION COMPUT.*, vol. 78, pp. 1–13, 2018. 49

[183] T. Wang, K.-C. Shu, C.-H. Chang, and Y.-F. Chen, "On the effect of data imbalance for multi-label pedestrian attribute recognition," in *Proc. TAAI*. IEEE, 2018, pp. 74–77. 49

[184] K.-H. Y. Chiat-Pin Tay, Sharmili Roy, "Aanet: Attribute attention network for person re-identifications," in *Proc. CVPR (CVPR)*, 2019, pp. 7134–7143.

[185] M. Raza, C. Zonghai, S. Rehman, G. Zhenhua, W. Jikai, and B. Peng, "Part-wise pedestrian gender recognition via deep convolutional neural networks," in *2nd IET ICBISP*. Institution of Engineering and Technology, 2017. [Online]. Available: https://doi.org/10.1049/cp.2017.0102 49

[186] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc AAAI Conf*, 2020, pp. 0–0. 49

[187] E. Yaghoubi, D. Borza, P. Alirezazadeh, A. Kumar, and H. Proença, "Person re-identification: Implicitly defining the receptive fields of deep learning classification frameworks," *arXiv preprint arXiv:2001.11267*, 2020. 49

[188] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *Proc. ICANN*. Springer, 2018, pp. 270–279. 50

[189] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *J. Big Data*, vol. 3, no. 1, p. 9, 2016.

[190] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE T. KNOWL. DATA EN.*, vol. 22, no. 10, pp. 1345–1359, 2009. 50

[191] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, "Pose-guided feature alignment for occluded person re-identification," in *Proc. ICCV*, 2019. 51

[192] C. Corbiere, H. Ben-Younes, A. Ramé, and C. Ollion, "Leveraging weakly annotated data for fashion image retrieval and label prediction," in *Proc. ICCVW*, 2017, pp. 2268–2274. 53

[193] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. IEEE international workshop on performance evaluation for tracking and surveillance (PETS)*, vol. 3. Citeseer, 2007, pp. 1–7. 53

[194] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person re-identification," in *Proc. ECCV*. Springer, 2016, pp. 868–884. 54

[195] Z. Ji, Z. Hu, E. He, J. Han, and Y. Pang, "Pedestrian attribute recognition based on multiple time steps attention," *Pattern Recognit. Lett.*, 2020. 54

[196] J. Jia, H. Huang, W. Yang, X. Chen, and K. Huang, "Rethinking of pedestrian attribute recognition: Realistic datasets with efficient method," *arXiv preprint arXiv:2005.11909*, 2020. 55

[197] X. Bai, Y. Hu, P. Zhou, F. Shang, and S. Shen, "Data augmentation imbalance for imbalanced attribute classification," *arXiv preprint arXiv:2004.13628*, 2020. 55

[198] X. Ke, T. Liu, and Z. Li, "Human attribute recognition method based on pose estimation and multiple-feature fusion," *SIGNAL IMAGE VIDEO P.*, 2020. 55

[199] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Parsing clothing in fashion photographs," in *Proc. CVPR*. IEEE, 2012, pp. 3570–3577. 55

[200] J. Yang, J. Fan, Y. Wang, Y. Wang, W. Gan, L. Liu, and W. Wu, "Hierarchical feature embedding for attribute recognition," in *Proc. CVPR*, 2020, pp. 13 055–13 064. 55

Chapter 3

SSS-PR: A Short Survey of Surveys in Person Re-identification

Abstract. Person re-id addresses the problem of whether *"a query image corresponds to an identity in the database*" and is believed to play a fundamental role in security enforcement in the near future, particularly in crowded urban environments. Due to many possibilities in selecting appropriate model architectures, datasets, and settings, the performance reported by the state-of-the-art re-id methods oscillates significantly among the published surveys. Therefore, it is difficult to understand the mainstream trends and emerging research difficulties in person re-id. This paper proposes a multi-dimensional taxonomy to categorize the most relevant researches according to different perspectives and tries to unify the categorization of re-id methods and fill the gap between the recently published surveys. Furthermore, we discuss the open challenges with a focus on privacy concerns and the issues caused by the exponential increase in the number of re-id publications over the recent years. Finally, we discuss several challenging directions for future studies.

3.1 Introduction

Many countries consider video surveillance either as a primary tool to enforce security and prosecute criminals or simply as a crime deterrent tool. Following an incident, law enforcement authorities can review the available video footage, and identify a set of interest subjects, by matching the captured images/video to the enrolled IDs [1].
Given an input query, the person re-id systems compare and match the input data with the existing identities in the database (*gallery* set), probably captured from non-overlapping cameras and at different time intervals [2]. The goal is to retrieve an ordered list of the known identities with the most similarities to the query person. To this end, as outlined in Fig. 3.1 (a), three modules (detection, tracking, and retrieval) work together, each one requiring a supervised learning phase on data that represent system settings. In the computer vision community, the tasks of person detection and tracking are considered independent fields that –at the end– help to obtain the gallery set. Therefore, aligned with the previous researches, in this paper we regard the person re-id exclusively as a retrieval problem that includes four main tasks: a) data collection; b) annotation; c) model training; and d) inference (see Fig. 3.1 (b)).
Full-body person re-id methods are either based on gait (dynamic) or appearance features. While gait is a *unique* behavioral biometric trait that is hard to counterfeit, it is highly dependent on the body-joints motion and can be affected by the slope of the surfaces, subjects' shoes and illness [3]. On the other side, appearance-based approaches rely

on visual features such as edges, shape, color, texture, and expressiveness of the data. Therefore, being intrinsically different, the gait-based and visual-based approaches can be considered as disjoint tasks, both in terms of the existing databases and identification techniques. In this paper, for consistency purposes, we focus exclusively on the visual-based re-id approaches and refer the readers interested in gait-based re-id to [4] and [3].

a) Person re-id system in operation (known as end-to-end) b) Steps for constructing a retrieval module from scratch

Figure 3.1: An end-to-end re-id model detects and tracks the individuals in a video, and then retrieves the query person, while a typical re-id model focuses on the retrieval task.

Person re-id has attracted considerable interest in the last decade, with more than 53 papers published only in Conference on Computer Vision and Pattern Recognition (CVPR) 2019 and International Conference on Computer Vision (ICCV) 2019. Over the past decade, many review articles have been published to organize the methods available in the research literature, each one study the problem from different and often contradictory perspectives. As relevant examples, [5] and [6] discuss the open-world setting versus close-world re-id and analyze the discrepancies, while [7] and [8] survey the methods from the deep learning point of view and emphasis the effectiveness of deep neural network structures upon re-id models performance. [9] addresses the challenge of heterogeneous re-id, in which the query and gallery sets allocate to different domains, and [10] studies the importance of efficiency and computational complexity in deep re-id architectures. Totally, we identified more than 20 body-based person re-id surveys, 12 were published as journal papers, 3 as books, and the remaining are available on ArXiv. From these resources, 9 papers have been published since 2019. For the complete list of surveys and reading more information about each article, we refer the readers to the Appendix.

3.1.1 Contributions

a) As our first and foremost motivation, we propose a multi-dimensional taxonomy that distinguishes between the person re-id models, based on their main approach, type of learning, identification settings, strategy of learning, data modality, type of queries and context (Section 3.2).
b) We briefly discuss the privacy and security concerns in surveillance, with a focus on Privacy-Enhancing Technologies (PET)s, to encourage the research community to introduce privacy-by-design and default systems (Section 3.3).
c) We identify several emerging deviations caused by an evidently growing number of publications over the last few years and discuss the open issues and point out for future directions in this topic (Section 3.4).
However, the detailed analysis of the existing methods is out of the scope of our discussion, and this short survey of surveys should be regarded as a complement to the existing

primary surveys.

3.2 Person Re-identification Taxonomy

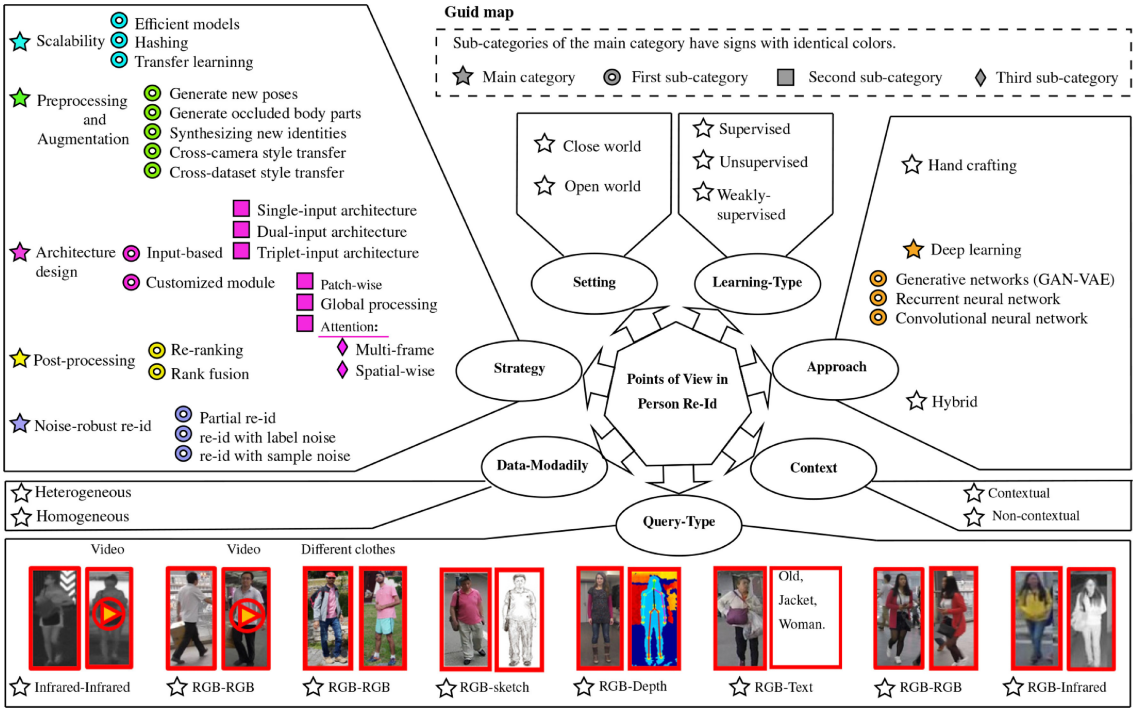


Figure 3.2: Multi-dimensional taxonomy (*Points-of-view*) of the person re-identification problem.

Generally, re-id models have several independent features that help to categorize the methods from different perspectives, as shown in Fig. 3.2. Here, we not only provide a multi-dimensional taxonomy as an overall insight into the existing research, but we explore novel ideas from various points of view as well. As an example, the challenges in a deep learning model based on a text-query with open-world setting are totally different from the challenges of a model designed for a close-world setting with an RGB video-query. Therefore, in the following subsections, after discussing how data-acquisition and data-domain affect the re-id methods, we review the existing strategies for designing a re-id model, followed by a short description of the most popular approaches for implementation of the strategies. Finally, we briefly explain the categorization in system settings, context, data-modality, and learning-type.

3.2.1 Query-type

Before developing any re-id technique, two main properties of the data should be analyzed with particular attention:

3.2.1.1 Data-domain

In image-based datasets, the model is trained on a few samples per individual, while in video-based benchmarks, for each person, several *sequence of images* (i.e, video segments) are available. The existing video-based datasets consist of either RGB or infrared data [11], and both the query and gallery data are from the same domain (i.e.,***infrared-infrared***, ***RGB-RGB***); whereas the image-based re-id datasets are classified into ***RGB-Depth***, ***RGB-infrared***, ***RGB-sketch***, ***RGB-text***, and ***RGB-RGB***. RGB-RGB image-based datasets are classified into short-term and long-term re-id –in which identical persons may appear with ***different clothes***. When retrieving a person from a gallery, the operator may input a query that comes from different domains, which results in large distances between the features extracted from gallery and query data. When dealing with different data modalities, developing methods for learning the gap between domains is critical, since typical similarity features (e.g., texture and color) may be misleading.

3.2.1.2 Data-content

Data acquisition protocols and conditions (which could be performed either by handheld devices or stationary cameras) strongly determine the properties of the resulting data and affect the kind of re-id techniques suitable for the problem. For instance, as shown in Fig. 3.3, some data variability factors such as pose, motion, and occlusions heavily depend on the camera view angle and constraint the model's performance.

Figure 3.3: Examples of how varying capturing angles affect the salient points in the data and demand specific re-id solutions to obtain acceptable performance.

3.2.2 Strategies

Upon our analysis to the problem and to the existing surveys, we suggest that the existing re-id strategies can be broadly grouped according to five perspectives: scalability, pre-processing and augmentation, model architecture design, post-processing strategies, and robustness to noise.

3.2.2.1 Scalability

Speed, accuracy, and on-board processing are critical factors of a real-world person re-id system. The process of retrieving from large-size gallery sets is a time-demanding

task, as a solution of which, designing ***efficient models*** and using ***hashing*** techniques have been effective. The unnecessary parts and parameters of the network are removed using pruning or distillation techniques [12] to increase the efficiency and build light-weighted models. Subsequently, the captured data can be processed on-board instead of transferring it to the operation center. Hashing [13] is the transformation of the features to a compressed form, which not only accelerates the searching process (matching) but occupies less area for storage as well. To tackle the problem of scalability in training phase and learn from huge volume of unlabeled data, a common solution is to apply ***transfer learning*** that is sometimes referred to as domain adaptation, in which we use an annotated source domain to learn the discriminative representation of the unlabeled target domain.

3.2.2.2 Pre-processing and augmentation

Apart from the basic pre-processing techniques (such as channel-wise color alteration or random erasing) that increase the volume of the labeled data, most of the methods in this category use Generative Adversarial Networks (GANs) to synthesize new data or edit the existing ones. ***Generate new poses*** for the existing identities is a technique that allows the network to learn a comprehensive presentation of individuals, while ***generating occluded body-parts*** provides the model with new sets of features. Moreover, ***synthesizing new identities*** can be seen as a data augmentation technique that contributes to the re-id models' performance if the synthetic data follows a similar distribution to the original dataset.
The data undergoes substantial changes in color-style if we collect them from multiple cameras. However, a ***cross-camera style transfer*** can cross-transforms the color and illumination between cameras, which can strongly improve the model performance. Performing style transfer over multiple datasets (***cross-dataset style transfer***) is also used to increase the volume of the training data in the desired domain (e.g., transferring the style of night images to RGB images).

3.2.2.3 Architecture design

The quality of the extracted features from the query and gallery sets is a factor that significantly determines the system's performance. Generally, there are two overlapped perspectives to design a novel architecture for extracting discriminative representation from the data:
1) Design ***stream-based*** models, which could be investigated from two points of view: a) in the first perspective, the main objective is to learn suitable metrics learning (using the loss function) to reduce the intra-class variations and increase the inter-class variations [14]. Different from typical re-id models that use ***single-stream architectures***, some novel models propose to use ***dual-stream architectures*** to focus on the inputs' similarity degree. Moreover, ***triplet/quadruplet-stream architectures*** use the images of the other identities as negative inputs and the images of the target person as positive and anchor inputs [15]. It worth mentioning that usually the

weights and parameters are sheared between streams of the model, leading to a popular architecture called Siamese networks [16]. b) the second perspective to design stream-based models is to extract various features from one identity using multiple streams and fuse them together (e.g., fusing extracted information from motion, semantic attributes, handcrafting techniques, and CNN-based methods).
2) Design ***customized modules*** to perform specific processes for extracting robust discriminative features from data. When discussing the customized-design, there are many possibilities; therefore, we sub-categorize them into three groups: a) ***Patch-wise techniques***. Patch-based analysis helps to extract minutiae information (known as fine-grained features) from the data, which helps to discriminate between inter-class samples that are visually similar to each other. Not only can the patch-wise techniques use various ways of patching (illustrated in Fig 3.4), but they use different approaches to analyze each patch as well. For example, when using a simple Long Short-Term Memory (LSTM) architecture, the comprehensive feature representation is obtained by processing all the patches one after another, while in a multi-input architecture, one can perform a cross-analysis –e.g., to extract shareable features from head-patches of two images. b) ***Global-based processing techniques*** focus on the topology of the cameras and network consistency [13]. Three widely-used datasets (i.e., Market1501, DukeMTMC, and underGround Re-IDentification (GRID)) have provided the locations (aerial map), where each camera covers, to allow studying the effects of cameras' topology on the model efficiency. As a vivid example, suppose two cameras cover the entrance and exit sides of a narrow street; thus, a person that is firstly captured in frontal-view probably appears in rear-view on the next camera. c) ***Attention-based techniques***. By capturing images from different angles, some parts of the input-data undergo substantial changes in appearance, texture, shape, occlusion, and illumination. Fundamentally, this is a misalignment problem, in which the model aims to find the target person by matching the corresponding regions of the body (e.g., head with head) in query and gallery data. The existing solutions are typically divided into: i) ***special-wise*** attention; and ii) ***multi-frame*** attention. Generally, special-wise techniques search for salient pixels/regions on the image, which could be accomplished by performing a channel-wise operation, learning hard-masks, developing modules for regional selection or by designing multi-input networks. In the multi-frame attention architecture, the aim is to provide *one* feature representation from a *sequence of images*.

3.2.2.4 Post-processing

The output of a re-id model is an ordered list of gallery identities, according to the similarity between the gallery and query data. This list is called ranking-list, and any further processes for re-ordering the results are known as ***re-ranking***. Many intuitive scenarios could help refine this ranking list. For example, in case of being ranked particularly high for one query, a gallery image should be ranked low for any other queries. Also, if the query person has dark-skin, individuals with light-skins should not be ranked high. Another frequent post-processing approach is the *rank fusion* (fusion of ranking-

Figure 3.4: Some of patching strategies used to obtain fine-grained local representations of the input data.

lists) of multiple re-id methods, which is particularly suitable when accuracy is much more important than speed and computational cost.

3.2.2.5 Robustness to Noise

Whether we use automatic human detection and tracking or perform it manually, errors, misalignment, and inconsistency in bounding-box detection are inevitable. Furthermore, the annotation process is a human-biased step that is mostly accompanied by some percentage of errors that may affect the quality of the learning process. There are three general approaches to tackle these challenges [6]. ***Partial re-id*** techniques construct models capable of extracting shareable features from unoccluded body parts, while outlaid bounding boxes and inaccurate tracking are studied under the ***sample-noise reduction***. ***Label-noise*** topic addresses the annotation errors by limiting the model not to be overfilled on the labels.

3.2.3 Approaches

The discussed strategies (in section 3.2.2) could be taken in to account by three approaches: ***deep learning***, ***hand-crafting***, and the combination of both (***hybrid***).
In the last decade, re-id systems were usually implemented based on knowledge-based feature extractors, which could be classified into four main groups: camera geometry/calibration, color calibration, descriptor learning, and distance metric learning. As most of the traditional techniques were built upon appearance-based similarities, designing discriminative visual descriptors and learning distance metrics upon person clothes were more popular than other methods [17].
Many studies focused on deep structures or a combination of deep neural networks and traditional methods after the advent of deep learning approaches. In the context of deep learning, ***Convolutional Neural Networks*** (CNNs) analyze the input data at a single instance, while in ***Recurrent Neural Networks*** (RNNs) the data is treated as a sequence of inputs; then, taking advantage of an internal state (memory), the critical information of each sequence is accumulated to construct the final feature representative of the input. Finally, ***generative networks*** are classified into Variational Auto-Encoder (VAE) and GAN, each aiming to find the distribution of the original dataset to generate

new data. In re-id, GAN-based approaches have shown promising results with both augmenting the dataset, and editing the samples (e.g., style transferring, completing the occluded body-parts, etc.)

3.2.4 Identification Settings

Re-id model are either classified into the ***open-world*** or ***closed-world*** settings. The closed-world assumption deals with matching one-to-many samples, so that the query image is surely corresponding to one of the gallery individuals. On the other hand, there are different interpretations for the open-world setting: 1) it might regard a multi-camera problem in which the gallery evolves over time, and the ever-changing query may not be presented in the gallery. Moreover, the system could re-identify multi-subjects at once [18]; 2) it might regard a group-based verification task aiming to determine whether the query appears in the gallery or not, without the necessity of retrieving matched person(s) [19]; and 3) any real-world application that excludes the close-world setting could be considered as an open-world problem. For example, in [6], researches that deal with heterogeneous data, raw images/videos, limited labels, and noisy annotations have been considered as open-world studies [20].

3.2.5 Context

Context is another point of view towards re-id problems so that if the system relies on the external contextual information (e.g., camera/geometric information) rather than using the data itself, it is considered as a contextual system [2]. However, after the advent of deep learning technologies, only a small proportion of works consider person re-id from the contextual perspective [13]. Meanwhile, contextual based re-id datasets should provide extra information such as full-frame data, cameras' locations and capturing angles e.g., using an aerial map.

3.2.6 Data-modality

Given the various data modalities for the query and gallery sets, the re-id task can be regarded either as a Heterogeneous re-id (He-Reid) or Homogeneous re-id (Ho-Reid) problem. In a Ho-Reid perspective, the query and gallery data have similar modalities, while in the He-Reid the query is from another domain (for example, if the gallery consists of RGB-images, the query could be a verbal description of the target person). Therefore, in He-Reid, discrepancies between the query-domain and the gallery-domain are huge so that the methods developed for Ho-Reid cannot be directly applied to these problems. Dealing with two different data modalities, He-Reid techniques aim to bridge the gap between domains and decrease the inter-modality discrepancy, for which there are several methods [9]: 1) learning a metric to decrease the gap between features of each domain; 2) learning shared features; and 3) unifying modalities before feature extraction by transferring both domains to a latent domain. So far, owing to Generative Adversarial

Networks (GAN), unifying the modalities has shown better results that are discussed at the end of this section.

3.2.7 Learning-type

Supervised, semi supervised, weakly supervised, and unsupervised learning [21] are the annotation-based learning types. Due to leveraging the manually annotated data, ***supervised*** methods achieve superior accuracy than other methods. However, some works develop ***weakly-supervised*** or ***unsupervised methods*** to not only ease the process of data annotation but also train the model on an excessive amount of unlabeled data. The main categories in unsupervised learning are *domain adaption*, *dictionary learning*, *feature representation extraction*, *distance measurement*, and *clustering* [13], from which *Unsupervised Domain Adaptation (UDA)* has attracted the most attention. In UDA, taking advantage of a labeled dataset (source domain), the model learns the discriminative representation of the unlabeled data (target domain). Therefore, the distance between the data distribution of domains is minimized, so that target-domain data can be treated as the source-domain data for training purposes. Different from the time-consuming annotation process for supervised methods (all people in the video are annotated one-by-one), weakly-supervised annotation is a video-level process, in which each video needs one label, indicating the IDs appeared in that video.

3.2.8 State-of-the-Art Performance Comparison

Table 3.1 shows the performance (rank-1 and mean Average Precision (mAP)) of the state-of-the-art techniques, most published in 2019 and 2020. In these works, the gallery set is always composed of RGB images/videos, except in [11] (with 14.3 % accuracy for rank-1 retrieval), where both the gallery and query sets contain infrared images captured at night.

[9] reported that the performance of all the He-Reid works is lower than 40%, whereas the latest papers have claimed 56.7%, 49.0%, 49.%9, and 70.0% rank-1 accuracy for RGB-text, RGB-sketch, RGB-infrared, and RGB-thermal, respectively, pointing for a fast improvement in performance in this field.

Table 3.1 enables to conclude that He-Reid and long-term re-id are the least matured fields of study, respectively with 70% [6] and 65.7% [22] rank-1 accuracy, while [23] is an unsupervised person re-id work that is close to the hopeful boundary, with 86.2 % and 76% rank-1 accuracy on the Market-1501 and DukeMTMC datasets, respectively.

On the other hand, even though studies based on RGB images and RGB videos have achieved higher results, their performance is highly dependent on the dataset, such that rank-1 accuracy in RGB video-based studies is in a rage from 63.1 % [24] to 96.2% [25] for the LS-VID and DukeMTMC-VideoReID datasets, respectively; similarly, in RGB image-based researches, [26] has achieved 95.7% rank-1 accuracy on the Market-1501 dataset, while [6] reports this number around 63.6% for their experiments on the CUHK03 dataset.

Field of study	Dataset	Method	R 1	mAP
RGB-Thermal	RegDB	[6]	70.0	66.4
RGB-infrared	SYSU-MM01	[27]	49.9	50.7
RGB-Sketch	Sketch Re-ID	[28]	49.0	-
RGB-Text	CUHK-PEDES	[29]	56.7	-
Infrared-infrared	KnightReid	[11]	14.3	10.2
RGB-D	KinectReID	[30]	99.4	-
	RGBD-ID	[30]	76.7	-
Unsupervised	Market-1501	[23]	86.2	68.7
	DukeMTMC*	[23]	76.0	60.3
RGB image-based	Market-1501	[26]	95.7	89.0
	CUHK03	[6]	63.6	62.0
	MSMT17	[6]	68.3	49.3
	DukeMTMC*	[26]	91.1	81.4
RGB video-based	3DPeS	[31]	78.9	-
	PRID2011	[24]	95.5	-
	iLDS-VID	[25]	88.9	93.0
	MARS	[32]	90.0	82.8
	DukeMTMC-VideoReID*	[25]	96.2	95.4
	LS-VID	[24]	63.1	44.3
	PRW	[33]	73.6	33.4
Long-term	Motion-ReID*	[22]	65.7	-
	Celeb-reID	[34]	51.2	9.8

*Not publicly available.

Table 3.1: Performance of the state-of-the-art re-id methods.

3.3 Privacy Concerns

IAPP, the International Association of Privacy Professionals, defines that *privacy* is the right to be free from interference or intrusion and to remain anonymous, and *information privacy* regards the control over our own personal information. Among the possible ways of privacy violation [35] (i.e., watching, listening, locating/tracking, detecting/sensing, personal data monitoring, and data analytics), we pay attention to the visual monitoring that has recently engaged the research community, due to the sensitiveness of monitoring people or collecting their personal visual data (from the Internet) without their consent. In this scope, several well-known benchmarks (e.g., *Brainwash*, *DukeMTMC*, and *MS-Celeb-1M*) were permanently suspended by their authors[36], in most cases due to the absence of explicit authorization from the subjects in the dataset to have their data collected and disseminated for research purposes.

Overall, there are two solutions to reduce the privacy concern in person re-id models: ***privacy-by-design principles*** and ***Privacy-Enhancing Technologies*** (PET).

Privacy-by-design principles are some standards to protect data through technology design, published by the law enforcement agencies[1,2] and enforce companies to respect the privacy of their customers. In these standards, information tracking is defined as a principle that allows people to manage and track whom they have access to their private

[1]https://gdpr-info.eu/

[2]https://www.priv.gc.ca/en/report-a-concern/

information (and to what extend). In contrast, the data minimization principle states that enterprises should only process the minimum necessary data. For example, a visual surveillance panel that processes the crowd for displaying related advertisements may need to recognize the human semantic attributes (e.g., gender, clothing styles, etc.), but should avoid designing a system that detects faces, analyzes the skin color, and people's race.

PET are methods of protecting data, including anonymization, perturbation, and encryption [37]. In anonymization, the sensitive information is removed to perform a complete de-identification, generally accomplished by masking, while in perturbation, the sensitive attributes of the data are replaced with noisy or otherwise altered data. On the other hand, security techniques *reversibly* disguise the identifying information. Examples of PET in person re-id could be disguising pedestrian's faces in the gallery set using generative networks to reduces the risk of privacy intrusion; however, it indicates the need for methods that are able to perform the re-id task on anonymized data and possibly reconstruct the true faces if asked by the authorities [38]. As a method for developing fast re-id, *hashing* could be used to design a re-id model that works with encrypted data and reduces the risk of hacking.

3.4 Discussion and Future Directions

3.4.1 Biases and Problems

The number of methods in person re-id has considerably increased in recent years, leading to some biases and problems such as unfair comparisons, low originality in techniques, and insufficient attention to some of the important perspectives in the problem.

3.4.1.1 Unfair comparisons

Based on the re-implementation of several state-of-the-art re-id methods, a recent baseline [39] explicitly concluded that the improvements reported in some works were mainly due to training tricks rather than to any conceptual advancement of the method itself, which has led to an exaggeration of the success of such techniques. Therefore, to show the effectiveness of the model, we suggest to perform an ablation study on the proposed method, such that the basic model is first evaluated, and each proposed component is added one by one over the baseline to show the effectiveness of the idea. Further, to show the superiority of the method over the existing state of the arts, authors should remain the architecture and parameters constant as much as possible, so that we are certain that the improvement is caused by the idea [40].

3.4.1.2 Low originality

Although using the power of other fields in person re-id is valuable and improves the performance of state of the art, in recent years, excessive attention to these kinds of contributions has decreased the number of original works with significant contributions.

In the literature, we repeatedly face with re-implementation of other fields' ideas as original re-id methods, creating competition for a mere copy of outside ideas into re-id problems. For example, as confirmed by [40], after the success of LSTM, GAN, Siamese network, backbone networks (ResNet, Inception, GoogleNet), various loss functions, etc., many authors repeated the same ideas on the re-id datasets.

3.4.1.3 Insufficient attention to some perspectives

A long-term re-id model capable of retrieving multi-modality queries is much more realistic and useful than a close-world, single-modality retrieval system. Nevertheless, why does exist more researches in the second scenario?. Understanding the nature of the deep neural network is the answer to this question. It is known that deep neural networks are efficient in feature extraction, and they have shown promising results specifically in problems dealing with appearance-based features. Thereby, re-id scenarios under close-world setting and homogeneous RGB data-modality have shown considerable performance improvement. On the other hand, there is little attention to some challenges such as open-world setting, long-time re-id, heterogeneous modality, and non-contextual tasks.

3.4.2 Open Issues

In this section, we discuss the major open issues in the re-id problem and point out for some possible further directions.

Person re-id performance has several important covariates, such as variations in background, illumination, occlusion, body-pose, and other view-dependent variables [5], [41]. In particular, we emphasize the role of data annotation: when training deep neural networks, the more the annotated data are available, the better the performance would be. However, data preparation for re-id is an expensive, tedious, and time-consuming process, opening the space for developing novel semi-supervised, weakly supervised or even unsupervised solutions for training the models [6].

Affected by similar covariates, other pattern recognition tasks (e.g., iris recognition, cross-domain clothing analysis, multi-object tracking) have significantly helped the person re-id in several directions such as unsupervised learning, extraction of discriminative feature sets, and application of robust metric learning techniques. Nevertheless, some challenges are related explicitly to the re-id task: for example, by increasing the volume of the re-id datasets, the matching process (for retrieving the query person from a large-scale gallery set) takes substantially more time, indicating the need for fast re-id methods [6].

Furthermore, for a real-world re-id system, it is necessary to search the query person independent of its data-type. However, due to the lack of large datasets consisted of multi-modal data, current heterogeneous works are limited to single cross-modality searches, and the gallery set often consists of RGB images. Unifying modalities of the query set and gallery set is another open issue in heterogeneous re-id that could be fulfilled by mapping the modality of both sets either to each other interchangeably or to a latent space [9].

Apart from most of researches in the literature that are based on appearance, long-term re-id solves the issue of retrieving the same person with different appearance and clothing style [7]. Therefore, studies in this area should consider challenges such as: 1) going beyond appearance-based features and extract discriminative features from hard-biometrics (face and gait) and more robust soft-biometrics (height, body volume, body contours). Meanwhile, recent facial recognition techniques that typically are trained on high-resolution data (with controlled pose-variation) may not increase the overall performance when dealing with low-quality faces in the wild; 2) long-term re-id in real applications is often tied to open-world setting challenges such as scalability (how to deal with large databases) and generalization (adding new cameras to the existing system) [5]. It worth mentioning that person search is a slightly different research area that aims to locate the prob person within a whole frame containing one/several persons [42].
Currently, a plethora of human detection and tracking techniques are available for different platforms. By generalizing them for the handheld devices –thanks to high-speed internet connections–, mobile person re-id can quickly become a trivial task, which raises many privacy and security concerns. Thus, both secure storage of the gallery set and proposing re-id methods that conform privacy concerns *by design and default* are of the utmost challenges.

3.5 Conclusion

Person re-id aims to retrieve an ordered list of the identities from a database, with respect to query images taken from one or multiple non-overlapping cameras. In result of the extensive research carried out over the last years for solving the primary pattern recognition challenges (e.g., pose variations, partial occlusions and dynamic data acquisition conditions), re-id systems have successfully passed the human accuracy-level in easy scenarios (i.e., when the model is trained based on supervised learning and close-world setting in RGB heterogeneous modality). In this paper, we proposed a multi-view taxonomy that considers the different categorizations available in the re-id literature to ease the discovery of realistic and feasible scenarios for future directions. Furthermore, we discussed the importance of the concept of privacy in this field and briefly reviewed several strategies to improve systems' security and privacy by default. Finally, after discussing some of the issues caused by an evidently growing number of publications in recent years, we pointed out for some of the open issues in this extremely challenging problem.

Bibliography

[1] M. Gou, Z. Wu, A. Rates-Borras, O. Camps, R. J. Radke *et al.*, "A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets," *IEEE TPAMI*, vol. 41, no. 3, pp. 523–536, 2018. 71

[2] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *IMAGE VISION COMPUT*, vol. 32, no. 4, pp. 270–286, 2014. 71, 78

[3] A. Nambiar, A. Bernardino, and J. C. Nascimento, "Gait-based person re-identification: A survey," *ACM Comput. Surv.*, vol. 52, no. 2, Apr. 2019. [Online]. Available: https://doi.org/10.1145/3243043 71, 72

[4] P. Connor and A. Ross, "Biometric recognition by gait: A survey of modalities and features," *Computer Vision and Image Understanding*, vol. 167, pp. 1–27, 2018. 72

[5] Q. Leng, M. Ye, and Q. Tian, "A survey of open-world person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 4, pp. 1092–1108, 2020. 72, 82, 83

[6] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *arXiv preprint arXiv:2001.04193*, 2020. 72, 77, 78, 79, 80, 82

[7] M. O. Almasawa, L. A. Elrefaei, and K. Moria, "A survey on deep learning-based person re-identification systems," *IEEE Access*, vol. 7, pp. 175 228–175 247, 2019. 72, 83

[8] D. Wu, S.-J. Zheng, X.-P. Zhang, C.-A. Yuan, F. Cheng, Y. Zhao, Y.-J. Lin, Z.-Q. Zhao, Y.-L. Jiang, and D.-S. Huang, "Deep learning-based methods for person re-identification: A comprehensive review," *Neurocomputing*, vol. 337, pp. 354–371, Apr. 2019. [Online]. Available: https://doi.org/10.1016/j.neucom.2019.01.079 72

[9] Z. Wang, Z. Wang, Y. Wu, J. Wang, and S. Satoh, "Beyond intra-modality discrepancy: A comprehensive survey of heterogeneous person re-identification," *arXiv preprint arXiv:1905.10048*, 2019. 72, 78, 79, 82

[10] H. Masson, A. Bhuiyan, L. T. Nguyen-Meidine, M. Javan, P. Siva, I. B. Ayed, and E. Granger, "A survey of pruning methods for efficient person re-identification across domains," *arXiv preprint arXiv:1907.02547*, 2019. 72

[11] J. Zhang, Y. Yuan, and Q. Wang, "Night person re-identification and a benchmark," *IEEE Access*, vol. 7, pp. 95 496–95 504, 2019. 74, 79, 80

[12] I. Ruiz, B. Raducanu, R. Mehta, and J. Amores, "Optimizing speed/accuracy trade-off for person re-identification via knowledge distillation," *Engineering Applications of Artificial Intelligence*, vol. 87, p. 103309, Jan. 2020. [Online]. Available: https://doi.org/10.1016/j.engappai.2019.103309 75

[13] H. Wang, H. Du, Y. Zhao, and J. Yan, "A comprehensive overview of person re-identification approaches," *IEEE Access*, vol. 8, pp. 45 556–45 583, 2020. 75, 76, 78, 79

[14] B. Lavi, I. Ullah, M. Fatan, and A. Rocha, "Survey on reliable deep learning-based person re-identification models: Are we there yet?" *arXiv preprint arXiv:2005.00355*, 2020. 75

[15] A. Khatun, S. Denman, S. Sridharan, and C. Fookes, "A deep four-stream siamese convolutional neural network with joint verification and identification loss for person re-detection," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 1292–1301. 75

[16] S. K. Roy, M. Harandi, R. Nock, and R. Hartley, "Siamese networks: The tale of two manifolds," in *Proc. ICCV*, 2019, pp. 3046–3055. 76

[17] R. Satta, "Appearance descriptors for person re-identification: a comprehensive review," *arXiv preprint arXiv:1307.5748*, 2013. 77

[18] M. A. Saghafi, A. Hussain, H. B. Zaman, and M. H. M. Saad, "Review of person re-identification techniques," *IET Computer Vision*, vol. 8, no. 6, pp. 455–474, 2014. 78

[19] S. Chan-Lang, "Closed and open world multi-shot person re-identification," Ph.D. dissertation, Paris 6, 2017. 78

[20] H. Wang, X. Zhu, T. Xiang, and S. Gong, "Towards unsupervised open-set person re-identification," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 769–773. 78

[21] Y. Lin, "Deep learning approaches to person re-identification," Ph.D. dissertation, University of Technology Sydney, 2019. 79

[22] P. Zhang, Q. Wu, J. Xu, and J. Zhang, "Long-term person re-identification using true motion from videos," in *Proc. WACV*. IEEE, 2018, pp. 494–502. 79, 80

[23] Y. Fu, Y. Wei, G. Wang, Y. Zhou, H. Shi, and T. S. Huang, "Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification," in *Proc. ICCV*, October 2019, pp. 6112–6121. 79, 80

[24] J. Li, J. Wang, Q. Tian, W. Gao, and S. Zhang, "Global-local temporal representations for video person re-identification," in *Proc. ICCV*, October 2019, pp. 3958–3967. 79, 80

[25] M. Li, H. Xu, J. Wang, W. Li, and Y. Sun, "Temporal aggregation with clip-level attention for video-based person re-identification," in *The IEEE Winter Conference on Applications of Computer Vision*, March 2020, pp. 3376–3384. 79, 80

[26] H. Chen, B. Lagadec, and F. Bremond, "Learning discriminative and generalizable representations by spatial-channel partition for person re-identification," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020, pp. 2472–2481. 79, 80

[27] D. Li, X. Wei, X. Hong, and Y. Gong, "Infrared-visible cross-modal person re-identification with an x modality," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 4610–4617. 80

[28] S. Gui, Y. Zhu, X. Qin, and X. Ling, "Learning multi-level domain invariant features for sketch re-identification," *Neurocomputing*, vol. 403, pp. 294–303, Aug. 2020. [Online]. Available: https://doi.org/10.1016/j.neucom.2020.04.060 80

[29] S. Aggarwal, V. B. RADHAKRISHNAN, and A. Chakraborty, "Text-based person search via attribute-aided matching," in *The IEEE Winter Conference on Applications of Computer Vision*, March 2020, pp. 2617–2625. 80

[30] L. Ren, J. Lu, J. Feng, and J. Zhou, "Uniform and variational deep learning for rgb-d object recognition and person re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 10, pp. 4970–4983, 2019. 80

[31] S. Zhou, J. Wang, D. Meng, Y. Liang, Y. Gong, and N. Zheng, "Discriminative feature learning with foreground attention for person re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 9, pp. 4671–4684, 2019. 80

[32] C.-T. Liu, C.-W. Wu, Y.-C. F. Wang, and S.-Y. Chien, "Spatially and temporally efficient non-local attention network for video-based person re-identification," *arXiv preprint arXiv:1908.01683*, 2019. 80

[33] Y. Yan, Q. Zhang, B. Ni, W. Zhang, M. Xu, and X. Yang, "Learning context graph for person search," in *Proc. CVPR*, June 2019, pp. 2158–2167. 80

[34] Y. Huang, J. Xu, Q. Wu, Y. Zhong, P. Zhang, and Z. Zhang, "Beyond scalar neuron: Adopting vector-neuron capsules for long-term person re-identification," *IEEE T CIRC SYST VID*, vol. 30, no. 10, pp. 3459–3471, 2020. 80

[35] C. D. Raab, "Privacy, security, surveillance and regulation," 2017. [Online]. Available: http://www.inf.ed.ac.uk/teaching/courses/pi/2017_2018/slides/RaabProfIssuesInformaticsCourse2017FINALppt.pdf 80

[36] J. Harvey, Adam. LaPlace. (2019) Megapixels.cc: Origins, ethics, and privacy implications of publicly available face recognition image datasets. [Online]. Available: https://megapixels.cc/ 80

[37] J. Curzon, A. Almehmadi, and K. El-Khatib, "A survey of privacy enhancing technologies for smart cities," *Pervasive and Mobile Computing*, vol. 55, pp. 76–95, Apr. 2019. [Online]. Available: https://doi.org/10.1016/j.pmcj.2019.03.001 81

[38] H. Proença, "The uu-net: Reversible face de-identification for visual surveillance video footage," *arXiv preprint arXiv:2007.04316*, 2020. 81

[39] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu, "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Trans Multimedia*, vol. 22, no. 10, pp. 2597–2609, 2020. 81

[40] K. Musgrave, S. Belongie, and S.-N. Lim, "A metric learning reality check," *arXiv preprint arXiv:2003.08505*, 2020. 81, 82

[41] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, and S. Z. Li, "Embedding deep metric for person re-identification: A study against large variations," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 732–748. 82

[42] K. Islam, "Person search: New paradigm of person re-identification: A survey and outlook of recent works," *IMAGE VISION COMPUT*, vol. 101, p. 103970, Sep. 2020. [Online]. Available: https://doi.org/10.1016/j.imavis.2020.103970 83

Chapter 4

Region-Based CNNs for Pedestrian Gender Recognition in Visual Surveillance Environments

Abstract. Inferring soft biometric labels in totally uncontrolled outdoor environments, such as surveillance scenarios, remains a challenge due to the low resolution of data and its covariates that might seriously compromise performance (e.g., occlusions and subjects pose). In this kind of data, even state-of-the-art deep-learning frameworks (such as ResNet) working in a holistic way, attain relatively poor performance, which was the main motivation for the work described in this paper. In particular, having noticed the main effect of the subjects' "*pose*" factor, in this paper we describe a method that uses the body keypoints to estimate the subjects pose and define a set of regions of interest (e.g., *head*, *torso*, and *legs*). This information is used to learn appropriate classification models, specialized in different poses/body parts, which contributes to solid improvements in performance. This conclusion is supported by the experiments we conducted in multiple *real-world* outdoor scenarios, using the data acquired from advertising panels placed in crowded urban environments.

4.1 Introduction

Being often the first mentioned attribute to describe a person, gender estimation is useful in many areas of computer vision, such as surveillance, forensic affairs, marketing, and human-robot interaction. In the first decade of this century, datasets were small and most approaches were based on handcrafted features such as Histogram of Oriented Gradients (HOG). However, after the advent of deep learning frameworks, scholars focused on collecting extensive labeled data and developing deeper networks.

In the literature, gender estimation from facial images has received more attention than whole-body. However, in this paper, we use full-body images since in Pedestrian Attribute Recognition (PAR) scenarios not only the quality of facial regions decreases, but also the body features are more robust to far distances.

[1] proposes a fine-tuned CNN model to predict the gender from the "front", "back" and "both" views. They employ a parsing mechanism via the Decompositional Neural Network (DNN) to remove the background. The foreground is then parsed in the upper and lower bodies so that the two CNNs are fine-tuned. As a conclusion, feeding upper-body images to the network slightly improves the results. However, they have gray scaled and forced-squared the images which cause the loss of color-based features and data deformation.

In [2], authors apply HOG alongside a CNN and concatenate the extracted features that are used as the input of a Softmax classifier. Although the expressiveness of the data is

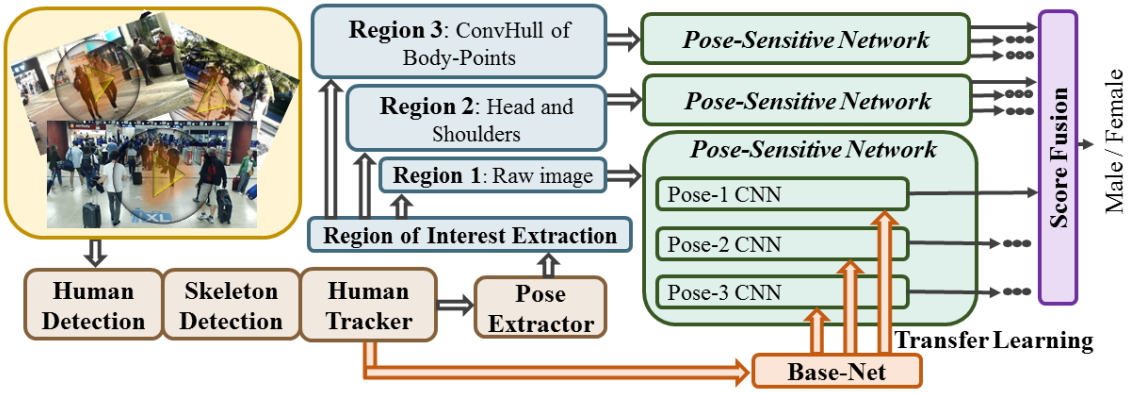


Figure 4.1: Overview of the proposed algorithm called Pedestrian Gender Recognition Network (PGRN). Taking advantage of the human detector, skeleton detector, and human tracker, we extract the bounding boxes alongside 16 body keypoints for each person. Afterward, the training set is split into three subsets corresponding to the desired poses (i.e. frontal, rear, and lateral). The RoIs are then extracted and fed to a Pose-Sensitive Network (PSN) which is constructed from three specialized ResNet50 networks. The weights of a pre-trained network (i.e. Base-Net) are shared with each of these PSNs to reduce the time of training. Finally, the most confident score from the RoIs is considered as the final score for recognition.

protected in this method, feature redundancy in the last layer can lead to a biased model that degrade the performance in real-world applications. [3] presents another work that adopts an extra thermal camera for data acquisition. Using CNN methods, they extract the features from visible images and thermal maps and fuse them in score level by exploiting Support Vector Machine (SVM) learner. As they apply thermal images for recognition, the algorithm can fail in crowded places with occlusion, which is a real and critical scenario.
In addition to the mentioned weaknesses, the datasets in previous works are mainly collected from one location which can cause some easiness such as: monotonous illumination, stable camera settings, controlled occlusion, similar background, and controlled distance acquisition. While in this paper, we collect a dataset from outdoor and indoor advertisement panels in more than 100 cities of Portugal and Brazil[1].
Further, we propose a Pedestrian Gender Recognition Network (PGRN) which provides several decisions based on the subject pose and some Regions of Interest (RoI) so that the decision with maximum certainty is reported as the final recognition (Fig. 4.1). The performed experiments on three datasets show the superiority of the proposed algorithm in comparison with the state-of-the-art methods, as detailed in section 3.

4.2 Pedestrian Gender Recognition Network (PGRN)

Regarding the impact of pose variation on the biometric system performance, we develop our proposed algorithm on a human body keypoint detection and tracker platform. In general, the suggested PGRN is divided into the following steps: training the baseline network called Base-Net, key point detection and tracking, pose extraction, RoI extraction, fine-tuning PSN, and score fusion.

[1]https://tomiworld.com/locations/

4.2.1 Base-Net

Although the pre-trained CNNs on the ImageNet dataset have shown promising results on various recognition tasks, it is interesting to note that training from scratch or updating the weights of all layers necessarily leads to better results upon the availability of sufficient data. As we have collected a large proprietary dataset (i.e. Biometria e Deteção de Incidentes (BIODI)), the weights of the network trained on the ImageNet dataset are considered as the initial weights for our model. Afterward, the whole layers of the network are trained on raw images of the BIODI. This network is named as Base-Net that later will be used for transferring the knowledge to the PSNs.

4.2.2 Body Key-Point Detection and Tracking

BIODI is composed of 216 video clips of wild visual surveillance environments taken from different countries. We started by analyzing each video using a state-of-the-art approach called Alphapose [4] that is an accurate real-time and multi-person skeleton detector based on an object detection method named Faster-RCNN [5]. This object detector provides the Bounding Boxes (BBs) of multiple humans in each frame. Then, the human BBs are fed to the Spatial Transformer Network (STN) [6], which yields high quality dominant human proposals. In other words, the out put of the STN are some transformed human proposals, therefore, after estimating the skeleton of each person using the Single Person Pose Estimator (SPPE) [7], each set of the body keypoints needs to be mapped to the original image coordinate using a de-transformer network.
So far, the detection of BBs and skeleton of each person in each frame is done. To perform the tracking, the straight forward approach is to connect the current skeletons to the closest skeletons in the next frame. However, this method produces errors when there are several poses close to each other. Therefore, we apply Poseflow [8] that works based on a small inter-frame skeleton distance (d_c) and a large intra-frame skeleton distance (d_f) of the form Eq. 4.1. Finally, we storage all the BBs and body keypoints related to each human subject to the disk for the next step.

$$
\begin{aligned}
& d_c(S^{(1)},\, S^{(2)}) = \sum_{n=1}^{N} \frac{f_2^n}{f_1^n}, \\
& d_f(S_1,\, S_2|\, \{\sigma_1, \sigma_2, \lambda\} = \frac{1}{K_{sim}(S_1,\, S_2|\sigma_1)} + \frac{\lambda}{H_{sim}(S_1,\, S_2|\sigma_2)}\,, \\
& s.t.\;\; K_{sim}(S_1,\, S_2|\sigma_1) = \begin{cases} \displaystyle\sum_{n=1}^{N} \tanh\frac{c_1^n}{\sigma_1}.\tanh\frac{c_2^n}{\sigma_1} \;\; ;\text{ if }\; S_2^n \;\text{ in }\; B(S_1^n) \\ 0\,;\text{ Otherwise}\,, \end{cases} \\
& s.t.\;\; H_{sim}(S_1,\, S_2|\sigma_2) = \sum_{n=1}^{N} e^{-\frac{(S_1^n - S_2^n)}{\sigma_2}}\,,
\end{aligned}
\tag{4.1}
$$

where S_1 and S_2 are two skeletons related to two different individuals in a frame in $B(S_1^n)$ and $B(S_2^n)$ bounding boxes, respectively. f_1^n and f_2^n are extracted features of these boxes

and $n \in \{1, ..., N\}$ in which N represents the number of body keypoints, and σ_1, σ_2, and λ can be determined in a data-driven manner.

4.2.3 Pose Inference

For a biometric system specialized in specific human body-pose, various body gestures provide different features, therefore, unseen poses in the test set highly impact its performance. On the other hand, pose-specialized networks are not able to learn the important features if we split the train set to many subsets of different poses. Regarding this matter and number of images of our dataset, we considered only the three most common poses of pedestrians, including "frontal", "rear", and "lateral" views.
As the BBs are extracted using an object detector, the aspect ratio (width/height) of each BB is 1.75. We visualized quite a few numbers of body keypoints (see Fig. 4.2(a)) on the resized images (175×100) and discovered that individuals with shoulder-width lower than nine pixels (out of 100 pixels) in the invariant-scale RoIs can be a nominate for lateral view images. It worth mentioning that, we considered the other body keypoints to perform this experiment, however, the best results are obtained using the shoulder-width points. If $p_i = (x_i, y_i)$ represents the coordinates of body points, the desired poses are:

$$\text{Pose} \equiv \begin{cases} \text{Frontal view}; & \text{if } x_v - x_z < -9 \\ \text{Rear view}; & \text{if } x_v - x_z > 9 \\ \text{Lateral view}; & \text{if } |x_v - x_z| =< 9 \text{ pixels}, \end{cases} \tag{4.2}$$

where (x_v, y_v) and (x_z, y_z) respectively are 13^{th} and 14^{th} body-point coordinates illustrated in Fig. 4.2(a).

4.2.4 RoI: Segmentation and Cropping Strategies

By joining the exterior body points p we obtain a polygon, we find it useful to create a mask by applying Convex-Hull on this set. For N points $p_1, ..., p_N$, the Convex-Hull is the set of all convex combinations of its points such that in a convex combination each point has a positive weight w_i. These weights are used to compute a weighted average of the points. For each choice of weights, the obtained convex combination is a point in the Convex-Hull. Therefore, choosing weights in all possible ways, we can form a black polygon-shape as Fig. 4.2(b). In a single equation, the Convex-Hull is the set:

$$CH(N) = \left\{ \sum_{i=1}^{N} w_i p_i \ : w_i \geq 0 \ \text{ for all } \ i, \ \text{ and } \sum_{i=1}^{N} w_i = 1 \right\}. \tag{4.3}$$

Figure 4.2 illustrates this process for a sample image. To avoid information lost when performing the Convex-Hull algorithm, we consider two extra points (x_l, y_l) and (x_r, y_r) near the ears. Therefore, $y_l = y_r = \frac{y_n + y_h}{2}$ and $x_l = x_n - y_l$, $x_r = x_n - y_r$, where (x_n, y_n) and (x_h, y_h) are 9^{th} and 10^{th} body-point coordinates illustrated in Fig. 4.2 (a), respectively.

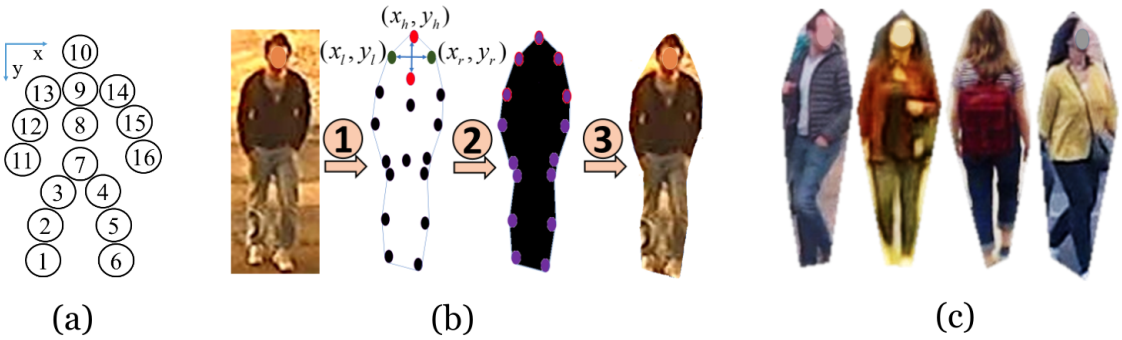


Figure 4.2: Foreground segmentation process. After determining the exterior border using the Convex-Hull, a mask is created and the foreground is cropped. (a) Body keypoints (b) Red points are considered as a reference for adding two green points near the head so that the polygon-crop contains the head and hair (c) Samples of segmented images which will have a black background in training phase.

The polygon-mask is then produced by painting inside of the obtained Convex-Hull with black, and this mask is employed to segment the raw images.
Considering that the facial region carries information about most human traits, including gender, we used different sets of body points such as the elbow, chest-bone, head, neck, and shoulders to crop the head. Under visual inspection, the best results are obtained using the head, chest-bone, and shoulders' points that have been shifted out ten pixels.

4.2.5 PSN and Score Fusion

The PSN is composed of three sub-networks, specialized in three poses (i.e. frontal, rear, and lateral poses). Using weight-sharing, the knowledge of the Base-Net is transferred to these sub-nets. For each image, there are three patches (i.e. head, polygon, whole image) corresponding to three PSNs (see Fig. 4.1). The obtained scores for each patch are then concatenated, and the highest one is selected as the final score of the image, which means that the model decides based on a optimistic perspective. For example, in case of partial body-occlusion and low score recognition for the full-body image, the model presumably decides based on the head-crop region.

4.3 Experiments and Discussion

First, we describe the strategy of the data collection and discuss the unique features of the collected dataset. We then briefly explain the two public datasets for which we evaluated our model. Finally, after describing the experimental settings, we provide the results.

4.3.1 Datasets

In general, deep-learning-based biometric systems are sensitive to data variability. Due to the environment and subject dynamics, a biometric system trained in a specific place cannot produce the best results in unseen places. This even becomes more critical in universal systems dealing with humans as the subject of interest, because not only the

Factors	Statistics
No. of videos, subjects, and BBs	216; 13,876; 503,433
Length of videos	7 minutes
Frame rate extraction	7 frames/sec.
Aspect ratio of BBs (Height/Width)	1.75
No. of BBs with frontal, rear, and lateral view	256,485; 235,564; 11,384

Table 4.1: Statistics of the BIODI dataset

environment alters, but the styles of clothing and body pose differ in various situations. For instance, the recognition rate will be highly affected in a cold and rainy night as people usually cover their bodies, heads, and faces while carrying an umbrella which has occluded the upper body. Therefore, regarding the lack of datasets that cover a wide range of variations in the environment and pedestrian, we collected the BIODI dataset from 36 advertisement panels in Portugal and Brazil at indoor and outdoor locations; different moments of the day including morning, noon, evening and night; and various weathers. Table 4.1 summarizes the statistics of this dataset. Each panel has one embedded camera with 1.5-meter vertical distance from the ground. Table 4.2 shows several samples of the BIODI dataset. It worth mentioning that this private dataset has been annotated manually for 16 soft biometric labels including gender, age, weight, race, height, hair color, hair style, beard, mustache, glasses, head attachments, upper-body cloths, lower-body clothes, shoes, accessories, and action.

To make our results reproducible, we report the performance of our method on public datasets such as PETA (excluding MIT) and MIT. Briefly, the MIT pedestrian dataset consists of 888 outdoor images with 64x128 pixels annotated for frontal and rear views. Approximately, half of the images are in frontal view, and female's share is one-third of the dataset. PETA is a collection of 19000 images consisting of 10 different datasets, including the MIT dataset. However, MIT is excluded from PETA since the proposed model will be evaluated on it, separately. It is worth mentioning that, in PETA benchmark, the number of males and females are almost the same and there is no view-wise annotation.

4.3.2 Experimental Settings

In our experiments, we use Python 3.5 and Keras 2.1.2 API on top of the Tensorflow 1.13. In order to avoid over fitting, we add the batch normalization, max pooling, and

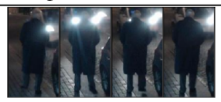
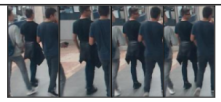
Description	Samples	Description	Samples	Description	Samples
Outdoor, Noon, Occlusion		Outdoor, Summer, Night		Outdoor, Winter, Night	
Outdoor, Fall, Evening		Outdoor, Summer, morning		Indoor, Spring, Occlusion	

Table 4.2: Sample images of the BIODI dataset that guarantees a wide spectrum of subject and environment changes.

Images	Network	BIODI	Frontal	Rear	Lateral	PETA	Frontal	Rear	Lateral
Raw	Base-Net	85.68	85.96	84.49	79.70	86.77	89.18	89.94	75.99
	Frontal-Net	-	87.53	-	-	-	90.56	-	-
	Rear-Net	-	-	85.18	-	-	-	93.06	-
	Lateral-Net	-	-	-	79.87	-	-	-	77.20
Head	Frontal-Net	-	88.42	-	-	-	88.73	-	-
	Rear-Net	-	-	85.13	-	-	-	90.15	-
	Lateral-Net	-	-	-	78.09	-	-	-	77.37
Polygon	Frontal-Net	-	90.44	-	-	-	91.29	-	-
	Rear-Net	-	-	87.44	-	-	-	91.44	-
	Lateral-Net	-	-	-	80.99	-	-	-	76.06
Fusion	Frontal-Net	-	**92.19**	-	-	-	**92.15**	-	-
	Rear-Net	-	-	**88.86**	-	-	-	**93.58**	-
	Lateral-Net	-	-	-	**84.16**	-	-	-	**80.16**

Table 4.3: Accuracy for the experiments on BIODI and PETA in percentage. The experiments on raw, head-crop, and polygon-crop images suggest that head-crop images provide the weakest results and confirm the fact that in surveillance scenarios, full-body recognition is more robust. Secondly, we perceived that as BIODI contains various environments, polygon segmentation provides better results while this is not true for PETA dataset. Finally, the last row of the table indicates that the adopted strategy for score fusion produces the highest score and accuracy among other approaches.

drop out layers to the ResNet50. The learning rate is set to 0.005 for the Stochastic Gradient Descent (SGD) optimizer. It is worth mentioning that we resized the images to 175 x 100 pixels, applied standardization per image, and performed horizontal mirror augmentation.

We evaluate the proposed model on three datasets BIODI, MIT, and PETA such that 70% of the BIODI (i.e. 352400 images) and PETA (i.e. 12680 images) datasets are allocated to the training phase. As MIT is a small dataset with 888 images, we used 50% of the data for test phase to have stable results because in each test-run the recognition rate have some variations.

4.3.3 Results and Discussion

Considering the explanations in the previous section, the experiments were conducted in three forms: raw images, head-cropped regions, and polygon-shape regions. Afterward, each trained model is tested. Table 4.3 shows the results of the proposed model on the RoIs which indicates that lateral-view state is the most difficult recognizable pose with around 84% and 80% accuracy for the BIODI and PETA datasets, respectively. Furthermore, Frontal-Net outperformed the Base-Net by 1.6% while Rear-Net improved the results from 84.49% to 85.18%, and Lateral-Net estimated the gender slightly better. Moreover, the increase of the 2% accuracy in polygon-crop images shows that the background negatively affects the performance of the networks. Hence, developing the powerful segmentation algorithms for human full-body is suitable for further studies.

Table 4.4 shows the evaluation of the proposed approach on MIT dataset. Notably, we achieve an average accuracy of 90.0%, 87.9%, and 89.0% for the frontal, rear, and mixed-

View	[9]	[10]	[11]	[12]	[1]	Proposed Method
Front	76.0	79.5	81.0	82.1	82.9	**90.0**
Back	74.6	84.0	82.7	81.3	81.8	**87.9**
Mixed	-	78.2	80.1	82.0	82.4	**89.0**

Table 4.4: Results on MIT test set in percentage.

view images, respectively, that are outperforming the results obtained by other methods.

4.4 Conclusions and Future Works

Regarding the ubiquitous surveillance cameras and the low-quality facial acquisitions, it is necessary to develop methods that deal with full-body images, occlusions, pose variation, and various illuminates. To this end, we proposed an algorithm for pedestrian gender recognition in crowded urban environments so that the output of a body-joints detector is applied for splitting the images into three common poses. Further, taking advantage of transfer learning, the specialized networks were fine-tuned for extracted RoIs. Extensive experiments on multiple challenging datasets showed that proposed PGRN can effectively estimate the gender and consistently outperform the state-of-the-art methods. As the next step, we have focused on developing an end-to-end network capable of estimating body related soft biometric traits such as weight, age, height, and race.

Bibliography

[1] M. Raza, M. Sharif, M. Yasmin, M. A. Khan, T. Saba, and S. L. Fernandes, "Appearance based pedestrians' gender recognition by employing stacked auto encoders in deep learning," *FUTURE GENER COMP SY*, vol. 88, pp. 28–39, 2018. 89, 96

[2] L. Cai, J. Zhu, H. Zeng, J. Chen, C. Cai, and K.-K. Ma, "Hog-assisted deep feature learning for pedestrian gender recognition," *Journal of the Franklin Institute*, vol. 355, no. 4, pp. 1991–2008, 2018. 89

[3] D. Nguyen, K. Kim, H. Hong, J. Koo, M. Kim, and K. Park, "Gender recognition from human-body images using visible-light and thermal camera videos based on a cnn for image feature extraction," *Sensors*, vol. 17, no. 3, p. 637, mar 2017. [Online]. Available: https://doi.org/10.3390/s17030637 90

[4] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in *Proc. ICCV*, 2017, pp. 2334–2343. 91

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99. 91

[6] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc NIPS - Volume 2*, ser. NIPS'15. Cambridge, MA, USA: MIT Press, 2015, p. 2017–2025. 91

[7] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *ECCV*. Springer, 2016, pp. 483–499. 91

[8] Y. Xiu, J. Li, H. Wang, Y. Fang, and C. Lu, "Pose flow: Efficient online pose tracking," *arXiv preprint arXiv:1802.00977*, 2018. 91

[9] L. Cao, M. Dikmen, Y. Fu, and T. S. Huang, "Gender recognition from body," in *Proceedings of the 16th ACM international conference on Multimedia*. ACM, 2008, pp. 725–728. 96

[10] G. Guo, G. Mu, and Y. Fu, "Gender from body: A biologically-inspired approach with manifold learning," in *ACCV*. Springer, 2009, pp. 236–245. 96

[11] C. D. Geelen, R. G. Wijnhoven, G. Dubbelman *et al.*, "Gender classification in low-resolution surveillance video: in-depth comparison of random forests and svms," in *VSTIA2015*, vol. 9407. International Society for Optics and Photonics, 2015, p. 94070M. 96

[12] M. Raza, C. Zonghai, S. Rehman, G. Zhenhua, W. Jikai, and B. Peng, "Part-wise pedestrian gender recognition via deep convolutional neural networks," in *2nd IET ICBISP*. Institution of Engineering and Technology, 2017. [Online]. Available: https://doi.org/10.1049/cp.2017.0102 96

Chapter 5

An Attention-Based Deep Learning Model for Multiple Pedestrian Attributes Recognition

Abstract. The automatic characterization of pedestrians in surveillance footage is a tough challenge, particularly with data acquisition conditions that are extremely diverse, cluttered backgrounds and subjects imaged from varying distances, under multiple poses, and partially occluded. Having observed that the state-of-the-art performance is still unsatisfactory, this paper provides a novel solution to the problem, with two-fold contributions: 1) considering the strong semantic correlation between the different full-body attributes, we propose a multi-task deep model that uses an element-wise multiplication layer to extract more comprehensive feature representations. In practice, this layer serves as a filter to remove irrelevant background features, and is particularly important to handle complex, cluttered data; and 2) we introduce a weighted-sum term to the loss function that not only relativizes the contribution of each task but also is crucial for performance improvement in multiple-attribute inference settings. Our experiments were performed in two well-known datasets (RAP and PETA) and point for the superiority of the proposed method with respect to the state-of-the-art. The code is available at `https://github.com/Ehsan-Yaghoubi/MAN-PAR-`.

5.1 Introduction

The automated inference of pedestrian attributes is a long-lasting goal in video surveillance and has been the scope of various research works [1] [2]. Commonly known as *pedestrian attribute recognition* (PAR), this topic is still regarded as an open problem, due to extremely challenging variability factors such as occlusions, viewpoint variations, low-illumination and low-resolution data (Fig. 5.1).

Deep learning frameworks have been repeatedly improving the state-of-the-art in many computer vision tasks, such as object detection and classification, action recognition and soft biometrics inference. In the PAR context, several models have been also proposed [3], [4], with most of these techniques facing particular difficulties to handle the heterogeneity of visual surveillance environments.

Researchers have been approaching the PAR problem from different perspectives [5]: [6], [7], [8] proposed deep learning models based on *full-body* images to address the data variation issues, while [9], [10], [11], [12] described *body-part* deep learning networks to consider the fine-grained features of the human body parts. Other works focused particularly the *attention mechanism* [13], [14], [11], and typically performed additional operations in the output of the mid-level and high-level convolutional layers. However, learning a comprehensive feature representation of pedestrian data, as the backbone for

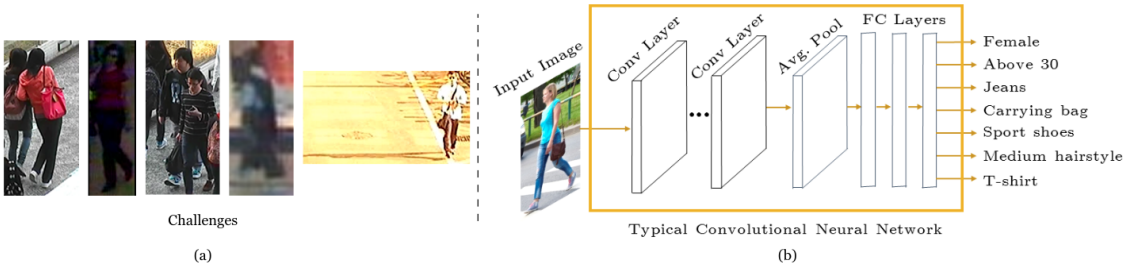


Figure 5.1: (a) Examples of some of the challenges in the PAR problem: crowded scenes, poor illumination conditions, and partial occlusions. (b) Typical structure of PAR networks, which receive a single image and perform labels inference.

all those approaches, still poses remarkable challenges, mostly resulting from the multi-label and multi-task intrinsic properties of PAR networks.

In opposition to previous works that attempted to jointly extract local, global and fine-grained features from the *input image*, in this paper, we propose a multi-task network that processes the *feature maps* and not only considers the correlation among the attributes but also captures the foreground features using a hard attention mechanism. The attention mechanism yields from the element-wise multiplication between the feature maps and a foreground mask that is included as a layer on top of the backbone feature extractor. Furthermore, we describe a weighted binary cross-entropy loss, where the weights are determined based on the number of categories (e.g., gender, ethnicity, age, ...) in each task. Intuitively, these weights control the contribution of each category during training, and are the key to avoid that some of the labels predominate over others, which was one of the major problems we identified in our evaluation on the previous works. In the empirical validation of the proposed method, we used two well-known PAR datasets (PETA and RAP) and three baseline methods considered to represent the state-of-the-art. The contributions of this work can be summarized as follows:

1. We propose a multi-task classification model for PAR that its main feature is to focus on the foreground (human body) features, attenuating the effect of background regions in the feature representations (Fig. 5.2);

2. We describe a weighted sum loss function that effectively handles the contribution of each category (e.g., gender, body figure, age, etc.) in the optimization mechanism, avoiding that some of the categories predominate over others during the inference step;

3. Inspired by the attention mechanism, we implement an element-wise multiplication layer that simulates a hard attention in the output of the convolutional layers, which particularly improves the robustness of feature representations in highly heterogeneous data acquisition environments.

The remainder of this paper is organized as follows: Section 5.2 summarises the PAR-related literature, and section 5.3 describes our method. In section 5.4, we provide the empirical validation details and discuss the obtained results. Finally, conclusions are provided in section 5.5.

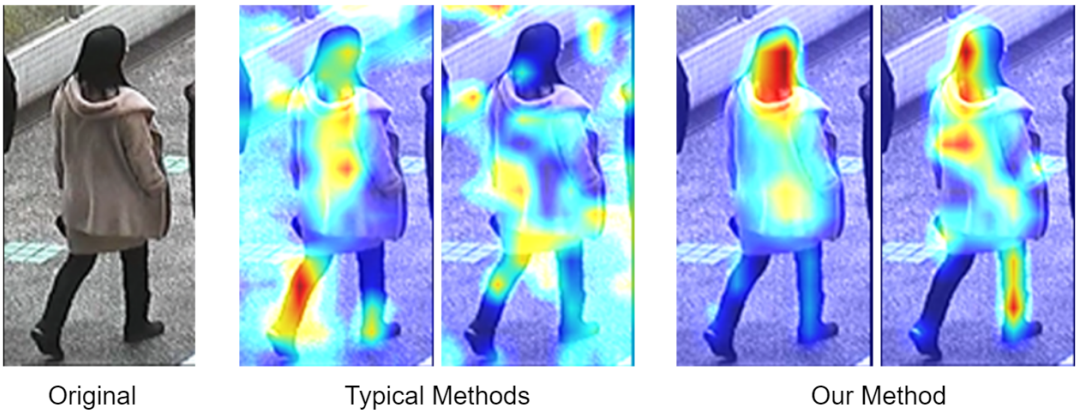


Figure 5.2: Comparison between the attentive regions obtained typically by previous methods [15], [16] and ours solution, while inferring the *Gender* attribute. Note the less importance given to background regions by our solution with respect to previous techniques.

5.2 Related Work

The ubiquity of CCTV cameras has been rising the ambition of obtaining reliable solutions for the automated inference of pedestrian attributes, which can be particularly hard in case of crowded urban environments. Given that face close-shots are rarely available at far distances, PAR upon full-body data is of practical interest. In this context, the earlier PAR methods focused individually on a single attribute and used handcrafted feature sets to feed classifiers such as SVM or AdaBoost[17], [18] [19]. More recently, most of the proposed methods were based on deep learning frameworks, and have been repeatedly advancing the state-of-the-art performance [20], [21], [22], [23].

In the context of deep learning, [24] proposed a multi-label model composed of several CNNs working in parallel, and specialized in segments of the input data. [6] compared the performance of single-label versus multi-label models, concluding that the semantic correlation between the attributes contributes to improve the results. [7] proposed a parameter sharing scheme over independently trained models. Subsequently, inspired by the success of Recurrent Neural Networks, [25] proposed a Long Short-Term Memory (LSTM) based model to learn the correlation between the attributes in low-quality pedestrian images. Other works also considered information about the subjects pose [26], body-parts [27] and viewpoint [9], [14], claiming to improve performance by obtaining better feature representations. In this context, by aggregating multiple feature maps from low, mid and high-level layers of the CNN, [28] enriched the obtained feature representation. For a comprehensive overview of the existing human attribute recognition approaches, we refer the readers to [5].

5.3 Proposed Method

As illustrated in Fig. 5.2, our main motivation is to provide a PAR pipeline that is robust to background-based irrelevant features, which should contribute for improvements

in performance, particularly in crowded scenes that partial occlusions of human body silhouettes occur (Fig. 5.1 (a) and Fig. 5.2).

5.3.1 Overall Architecture

Fig. 5.3 provides an overview of the proposed model, inferring the complete set of attributes of a pedestrian at once, in a single-shot paradigm. Our pipeline is composed of four main stages: 1) the *convolutional layers*, as general feature extractors; 2) the *body segmentation module*, that is responsible to discriminate between the foreground/background regions; 3) the *multiplication layer*, that in practice implements the attention mechanism; and 4) the *task-oriented branches*, that avoid the predominance of some of the labels over others in the inference step.
At first, the input image feeds a set of convolutional layers, where the local and global features are extracted. Next, we use the body segmentation module to obtain the binary mask of the pedestrian body. This mask is used to remove the background features, by an element-wise multiplication with the feature maps. The resulting features (that are free of background noise) are then compressed using an average pooling strategy. Finally, for each *task*, we add different fully connected layers on top of the network, not only to leverage the useful information from other tasks but also to improve the generalization performance of the network. We have adopted a multi-task network, because the shared convolutional layers extract the *common* local and global features that are necessary for all the tasks (i.e., behavioral attributes, regional attributes, and global attributes) and then, there are separate branches that allow the network to focus on the most important features for each task.

5.3.2 Convolutional Building Blocks

The implemented convolution layers are based on the concept of residual block. Considering $\boldsymbol{x}$ as the input of a conventional neural network, we want to learn the true distribution of the output $H(\boldsymbol{x})$. Therefore, the difference (residual) between the input and output is $R(\boldsymbol{x}) = H(\boldsymbol{x}) - \boldsymbol{x}$, and can be rearranged to $H(\boldsymbol{x}) = R(\boldsymbol{x}) + \boldsymbol{x}$. In other words, traditional network layers learn the true output $H(\boldsymbol{x})$, whereas residual network layers learn the residual $R(\boldsymbol{x})$. It is worth mentioning that it is easier to learn the residual of the output and input, rather than only the true output [29]. In fact, residual-based networks have the degree of freedom to train the layers in residual blocks or skip them. As the optimal number of layers depends on the complexity of the problem under study, adding skip connections makes the neural network active in training the useful layers.
There are various types of residual blocks made of different arrangements of the Batch Normalization (BN) layer, activation function, and convolutional layers. Based on the analysis provided in [30], the forward and backward signals can directly propagate between two blocks, and optimal results will be obtained when the input $\boldsymbol{x}$ is used as skip connection (Fig. 5.4).

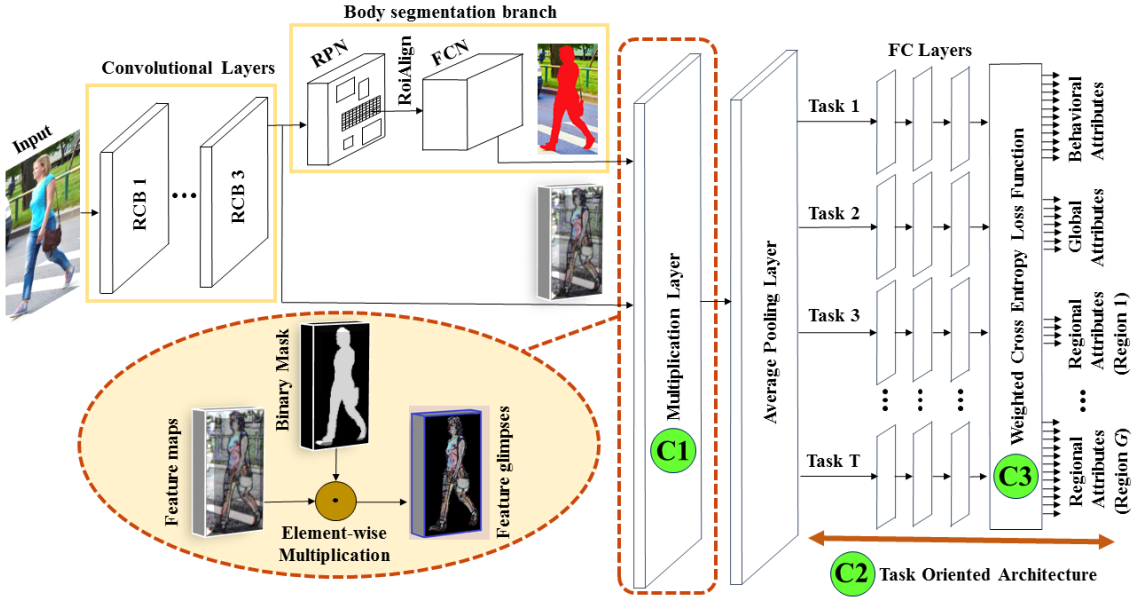


Figure 5.3: Overview of the major contributions (Ci) in this paper. C1) the element-wise multiplication layer receives a set of feature maps $\boldsymbol{F}_{H\times W\times D}$ and a binary mask $\boldsymbol{M}_{H\times W\times D}$, and outputs a set of *attention glimpses*. C2) The multitask-oriented architecture provides to the network the ability to focus on the local (e.g., head accessories, types of shoes), behavioral (e.g., talking, pushing), and global (e.g., age, gender) features (visual results are given in Fig. 5.7). C3) a weighted cross-entropy loss function not only considers the interconnection between the different attributes, but also handles the contribution of each label in the inference step. Residual Convolutional Block (RCB) is the abbreviation for Residual Convolutional Block, illustrated in Fig. 5.4. Region Proposal Network (RPN), Fully Connected Network (FCN), and Fully Connected Layer (FCL) stand for Region Proposal Network, Fully Connected Network, and Fully Connected Layer, respectively.

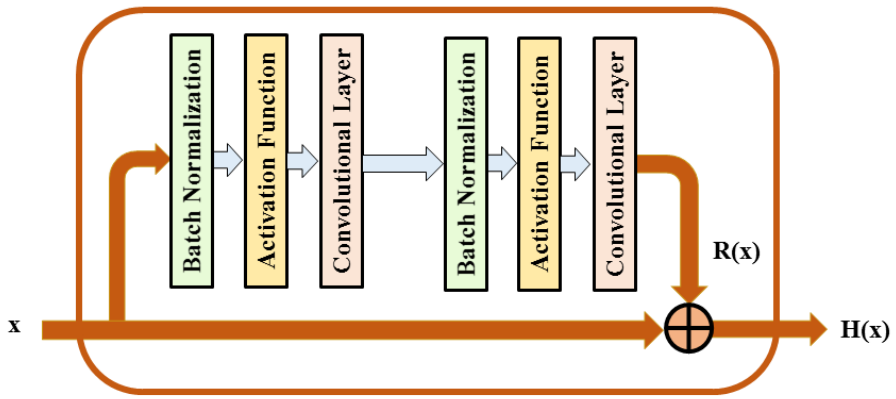


Figure 5.4: Residual convolutional block in which the input x is considered a skip connection.

5.3.3 Foreground Human Body Segmentation Module

We used the Mask R-CNN [31] model to obtain the full body human masks. This method adopts a two-stage procedure after the convolutional layers: i) a RPN [32] that provides several possibilities for the object bounding boxes, followed by an alignment layer; and ii) a FCN [33] that infers the bounding boxes, class probabilities, and the segmentation masks.

5.3.4 Hard Attention: Element-wise Multiplication Layer

The idea of an attention mechanism is to provide the neural network with the ability to focus on a feature subset. Let $\boldsymbol{I}$ be an input image, $\boldsymbol{F}$ the corresponding feature maps, $\boldsymbol{M}$ an attention mask, $f_\phi(\boldsymbol{I})$ an attention network with parameters ϕ, and $\boldsymbol{G}$ an attention glimpse (*i.e.*, the result of applying an attention mechanism to the image $\boldsymbol{I}$). Typically, the attention mechanism is implemented as $F = f_\phi(\boldsymbol{I})$, and $\boldsymbol{G} = \boldsymbol{M} \odot \boldsymbol{F}$, where $\odot$ is an element-wise multiplication. In soft attention, features are multiplied with a mask of values between zero and one, while in the hard attention variant, values are binarized and - hence - they should be fully considered or completely disregarded.
In this work, as we produce the foreground binary masks, we applied a hard attention mechanism on the output of the convolutional layers. To this end, we used an element-wise multiplication layer that receives a set of feature maps $\boldsymbol{F}_{H\times W\times D}$ and a binary mask $\boldsymbol{M}_{H\times W\times D}$, and returns a set of attention glimpses $\boldsymbol{G}_{H\times W\times D}$, in which H, W, and D are the height, weight, and the number of the feature maps, respectively.

5.3.5 Multi-Task CNN Architecture and Weighted Loss Function

We consider multiple soft label *categories* (e.g., gender, age, lower-body clothing, ethnicity and hairstyle), with each of these including two or more *classes*. For example, the category of *lower-body clothing* is composed of 6 classes: {'pants', 'jeans', 'shorts', 'skirt', 'dress', 'leggings'}. As stated above, there are evident semantic dependencies between most of the labels (e.g., it is not likely that someone uses a 'dress' and 'sandals' at the same time). Hence, to model these relations between the different categories, we use a hard parameter sharing strategy[34] in our multi-task residual architecture. Let T, C_t, K_c, N_k be the number of tasks, the number of categories (labels) in each task, the number of classes in each category, and the number of samples in each class, respectively.
During the learning phase, the model $\mathcal{H}$ receives one input image $\boldsymbol{I}$, its binary mask $\boldsymbol{S}$, the ground truth labels $\boldsymbol{Y}$, and returns $\hat{\boldsymbol{Y}}$ as the predicted attributes (labels):

$$\hat{\boldsymbol{Y}} = \left\{ \begin{array}{c} \hat{y}_{t,c_t,k_t} | t \in \left\{1,...,T\right\}, c \in \left\{1,...,C_t\right\}, k \in \left\{1,...,K_c\right\}, \\ T, C_t, K_c \in \mathbb{N}, \hat{y}_i \in \left\{1,0\right\} \end{array} \right\}, \quad (5.1)$$

in which $\hat{\boldsymbol{y}}_{t,c,k}$ denotes the predicted attributes.
The key concept of the learning process is the loss function. In the single attribute

recognition[35] setting, if the n-th image $\boldsymbol{I}_n$, $(n = 1, ..., N)$ is characterized by the m-th attribute, $(m = 1, ..., M)$, then $\boldsymbol{y}_{nm} = 1$; otherwise, $\boldsymbol{y}_{nm} = 0$. In case of having multiple attributes (multi-task), the predicting functions are in the form of $\Phi = \Big\{\Phi_1, \Phi_2, ..., \Phi_m, ..., \Phi_M\Big\}$, and $\Phi_m(\boldsymbol{I'}) \in \Big\{1, 0\Big\}$. We define the minimization of the loss function over the training samples for the mth attribute as:

$$\Psi_m = argmin_{\Psi_m} \sum_{n=1}^{N} \mathcal{L}\Big(\Phi_m(I_n, \Psi_m), y_{nm}\Big), \quad (5.2)$$

where Ψ_m contains a set of optimized parameters related to the m-th attribute, while $\Phi_m(I_n, \Psi_m)$ returns the predicted label ($\hat{\boldsymbol{y}}_{\boldsymbol{nm}}$) for the m-th attribute of the image I_n. Besides, $\mathcal{L}(.)$ is the loss function that measures the difference between the predictions and ground-truth labels.

Considering the interconnection between attributes, one can define a unified multi-attribute learning model for all the attributes. In this case, the loss function jointly considers all the attributes:

$$\Psi = argmin_{\Psi} \sum_{m=1}^{M} \sum_{n=1}^{N} \mathcal{L}\Big(\Phi_m(I_n, \Psi_m), y_{nm}\Big), \quad (5.3)$$

in which Ψ contains the set of optimized parameters related to all attributes.

In opposition to the above mentioned functions, in order to consider the contribution of each category in the loss value, we define a weighted sum loss function:

$$\Psi = argmin_{\Psi} \sum_{t=1}^{T} \sum_{c=1}^{C_t} \sum_{k=1}^{K_c} \sum_{n=1}^{N_k} \frac{1}{\mathcal{R}_c} \mathcal{L}\Big(\Phi_{tck}(I_n, \Psi_{tck}), y_{tckn}\Big), \quad (5.4)$$

where $\mathcal{R}_c \in \{R_1, ..., R_{C_t}\}$ are scalar values corresponding to the number of classes in the categories $1, ..., C_t$.

Using the *sigmoid* activation function for all classes in each category, we can formulate the *cross-entropy* loss function as:

$$Loss = -\sum_{t=1}^{T} \sum_{c=1}^{C_t} \sum_{k=1}^{K_c} \sum_{n=1}^{N_k} \frac{1}{n\mathcal{R}_c} \Big(y_{tckn} log(\hat{p}_{tckn}) + (1 - y_{tckn}) log(1 - \hat{p}_{tckn})\Big), \quad (5.5)$$

where $\boldsymbol{y}_{tckn}$ is the binary value that relates the class label k in category c. The ground-truth label for observation n and $\hat{\boldsymbol{p}}_{tckn}$ is the predicted probability of the observation n.

5.4 Experiments and Discussion

The proposed PAR network was evaluated on two well-known datasets: the PETA [17] and the Richly Annotated Pedestrian (RAP) [15], with both being among the most frequently

Table 5.1: RAP dataset annotations

Branch	Annotations
Soft Biometrics	Gender, Age, Body figure, Hairstyle, Hair color
Clothing Attributes	Hat, Upper body clothes style and color, Lower body clothes style and color, Shoe style
Accessories	Glasses, Backpack, Bags, Box
Action	Telephoning, Talking, Pushing, Carrying, Holding, Gathering

used benchmarks in PAR experiments.

5.4.1 Datasets

RAP [15] is the largest and the most recent dataset in the area of surveillance, pedestrian recognition, and human re-identification. It was collected at an indoor shopping mall with 25 High Definition (HD) cameras (spatial resolution $1,280 \times 720$) during one month. Benefiting from a *motion detection and tracking algorithm*, authors have processed the collected videos, which resulted in 84,928 human full-body images. The resulting bounding boxes vary in size from 33×81 to 415×583. The annotations provide information about the viewpoint ('front', 'back', 'left-side', and 'right-side'), body occlusions and body-part pose, along with a detailed specification of the train-validation-test partitions, person ID, and 111 binary human attributes. Due to the unbalanced distribution of the attributes and insufficient data for some of the classes, only 55 of these binary attributes were selected [15]. Table 5.1 shows the categories of these attributes. It is worth mentioning that, as the annotation process is performed per subject instance, the same identity may have different attribute annotations in distinct samples.
PETA [17] contains ten different pedestrian image collections gathered in outdoor environments. It is composed of 19,000 images corresponding to 8,705 individuals, each one annotated with 61 binary attributes, from which 35 were considered with enough samples and selected for the training phase. Camera angle, illumination, and the resolution of images are the particular variation factors in this set.

5.4.2 Evaluation Metrics

PAR algorithms are typically evaluated based on the standard classification accuracy per attribute, and on the mean accuracy ($\overline{mA}$) of the attribute. Further, the mean accuracy over all attributes was also used [36], [37]:

$$\overline{mA} = \frac{1}{2M} \sum_{m=1}^{M} \left(\frac{\hat{\mathcal{P}}_m}{\mathcal{P}_m} + \frac{\hat{\mathcal{N}}_m}{\mathcal{N}_m} \right), \quad (5.6)$$

where m denotes one attribute and M is the total number of attributes. For each attribute m, $\mathcal{P}_m$, $\mathcal{N}_m$, $\hat{\mathcal{P}}_m$, and $\hat{\mathcal{N}}_m$ stand for the number of positive samples, negative samples, correctly recognized as positive samples, correctly recognized as negative samples.

Table 5.2: Parameter Settings for the performed experiments on the RAP dataset.

Parameter	Value
Image input shape	$256 \times 256 \times 3$
Mask input shape	$16 \times 16 \times 3$
Learning rate	$1 \times e^{-4}$
Learning decay	$1 \times e^{-6}$
Number of epochs	200
Drop-out probability	0.7
Batch size	8

5.4.3 Preprocessing

RAP and PETA samples vary in size, with each image containing exclusively one subject annotated. Therefore, to have constant ratio images, we first performed a zero padding and then, resized them into 256×256. It worth mentioning that, after each residual block, the input size is divided by 2. Therefore, as we have implemented the backbone with 4 residual stages, to multiply the binary mask and feature maps with a size of 16×16, the input size should be 256×256. Note that the sharp edges caused by these zero pads do not affect the network due to the presence of the *multiplication layer* before the classification layers.

To assure a fair comparison between the tested methods, we used the same train-validation-test splits as in [15]: $50,957$ images were used for learning, $16,986$ for validation purposes, and the remaining 16,985 images used for testing. The same strategy was used for the PETA dataset. Table 5.2 shows the parameter settings of our multi-task network.

5.4.4 Implementation Details

Our method was implemented using Keras $2.2.5$ with Tensorflow $1.12.0$ backend [38], and all the experiments were performed on a machine with an Intel Core $i5-8600$K CPU @ 3.60 GHz (Hexa Core | 6 Threads) processor, NVIDIA GeForce RTX 2080 Ti GPU, and 32 GB RAM.

The proposed CNN architecture was fulfilled as a dual-step network. At first, we applied the body segmentation network (*i.e.*, Mask R-CNN, explained in the next subsection) to extract the human full-body masks, and then trained a two-input multi-task network that receives the preprocessed masks and the input data. It is worth mentioning that, on account of the spreading or gathering nature of the attributes features in the full-body human images, we intuitively clustered all the binary attributes into 7 and 6 groups for the experiments on RAP and PETA, respectively, as given in Table 5.3.

As stated above, we used the pre-trained Mask R-CNN [39] to obtain all the foreground masks in our experiments. The used segmentation model was trained in the MS-COCO dataset [40]. Table 5.4 provides the details of our implementation settings.

By feeding the input images to the convolutional building blocks, we obtain a set of feature maps that will be multiplied by the corresponding mask, using the element-wise multiplication layer. This layer receives two inputs with the same shapes. Transferring the input data with shape of $256 \times 256 \times 3$ into a 4-residual block backbone, we obtain

Table 5.3: Task specification policy for the PETA and RAP datasets.

Dataset	Task 1 (Full Body)	Task 2 (Head)	Task 3 (Upper Body)	Task 4 (Lower Body)	Task 5 (Foot wears)	Task 6 (Accessories)	Task 7 (Action)
PETA	Female, Male, AgeLess30, AgeLess45, AgeLess60, AgeLarger60	Hat, LongHair, Scarf, Sunglasses, Nothing	Casual, Formal, Jacket, Logo, Plaid, ShortSleeves, Strip, Tshirt, Vneck, Other	Casual, Formal, Jeans, Shorts, ShortSkirt, Trousers	LeatherShoes, Sandals, FootwearShoes, Sneaker	Backpack , MessengerBag, PlasticBags, CarryingNothing, CarryingOther	-
RAP	Female, Male, AgeLess16, Age17-30, Age31-45, Age46-60, BodyFat, BodyNormal, BodyThin, Customer, Employee	BaldHead, LongHair, BlackHair, Hat, Glasses	Shirt, Sweater, Vest, TShirt, Cotton, Jacket, SuitUp, Tight, ShortSleeves, Others	LongTrousers, Skirt, ShortSkirt, Dress, Jeans, TightTrousers	Leather, Sports, Boots, Cloth, Casual, Other	Backpack, ShoulderBag, HandBag, Box, PlasticBag, PaperBag, HandTrunk, Other	Calling, Talking, Gathering, Holding, Pushing, Pulling, CarryingByArm, CarryingByHand

Table 5.4: Mask R-CNN parameter settings

Parameter	Value
Image input dimension	$1024 \times 1024 \times 3$
RPN anchor scales	32, 64, 128, 256, 512
RPN anchor ratio	0.5, 1, 2
Number of proposals per image	256

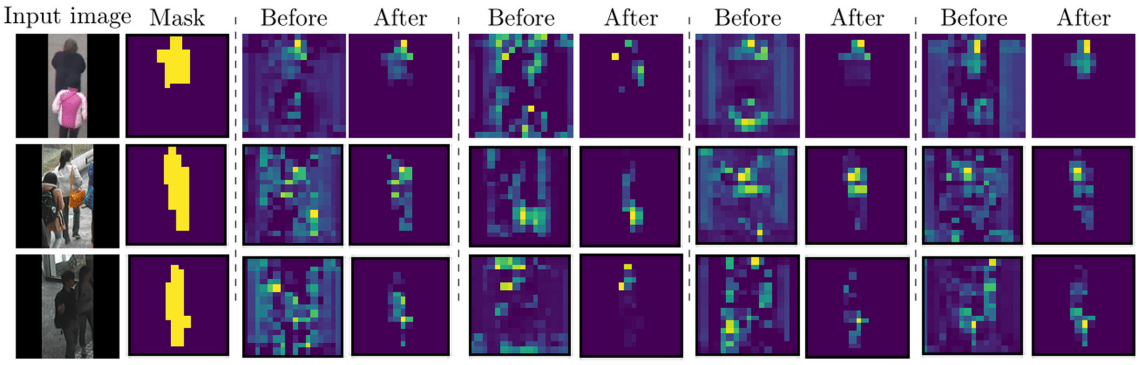


Figure 5.5: The effectiveness of the *multiplication layer* on filtering the background features from the feature maps. The far left column shows the input images to the network, the *Mask* column presents the ground truth binary mask (the first input of the multiplication layer), the columns with *Before* label (the second input of the multiplication layer) display the feature maps before applying the multiplication operation, and the columns with *After* label show the output of the multiplication layer.

a $16 \times 16 \times 1,024$-shaped output. Also, masks are resized to have the same size as the corresponding feature maps. Therefore, as a result of multiplying the binary mask and feature maps, we obtain a set of attention glimpses with the $16 \times 16 \times 1,024$ shape. These *glimpses* are down-sampled to $1,024$ features using a global average pooling layer to decrease the sensitivity of the locations of the features in the input image [41]. Afterward, in the interest of training one classifier for each task, a $Dense[ReLU] \rightarrow DropOut \rightarrow Dense[ReLU] \rightarrow DropOut \rightarrow Dense[ReLU] \rightarrow Dense[Sigmoid]$ architecture is stacked on top of the shared layers for each task.

5.4.5 Comparison with the State-of-the-art

We compared the performance attained by our method to three baselines, that were considered to represent the state-of-the-art: Attributes Convolutional Net (ACN) [7], DeepMar [15], and Multi-Label Convolutional Neural Network (MLCNN) [16] on the RAP and the PETA datasets. These methods have been selected for two reasons: 1- in a way similar to our method, ACN and DeepMar are global-based methods (i.e., they extract features from the full-body images) 2- Authors of these methods have reported the results for all the attributes in a separate way, assuring a fair comparison between the performance of all methods.

As the solution proposed in this paper, the ACN [7] method analyzes the full-body images and jointly learns all the attributes without relying on additional information. DeepMar [15] is a global-based end-to-end CNN model that provides all the binary labels for the input image, simultaneously. In [16], authors propose a MLCNN that divides the input

Table 5.5: Comparison between the results observed in the PETA dataset (mean accuracy percentage). The highest accuracy values per attribute among all methods appear in bold.

Attributes	DeepMar [15]	MLCNN [16]	Proposed
Male	89.9	84.3	**91.2**
AgeLess30	**85.8**	81.1	85.3
AgeLess45	81.8	79.9	**82.7**
AgeLess60	86.3	92.8	**93.9**
AgeLarger60	94.8	97.6	**98.6**
Head-Hat	91.8	96.1	**97.4**
Head-LongHair	88.9	88.1	**92.3**
Head-Scarf	96.1	97.2	**98.2**
Head-Nothing	85.8	86.1	**90.7**
UB-Casual	84.4	89.3	**93.4**
UB-Formal	85.1	91.1	**94.6**
UB-Jacket	79.2	92.3	**95.0**
UB-ShortSleeves	87.5	88.1	**93.4**
UB-Tshirt	83.0	90.6	**93.8**
UB-Other	**86.1**	82.0	84.8
LB-Casual	84.9	90.5	**93.7**
LB-Formal	85.2	90.9	**94.0**
LB-Jeans	85.7	83.1	**86.7**
LB-Trousers	**84.3**	76.2	78.9
Shoes-Leather	87.3	85.2	**89.8**
Shoes-Footwear	**80.0**	75.8	79.8
Shoes-Sneaker	78.7	81.8	**86.6**
Backpack	82.6	84.3	**89.2**
MessengerBag	82.0	79.6	**86.3**
PlasticBags	87.0	93.5	**94.5**
Carrying-Nothing	83.1	80.1	**85.9**
Carrying-Other	77.3	**80.9**	78.8
Average of 27 Attributes	85.4	86.6	**90.0**
Average of 35 Attributes	82.6	-	**91.7**

image into overlapped parts and fuses the features of each CNN to provide the binary labels for the pedestrians. Tables 5.5 and 5.6 provide the obtained results observed for the three methods considered in the PETA and RAP datasets.

Table 5.5 shows the evaluation results of the DeepMar and MLCNN methods, including our model on the PETA dataset. According to this table, our model shows superior recognition rates for 22 (out of 27) attributes, concluded to more than 3% improvement in total accuracy. If we consider 35 attributes, the proposed network achieves a 91.7% recognition rate while this value for the DeepMar approach is 82.6%.

The experiment carried out without considering image augmentation (*i.e.*, 5-degree rotation, horizontal flip, 0.02 width and height shift range, 0.05 shear range, 0.08 zoom range and changing the brightness in the interval [0.9,1.1]), showed 85.5% and 88.2% average accuracy for 27 and 35 attributes, respectively. We augmented the images randomly, and after the visualization of some images, we determined the values in augmentations.

As shown in Table 5.6, the average recognition rates for the ACN and DeepMar methods respectively were 68.92% and 75.54%, while our approach achieved more than 92%. In particular, excluding five attributes (*i.e.*, *Female*, *Shirt*, *Jacket*, *Long Trousers*, and *Other*

class in attachments category), our PAR model provides notoriously better results than the DeepMar method, and better than the ACN model in all cases.
The proposed method shows superior results in both datasets; however, in 22 attributes of the RAP benchmark, the recognition percentage is yet less than 95%, and in 7 cases, this rate is even less than 80%. The same interpretation is valid for the PETA dataset as well, which indicates the demands of more research works in the PAR field of study.

5.4.6 Ablation Studies

In this section, we study the effectiveness of the mentioned contributions in Fig. 5.3. To this end, we trained and tested a light version of the network (with three residual blocks and input image size 128×128) on the PETA dataset with similar initialization, but different settings (Table 5.7). The first row of Table 5.7 shows the performance of a network, constructed from three residual blocks with four shared fully connected layers on top, plus one fully connected layer for each attribute. In this architecture, as the system cannot decide on each task independently, the performance is poor (81.11%), and the network cannot predict the uncorrelated attributes (e.g., behavioral attributes versus appearance attributes) effectively. However, the results in the second row of Table 5.7 show that repeating the fully connected layers for each task independently (while keeping the rest of the architecture unchanged), improves the results by around 8%. Further, equipping the network with the proposed weighted loss function (Table 5.7, row 3) and adding the *Multiplication layer* (Table 5.7, row 4) showed further improvements in the performance to 89.35% and 89.73%, respectively.
Feature map visualization. Neural networks are known as poorly interpretable models. However, as the internal structures of the CNNs are designed to operate upon two-dimensional images, they preserve the spatial relationships for what it is being learned [42]. Hence, by visualizing the operations on each layer, we can understand the behavior of the network. As a result of slicing the small linear filters over the input data, we obtain the activation maps (feature maps). To analyze the behavior of the proposed *multiplication layer* (Fig. 5.3), we visualized the input and output feature maps in Fig. 5.5, such that the columns labeled as *Mask* and *Before* refer to the inputs of the layer, and the columns labeled as *After* show the multiplication results of the two inputs. As it is evident, unwanted features resulting from the partial occlusions were filtered from the feature map, which improved the overall performance of the system.
Where is the network looking at? As a general behavior, CNNs infer what could be the optimal local/global features of a training set and generalize them to decide on unseen data. Here, partial occlusions can easily affect this behavior and decrease the performance, being helpful to understand where the model are actually looking at in the prediction phase. To this end, we plot some heat maps to investigate the effectiveness of the proposed *multiplication layer* and *task-oriented architecture*. Heat maps are easily understandable and highlight the regions on which the network focuses while making a prediction.
Fig. 5.6 shows the behavior of the system under examples with partial occlusions. As

Table 5.6: Comparison of the results observed in the RAP dataset (mean accuracy percentage).

Attributes	ACN [7]	DeepMar [15]	Proposed
Female	94.06	**96.53**	96.28
AgeLess16	77.29	77.24	**99.25**
Age17-30	69.18	69.66	**69.98**
Age31-45	66.80	66.64	**67.19**
Age46-60	52.16	59.90	**96.88**
BodyFat	58.42	61.95	**87.24**
BodyNormal	55.36	58.47	**78.20**
BodyThin	52.31	55.75	**92.82**
Customer	80.85	82.30	**96.98**
Employee	85.60	85.73	**97.67**
BaldHead	65.28	80.93	**99.56**
LongHair	89.49	92.47	**94.67**
BlackHair	66.19	79.33	**94.94**
Hat	60.73	84.00	**99.02**
Glasses	56.30	84.19	**96.76**
UB-Shirt	81.81	**85.86**	83.93
UB-Sweater	56.85	64.21	**92.66**
UB-Vest	83.65	89.91	**96.91**
UB-TShirt	71.61	75.94	**77.17**
UB-Cotton	74.67	79.02	**89.48**
UB-Jacket	78.29	**80.69**	71.93
UB-SuitUp	73.92	77.29	**97.18**
UB-Tight	61.71	68.89	**96.10**
UB-ShortSleeves	88.27	90.09	**90.79**
UB-Others	50.35	54.82	**97.91**
LB-LongTrousers	86.60	**86.64**	84.88
LB-Skirt	70.51	74.83	**97.37**
LB-ShortSkirt	73.16	72.86	**98.10**
LB-Dress	72.89	76.30	**97.34**
LB-Jeans	90.17	89.46	**91.56**
LB-TightTrousers	86.95	87.91	**94.71**
Backpack	68.87	80.61	**98.03**
ShoulderBag	69.30	82.52	**93.29**
HandBag	63.95	76.45	**97.64**
Box	66.72	76.18	**96.30**
PlasticBag	61.53	75.20	**97.78**
PaperBag	52.25	63.34	**99.07**
HandTrunk	79.01	84.57	**97.74**
Other	66.14	**76.14**	71.54
Calling	74.66	86.97	**97.13**
Talking	50.54	54.65	**97.54**
Gathering	52.69	58.81	**95.47**
Holding	56.43	64.22	**97.71**
Pushing	80.97	82.58	**99.15**
Pulling	69.00	78.35	**98.24**
CarryingByArm	53.55	65.40	**97.77**
CarryingByHand	74.58	82.72	**87.57**
Other	54.83	58.79	**99.13**
Average	68.92	75.54	**92.23**

Table 5.7: Ablation studies. The first row shows our baseline system with a multi-label architecture and binary-cross-entropy loss function, while the other rows indicate the proposed system with various settings.

Multi-task architecture	Multiplication Layer	Weighted Loss (Binary-cross-entropy)	mAP (%)
-	-	-	81.11
✓	-	-	89.18
✓	-	✓	89.35
✓	✓	-	89.73

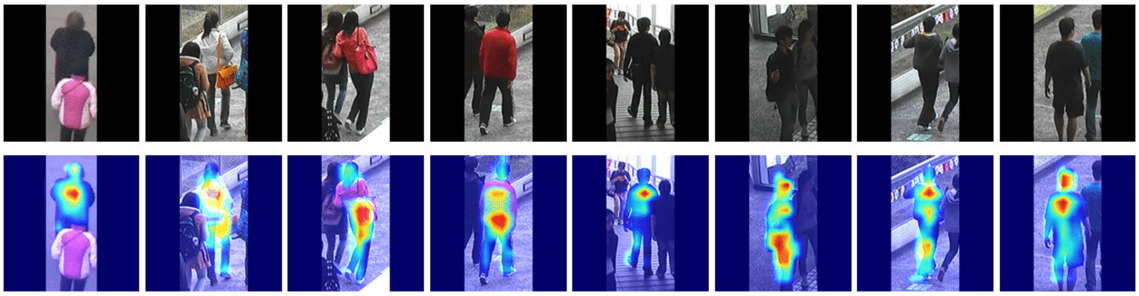


Figure 5.6: Illustration of the effectiveness of the *multiplication layer* upon the focus ability of the proposed model in case of partial occlusions. Samples regard the PETA dataset, with the network predicting the *age* and *gender* attributes.

it is seen, the proposed network is able to filter the harmful features of the distractors effectively, while focusing on the target subject. Moreover, Fig. 5.7 shows the model behavior during the attribute recognition in each task.

Loss Function. Table 5.8 provides the performance of the proposed network, when using different loss functions suitable for binary classification. Focal loss [43] forces the network to concentrate on hard samples, while the weighted Binary Cross-Entropy (BCE) loss [6] allocates a specific binary weight to each class. Training the network using binary focal loss function showed 79.30% accuracy in the test phase, while this number was 90.19% for the weighted BCE loss (see Table 5.8).

The proposed weighted loss function uses the BCE loss function, while recommends different weights for each *class*. We further trained the proposed model with the binary focal loss function using the proposed weights. The results in Table 5.8 indicate a slight improvement in the performance when we train the network using the proposed weighted loss function with BCE (90.34%).

Table 5.8: Performance of the network trained with different loss functions on the PETA dataset.

Loss Function	mAP (%)
Binary focal loss function [43]	79.30
Weighted BCE loss function [6]	90.19
Proposed weighted loss function (with BCE)	90.34
Proposed weighted loss function (with binary focal loss)	89.27

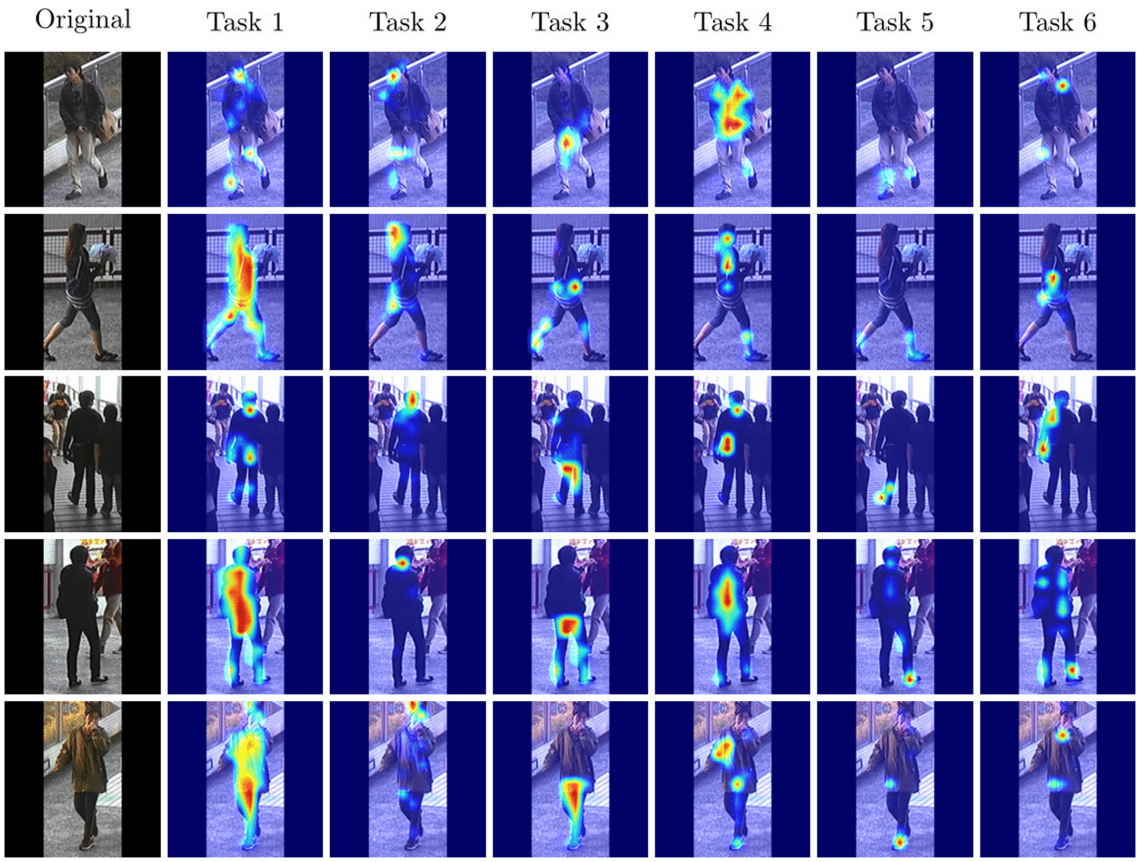


Figure 5.7: Visualization of the heat maps resulting of the proposed multi-task network. Sample regard the PETA dataset. The leftmost column shows the original samples, the column *Task 1* (i.e., recognizing *age* and *gender*) presents the effectiveness of the network focus on the human full-body, and the remaining columns display the ability of the system on region-based attribute recognition. The task policies are given in Table 5.3.

5.5 Conclusions

Complex background clutter, viewpoint variations and occlusions are known to have a noticeable negative effect on the performance of person attribute recognition (PAR) methods. According to this observation, in this paper, we proposed a deep-learning framework that improves the robustness of the obtained feature representation by directly discarding the background regions in the fully connected layers of the network. To this end, we described an element-wise multiplication layer between the output of the residual convolutional layers and a binary mask representing the human full-body foreground. Further, the refined feature maps were down-sampled and fed to different fully connected layers, that each one is specialized in learning a particular task (i.e., a subset of attributes). Finally, we described a loss function that weights each category of attributes to ensure that each attribute receives enough attention, and there are not some attributes that bias the results of others. Our experimental analysis in the PETA and RAP datasets pointed for solid improvements in performance of the proposed model with respect to the state-of-the-art.

Bibliography

[1] A. B. Mabrouk and E. Zagrouba, "Abnormal behavior recognition for intelligent video surveillance systems: A review," *Expert Systems with Applications*, vol. 91, pp. 480–491, 2018. 99

[2] J. Kumari, R. Rajesh, and K. Pooja, "Facial expression recognition: A survey," *Procedia Computer Science*, vol. 58, pp. 486–491, 2015. 99

[3] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015. 99

[4] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, 2017. 99

[5] X. Wang, S. Zheng, R. Yang, B. Luo, and J. Tang, "Pedestrian attribute recognition: A survey," *arXiv preprint arXiv:1901.07474*, 2019. 99, 101

[6] D. Li, X. Chen, and K. Huang, "Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios," in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*. IEEE, 2015, pp. 111–115. 99, 101, 113

[7] P. Sudowe, H. Spitzer, and B. Leibe, "Person attribute recognition with a jointly-trained holistic cnn model," in *Proc. ICCVW*, 2015, pp. 87–95. 99, 101, 109, 112

[8] A. H. Abdulnabi, G. Wang, J. Lu, and K. Jia, "Multi-task cnn model for attribute prediction," *IEEE Trans. Multimed.*, vol. 17, no. 11, pp. 1949–1959, 2015. 99

[9] P. Liu, X. Liu, J. Yan, and J. Shao, "Localization guided learning for pedestrian attribute recognition," *arXiv preprint arXiv:1808.09102*, 2018. 99, 101

[10] G. Gkioxari, R. Girshick, and J. Malik, "Actions and attributes from wholes and parts," in *Proc. ICCV*, 2015, pp. 2470–2478. 99

[11] Y. Li, C. Huang, C. C. Loy, and X. Tang, "Human attribute recognition by deep hierarchical contexts," in *Proc. ECCV*. Springer, 2016, pp. 684–700. 99

[12] Y. Chen, S. Duffner, A. STOIAN, J.-Y. Dufour, and A. Baskurt, "Pedestrian attribute recognition with part-based CNN and combined feature representations," in *VISAPP2018*, Funchal, Portugal, Jan. 2018. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01625470 99

[13] N. Sarafianos, X. Xu, and I. A. Kakadiaris, "Deep imbalanced attribute classification using visual attention aggregation," in *Proc. ECCV*, 2018, pp. 680–697. 99

[14] M. S. Sarfraz, A. Schumann, Y. Wang, and R. Stiefelhagen, "Deep view-sensitive pedestrian attribute inference in an end-to-end model," *arXiv preprint arXiv:1707.06089*, 2017. 99, 101

[15] D. Li, Z. Zhang, X. Chen, and K. Huang, "A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios," *IEEE T IMAGE PROCESS*, vol. 28, no. 4, pp. 1575–1590, 2018. 101, 105, 106, 107, 109, 110, 112

[16] J. Zhu, S. Liao, Z. Lei, and S. Z. Li, "Multi-label convolutional neural network based pedestrian attribute classification," *IMAGE VISION COMPUT*, vol. 58, pp. 224–229, 2017. 101, 109, 110

[17] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," in *Proceedings of the 22Nd ACM International Conference on Multimedia*, ser. MM '14. New York, NY, USA: ACM, 2014, pp. 789–792. [Online]. Available: http://doi.acm.org/10.1145/2647868.2654966 101, 105, 106

[18] J. Zhu, S. Liao, Z. Lei, D. Yi, and S. Li, "Pedestrian attribute classification in surveillance: Database and evaluation," in *Proc. ICCVW*, 2013, pp. 331–338. 101

[19] R. Layne, T. M. Hospedales, and S. Gong, "Attributes-based re-identification," in *Person Re-Identification*. Springer, 2014, pp. 93–117. 101

[20] Z. Tan, Y. Yang, J. Wan, H. Wan, G. Guo, and S. Z. Li, "Attention based pedestrian attribute analysis," *IEEE transactions on image processing*, 2019. 101

[21] Q. Li, X. Zhao, R. He, and K. Huang, "Visual-semantic graph reasoning for pedestrian attribute recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8634–8641. 101

[22] X. Zhao, L. Sang, G. Ding, J. Han, N. Di, and C. Yan, "Recurrent attention model for pedestrian attribute recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 9275–9282. 101

[23] M. Lou, Z. Yu, F. Guo, and X. Zheng, "Mse-net: Pedestrian attribute recognition using mlsc and se-blocks," in *International Conference on Artificial Intelligence and Security*. Springer, 2019, pp. 217–226. 101

[24] J. Zhu, S. Liao, D. Yi, Z. Lei, and S. Z. Li, "Multi-label cnn based pedestrian attribute learning for soft biometrics," in *2015 International Conference on Biometrics (ICB)*. IEEE, 2015, pp. 535–540. 101

[25] J. Wang, X. Zhu, S. Gong, and W. Li, "Attribute recognition by joint recurrent learning of context and correlation," in *Proc. ICCV*, 2017, pp. 531–540. 101

[26] D. Li, X. Chen, Z. Zhang, and K. Huang, "Pose guided deep model for pedestrian attribute recognition in surveillance scenarios," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2018, pp. 1–6. 101

[27] L. Yang, L. Zhu, Y. Wei, S. Liang, and P. Tan, "Attribute recognition from adaptive parts," *arXiv preprint arXiv:1607.01437*, 2016. 101

[28] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang, "Hydraplus-net: Attentive deep features for pedestrian analysis," in *Proc. ICCV*, 2017, pp. 350–359. 101

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778. 102

[30] ——, "Identity mappings in deep residual networks," in *Proc. ECCV*. Springer, 2016, pp. 630–645. 102

[31] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proc. IEEE ICCV*, 2017, pp. 2961–2969. 104

[32] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99. 104

[33] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. CVPR*, 2015, pp. 3431–3440. 104

[34] S. Ruder, "An overview of multi-task learning in deep neural networks," *CoRR*, vol. abs/1706.05098, 2017. [Online]. Available: http://arxiv.org/abs/1706.05098 104

[35] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. ICCV*, 2015, pp. 3730–3738. 105

[36] K. He, Z. Wang, Y. Fu, R. Feng, Y.-G. Jiang, and X. Xue, "Adaptively weighted multi-task deep network for person attribute classification," in *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 2017, pp. 1636–1644. 106

[37] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang, "Improving person re-identification by attribute and identity learning," *Pattern Recognition*, vol. 95, pp. 151–161, 2019. 106

[38] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 2016, pp. 265–283. 107

[39] W. Abdulla, "Mask r-cnn for object detection and instance segmentation on keras and tensorflow," https://github.com/matterport/Mask_RCNN, 2017. 107

[40] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proc. ECCV*. Springer, 2014, pp. 740–755. 107

[41] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013. 109

[42] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016. 111

[43] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. ICCV*, 2017, pp. 2980–2988. 113

Chapter 6

Person Re-identification: Implicitly Defining the Receptive Fields of Deep Learning Classification Frameworks

Abstract. The *receptive fields* of deep learning models determine the most significant regions of the input data for providing correct decisions. Up to now, the primary way to learn such receptive fields is to train the models upon masked data, which helps the networks to ignore any unwanted regions, but also has two major drawbacks: 1) it yields edge-sensitive decision processes; and 2) it augments considerably the computational cost of the inference phase. Having theses weaknesses in mind, this paper describes a solution for implicitly enhancing the inference of the networks' receptive fields, by creating synthetic learning data composed of interchanged segments considered *apriori* important or irrelevant for the network decision. In practice, we use a segmentation module to distinguish between the foreground (important) versus background (irrelevant) parts of each learning instance, and randomly swap segments between image pairs, while keeping the class label exclusively consistent with the label of the segments deemed important. This strategy typically drives the networks to interpret that the identity and clutter descriptions are not correlated. Moreover, the proposed solution has other interesting properties: 1) it is parameter-learning-free; 2) it fully preserves the label information; and 3) it is compatible with the data augmentation techniques typically used. In our empirical evaluation, we considered the person re-identification problem, and the well known RAP, Market1501 and MSMT-V2 datasets for two different settings (*upper-body* and *full-body*), having observed highly competitive results over the state-of-the-art. Under a reproducible research paradigm, both the code and the empirical evaluation protocol are available at `https://github.com/Ehsan-Yaghoubi/reid-strong-baseline`.

6.1 Introduction

Person re-identification (re-id) refers to the cross-camera retrieval task, in which a query from a target subject is used to retrieve identities from a gallery set. This process is tied to many difficulties, such as variations in human pose, illumination, partial occlusion, and cluttered background. The primary way to address these challenges is to provide large-scale *labeled* learning data (which are not only hard to collect, but particularly costly to annotate) and expect that the deep model learns the critical parts of the input data autonomously. This strategy is supposed to work for any problem, upon the existence of enough learning data, which might correspond to millions of learning instances in hard problems.

To skim the costly annotation step, various works propose to augment the learning data using different techniques [1]. They either use the available data to synthesize new images or generate new images by sampling from the learned distribution. In both cases, the main objective is to increase the quantity of data, without assisting the model in finding the input regions, so that often the networks find spurious patterns in the background regions that –yet– are matched with the ground truth labels. This kind of techniques shows positive effects in several applications; for example, [2] proposes an object detection model, in which the objects are cut out from their original background and pasted to other scenes (e.g., a plane is pasted between different sky images). On the contrary, in the pedestrian attribute recognition and re-identification problems, the background clutter is known as a primary obstacle to the reliability of the inferred models.

Holistic CNN-based re-id models extract global features, regardless of any critical regions in the input data, and typically fail when the background covers most of the input. In particular, when dealing with limited amounts of learning data, three problems emerge: 1) holistic methods may not find the foreground regions automatically; 2) part-based methods [3], [4] typically fail to detect the appropriate critical regions; and 3) attention-based models (e.g., [5] and [6]) face difficulties in case that multiple persons appear in a single bounding box. As an attempt to reduce the classification bias due to the background clutter (caused by inaccurate person detection or crowded scenes), [7] proposes an alignment method to refine the bounding boxes, while [8] uses a local feature matching technique. As illustrated in Fig. 6.1, although the alignment-based re-id approaches reduce the amounts of clutter in the learning data, the networks still typically suffer from the remaining background features, particularly if some of the IDs always appear in the same scene (background).

To address the above-described problems, this paper introduces a receptive field implicit definition method based on data augmentation that could be applied to the existing re-id methods as a complementary step. The proposed solution is 1) mask-free for the *test* phase, i.e., it does not require any additional explicit segmentation in test time; and 2) contributes to foreground-focused decisions in the inference phase. The main idea is to generate synthetic data composed of interleaved segments from the original learning set, while using class information only from specific segments. During the learning phase, the newly generated samples feed the network, keeping their label exclusively consistent with the identity from where the region-of-interest was cropped. Hence, as the model receives images of each identity with inconsistent unwanted areas (e.g., background), it naturally pays the most attention to the regions consistent with ground truth labels. We observed that this pre-processing method is equivalent to only learn from the effective receptive fields and ignore the destructive regions. During the test phase, samples are provided without any mask, and the network naturally disregards the detrimental information, which is the insight for the observed improvements in performance.

In particular, when compared to [9] and [10], this paper can be seen as a data augmentation technique with several singularities: 1) we not only enlarge the learning data but also implicitly provide the inference model with an attentional decision-making

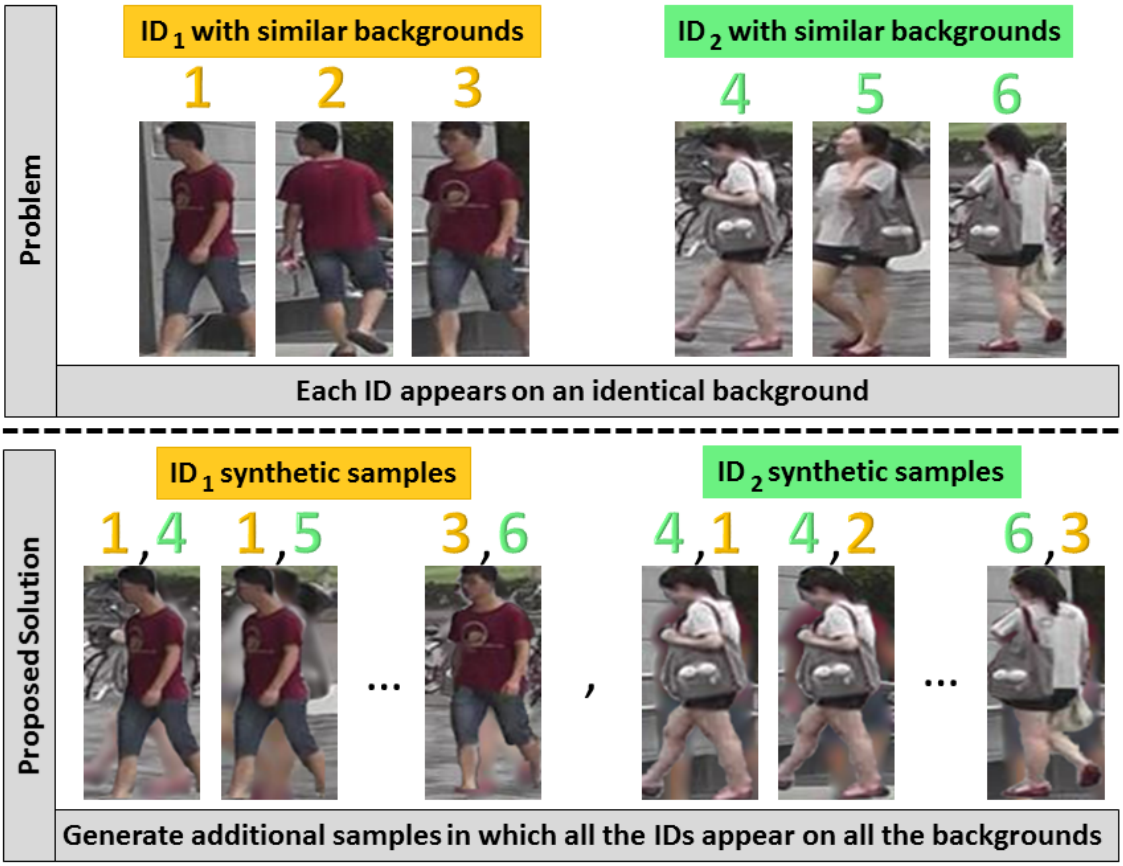


Figure 6.1: The main challenge addressed in this paper: during the learning phase, if the model sees all samples of one ID in a single scene, the final feature representation of that subject might be entangled with spurious (background) features. By creating synthetic samples with multiple backgrounds, we implicitly *guide* the network to focus on the deemed important (foreground) features.

skill, contributing to *ignore* irrelevant image features during the test phase; 2) we generate highly representative samples, making it possible to use our solution along with other data augmentation methods; and 3) our solution allows the on-the-fly data generation, which makes it efficient and easy to be implemented beside the common data augmentation techniques. Our evaluation results point for consistent improvements in performance when using our solution over the state-of-the-art person re-id method.

6.2 Related Work

Data Augmentation. Data augmentation targets the root cause of the over-fitting problem by generating new data samples and preserving their ground truth labels. *Geometrical transformation* (scaling, rotations, flipping, etc.), *color alteration* (contrast, brightness, hue), *image manipulation* (random erasing [10], kernel filters, image mixing [9]), and *deep learning approaches* (neural style transfer, generative adversarial networks) [1] are the common augmentation techniques.

Recently, various methods have been proposed for image synthesizing and data augmentation [1]. For example, [9] generates n^2 samples from an n-sized dataset by

using a sample pairing method, in which a random couple of images are overlaid based on the average intensity values of their pixels. [10] presents a *random erasing* data augmentation strategy that inflates the learning data by randomly selecting rectangular regions and changing their pixels values. As an attempt to robustify the model against occlusions, increasing the volume of the learning data turned the concept of *random erasing* into a popular data augmentation technique. [2] addressed the problem of object detection, in which the background has helpful features for detecting the objects; therefore, authors developed a context-estimator network that places the instances (i.e., cut out objects) with meaningful sizes on the relevant backgrounds.

Person Re-ID. In general, early person re-id works studied either the descriptors to extract more robust feature representations or metric-based methods to handle the distance between the inter-class and intra-class samples [11]. However, recent re-id studies are mostly based on deep learning neural networks that can be classified into three branches [12]: Convolutional Neural Network (CNN), CNN-Recurrent neural network, and Generative Adversarial Network (GAN).

Among the CNN and CNN-RNN methods, those based on attention mechanisms follow a similar objective to what we pursue in this paper; i.e., they ignore background features by developing attention modules in the backbone feature extractor. Attention mechanism may be developed for either single-shot or multi-shot (video) [13], [14], [15] scenarios, both of them aim to learn a distinctive feature representation that focuses on the critical regions of the data. To this end, [16] use the body-joint coordinates to remove the extra background and divide the image into several horizontal pieces to be processed by separate CNN branches. [5] and [6] propose a body-part detector to re-identify the probe person with matching the bounding boxes of each body-part, while [17] uses the masked out body-parts to ignore the background features in the matching process. In contrast to these works that explicitly implement the attentional process in the structure of the neural network [18], we provide an attentional control ability based on receptive field augmentation detailed in section 6.3. Therefore, in some terms, our work is similar to the GAN-based re-id techniques, which usually aim to either increase the quantity of the data [19] or present novel poses of the existing identities [20], [21] or transfer the camera style [22], [23]. Although GAN-based works present novel features for each individual, they generate some destructive features that are originated from the new backgrounds. Furthermore, these works ignore to handle the problem of co-appearance of multiple identities in one shot.

6.3 Proposed Method

Figure 6.2 provides an overview of the proposed image synthesis method, in this case, considering the full-body as the region of interest (ROI). We show the first synthesize subset, in which the new samples comprise of the ROI of the 1^{st} sample and the background of the other samples.

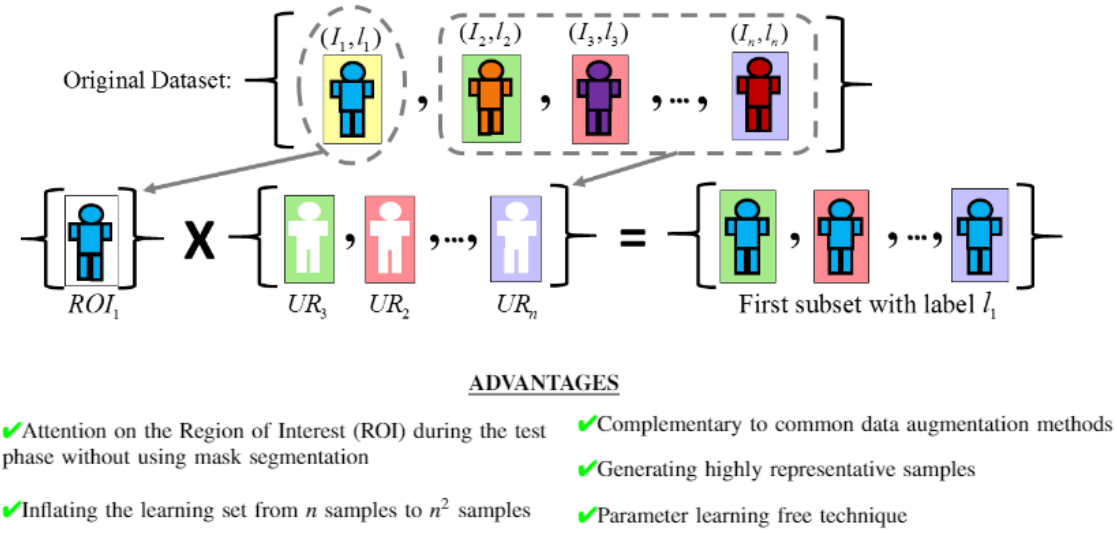


Figure 6.2: The proposed full-body attentional data augmentation (best viewed in color). Blue, orange, purple, and red denote the samples 1, 2, 3, and N, respectively. The pale-yellow, green, pink, and purple colors represent their cluttered (background) regions, which should be irrelevant for the inference process. Therefore, all the synthetic images labeled as 1 share the blue body region but have different backgrounds, which provides a strong cue for the network to disregard such segments from the decision process.

6.3.1 Implicit Definition of Receptive Fields

As an intrinsic behavior of CNNs, in the learning phase, the network extracts a set of essential features in accordance with the image annotations. However, extracting relevant and compressed features is an ongoing challenge, especially when the background[1] changes with person ID. Intuitively, when a person's identity appears with an identical background, some background features are entangled with the useful foreground features and reduce the inference performance. However, if the network sees one person with different backgrounds, it can automatically discriminate between the relevant regions of the image and the ground truth labels. **Therefore, to help the *inference model* automatically distinguish between the unwanted features and foreground features, in the *learning phase*, we repeatedly feed synthetically generated, fake images to the network that has been composed of two components:**

1. critical parts of the current input image that describe the ground truth labels (i.e., person's identity), and we would like to have an attention on them, and
2. parts of the other real samples that intuitively are uncorrelated with the current identity –i.e., background and possible body parts (if any) that we would like the network to ignore them.

Thus, the model looks through each region of interest, juxtaposed with different unwanted regions –of all the images– enabling the network to

[1]The terms *(unwanted region/region-of-interest)*, *(undesired/desired) boundaries*, *(background/foreground) areas*, and *(unwanted/wanted) areas* refer to the data segments that are deemed to be irrelevant/relevant to the ground truth label. For example, in a hair color recognition problem, the region-of-interest is the hair area, which can be defined by a binary mask

learn where to look at in the image and ignores the parts that are changing arbitrarily and are not correlated with ground truth labels. Consequently, during the test phase, the model explores the region of interest and discards the features of unwanted regions that have been trained for.

Formally, let $\boldsymbol{I}_i$ represent the ith image in the learning set, $\boldsymbol{l}_i$ its ground truth label (ID) and $\boldsymbol{M}_j$ the corresponding ground-truth binary mask that discriminates between the foreground/background regions. As the available re-id datasets do not provide ground-truth human body masks, we use the Mask R-CNN [24] to obtain such masks (see Section 6.4). Considering that $\boldsymbol{ROI}_.$ refers the region of interest and $\boldsymbol{UR}_.$ the unwanted regions, the goal is to synthesis the artificial sample $\boldsymbol{S}_{i\neg j}$, using label $\boldsymbol{l}_i$: $\boldsymbol{S}_{i\neg j}(x,y) = \boldsymbol{ROI}_i \cup \boldsymbol{UR}_j$, where $\boldsymbol{ROI}_. = \boldsymbol{I}_.(x,y)$ such that $\boldsymbol{M}_.(x,y) = 1$, $\boldsymbol{UR}_. = \boldsymbol{I}_.(x,y)$ such that $\boldsymbol{M}_.(x,y) = 0$, and (x,y) are the coordinates of the pixels.

Therefore, for an n-sized dataset, the *maximum* number of generated images is equal to $n^2 - n$. However, to avoid losing the expressiveness of the generated samples, we consider several constraints. Hence, a combination of the common data transformations (e.g., flipping, cropping, blurring) can be used along with our method. Obviously, since we utilize the ground truth masks, our technique should be used in the first place, before any other augmentation transformation, to avoid extra processing on the binary masks.

6.3.2 Synthetic Image Generation

To ensure that the synthetically generated images have a natural aspect, we impose the following constraints:

6.3.2.1 Size and shape constraint

Considering that human bodies are deformable objects of varying size and alignment within the bounding boxes, any blind image generation process will yield unrealistic results. Therefore, we added a constraint that avoids combining images with significant differences in their aspect ratios of the ROIs to circumvent the unrealistic stretching/shrinking of the replaced content in the generated images. To this end, the ratio between the foreground areas defined by masks $\boldsymbol{M}_j$ and $\boldsymbol{M}_i$ should be more than the threshold $\boldsymbol{T}_s$ (we considered $\boldsymbol{T}_s = 0.8$ in our experiments). Let $\boldsymbol{A}_.$ be the area of the foreground region (i.e., mask $\boldsymbol{M}_.$): $\boldsymbol{A}_j = \sum_{x=0}^{w}\sum_{y=0}^{h} M_j(x,y)$, where w and h are the width and height of the image, respectively.

This constraint translates to $\min(\boldsymbol{A}_i, \boldsymbol{A}_j)/\max(\boldsymbol{A}_i, \boldsymbol{A}_j) > \boldsymbol{T}_s$. Moreover, to ensure the shape similarity, we calculate the Intersection over Union metric (IoU) for masks $\boldsymbol{M}_i$ and $\boldsymbol{M}_j$: $IoU(\boldsymbol{M}_i, \boldsymbol{M}_j) = (\boldsymbol{M}_i \cap \boldsymbol{M}_j)/(\boldsymbol{M}_i \cup \boldsymbol{M}_j)$.

For the IoU calculation, we ought to consider only the rectangular area around the masks (instead of the whole image area); moreover, when calculating the IoU, the size of the masks must match, and in case of resizing the masks, the aspect ratios should be preserved. To fulfill these conditions, we find the contours in the binary masks using [25] and calculate the minimal up-right bounding rectangle of the masks. The width of

the rectangular masks in all images is set to a fixed size and, afterwards, we apply zero padding to the height of the smaller mask to match the sizes. Finally, if the $IoU(\boldsymbol{M}_i, \boldsymbol{M}_j)$ is higher than a threshold $\boldsymbol{T}_i$, we consider those images for the merging process ($\boldsymbol{T}_i = 0.5$ was used in our experiments).

6.3.2.2 Smoothness constraint

The transition between the source image and the replaced content should be as smooth as possible to prevent from strong edges. One challenge is that $\boldsymbol{M}_i$ and the body silhouette of the *j*-th person do not match perfectly. To overcome this issue, we enlarge the mask $\boldsymbol{M}_j$ by using the morphological dilation operator with a 5×5 kernel: $\boldsymbol{M}_d = \boldsymbol{M}_j \oplus K_{5\times5}$. Next, to guarantee the continuity between the background and the newly added content, we use the image in-painting technique in [26] to remove the undesired area from the source image, as it has been dictated by the enlarged mask $\boldsymbol{M}_d$.

6.3.2.3 Viewpoint constraint

The proposed method can be used for focusing on a specific region of the body. For example, supposing that the upper-body should be considered the RoI, the generated images will be composed of the 1^{st} sample's upper-body and the remaining segments (background and lower-body regions) of the other images, while keeping the label of the 1^{st} sample. When defining the receptive fields of specific regions (e.g., upper body in Fig. 6.3), it is important to generate high representative samples. Hence, we consider the body poses of samples and only combine images with the same viewpoint annotations causing to prevent from generating images composed of the anterior upper-body of the *i*-th person and posterior lower-body (and background) of the *j*-th person. One can apply Alphapose [27] to any pedestrian dataset to estimate the body poses and then, uses a clustering method such as [28], [29], [30], or [31] to create clusters of poses as the viewpoint label. The detailed information for the two experiments carried out is given in subsection 6.5.3. Figure 6.3 shows some examples generated by our technique, providing attention to the upper-body or full-body region. When defining the CNN's receptive fields on the upper-body region, fake samples are different in the human lower body and the environment, while they resemble each other in the person's upper body and identity label. By selecting the full-body as the ROI, the generated images will be composed of similar body silhouettes with different surroundings.

6.4 Implementation Details

As the settings and configurations on all the datasets are identical, in the following we only mention the details for the RAP dataset. We based our method on the baseline [32] and selected similar model architecture, parameter settings, and optimizer. In this baseline , authors resized images on-the-fly into 128×128 pixels. As the RAP images vary in resolution (from 33×81 to 415×583), to avoid any data deformation, we first mapped the

Figure 6.3: Examples of synthetic data generated for upper-body (center columns) and full-body (rightmost columns) receptive fields. The leftmost column shows the original images. Additional examples are provided at *https://github.com/Ehsan-Yaghoubi/reid-strong-baseline*.

images to a squared shape, using a *replication* technique, in which the row or column at the very edge of the original image is replicated to the extra border of the image.
The RAP dataset does not provide human body segmentation annotations. To generate the segmentation masks, we first fed the images to Mask-R-CNN model [24] (using its default parameter settings described in `https://github.com/matterport/Mask_RCNN`). Next, as described in subsection 6.3.2, we generated the synthetic images.
To provide the train and test splits for our model, we followed the instructions of the dataset publishers in [23; 33; 34]. Furthermore, following the configurations suggested in [32], we used the state-of-the-art tricks such as warm-up learning rate [35], random erasing data augmentation [10], label smoothing [36], last stride [37], and BNNeck [32], alongside the conventional data augmentation transformations (i.e., random horizontal flip, random crop, and 10-pixel-padding and original-size-crop).

6.5 Experiments and Discussion

We evaluate the proposed method under two settings: (1) by defining the upper-body receptive fields, assuming that most of the identity information lies in upper body. In this setting, we generate the synthetic data by modifying the lower-body parts of the subject images. This setting requires both segmentation masks and viewpoint annotations, as the perspective/viewpoint of the upper-body region should be consistent with the perspective of the lower body. In practice, this strategy assures that we do not combine a front-view upper body with a rear-view lower body. (2) by defining the full-body receptive fields, in which the attention of the network is "oriented" towards the entire body. The notion of viewpoint does not apply here, since the method can be seen as a simple background swapping process, where the person is placed in a different environment. In our experiments, we evaluate our model on the earlier setting and RAP dataset for two modes: (a) when human-based annotations are available for four viewpoints, and (b) when the subjects' viewpoint is inferred using a clustering method. Furthermore, we tested our method with the later setting over the RAP, Market1501, and MSMT17 datasets.

Table 6.1: Results comparison between the baseline (top row) and our solutions for defining receptive fields, particularly tuned for the *upper body* and *full body*, on the RAP benchmark. mAP and Ranks 1, 5, and 10 are given, for the *softmax* and *triplet-softmax* samplers. Ours-1 shows the results for setting 1, mode 1: upper body with viewpoint annotations. Ours-2 shows the results for setting 1, mode 2: upper body without viewpoint annotations. Ours-3 shows the results for setting 2: full body. The **best possible results** for Luo *et* al. [32] occurred using *triplet-softmax* sampler in epoch 1120, whereas our models were trained for 280 epochs which lasted around 20 hours. The best results appear in bold.

	softmax sampler				triplet-softmax sampler			
Model	rank=1	rank=5	rank=10	mAP	rank=1	rank=5	rank=10	mAP
Luo *et* al. [32]	64.1	81.5	86.8	**45.8**	66.1	81.9	86.3	45.9
Ours-1	64.4	80.5	85.6	42.5	66.5	81.5	86.0	43.0
Ours-2	65.1	81.4	86.2	43.3	66.8	82.0	86.5	43.8
Ours-3	**65.7**	**82.2**	**87.2**	45.0	**69.0**	**83.6**	**88.1**	**46.3**

6.5.1 Datasets

The *Richly Annotated Pedestrian* (RAP) benchmark [33] is one of the largest well-known pedestrian dataset composing of around 85,000 samples, from which 41,585 images have been selected manually for identity annotation. The RAP re-id set includes 26,638 images of 2,589 identities and 14,947 samples as distractors that have been collected from 23 cameras in a shopping mall. The provided human bounding boxes have different resolutions ranging from 33×81 to 415×583. In addition to human attributes, the RAP dataset is annotated for camera angle, body-part position, and occlusions. The MSMT17-V2 re-id dataset [23] consists of 4101 identities captured with 15 cameras in outdoor and indoor environment. The total number of person bounding boxes are $126,441$ which have been detected using Faster RCNN [38]. The Market1501 dataset [34] used the Deformable Part Model (DPM) detector [39] to extract 32,668 person bounding boxes from 1105 identities using 6 cameras in outdoor scenes. The Market1501 dataset images were normalized to 128×64 pixel resolution.

6.5.2 Baseline

A recent work by Facebook AI [40] mentions that upgrading factors such as the learning method (e.g., [41], [42]), network architecture (e.g., ResNet, GoogleNet, BN-Inception), loss function (e.g., embedding losses [43], [44] and classification losses [45], [46]), and parameter settings may improve the performance of an algorithm, leading to unfair comparison. This way, to be certain that the proposed solution actually contributes to performance improvement, our empirical framework was carefully designed in order to keep constant as many factors as possible with a recent re-id baseline [32] This baseline has advanced the state-of-the-art performance with respect to several techniques such as [47],[48], and [49]. In summary, it is a holistic deep learning-based framework that uses a bag of tricks that are known to be particularly effective for the person re-id problem. Authors employ the ResNet-50 model as the backbone feature extractor

Table 6.2: Results of the proposed receptive field definer solution for upper-body and full-body models. Bold and underline styles denote the best and runner-up results. "Aug. Prob." stands for *augmentation probability*.

Model	Aug. Prob.	Rank 1	Rank 5	Rank 10	Rank 50	mAP
Upper-body	0.1	53.4	72.3	78.9	90.6	34.8
	0.3	63.1	79.8	84.8	**93.2**	**41.1**
	0.5	**64.4**	**80.5**	**85.6**	92.7	42.5
	0.7	62.1	78.3	83.0	91.6	37.7
	0.9	59.0	75.3	80.6	90.2	34.8
Full-body	0.3	**69.0**	**83.6**	**88.1**	**94.8**	**46.3**
	0.5	68.0	82.6	87.0	94.3	44.6

6.5.3 Re-ID Results

6.5.3.1 Experiments on the RAP dataset

As stated before, the proposed method with the upper-body setting requires viewpoint labels; however, not all pedestrian datasets provide this ground truth information. As annotating a large dataset with this information would be extremely time consuming, we suggested that state of the art pose detectors are used to automatically infer the subjects viewpoint. To test this hypothesis, we have chosen the RAP dataset since it includes manual annotations for the samples viewpoint. Hence, we evaluated our upper-body-based model for two different modes: (1) by considering the human-based viewpoint annotations; and (2) by using Alphapose followed by a clustering method (Balanced Iterative Reducing and Clustering using Hierarchies [28]) to automatically estimate human poses. In the latter case, we used Alphapose with its default settings to extract the body key-points of all the persons in the dataset; next, we applied the BIRCH clustering method and created 8 clusters of body poses. Finally, to swap the unwanted regions in the original image with another sample, the candidate image is selected from the same cluster where the original image is located. In both modes, the network configuration and the hyper-parameters were exactly the same.

Table 6.1 provides the overall performances based on the mean Average Precision (mAP) metric and Cumulative Match Characteristic (CMC) for ranks 1, 5, and 10, denoting the possibility of retrieving at least one true positive in the top-1, 5, and 10 ranks. We evaluated the proposed method using two sampling methods and observed a slight improvement in the performance of both methods when using the *triplet-softmax* over *softmax* sampler. As previously mentioned, our method could be treated as an augmentation method that requires a paired-process (i.e., exchanging the foreground and background of each pair of images), imposing a computational cost only to the *learning phase*. Moreover, due to increasing the learning samples from n to less than n^2, the network needs more time and the number of epochs to converge. Therefore, learning our method (using *triplet-softmax* sampler) for 280 epochs lasted around 20 hours with loss value 1.3, while the baseline method completed 2000 epochs after 37 hours of learning with loss value 1.0.

The experimental results of upper-body setting are given in rows 2 and 3 of Table 6.1,

pointing for an optimal performance, when we use 8 cluster of poses instead of the ground truth viewpoint labels; therefore, our method could be used in conjunction with viewpoint estimation models to boost the performance, without requiring viewpoint annotations.
Comparison of the first and second rows of Table 6.1 shows that our technique with an attention on the human upper-body achieves competitive results, such that retrieval accuracy in rank 1 is 0.3% better than the baseline. However, in higher ranks and mAP metrics, the baseline has better performance.
The fourth row of Table 6.1 provides the performance of the proposed method with an attention on the human full-body and –not surprisingly– indicates that concentration on the full-body (rather than upper-body) yields more useful features for short-time person re-id. However, comparing four rows of the result table together, we could perceive how much is the lower-body important –as a body-part with most background region? For example, when using full-body region (over the upper-body) with *triplet-softmax* sampler, the rank 1 accuracy improves from 66.8 to 69.0 (i.e., 2.2% improvement), while the accuracy difference of rank 1 between the holistic baseline and full-body method is 2.9%, indicating that 2.2 of our improvement (in rank 1) over the baseline is because of attention on the lower-body and the rest (0.7%) is due to focusing on the upper-body.
During the learning phase, each synthesized sample is generated with a probability between $[0, 1]$, with 0 meaning that no changes will be done in the dataset (i.e., we use the original samples) and 1 indicates that all samples will be transformed (augmented). We studied the effectiveness of our method for different probabilities (from 0.1 to 0.9) and gave the obtained results in Table 6.2. Overall, the optimal performance of the proposed technique is attained when the augmentation probability lies in the $[0.3, 0.5]$ interval. This leads us to conclude that such intermediate probabilities of augmentation keep the discriminating information of the original data while also guarantee the transformation of enough data for yielding an effective attention mechanism.

6.5.3.2 Experiments on the Market1501 dataset

Table 6.3 compares the performance of our method with respect to several state-of-the-art techniques on the Market1501 set [34], and supports the superiority of our method, with 0.4 of rank 1 accuracy over [50] and 1.1 of mAP over [51]. Additionally, we post-processed our results on the Market1501 using the re-ranking method proposed by [52]. [52] post-processes the global features of the gallery set and the probe person. This method indicates that the k-reciprocal nearest neighbors to the probe image should have more priorities in the ranking list. Using this technique with settings $k1 = 20$, $k2 = 6$, and $\lambda = 0.3$, the rank 1 and mAP results were improved from 95.1 and 86.5 to 95.8 and 94.3, respectively.

6.5.3.3 Experiments on the MSMT17-V2 dataset

The empirical results over the MSMT17-v2 benchmark [23] are given in Table 6.4. Results show that the proposed method advances the state-of-the-art methods in ranks 1, 5, and

Table 6.3: Results comparison on the Market1501 benchmark. The top two results are given in **bold**.

Model	Rank 1	Rank 5	Rank 10	mAP
[53]	88.5	–	–	71.5
[54]	84.7	94.2	96.6	64.7
[55]	90.5	–	–	77.7
[56]	91.3	–	–	76.0
[57]	91.4	96.6	97.7	76.7
[51]	91.5	96.8	97.3	85.4
[58]	92.1	96.5	98.6	81.9
[59]	92.3	–	–	78.2
[60]	93.3	–	–	76.8
[61]	93.3	97.5	98.4	81.3
[62]	93.4	97.6	98.5	82.2
[63]	93.9	–	–	84.5
[50]	94.7	95.7	98.5	–
Ours	**95.1**	**98.2**	**99.0**	**86.5**
Ours + re-ranking	**95.8**	**98.0**	**98.5**	**94.3**

10 by more than 2 percent, while - based on the mAP metric - our method (45.9%) ranks second best, after [64].

6.6 Conclusions

CNNs are known to be able to autonomously find the critical regions of the input data and discriminate between foreground-background regions. However, to accomplish such a challenging goal, they demand large volumes of learning data, which can be hard to collect and particularly costly to annotate, in case of supervised learning problems. In this paper, we described a solution based on data segmentation and swapping, that interchanges segments *apriori* deemed to be important or irrelevant for the network responses. The proposed method can be seen as a data augmentation solution that implicitly empowers the network to improve its *receptive fields inference skill*. In practice, during the learning phase, we provide the network with an attentional mechanism derived from prior information (i.e., annotations and body masks), that determines not only the critical regions of the input data but also provides important cues about any useless input segments that should be disregarded from the decision process. Finally, it is important to stress that, in *test* time, samples are provided without any segmentation mask, which lowers the computational burden with respect to previously proposed explicit attention mechanisms. As a proof-of-concept, our experiments were carried out in

Table 6.4: Results comparison on the MSMT17 benchmark. The best results are given in **bold**.

Model	Rank 1	Rank 5	Rank 10	mAP
[64]	68.3	–	–	**49.3**
[59]	68.8	–	–	41.0
[57]	69.4	81.5	85.6	39.2
Ours	**71.7**	**83.6**	**87.4**	45.9

the highly challenging pedestrian re-identification problem, and the results show that our approach –as a complementary data augmentation technique– could contribute to significant improvements in the performance of the state-of-the-art.

Bibliography

[1] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, p. 60, 2019. 120, 121

[2] N. Dvornik, J. Mairal, and C. Schmid, "On the importance of visual context for data augmentation in scene understanding," *IEEE TPAMI*, 2019. 120, 122

[3] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A siamese long short-term memory architecture for human re-identification," in *Proc. ECCV*. Springer, 2016, pp. 135–153. 120

[4] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proc. CVPR*, 2017, pp. 384–393. 120

[5] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," in *Proc. CVPR*, 2018, pp. 2119–2128. 120, 122

[6] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *Proc. ICCV*, 2017, pp. 3219–3228. 120, 122

[7] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 10, pp. 3037–3045, 2018. 120

[8] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun, "Alignedreid: Surpassing human-level performance in person re-identification," *arXiv preprint arXiv:1711.08184*, 2017. 120

[9] H. Inoue, "Data augmentation by pairing samples for images classification," *arXiv preprint arXiv:1801.02929*, 2018. 120, 121

[10] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc AAAI Conf*, 2020, pp. 0–0. 120, 121, 122, 126

[11] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *IMAGE VISION COMPUT*, vol. 32, no. 4, pp. 270–286, 2014. 122

[12] D. Wu, S.-J. Zheng, X.-P. Zhang, C.-A. Yuan, F. Cheng, Y. Zhao, Y.-J. Lin, Z.-Q. Zhao, Y.-L. Jiang, and D.-S. Huang, "Deep learning-based methods for person re-identification: A comprehensive review," *Neurocomputing*, vol. 337, pp. 354–371, Apr. 2019. [Online]. Available: https://doi.org/10.1016/j.neucom.2019.01.079 122

[13] G. Chen, J. Lu, M. Yang, and J. Zhou, "Spatial-temporal attention-aware learning for video-based person re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 9, pp. 4192–4205, 2019. 122

[14] L. Zhang, Z. Shi, J. T. Zhou, M.-M. Cheng, Y. Liu, J.-W. Bian, Z. Zeng, and C. Shen, "Ordered or orderless: A revisit for video based person re-identification," *IEEE TPAMI*, 2020. 122

[15] L. Cheng, X.-Y. Jing, X. Zhu, F. Ma, C.-H. Hu, Z. Cai, and F. Qi, "Scale-fusion framework for improving video-based person re-identification performance," *Neural Computing and Applications*, pp. 1–18, 2020. 122

[16] F. Yang, K. Yan, S. Lu, H. Jia, X. Xie, and W. Gao, "Attention driven person re-identification," *Pattern Recognition*, vol. 86, pp. 143–155, 2019. 122

[17] C. Zhou and H. Yu, "Mask-guided region attention network for person re-identification," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2020, pp. 286–298. 122

[18] M. Denil, L. Bazzani, H. Larochelle, and N. de Freitas, "Learning where to attend with deep architectures for image tracking," *Neural Comput.*, vol. 24, no. 8, pp. 2151–2184, 2012. 122

[19] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *Proc. ICCV*, 2017, pp. 3754–3762. 122

[20] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu, "Pose transferrable person re-identification," in *Proc. CVPR*, 2018, pp. 4099–4108. 122

[21] A. Borgia, Y. Hua, E. Kodirov, and N. Robertson, "Gan-based pose-aware regulation for video-based person re-identification," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1175–1184. 122

[22] Y. Lin, Y. Wu, C. Yan, M. Xu, and Y. Yang, "Unsupervised person re-identification via cross-camera similarity exploration," *IEEE Transactions on Image Processing*, vol. 29, pp. 5481–5490, 2020. 122

[23] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proc. CVPR*, 2018, pp. 79–88. 122, 126, 127, 129

[24] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proc. IEEE ICCV*, 2017, pp. 2961–2969. 124, 126

[25] S. Suzuki *et al.*, "Topological structural analysis of digitized binary images by border following," *Computer vision, graphics, and image processing*, vol. 30, no. 1, pp. 32–46, 1985. 124

[26] A. Telea, "An image inpainting technique based on the fast marching method," *J. Graph. Tools*, vol. 9, no. 1, pp. 23–34, 2004. 125

[27] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in *Proc. ICCV*, 2017, pp. 2334–2343. 125

[28] T. Zhang, R. Ramakrishnan, and M. Livny, "Birch: an efficient data clustering method for very large databases," *ACM sigmod record*, vol. 25, no. 2, pp. 103–114, 1996. 125, 128

[29] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297. 125

[30] D. Sculley, "Web-scale k-means clustering," in *Proceedings of the 19th international conference on World wide web*, 2010, pp. 1177–1178. 125

[31] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in neural information processing systems*, 2002, pp. 849–856. 125

[32] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. CVPRW*, 2019, pp. 0–0. 125, 126, 127

[33] D. Li, Z. Zhang, X. Chen, and K. Huang, "A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios," *IEEE T IMAGE PROCESS*, vol. 28, no. 4, pp. 1575–1590, 2018. 126, 127

[34] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE ICCV*, 2015, pp. 1116–1124. 126, 127, 129

[35] X. Fan, W. Jiang, H. Luo, and M. Fei, "Spherereid: Deep hypersphere manifold embedding for person re-identification," *J VIS COMMUN IMAGE R.*, vol. 60, pp. 51–58, 2019. 126

[36] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned cnn embedding for person reidentification," *ACM TOMM*, vol. 14, no. 1, p. 13, 2018. 126

[37] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. ECCV*, 2018, pp. 480–496. 126

[38] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99. 127

[39] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE TPAMI*, vol. 32, no. 9, pp. 1627–1645, 2009. 127

[40] K. Musgrave, S. Belongie, and S.-N. Lim, "A metric learning reality check," *arXiv preprint arXiv:2003.08505*, 2020. 127

[41] K. Roth, B. Brattoli, and B. Ommer, "Mic: Mining interclass characteristics for improved metric learning," in *Proc. ICCV*, 2019, pp. 8000–8009. 127

[42] W. Kim, B. Goyal, K. Chawla, J. Lee, and K. Kwon, "Attention-based ensemble for deep metric learning," in *Proc. ECCV*, 2018, pp. 736–751. 127

[43] F. Cakir, K. He, X. Xia, B. Kulis, and S. Sclaroff, "Deep metric learning to rank," in *Proc. CVPR*, 2019, pp. 1861–1870. 127

[44] X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott, "Multi-similarity loss with general pair weighting for deep metric learning," in *Proc. CVPR*, 2019, pp. 5022–5030. 127

[45] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proc. CVPR*, 2018, pp. 5265–5274. 127

[46] Q. Qian, L. Shang, B. Sun, J. Hu, H. Li, and R. Jin, "Softtriple loss: Deep metric learning without triplet sampling," in *Proc. ICCV*, 2019, pp. 6450–6458. 127

[47] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *Proc IEEE CVPR*, 2018, pp. 1062–1071. 127

[48] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camstyle: A novel data augmentation method for person re-identification," *IEEE T IMAGE PROCESS*, vol. 28, no. 3, pp. 1176–1190, 2018. 127

[49] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. CVPR*, 2018, pp. 2285–2294. 127

[50] A. Khatun, S. Denman, S. Sridharan, and C. Fookes, "Semantic consistency and identity mapping multi-component generative adversarial network for person re-identification," in *The IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 2267–2276. 129, 130

[51] Q. Zhou, B. Zhong, X. Lan, G. Sun, Y. Zhang, B. Zhang, and R. Ji, "Fine-grained spatial alignment model for person re-identification with focal triplet loss," *IEEE Transactions on Image Processing*, vol. 29, pp. 7578–7589, 2020. 129, 130

[52] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proc. CVPR*, 2017, pp. 1318–1327. 129

[53] Z. Zeng, Z. Wang, Z. Wang, Y. Zheng, Y.-Y. Chuang, and S. Satoh, "Illumination-adaptive person re-identification," *IEEE Trans. Multimed.*, 2020. 130

[54] Y.-S. Chang, M.-Y. Wang, L. He, W. Lu, H. Su, N. Gao, and X.-A. Yang, "Joint deep semantic embedding and metric learning for person re-identification," *Pattern Recognit. Lett.*, vol. 130, pp. 306–311, 2020. 130

[55] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang *et al.*, "Fd-gan: Pose-guided feature distilling gan for robust person re-identification," in *Advances in neural information processing systems*, 2018, pp. 1222–1233. 130

[56] Z. Wang, J. Jiang, Y. Wu, M. Ye, X. Bai, and S. Satoh, "Learning sparse and identity-preserved hidden attributes for person re-identification," *IEEE Transactions on Image Processing*, vol. 29, no. 1, pp. 2013–2025, 2019. 130

[57] Y. Yuan, W. Chen, Y. Yang, and Z. Wang, "In defense of the triplet loss again: Learning robust person re-identification with fast approximated triplet loss and label distillation," in *Proc. CVPRW*, 2020, pp. 354–355. 130

[58] Z. Chang, Z. Qin, H. Fan, H. Su, H. Yang, S. Zheng, and H. Ling, "Weighted bilinear coding over salient body parts for person re-identification," *Neurocomputing*, vol. 407, pp. 454–464, 2020. 130

[59] M. Jiang, C. Li, J. Kong, Z. Teng, and D. Zhuang, "Cross-level reinforced attention network for person re-identification," *Journal of Visual Communication and Image Representation*, p. 102775, 2020. 130

[60] S. Liu, T. Si, X. Hao, and Z. Zhang, "Semantic constraint gan for person re-identification in camera sensor networks," *IEEE Access*, vol. 7, pp. 176 257–176 265, 2019. 130

[61] W. Zhang, L. Huang, Z. Wei, and J. Nie, "Appearance feature enhancement for person re-identification," *Expert Systems with Applications*, p. 113771, 2020. 130

[62] Y. Tang, X. Yang, N. Wang, B. Song, and X. Gao, "Person re-identification with feature pyramid optimization and gradual background suppression," *Neural Networks*, vol. 124, pp. 223–232, 2020. 130

[63] F. Chen, N. Wang, J. Tang, D. Liang, and H. Feng, "Self-supervised data augmentation for person re-identification," *Neurocomputing*, 2020. 130

[64] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *arXiv preprint arXiv:2001.04193*, 2020. 130

Chapter 7

You Look So Different! Haven't I Seen You a Long Time Ago?

Abstract. Person re-identification (re-id) aims to match a query identity (ID) to an element in a gallery, collected from multiple cameras. Most of the existing re-id methods are trained and evaluated under short-term settings, where the query subjects appear with the same clothes in the gallery. In this setting, the learned feature representations are dominated by the visual appearance of clothes, which considerably drops the identification accuracy for long-term settings. To alleviate this problem, we propose a model that learns the long-term representations of persons by ignoring the features previously learned by a short-term re-id method and naturally makes it invariant to clothing styles. We first synthesize a set in which we distort the most relevant biometric information of people (face, body shape, height, and weight) and keep the short-term cues (color and texture of clothes) unchanged. This way, while the original data expresses both the ID-related and all varying features, the synthesized representations are composed mostly of short-term attributes – e.g., color and texture of clothes. Following this idea, the key to obtaining stable long-term representations is to learn embeddings of the original data that maximize the dissimilarity with the short-term embeddings. In practice, we first use the synthetic data to learn a model that embeds the ID-unrelated features and then learn a second model from the original data, where the long-term embeddings are extracted in such a way to be independent of the previously obtained ID-unrelated features. Our experiments were performed on two challenging cloth-changing sets (LTCC and PRCC) and our results support the effectiveness of the proposed method, which advances the state-of-the-art for both short and long-term re-id.

7.1 Introduction

Retrieving a query identity from a gallery of people with consistent clothing-style, across a distributed camera network is known as short-term person re-identification (re-id) [1]. Being inherently a challenging task, short-term re-id has been the topic of substantial research for more than a decade, with several datasets announced [2; 3], methods proposed [4; 5], and multiple surveys published [5–8]. In this problem, the major challenges are the variations in body pose, varying illumination, occlusions, camera resolution, and viewing angle. Therefore, in *cloth-consistent* setting, the assumption is to obtain representations that are mostly based on the clothing textures and colors. However, re-identifying people form biological traits rather than any transient appearance characteristics is more challenging [9]. Short-term re-id methods are known to substantially degrade their performance under cloth-changing scenarios [10], which

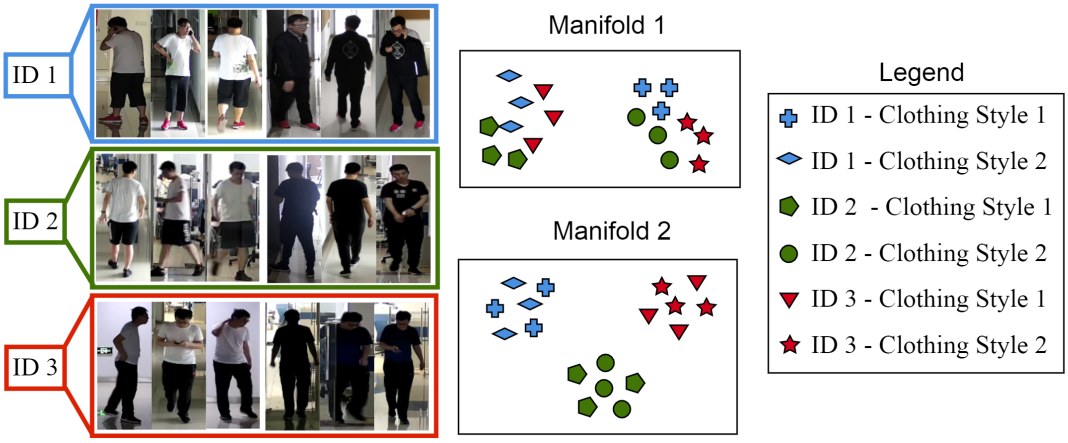


Figure 7.1: Main motivation of the proposed work. Short-term person re-id methods rely on appearance features that are likely to converge towards "Manifold 1", in which samples with similar clothes appear nearby. Instead, our goal is to obtain an embedding such as "Manifold 2", where samples of different persons appear together, regardless of their clothing styles. Best viewed in color.

provides the main motivation for this work: it is crucial to develop re-id models that are naturally invariant to clothing features such as colors, textures, shapes, and styles.

As illustrated in Fig. 7.1, in long-term person re-id settings, the model should recognize instances of the same person after long periods (several weeks or months), with the assumption that the query subject might be wearing different clothes than any instance of the gallery. Recently, some models proposed to learn cloth-independent features, by either generating people with different clothing patterns [10; 11] or extracting shape-based body features [12; 13]. Other authors assumed specific constraints (e.g., constant walking patterns [14], moderate clothing changes [13], and visible facial images [15]), attempting to learn ID-sensitive embeddings by changing the clothes colors/patterns. In opposition, other works exclusively focused in the body-shape or facial attributes [13; 15], most of which reported to have poor generalization capabilities.

Learning robust features is a key factor in long-term person re-id. Robustness refers to 1) the extraction of discriminative features from inter-person samples and 2) being invariant to intra-person attribute variations. Although the cross-entropy loss function optimizes the re-id model for these criteria, high variations in the intra-person samples hinder the model from learning useful long-term representations and lead to learning a manifold similar to "manifold 1" illustrated in Fig. 7.1.

Based on our analysis, we concluded that the key to mitigating the above problems is to keep the visual appearance information that are *useful* (face, body shape, body figure, height, gender) while disregarding any other ID-unrelated features (clothing styles and background features). This paper proposes a framework that firstly transforms the original learning data in a way to help the model to infer ID-unrelated features (i.e., short-term). At a second step, a long-term embedding is learned by minimizing the correlation between the inferred features and the previously obtained short-term feature representations, according to a cosine similarity loss.

The main contributions of this paper are as follows:

- We discuss the person re-id problem under the long-term scenario, which is up to the moment rarely seen in the literature.
- We propose an image transformation pipeline that helps the image-based re-id models to disregards background and clothing-based features.
- We propose a framework that re-identifies people based on their face and soft biometrics (e.g., body shape), while automatically disregarding any changeable visual appearance features (e.g., clothes). Moreover, at the inference time, our solution does not depend on any kind of additional labeling information, such as body masks or key-points.
- The proposed framework implicitly disentangles the short-term and long-term representations using the cosine similarity measure. Hence, the proposed strategy could be applied to other object recognition tasks.

7.2 Related work

Most of the prior person re-id studies assume that the query persons wear the same outfits in the gallery set [8]. However, this assumption is not always valid and leads to poor performance when applied to long-term re-id settings. In this paper, we focus on a real-world scenario, when people may appear with different clothes, and refer the readers to [5; 8] for discussions on the representative works of the short-term person re-id.
As an early study in the context of long-term re-id, Zhang *et al.* [14] proposes a video-based re-id technique based on the body motion to address the challenge of person appearance variations. In this work, the authors applied local descriptors (i.e., Histogram of Optical Flow and Motion Boundary Histogram) to capture the latent motion cues of a person's walking style and relative motion between feature points, based on the hypothesis that persons' movement follows a consistent pattern. Although this method captures some fine-grained gait features, it disregards the useful appearance features related to the body-shape and head area.
In [15], the authors focused solely on scenarios where the face is clearly visible. The proposed model processes two persons' pictures and uses the face area to yield the person ID and detects whether the subject has different clothes based on the body area or not. However, the high resolution face shots are rarely available in the surveillance data, which leads to an undesirable performance of the state-of-the-art face recognition models. So, coupling a short-term re-id model equipped to a face re-id branch cannot obtain satisfactory results [1; 13]. Later, [13] performed a case-study, in which the individuals change their clothes, such that the overall body shape is preserved. In other words, the authors proposed a re-id model based on the person's contour sketches to ignore the color-based features and demonstrate the importance of the body-shape in long-term person re-id. In order to enhance the performance of the deep-based long-term person re-id, one strategy would be to increase the learning data, such that each subject wears numerous different clothes. As collecting such a dataset on a large scale demands expensive

gathering and annotation processes, some studies proposed applying generative models. In this context, inspired by a pose-invariant generative re-id model [16], Yu *et al.* [10] proposed a clothing simulator model to synthesize more samples for each ID with several different clothing styles. The authors applied a body-parsing technique on the image to mask out the clothes area and trained a generative model to reconstruct the clothes area differently. Afterward, another model used both the original image and the reconstructed image to learn the differences (clothes area). Although this method has tried to decrease the clothing change effects, it has some drawbacks: 1) segmentation clothing area is itself a challenging task in computer vision and yet cannot yield reliable results on the real-world human surveillance data, 2) this method neglects the feature similarities in the background area, 3) the shape of the clothing styles (e.g., short dress and long dress) highly affects the final feature representation of the persons, which has been neglected.
In another generative-based study [12], the authors proposed an adversarial learning-based model to ignore the color features and focused solely on the body-shape features. To derive the body-shape representation, the authors extracted image features in RGB and grey-scale modes and fed them into a feature discriminator to distinguish between the RGB and grey-scale feature sets. Supposing that another image of the same person contains similar body-shape features, the authors concatenated the grey-scale features of a first body-pose with the RGB features of a second body-pose. Then, they trained a generator to reconstruct an RGB image with the first body-pose.
With an assumption that the body-shape is a reliable soft-biometric for long-term re-id scenarios, Qian *et al.* [1] used the human joint coordinates to model the relations among them by two scalar numbers in x-axis and y-axis directions. Next, these scalars were used to generate the shape-based features that their difference with the image-based features could result in a shape-sensitive feature representation of the input sample. [1] relies on capturing the information of the body-joints coordinates; however, [13] shows that the contour sketch of the body has useful information which cannot be inferred from the body key-points.
Based on the above-mentioned review on the recent studies, a long-term person re-id model may extract useful information from head-neck area, full-body soft biometrics, and body-shape characteristics. In the next section, we explain how our model captures these data and disregards the short-term features.

7.3 Proposed method

The proposed Long-term, Short-term features Decoupler (LSD) framework is an image-based person re-id network that extracts long-term discriminative representations of people that are invariant to clothes and background changes. The LSD framework is developed in four phases: pre-processing, learning short-term embeddings (ID-unrelated features), learning the long-term embeddings (ID-related features), and inferring the long-term feature representations of people. In the pre-processing phase, we generate a synthesized dataset, in which we apply several image transformations on each sample of

the original learning set to distort the visual identity cues such as facial area, body figure, height, weight, and gender (see Fig. 7.2 and Fig. 7.3). Then, in the first learning phase, we train an auxiliary model, named as Short-Term Embedding Convolutional Neural Network (STE-CNN), on the *synthesized data* to extract the ID-unrelated embeddings of each instance. In the next learning phase, we use a *cosine similarity loss function* in the learning phase of a second model, called Long-Term Embedding CNN (LTE-CNN), to learn from the *original images* such that the learned embeddings are dissimilar to the ID-unrelated embeddings. This way, the LTE-CNN model captures the embeddings of the identity cues that are unchangeable during long time intervals and disregards the attributes that are more prone to change e.g., clothing style, accessories and background. In the evaluation phase, we only use the LTE-CNN model to infer the long-term representations of people. This denotes that *training the STE-CNN model and generating synthesized data are auxiliary steps that enhances the learning quality of the LTE-CNN model and are skipped in the inference phase.* Meanwhile, the evaluation process of the LTE-CNN model is similar to the typical re-id models: the gallery samples are ranked based on the similarity between the long-term representations of the gallery and query instances.

It is worth noting that the STE-CNN and LTE-CNN are regular deep architectures (e.g., resnet-50) that extract the global features of the input data, and the given names are to provide the reader with a feeling about their functionality; therefore, both the STE-CNN and LTE-CNN may have an identical architecture, but are different in terms of the input data and loss function.

7.3.1 Pre-processing: Image Transformation Pipeline

In the proposed LSD model, the STE-CNN must learn the embeddings unrelated to the subject's ID, such as clothes and background features. This section describes the various image processing steps applied to the original learning set to remove the ID cues and generate the learning data for the STE-CNN model. Fig. 7.2 gives an overview of the image transformation pipeline and Fig. 7.3 shows some examples of several synthesized samples. The results show that as we intended, the robust soft biometrics (such as weight, height, and body shape) have been visually distorted in the transformed images, while the background area and accessories have been unchanged approximately.

The proposed pipeline requires the input image, the segmentation mask, and the body key-points of the subject. The latter data are extracted using the state-of-the-art methods, for instance, segmentation [17] and human body key-point localization [18]. It is worth noting that our approach does not require a perfect segmentation and localization of the body parts, as these data are used to roughly establish an irregular shaped region of interest (body contour) to be removed from the input image.

We hypothesize that the head area and the overall body contour (shape) contain the most ID-related cues, while background, accessories, clothes texture, and clothes color result in temporary features. Therefore, we apply several transformations on each input image to (1) remove the subject ID from the scene and create a plain background, (2) generate the

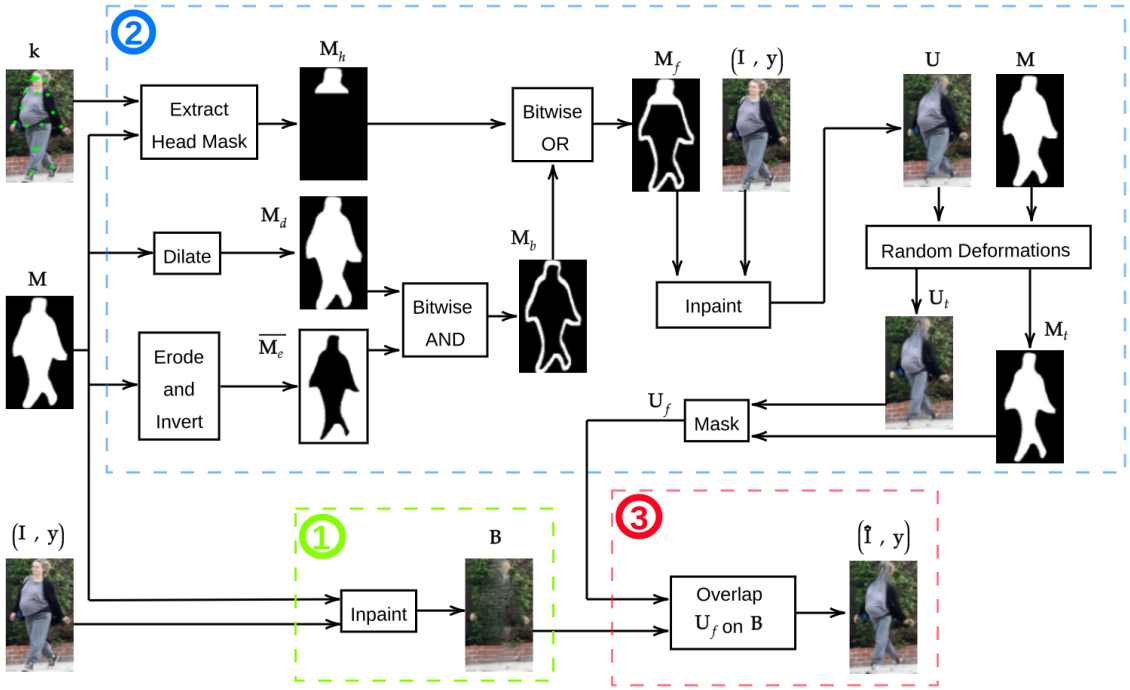


Figure 7.2: Overview of the image transformation pipeline for removing the ID-related cues. $\boldsymbol{k}$, $\boldsymbol{M}$, $\boldsymbol{I}$, y, $\boldsymbol{U}$, and $\boldsymbol{B}$ are respectively the body keypoints, binary mask, RGB image, ID label, transformed image, and reconstructed background of the person. (1) shows the reconstruction of plain background $\boldsymbol{B}$, (2) illustrates the steps to generate distorted foreground area $\boldsymbol{U}_f$, and (3) shows that ID-unrelated image $\hat{\boldsymbol{I}}$ is generated by overlapping $\boldsymbol{U}_f$ over $\boldsymbol{B}$. Best viewed in color.

ID-unrelated foreground, for which we distort the ID-related cues of the person body and face, (3) overlap the ID-unrelated foreground on the plain background. Fig. 7.2 presents an overview of our strategy for generating ID-unrelated images. In the remainder of this section, we explain each of these steps in detail. For simplicity, we skip the index i and use $\boldsymbol{I}$ to denote the i$^{\text{th}}$ original input image, $\boldsymbol{M}$ to refer its corresponding, original body mask, and $\boldsymbol{K} = \{(\boldsymbol{k}_{x1}, \boldsymbol{k}_{y1}), (\boldsymbol{k}_{x2}, \boldsymbol{k}_{y2}), ..., (\boldsymbol{k}_{x17}, \boldsymbol{k}_{y17})\}$ to show the body key-points for this image.

1. To generate a plain background image $\boldsymbol{B}$, we consider the foreground area (subject body) using the mask $\boldsymbol{M}$ as the missing pixels and apply the in-painting method [19] to restore the background area (see the green box in Fig. 7.2).

2. Next, we generate an ID-unrelated foreground area $\boldsymbol{U}_f$ that contains the short-term attributes (illustrated in a blue box in Fig. 7.2). To this end, (a) We use the body key-points $\boldsymbol{K}$ and the full-body mask $\boldsymbol{M}$ to select a head-neck mask $\boldsymbol{M}_h$ from the original mask $\boldsymbol{M}$. (b) In parallel, we should obtain a body contour mask $\boldsymbol{M}_b$, for which we use a method similar to the top-hat morphological transformation. The original body mask $\boldsymbol{M}$ is first expanded using a morphological dilation operator to obtain the mask $\boldsymbol{M}_d$: $\boldsymbol{M} \oplus B = \bigcup_{d \in B} \boldsymbol{M}_d$; (the size of the dilation kernel B is proportional to the size of the original mask. We used 3% of the width and height of the mask in our experiments). Then, we use an erosion morphological operator to shrink the body mask area: $\boldsymbol{M} \ominus B = \bigcap_{e \in B} \boldsymbol{M}_e$. Next, the body contour $\boldsymbol{M}_b$

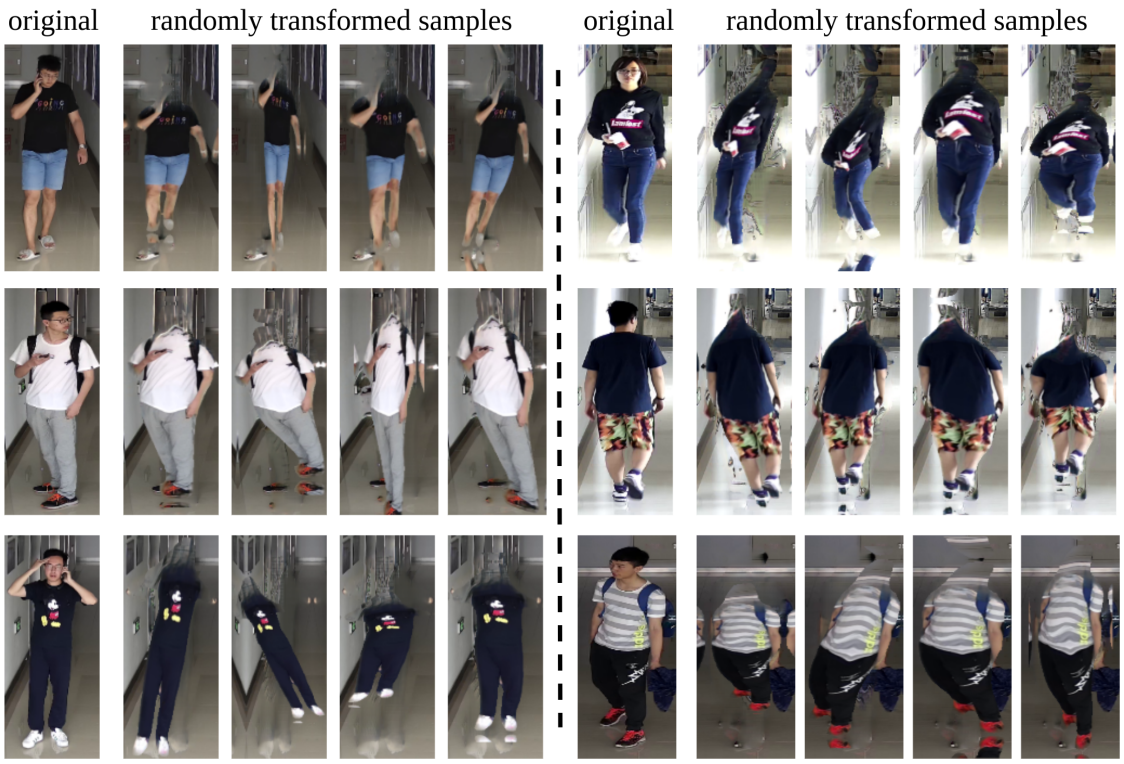


Figure 7.3: Samples of the synthesized data from several subjects in the LTCC dataset. As we intend, the visual identity cues such as face, height, weight, and body shape are distorted successfully.

is obtained by taking the intersection (bitwise AND operation) between the dilated mask $\boldsymbol{M}_d$ and the inversion (bitwise NOT operation) eroded body mask $\overline{\boldsymbol{M}_e}$. (c) A final mask $\boldsymbol{M}_f$ is obtained by adding (bitwise OR operation) the head-neck pixels with the body contour pixels: $\boldsymbol{M}_f = \boldsymbol{M}_b + \overline{\boldsymbol{M}_e}$. (d) ID-related pixels are then in-painted in the input image $\boldsymbol{I}$ using [19] to generate an image ($\boldsymbol{U}$) without any identity information. (e) It is important to deform the overall body shape of the person (by simulating random changes in weight, height, and clothes pattern). We apply this deformation to remove the remained ID-related features. However, to preserve the background area from deformation, we perform the same random transformations on the mask $\boldsymbol{M}$ and the in-painted image $\boldsymbol{U}$; so, in the next step, we could mask out the body area. We use [20] followed by a random stretching in height and width of the body area to apply some image deformations randomly. Precisely speaking, we impose a perturbation mesh on the mask $\boldsymbol{M}$ and image $\boldsymbol{U}$ to alter the subject's silhouette. Then, some points are selected on the mesh to distort the body shape by some random directions and strengths; this mesh deformation is applied by linear interpolation at a pixel-level on both $\boldsymbol{M}$ and $\boldsymbol{U}$. (f) Finally, the deformed foreground area $\boldsymbol{U}_f$ is obtained by masking out the image $\boldsymbol{U}_t$ with $\boldsymbol{M}_t$.

3. The last transformation step in the proposed pipeline overlaps the deformed foreground region $\boldsymbol{U}_f$ on the background $\boldsymbol{B}$, yielding ID-unrelated image $\hat{\boldsymbol{I}}$ (see the red box in Fig. 7.2).

Fig. 7.3 shows some examples of the long-term cloth-changing (LTCC) data set [1] that have been transformed by our pre-processing pipeline due to the removal of their ID-related cues.

7.3.2 Proposed Model: Learning Phase

Learning robust features is a key factor in long-term person re-id. In the context of this task, robustness refers to 1) the extraction of discriminative features from inter-person samples and 2) being invariant to intra-person attribute variations. Although the cross-entropy loss function optimizes these criteria, high variations in the intra-person samples and limited data hinder the model from learning useful long-term representations. *The key to enhance the quality and speed of the learning process of long-term representations of people is to focus on both distilling the identity-related features and disregarding the identity-unrelated features.*

Suppose that the learning set $\mathbb{G} = \{(\boldsymbol{I_i}, y_i, c_j)\}$ consists of n persons with m different clothing styles for each person, where $y_.$ denotes the person-ID label, $c_.$ refers to the clothing label, $i = 1, \ldots, n$ and $j = 1, \ldots, m$. By performing several image transformations on the learning set $\mathbb{G}$, we synthesize another learning data set $\hat{\mathbb{G}} = \{\hat{\boldsymbol{I}}_{\boldsymbol{i}}, y_i, c_j\}$ that excludes the ID-related visual features. This phase was described in the previous subsection.

As shown in the first learning phase in Fig. 7.4 (b), we feed the synthesized data $(\hat{\boldsymbol{I}}_{\boldsymbol{i}}, y_i, c_j)$ to the STE-CNN model $\hat{\phi}(\hat{\mathbb{G}}; \hat{\theta})$ and learn labels y_i, c_j with a cross-entropy loss function. The label y_i, c_j refers to the person i with the ID label y_i with the clothing label c_j; in other words, this network learns to distinguish between the outfits worn by person i. The extracted features of this person are denoted as short-term features $\hat{\boldsymbol{f}}_{\boldsymbol{ij}}$ and are frozen during the next learning phase, where we feed the original image of person i to a second model. Precisely, given the original data $(\boldsymbol{I_i}, y_i, c_j)$ and frozen short-term features $\hat{\boldsymbol{f}}_{\boldsymbol{ij}}$, the LTE-CNN model $\phi(\mathbb{G}, \theta)$ learns the long-term representations $\boldsymbol{f}_i$, such that it is *mathematically dissimilar* to the ID-unrelated feature vector $\hat{\boldsymbol{f}}_{\boldsymbol{ij}}$, while simultaneously learns the ID-related features, using an aggregation loss function:

$$\mathcal{L}_{LTE} = \sum_{i=1}^{n} \frac{\boldsymbol{f}_i . \hat{\boldsymbol{f}}_{ij}}{\|\boldsymbol{f}_i\| \|\hat{\boldsymbol{f}}_{ij}\|} + \sum_{i=1}^{n} t_i \log(s_i), \quad (7.1)$$

where n is the number of person IDs in the learning set, t_i is the ground-truth person ID (label), and s_i denotes the predicted probability score of person i. In equation 1, the cosine-similarity term minimizes the similarity between the short-term and long-term features, while the cross-entropy term helps the LTE-CNN learn the person ID.

Finally, in the inference phase, we only use LTE-CNN model $\phi(\mathbb{G}, \theta)$ to extract the long-term representations of the query and gallery data. Next, similar to the short-term person re-id methods, the gallery set is ordered based on the euclidean distances between the query and gallery samples. Then, the Cumulative Matching Characteristics (CMC) and Mean Average Precision (mAP) metrics are reported as the evaluation criteria.

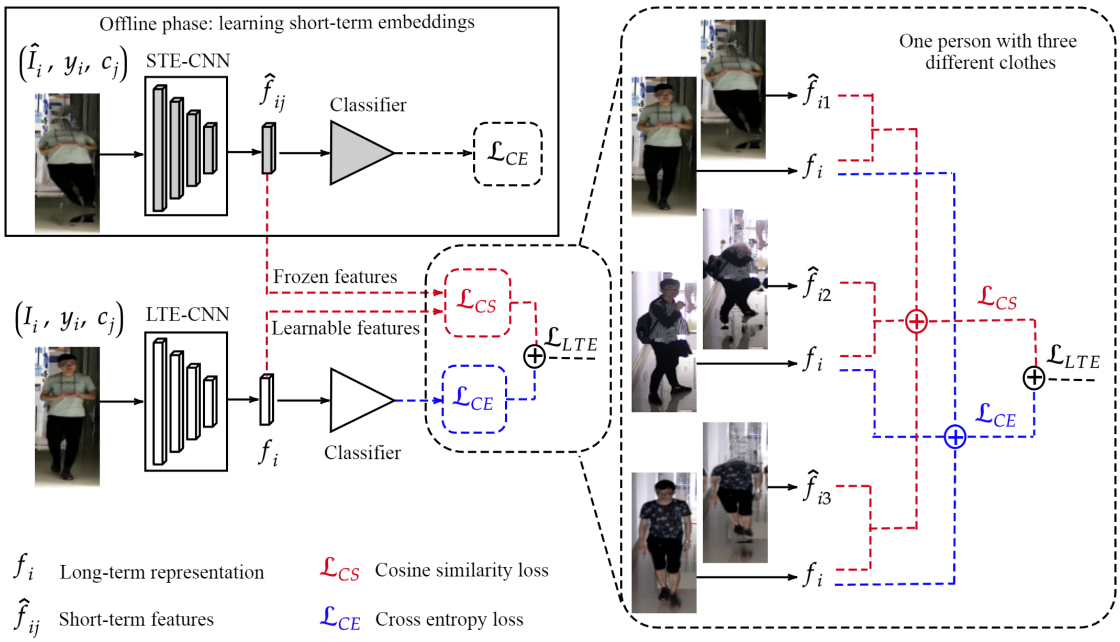


Figure 7.4: Overview of the learning phase of the proposed model. In the offline learning phase, the STE-CNN model receives a transformed image $\hat{\boldsymbol{I}}_{\boldsymbol{i}}$ and extracts its short-term embeddings (ID-unrelated) $\hat{\boldsymbol{f}}_{\boldsymbol{ij}}$. Then, the long-term representation (ID-related) of the original image $\boldsymbol{I}_{\boldsymbol{i}}$ is obtained by minimizing the similarity between the long-term feature vector $\boldsymbol{f}_{\boldsymbol{i}}$ and the frozen short-term embeddings $\hat{\boldsymbol{f}}_{\boldsymbol{ij}}$. The magnified box shows the images of one person with three different clothes and indicates that how LTE-CNN loss function helps to learn the identity of the person (blue traces) and disregard clothing features (red traces). $\boldsymbol{I}_{\boldsymbol{i}}$ refers to the original image of person i with clothing style j, and $\hat{\boldsymbol{I}}_{\boldsymbol{i}}$ is the ID-unrelated version of $\boldsymbol{I}_{\boldsymbol{i}}$. Best viewed in color.

7.4 Experiments and Discussion

7.4.1 Datasets

The Long-Term cloth-changing (LTCC) Person Re-identification dataset [1] was collected using a CCTV system with 12 cameras installed on different floors in an office building. It comprises 24 hours of video recording that were collected over two months. As a result, persons were appeared with substantial changes in lighting, viewing angle, and body pose. The authors used the Mask-RCNN framework [17] to extract the person bounding boxes from video frames and then annotated each bounding box with a person ID and clothing label. The LTCC dataset comprises 17,138 images from 152 identities with 478 outfits, and on average, each person appears with five different clothing outfits. The LTCC dataset is publicly available in two subsets: 1) training subset with 77 individuals, where 46 subjects are wearing different clothes and 31 elements appear with identical garments. 2) testing subset with 76 persons, where 46 people appear with different outfits and 30 individuals are wearing the same clothes.

The Person Re-identification by Contour Sketch (PRCC) dataset [13] was captured indoors using three cameras positioned in separate rooms. The PRCC dataset consists of 221 identities and a total of 33,698 images. In two camera views, the subjects wear the same clothes, while on the other camera, the garments change. Therefore, there are precisely

two different clothes-changes per subject.
We trained and evaluated our model on the LTCC [1] and PRCC [13] long-term re-id datasets, as both comprise real-world data recorded with cameras and are large enough to be suitable for deep architectures. These datasets are publicly available in train and test splits, and there is no overlap between the subjects in the test and train sets. We followed the same evaluation settings in the original papers [1; 13] to have a fair comparison.

Methods	Standard Setting					Cloth-Changing Setting				
	R-1	R-5	R-10	R-50	mAP	R-1	R-5	R-10	R-50	mAP
LOMO [21] + KISSME [22]	26.6	-	-	-	9.1	10.8	-	-	-	5.3
LOMO [21] + NullSpace [23]	34.8	-	-	-	11.9	16.5	-	-	-	6.3
resnet-50 [24]*	9.4	23.2	31.3	59.8	5.9	22.9	43.0	53.9	**77.7**	9.8
Luo *et al.* [25]*	25.8	47.5	57.2	80.6	10.2	11.7	23.8	33.4	62.9	5.9
resnet-50 [24]	49.7	64.9	70.4	86.6	19.7	18.1	32.4	38.8	59.2	8.1
se-resnext [26]	48.3	64.1	71.4	85.4	19.0	20.4	34.2	44.1	63.8	9.3
senet [26]	54.6	70.0	77.9	87.2	21.2	24.2	36.6	45.2	62.0	9.4
resnet50-ibn-a [27]	55.4	69.2	74.4	86.2	23.3	23.7	35.7	42.1	64.0	10.4
HACNN [28]	60.2	-	-	-	26.8	21.9	-	-	-	9.3
MuDeep [29]	61.9	-	-	-	27.5	23.5	-	-	-	10.2
Luo *et al.* [25]	60.2	74.0	80.1	88.8	25.6	24.2	40.6	51.5	71.2	11.3
Qian *et al.* [1]	71.4	-	-	-	34.3	26.2	-	-	-	12.4
Ours (LSD)	72.2	80.3	84.6	**91.9**	31.0	31.4	46.7	54.3	73.5	13.6
LSD + re-ranking [30]	**76.7**	**83.6**	**85.2**	**91.9**	**44.9**	**41.1**	**53.6**	**57.7**	74.0	**19.5**

Table 7.1: Results on the LTCC data set. The method performance on head patches is denoted by $*$ symbol.

7.4.2 Implementation Details

We processed the original image $\boldsymbol{I}$ using the off-the-shelf Mask R-CNN [17] and Alpha-Pose [18] models with default configurations[1] and prepared the inputs of the pre-processing pipeline i.e., $\boldsymbol{K}$ and $\boldsymbol{M}$. The dilation and erosion transformations were performed using a kernel (filter), with a size that is proportional to 3% of the image width. The in-painting technique [19] was also used in its default configurations[2] using the pre-trained weights on the Places2 dataset [31].
The proposed framework, including the STE-CNN and LTE-CNN, can be implemented using any CNN architecture as feature extractors. In this paper, we implemented the proposed model based on residual CNNs using the Pytorch library to evaluate the effectiveness of our method. We started the training phases by fine-tuning the ImageNet pre-trained weights, using the Adam optimizer [32], for 250 epochs. The input images were 256×128 for both networks, i.e., STE-CNN and LTE-CNN. For more implementation details, we refer the readers to the project page at `https://github.com/canarybird33/YouLookDifferent`.

[1] `https://github.com/matterport/Mask_RCNN`, `https://github.com/MVIG-SJTU/AlphaPose`
[2] `https://github.com/Atlas200dk/sample-imageinpainting-HiFill`

Methods	Cameras A and C (different clothes)				Cameras A and B (same clothes)			
	R-1	R-10	R-20	mAP	R-1	R-10	R-20	mAP
[21] + [22]	18.6	49.8	67.3	-	47.4	81.4	90.4	-
[33] + [34] + [21]	23.7	62.0	74.5	-	54.2	84.1	91.2	-
resnet-50 [24]	24.1±10.8	56.9±2.4	68.5±3.3	35.3±6.6	76.3±5.0	94.0±1.6	97.4±0.6	82.6±3.8
se-resnext101 [26]	27.7±1.7	57.6±6.6	70.3±3.5	37.8±2.1	69.1±8.9	94.4±4.0	97.6±2.2	78.4±5.1
senet [26]	27.2±4.7	54.5±4.9	66.9±1.7	36.6±3.2	76.7±4.6	96.0±1.7	97.9±0.6	83.9±2.6
resnet-ibn-a [27]	32.9±6.7	67.2±4.7	81.6±3.7	44.1±5.0	84.8±3.6	98.3±1.5	99.5±0.4	89.8±2.0
HACNN [28]	21.8	59.5	67.5	-	82.5	98.1	99.0	-
PCB [35]	22.9	61.2	78.3	-	86.9	98.8	99.6	-
DCN [36]	26.0	71.7	85.3	-	61.9	92.1	97.7	-
STN [37]	27.5	69.5	83.2	-	59.2	91.4	96.1	-
Yang *et al.* [13]	34.4	**77.3**	**88.1**	-	64.2	92.6	96.7	-
Ours (LSD)	37.2±6.7	68.7±2.0	80.5±4.1	47.6±3.4	93.6±1.7	99.5±0.6	99.8±0.1	95.8±1.1
Ours + re-ranking	**42.7**±4.2	71.2±3.5	81.5±2.4	**52.2**±2.2	**97.9**±0.4	**99.8**±0.0	**99.9**±0.0	**98.7**±0.1

Table 7.2: Results for two settings of the PRCC data set: 1) when the query person appears with different clothes in the gallery set (at left-side), 2) when the query's outfit is not changed in the gallery set (at left-side). The locally performed evaluations were repeated 10 times, and the variances from the mean values were shown by ±.

7.4.3 Results

7.4.3.1 LTCC Dataset

To evaluate our model on the LTCC dataset [1], we considered the two settings suggested in the original paper [1]: 1) standard setting, in which we ignore those images of the gallery that have captured from the same person and same camera. 2) cloth-changing setting, where the images of the same person with identical clothes captured with the same camera are discarded from the gallery before ranking the gallery elements based on the query person.

We provide a comparison between our model performance to several baselines in Table 7.1. In general, our model shows superior performance for both the evaluation metrics: mAP and CMC for ranks 1 to 50.

As shown in the middle column of Table 7.1, in standard evaluation setting, the hand-crafted based methods can extract better feature representations (from full-body images of persons) in comparison with simple baselines [24; 25], when simple baselines are learned based on the face/head patches. In the next performance level, resnet50-ibn-a [27] achieves 55.4% and 23.3% of rank-1 and mAP, respectively; these numbers improve by the short-term re-id baselines, specifically to 61.9% and 27.5% by [29]. As a long-term re-id framework, Qian *et al.* [1] presents competitive results (71.4%/34.3% of rank-1/mAP) compared to our method without re-ranking (72.2%/31.0% of rank-1/mAP). However, by applying the re-ranking technique [30] on our results, our method consistently outperforms the other competitors and achieves 76.7%/44.9% of rank-1/mAP.

In Table 7.1 section cloth-changing evaluation setting, it is noticeable that the performance of the short-term re-id methods [25; 28; 29] roughly degrades to their one-third, which denote that these methods heavily rely on the color and texture of the clothes to re-id people. It is also interesting that a resnet-50 model could extract more useful long-term

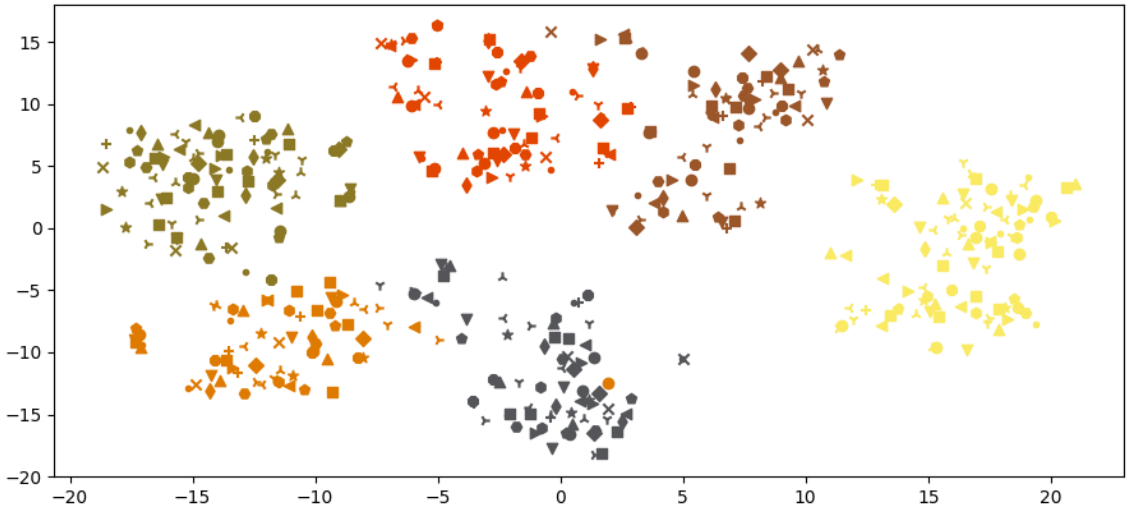


Figure 7.5: Visualization of the long-term representations, according to t-SNE [38], for six IDs with varying clothes (LTCC test set). The data related to each person are presented in a different color, and variety in outfits is denoted by different markers. Best viewed in color.

information from head-shots (22.9%/9.8% of rank-1/mAP) rather than full-body images 18.1%/8.1% of rank-1/mAP, whereas the short-term model [25] fails in the head-shot long-term re-id setting, by achieving 24.2%/11.3% of rank-1/mAP from the full-body images and obtaining 11.7%/5.9% of rank-1/mAP from the head patches. In the cloth-changing context, our method obtains better re-id results with 31.4%/13.6% of rank-1/mAP before the re-ranking process and 41.1%/19.5% of rank-1/mAP after re-ranking the re-id retrieval list, which indicates the superiority of our approach in comparison with all the other methods, specifically [1] that achieves 26.2%/12.4% of rank-1/mAP.
Fig. 7.5 shows t-SNE [38] visualization of long-term representations provided with our proposed method for several persons from the LTCC test set that are wearing various clothing outfits. The representations related to consecutive frames of the same person with the same clothes are not close to each other, indicating that our method does not rely on the appearance similarity to re-identify people.

7.4.3.2 PRCC Dataset

As previously mentioned, the PRCC dataset was collected using three cameras, namely A, B, and C, such that the individuals' clothes in cameras A and B are the same, while in the camera C, subjects wear different outfits. Following the evaluation protocol in [13], we select one image of each person from camera A and build a one-shot gallery, while samples captured by the other two cameras are considered to be as the queries for two different settings: evaluation on the cloth-changing and cloth-consistent settings.
Table 7.2 shows the performance of several baselines versus our method on the PRCC dataset. The baseline could be roughly divided to four groups: 1) methods based on the hand-crafted features [21; 22; 33; 34], 2) plain deep residual networks [24; 26; 27], 3) short-term person re-id techniques [28; 35–37], and 4) long-term re-id method [13].
In general, methods based on the hand-crafted features obtain the lowest recognition

results, with the rank-1 accuracy less than 24% and 55% in the cloth-changing and standard settings, respectively, whereas the second group of methods could achieve a rank-1/mAP approximately between 24%/35% to 33%/44% in the cloth-changing scenario and between 70%/78% to 85%/90% in the standard-setting. Interestingly, the short-term re-id techniques could improve the rank-1 results up to 86.9%, only when the inquiry person wears consistent clothing outfits in the gallery. When the query person appears with different clothing styles, our method achieves 37.2%/47.6% for rank-1/mAP (and 42.7%/52.2% after the re-ranking process), while the approach presented by Yang *et al.* [13] obtains a rank-1 accuracy of around 34.4%. Moreover, when people wear identical clothes in the query and gallery sets, our method still outperforms all the baselines with 93.6% rank-1 and 95.8% mAP; these numbers could even be improved to 97.9% and 98.7%, respectively, when we apply the re-ranking technique [30] on the obtained ranking list.

7.4.3.3 Discussion

As indicated in Tables 7.1 and 7.2, the proposed method has improved the long-term re-id accuracy, while it can provide reliable results for short-term re-id task. Our interpretation of the superior performance of our method in both tasks is that, holistic CNNs can provide discriminative representation based on the identity (rather than clothes and background) when we use an aggregation loss function, in which we learn the ID labels using a cross-entropy loss term and penalize the learning of the ID-unrelated features by a similarity loss term. In fact, learning the identity cues by an aggregation loss function *implicitly* prevents the model from predicting the identity of people based on their clothes and background. Whereas, architectural based design may *explicitly* limit the model, which results into better long-term re-id but worse short-term re-id accuracy.

7.5 Ablation Studies

We performed several experiments with different backbones and input image sizes to evaluate the performance of the proposed LSD model in various conditions and find the limits of our method. The experiments in this section were carried out on the LTCC dataset, and the LSD model was trained for 50 epochs, and results were reported after the re-ranking process. The other settings remained as same as the previous experiments.
Left section of Table 7.3 shows the experiment results of the LSD for five different image resolutions from 32×16 to 512×256 and indicates that the improvement of the rank-1 accuracy saturates when the size of the images is increased from 256×128 to 512×256. In contrast, the mAP increases sharply in cloth-changing settings, from 13.7% to 17.4%. The reason behind the variation of accuracy is that, when we reduce the size of the images, some critical information (details probably) are lost permanently, whereas when we resize the images to 512×256, no extra detail are induced from data, probably because of the limits imposed by the image-quality of the captured data from far distances by the surveillance cameras . Furthermore, we trained and evaluated our

Architecture	SS		CCS		Input Resolution	SS		CCS	
	R-1	mAP	R-1	mAP		R-1	mAP	R-1	mAP
resnet50	52.3	26.0	20.4	10.0	32×16	21.5	9.8	8.4	4.6
resnet101	47.9	24.9	17.1	10.0	64×32	43.2	23.1	15.3	8.8
resnet152	51.7	26.2	18.9	10.1	128×64	62.5	35.6	24.7	12.7
se-resnet101	56.2	29.6	22.4	11.6	256×128	70.0	39.5	35.2	13.7
se-resnet152	55.0	28.7	21.4	10.2	512×256	69.8	41.4	35.7	17.4
se-resnext101	55.8	27.9	23.0	11.4					
resnet50-ibn-a	57.8	30.0	23.7	11.5					
senet154	58.6	29.1	27.8	11.7					

Table 7.3: The performance of the proposed LSD model with different residual backbones and input resolutions, when trained for 50 epochs on the LTCC data set. When architecture is changing, the input resolution is fixed to 256×128, and when input resolution is changing, the senet154 architecture is used. SS and CCS stand for Standard Setting and Cloth-Changing Setting, respectively.

model with several different feature extraction backbones. As shown in the right section of Table 7.3, the se-resnet models achieve better results than plain resnet methods. The proposed framework achieves better results when implemented based on the resnet50-ibn-a, with 57.8%/30.0% and 23.7%/11.5% of rank-1/mAP for the standard and cloth-changing settings, respectively. Moreover, these numbers improve to 58.6%/29.1% and 27.8%/11.7%, when the senet154 model is used as the backbone feature extractor.

7.6 Conclusions

Long-term person re-id aims to retrieve a query ID from a gallery, where elements are expected to appear with different clothing, hairstyles, or additional accessories. This is an extremely ambitious identification setting, where the majority of the existing re-id methods still have poor performance. Hence, it is critical to find alternate feature representations that are naturally insensitive to short-term re-id features. Moreover, manually annotating large amounts of long-term instances for feeding supervised classification frameworks might be an insurmountable task, not only due to the lack of available data but also to the number of human resources required for the task. Based on these observations, this paper describes an LSD model, which its most innovative point is to naturally learn long-term representations of persons while ignoring the typically varying short-term attributes (clothing style, body shape, and background). To this end, we propose an image transformation pipeline over the ID-related regions (the head and the body shape) and create a model (STE-CNN) that identifies the most relevant short-term features. These representations are then separated from the long-term representation via the cosine similarity loss function. The experimental results on the state-of-the-art cloth-changing benchmarks confirmed the effectiveness of the proposed method by consistently advancing the performance of the best performing techniques.

7.7 Acknowledgments

This work was supported in part by the FCT/MEC through National Funds and Co-Funded by the FEDER-PT2020 Partnership Agreement under Project UIDB/50008/2020, Project POCI-01-0247-FEDER-033395 and in part by operation Centro-01-0145-FEDER-000019 - C4 - Centro de Competências em Cloud Computing, co-funded by the European Regional Development Fund (ERDF) through the Programa Operacional Regional do Centro (Centro 2020), in the scope of the Sistema de Apoio à Investigação Científica e Tecnológica - Programas Integrados de IC&DT. This research was also supported by 'FCT - Fundação para a Ciência e Tecnologia' through the research grant 'UI/BD/150765/2020'.

Bibliography

[1] X. Qian, W. Wang, L. Zhang, F. Zhu, Y. Fu, T. Xiang, Y.-G. Jiang, and X. Xue, "Long-term cloth-changing person re-identification," *arXiv preprint arXiv:2005.12633*, 2020. 137, 139, 140, 144, 145, 146, 147, 148

[2] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. IEEE ICCV*. Springer, 2016, pp. 17–35. 137

[3] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE ICCV*, 2015, pp. 1116–1124. 137

[4] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu, "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Trans Multimedia*, vol. 22, no. 10, pp. 2597–2609, 2020. 137

[5] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *arXiv preprint arXiv:2001.04193*, 2020. 137, 139

[6] B. Lavi, I. Ullah, M. Fatan, and A. Rocha, "Survey on reliable deep learning-based person re-identification models: Are we there yet?" *arXiv preprint arXiv:2005.00355*, 2020.

[7] M. O. Almasawa, L. A. Elrefaei, and K. Moria, "A survey on deep learning-based person re-identification systems," *IEEE Access*, vol. 7, pp. 175 228–175 247, 2019.

[8] E. Yaghoubi, A. Kumar, and H. Proença, "Sss-pr: A short survey of surveys in person re-identification," *Pattern Recognit. Lett.*, vol. 143, pp. 50–57, 2021. 137, 139

[9] J. Dietlmeier, J. Antony, K. McGuinness, and N. E. O'Connor, "How important are faces for person re-identification?" *arXiv preprint arXiv:2010.06307*, 2020. 137

[10] Z. Yu, Y. Zhao, B. Hong, Z. Jin, J. Huang, D. Cai, X. He, and X.-S. Hua, "Apparel-invariant feature learning for apparel-changed person re-identification," *arXiv preprint arXiv:2008.06181*, 2020. 137, 138, 140

[11] F. Wan, Y. Wu, X. Qian, Y. Chen, and Y. Fu, "When person re-identification meets changing clothes," in *Proc. CVPRW*, 2020, pp. 830–831. 138

[12] Y.-J. Li, Z. Luo, X. Weng, and K. M. Kitani, "Learning shape representations for clothing variations in person re-identification," *arXiv preprint arXiv:2003.07340*, 2020. 138, 140

[13] Q. Yang, A. Wu, and W.-S. Zheng, "Person re-identification by contour sketch under moderate clothing change," *IEEE TPAMI*, pp. 1–1, 2019. 138, 139, 140, 145, 146, 147, 148, 149

[14] P. Zhang, Q. Wu, J. Xu, and J. Zhang, "Long-term person re-identification using true motion from videos," in *Proc. WACV*. IEEE, 2018, pp. 494–502. 138, 139

[15] J. Xue, Z. Meng, K. Katipally, H. Wang, and K. van Zon, "Clothing change aware person identification," in *Proc. CVPRW*, 2018, pp. 2112–2120. 138, 139

[16] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, and X. Xue, "Pose-normalized image generation for person re-identification," in *Proc. IEEE ICCV*, 2018, pp. 650–667. 140

[17] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proc. IEEE ICCV*, 2017, pp. 2961–2969. 141, 145, 146

[18] Y. Xiu, J. Li, H. Wang, Y. Fang, and C. Lu, "Pose flow: Efficient online pose tracking," *arXiv preprint arXiv:1802.00977*, 2018. 141, 146

[19] Z. Yi, Q. Tang, S. Azizi, D. Jang, and Z. Xu, "Contextual residual aggregation for ultra high-resolution image inpainting," in *Proc. CVPR*, 2020, pp. 7508–7517. 142, 143, 146

[20] K. Ma, Z. Shu, X. Bai, J. Wang, and D. Samaras, "Docunet: Document image unwarping via a stacked u-net," in *Proc. CVPR*, 2018, pp. 4700–4709. 143

[21] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. CVPR*, 2015, pp. 2197–2206. 146, 147, 148

[22] A. Kittur, E. H. Chi, and B. Suh, "Crowdsourcing user studies with mechanical turk," in *Proc SIGCHI Conf Hum Factor Comput Syst*, 2008, pp. 453–456. 146, 147, 148

[23] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proc. CVPR*, 2016, pp. 1239–1248. 146

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778. 146, 147, 148

[25] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. CVPRW*, 2019, pp. 1487–1495. 146, 147, 148

[26] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, 2018, pp. 7132–7141. 146, 147, 148

[27] X. Pan, P. Luo, J. Shi, and X. Tang, "Two at once: Enhancing learning and generalization capacities via ibn-net," in *Proc. ECCV*, September 2018. 146, 147, 148

[28] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. CVPR*, 2018, pp. 2285–2294. 146, 147, 148

[29] X. Qian, Y. Fu, T. Xiang, Y.-G. Jiang, and X. Xue, "Leader-based multi-scale attention deep architecture for person re-identification," *IEEE TPAMI*, vol. 42, no. 2, pp. 371–385, 2019. 146, 147

[30] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proc. CVPR*, 2017, pp. 1318–1327. 146, 147, 149

[31] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE TPAMI*, vol. 40, no. 6, pp. 1452–1464, 2017. 146

[32] K. Da, "A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. 146

[33] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognit.*, vol. 29, no. 1, pp. 51–59, 1996. 147, 148

[34] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. CVPR*, vol. 1. IEEE, 2005, pp. 886–893. 147, 148

[35] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. ECCV*, 2018, pp. 480–496. 147, 148

[36] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. ICCV*, 2017, pp. 764–773. 147

[37] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc NIPS - Volume 2*, ser. NIPS'15. Cambridge, MA, USA: MIT Press, 2015, p. 2017–2025. 147, 148

[38] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008. 148

Chapter 8

Conclusions

8.1 Summary

Ubiquitous CCTV cameras have raised the desire for human attribute estimation and person re-identification in crowded urban environments. Given that face close-shots are rarely available at far distances, feature extraction from the body is of practical interest nowadays. However, full-body data is accompanied by a wide background area and has more complexity in terms of viewpoint variations and occlusions. The primary solution to tackle these general challenges is to provide large learning data because deep neural networks can automatically provide a discriminative and comprehensive feature representation from critical regions of the input data. As huge data collection and annotation is expensive, developing approaches to address data-dependent challenges such as imbalanced class data in PAR tasks and cloth-changing person re-id is important.
To study the above-mentioned difficulties, in the scopes of this research, we first reviewed existing PAR and person re-id approaches, including the state-of-the-art architectures, recent datasets, and future directions with a focus on deep learning methods. Then, we proposed several novel frameworks for both PAR and person re-id and evaluated the performance of our approaches on several well-known publicly available datasets and compared our experimental results with the recent existing methods.

8.2 Summary of Contributions

The main contributions of this thesis are as follows.

- We provide a comprehensive survey on the PAR approaches and benchmarks with an emphasis on deep learning methods. We study the typical pipeline of the HAR systems, which is started by data preparation and continues with designing a model to be trained and evaluated. We then highlight several factors that are required to be considered for developing an optimal HAR framework, pointing that 1) we should design an end-to-end model that predicts multiple attributes at once; 2) the model should extract a discriminate and comprehensive features representation from each instance of the dataset; 3) we should consider the location of each attribute on the body of the person; 4) model should deal with general challenges such as low-quality data, pose variation, illumination variation, cluttered background, and occlusion; 5) model should handle the class imbalanced data and avoid to over-fit or and under-fit on some classes; 6) model should manage the limited-data problem effectively, for example, by using data augmentation techniques or learning from synthesized data. Next, we propose a challenged-based taxonomy for HAR approaches and categorize the existing methods in five general groups, based on which, we conclude that the most recent HAR methods study the effects of the attribute localization and attribute correlations in the performance of the model. Finally, we provide a comprehensive study on the HAR benchmarks based on the data content: face, full-body, fashion style, and synthetic data.

- We conduct a short survey of surveys on person re-id methods and propose a multi-dimensional taxonomy that distinguishes between person re-id models, based on their main approach, type of learning, identification settings, the strategy of learning, data modality, type of queries and context. Most of the existing state-of-the-art methods could be studied under the strategy point-of-view that explains architecture based methods and data augmentation techniques. We then discuss some privacy and security concerns caused by processing people's personal data via surveillance systems. Finally, we explain some biases and problems in the literature of person re-id, such as unfair comparison of methods, low originality in techniques, and insufficient attention to some of the important perspectives in the problem.
- As gender attribute is often one of the primary properties of people, we present a multi-branch framework that provides a comprehensive and discriminative representation of persons. The proposed solution uses several pose-specialized CNNs to extract the features of different regions of interest and aggregates the output scores of CNN branches. To evaluate the performance of the model, we trained and tested the algorithm on the BIODI and PETA datasets. Our experimental results confirm that CNNs specialized in predicting the gender attribute from images cropped by the *convhull* of full-body keypoints can achieve better results than CNNs that work on the head crops or raw full-body images. Overall, surveillance data have low resolutions and predicting the gender on head crops yields poor accuracy, whereas predicting from raw images suffers from interference of background features in the final feature representation of person.
- Inspired by our previous observations about the adverse effects of background features in the model performance, we propose a multiplication layer that explicitly filters the background features. The proposed model works with full-body images of pedestrians that are captured in uncontrolled environments and has a multi-task architecture that yields multiple soft biometrics of persons at once. The task-oriented architecture is integrated with a weighted loss function that relativizes the importance of each class of attributes and handles the imbalanced PAR data. The evaluation of our method on the PETA and RAP datasets shows the superiority of the proposed framework with respect to the state of the arts.
- We propose an image transformation technique that helps to implicitly define the receptive fields of CNNs in the short-term person re-id task. The receptive fields determine the critical regions of the input data that are correlated with the label information. Therefore, to assist the inference model to find the important regions efficiently, we generate a synthesized learning dataset in which the irrelevant (e.g., background) and important (e.g., body area) regions of the original data are swapped, and the label of the synthesized data is inherited from the image that has shared its important region. This solution can be implemented as a data augmentation technique, which means that we can skip the computation expenses of the image transformation process during the inference phase. Further, our solution preserves the label information and is parameter-learning-free. The experimental results on several datasets such as RAP, Market1501, and MSMT-V2 datasets confirm the effectiveness of the proposed solution for the person re-id task from full-body images in the wild.
- CNNs are dominated by the texture-based feasters, resulting in a challenge to learn long-term person re-id, in which people appear with different clothes that have been seen before. To address this problem, we present a long-term, short-term decoupler model that, regardless of the cloth and background texture, captures the identity-based features resulting from height, weight, body shape, and head area. To this end, we propose an image

Figure 8.1: Comparison between synthesized data of face and full-body of persons. The two first rows show the face examples generated using StyleGan when trained on the celebaHQ dataset. The second row illustrates the instances that StyleGan has failed to produce flawless images. The other two rows illustrate the full-body examples generated by StyleGan with the same settings when trained on the RAP dataset.

transformation chain to synthesize some data from original images such that the identity characteristics of a person are distorted. We then train a CNN model on the synthesized data to obtain the ID-unrelated feature of each instance of the learning set. Latter, we train another CNN model on the original data and use the ID-unrelated features in a cosine similarity loss function to focus on learning the ID-related features. This way, in the training phase, the model learns that the background and clothing texture is not correlated to the identity of the person. Therefore, in the inference phase, we only use the second model (and skip the image transformation processes) to predict the identity of the query person. The experimental results on the cloth-changing benchmarks (PRCC and LTCC) confirm the superiority of the proposed solution compared to the state of the arts.

8.3 Future Research Directions

PAR and person re-id fields of study are at early stages, and there are many possibilities for future works. In the following, we enumerate some future directions that are rarely discussed in the literature.

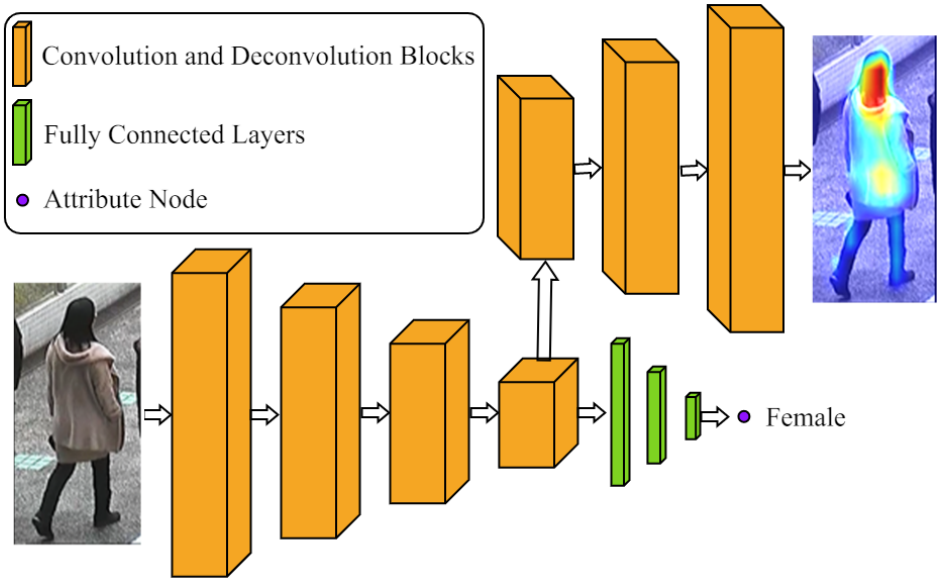


Figure 8.2: A rough example of a visually interpretable PAR model with an extra head to show the active receptive fields when making a prediction.

8.3.1 Limited Data

Deep neural networks require massive learning data to improve their performance. However, the process of data collection and annotation is costly and time-consuming. Recently, generative models have shown impressive evolution in synthesizing high-quality human face data (see Fig. 8.1). However, existing full-body generative models produce unsatisfactory results, mainly because of the wide variations in the full-body data and small learning sets. As shown in Fig. 8.1, details in the generated full-body images (e.g., facial attributes) are mainly missed, and there are samples without hands or do not follow the logical structure of the humans such that the frontal upper body has been attached to a backward lower body. There are also some examples the model has failed to build the general structure of the body. To overcome these challenges, future works may study either novel generative architectures to create visually pleasant full-body data or propose rich datasets to enhance the quality of the learning phase of existing models. For instance, adding a constrain term to the loss function of the generator model –that could be based on the body pose information of the real data– can help the model converge sooner and prevent from generating illogical body structures.

8.3.2 Explainable Architectures

The performance of the state-of-the-art deep models is impressive, yet most of them cannot vividly mention the reasons behind the decisions made. The reliability of PAR and person re-id systems enhances when we highlight the essential information that results in the final predictions. For example, PAR frameworks that estimate, e.g., hair color and style attributes, could have an extra output to highlight that the estimation is based on the information extracted from the people's head region. A rough example of an explainable PAR model is illustrated in Fig. 8.2, where the model has one extra head that yields a heatmap to show the pixels that have led to the classification result.

8.3.3 Prior-Knowledge Based Learning

Providing prior human knowledge to the person re-id and PAR models can help mimic human recognition ability in some aspects. For example, in an outdoor environment in the winter season, it is hardly expected that people appear with summer clothing style. Similarly, while it is natural for someone with sport suits to work out, it is unexpected that someone with formal clothes has quick motions. Therefore, the accumulation of useful information such as scene understanding, human-environment interaction estimation, and activity recognition may improve the performance of person re-id and PAR systems.

Chapter 9

Anexos

Some other publications that extend the objectives of this thesis and are resulted from this doctoral research program are as follows. These research articles have not been included in the main body of the manuscript.

1696 IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, VOL. 16, 2021

The P-DESTRE: A Fully Annotated Dataset for Pedestrian Detection, Tracking, and Short/Long-Term Re-Identification From Aerial Devices

S. V. Aruna Kumar, Ehsan Yaghoubi, *Member, IEEE*, Abhijit Das, *Member, IEEE*, B. S. Harish, and Hugo Proença, *Senior Member, IEEE*

***Abstract*— Over the years, unmanned aerial vehicles (UAVs) have been regarded as a potential solution to surveil public spaces, providing a cheap way for data collection, while covering large and difficult-to-reach areas. This kind of solutions can be particularly useful to detect, track and identify subjects of interest in crowds, for security/safety purposes. In this context, various datasets are publicly available, yet most of them are only suitable for evaluating detection, tracking and *short-term re-identification* techniques. This paper announces the free availability of the P-DESTRE dataset, the first of its kind to provide video/UAV-based data for *pedestrian long-term re-identification* research, with ID annotations consistent across data collected in different days. As a secondary contribution, we provide the results attained by the state-of-the-art pedestrian detection, tracking, short/long term re-identification techniques in well-known surveillance datasets, used as baselines for the corresponding effectiveness observed in the P-DESTRE data. This comparison highlights the discriminating characteristics of P-DESTRE with respect to similar sets. Finally, we identify the most problematic data degradation factors and co-variates for UAV-based automated data analysis, which should be considered in subsequent technologic/conceptual advances in this field. The dataset and the full specification of the empirical evaluation carried out are freely available at http://p-destre.di.ubi.pt/.**

***Index Terms*— Visual surveillance, aerial data, pedestrian detection, object tracking, pedestrian re-identification, pedestrian search.**

I. Introduction

VIDEO-BASED surveillance refers *the act of watching a person or a place, esp. a person believed to be involved with criminal activity or a place where criminals gather*.[1] Over the years, this technology has been used in far more applications than its roots in crime detection, such as traffic control and management of physical infrastructures. The first generation of video surveillance systems was based in closed-circuit television (CCTV) networks, being limited by the stationary nature of cameras. More recently, unmanned aerial vehicles (UAVs) have been regarded as a solution to overcome such limitations: UAVs provide a fast and cheap way for data collection, and can easily assess confined spaces, producing minimal noise while reducing the staff demands and cost. UAV-based surveillance of crowds can host crime prevention measures throughout the world, but it also raises a sensitive debate about faithful balances between security/privacy issues. In this context, it is important that legal authorities strictly define the cases where this kind of solutions can be used (e.g., missing child or disoriented elderly? Criminal seek?).

Being at the core of video surveillance, many efforts have been concentrated in the development of video-based pedestrian analysis methods that work in real-world conditions, which is seen as a *grand challenge*.[2] In particular, the problem of identifying pedestrians in crowds is especially difficult when the time elapsed between consecutive observations denies the use of clothing-based features (bottom row of Fig. 1).

To date, the research on pedestrian analysis has been mostly conducted on databases (e.g., [11], [17], and [30]) that provide data with short lapses of time between consecutive observations of each ID (typically within a single day), which allows to use clothing-based appearance features for identification (top row of Fig. 1). Also, datasets related to other problems are used (e.g., gait recognition [38]), where the data acquisition conditions are evidently different of the seen in surveillance environments.

As a tool to support further advances in video/UAV-based pedestrian analysis, the P-DESTRE is a joint effort from research groups in two universities of Portugal and India. It is a multi-session set of videos, taken in outdoor crowded environments. "*DJI Phantom 4*"[3] drones controlled by human

Manuscript received April 7, 2020; revised September 21, 2020 and October 30, 2020; accepted November 12, 2020. Date of publication November 26, 2020; date of current version December 21, 2020. This work was supported in part by the FCT/MEC through National Funds and Co-Funded by the FEDER-PT2020 Partnership Agreement under Project UIDB/EEA/50008/2020, Project POCI-01-0247-FEDER-033395 and in part by the C4: Cloud Computing Competence Centre. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Siwei Lyu. *(Corresponding author: Hugo Proença.)*

S. V. Aruna Kumar is with the Department of Computer Science and Engineering, Ramaiah University of Applied Sciences, Bengaluru 560054, India (e-mail: arunkumarsv55@gmail.com).

Ehsan Yaghoubi and Hugo Proença are with the IT: Instituto de Telecomunicações, Department of Computer Science, University of Beira Interior, 6201-001 Covilhã, Portugal (e-mail: d2389@ubi.pt; hugomcp@di.ubi.pt).

Abhijit Das is with the Indian Statistical Institute, Kolkata 700108, India (e-mail: abhijitdas2048@gmail.com).

B. S. Harish is with the Department of Information Science and Engineering, JSS Science and Technology University, Mysuru 570006, India (e-mail: bsharish@jssstuniv.in).

Digital Object Identifier 10.1109/TIFS.2020.3040881

[1] https://dictionary.cambridge.org/dictionary/english/surveillance
[2] https://en.wikipedia.org/wiki/Grand_Challenges
[3] https://www.dji.com/pt/phantom-4

Fig. 1. Key difference between the pedestrian *short-term re-identification* (upper row) and *long-term re-identification* problems (bottom row). In the former case, it is assumed that subjects keep the same clothes between consecutive observations, which does not happen in the *long-term* problem. Matching IDs across long-term observations is highly challenging, as the state-of-the-art re-identification techniques rely in clothing appearance-based features. **The P-DESTRE set is the first to supply video/UAV-based data for pedestrian *long-term re-identification***.

operators flew over various scenes of both universities *campi*, with the data acquired simulating the everyday conditions in surveillance environments. All subjects offered explicitly as volunteers and they were asked to act normally and ignore the UAVs. Moreover, the P-DESTRE set is fully annotated at the frame level by human experts, providing four families of meta-data:

- **Bounding boxes**. The position of each pedestrian at every frame is given as a bounding box, to support object detection, tracking and semantic segmentation experiments;
- **IDs**. Each pedestrian has a unique identifier that is kept consistent over all the data acquisition days/sessions. This is a singular characteristic that turns the P-DESTRE suitable for various kinds of identification problems. The *unknown* identities are also annotated, and can be used as distractors to increase the identification challenges;
- **Soft biometrics labels**. Each pedestrian is fully characterised by 16 labels: {'gender', 'age', 'height', 'body volume', 'ethnicity', 'hair colour', 'hairstyle', 'beard', 'moustache', 'glasses', 'head accessories', 'body accessories', 'action' and 'clothing information' (x3)}, which allows to perform soft biometrics and action recognition experiments.
- **Head pose**. 3D head pose angles are given in terms of *yaw*, *pitch* and *roll* values for all the bounding boxes, except backside views. This information was automatically obtained according to the Deep Head Pose [29] method.

As a consequence of its annotation, the P-DESTRE is the first suitable for evaluating video/UAV-based *long-term re-identification* methods. Using data collected over large periods of time (days/weeks), the re-identification techniques cannot rely in clothing-based features, which is the key characteristic that distinguishes between the *long-term* and the *short-term* re-identification problems (Fig. 1).

In summary, this paper offers the following contributions:

1) we announce the free availability of the P-DESTRE dataset, the first of its kind that is fully annotated at the frame level and was designed to support the research on video/UAV-based long-term re-identification. Moreover, the P-DESTRE set can be used in pedestrian detection, tracking, short-term re-identification and soft biometrics experiments;
2) we provide a systematic review of the related work in the scope of the P-DESTRE set, comparing its main discriminating features with respect to the related sets;
3) based in our own empirical evaluation, we report the results that state-of-the-art methods attain in the pedestrian detection, tracking and short-term re-identification tasks, when considering well-known surveillance datasets. The comparison between such results and those attained in P-DESTRE supports the originality of the novel dataset.

The remainder of this paper is organized as follows: Section II summarizes the most relevant research in the scope of the novel dataset. Section III provides a detailed description of the P-DESTRE data. Section IV discusses the results observed in our empirical evaluation, and the conclusions are given in Section V.

II. Related Work

This section describes the most relevant UAV-based datasets and also pays special attention to datasets that focus the problems of pedestrian detection, tracking, re-identification and search.

A. UAV-Based Datasets

Various datasets of UAV-based data are available to the research community, most of them serving for object detection and tracking purposes. The 'Object deTection in Aerial images' [35] set supports research on multi-class object detection, and has 2,806 images, with 188K instances of 15 categories. The 'Stanford drone dataset' [28] provides video data for object tracking, containing 60 videos from 8 scenes, annotated for 6 classes. Similarly, the 'UAV123' [24] set provides 123 video sequences from aerial viewpoints, containing over 110K frames, annotated for object detection/tracking. The 'VisDrone' [40] consists of 288 videos/261,908 frames, with over 2.6M bounding boxes covering pedestrians, cars, bicycles, and tricycles. Finally, the largest freely available source is the 'Multidrone' [23], providing data for multiple category object detection and tracking. It contains videos of various actions, collected under various weather conditions and in different places, yet not all the data are annotated. The 'UAVDT' [9] is an image-based dataset that supports research on vehicle detection and tracking. It has 80K frames/ 841.5K bounding boxes, selected from 10 hours raw videos, that were manually annotated for 14 attributes (e.g., *weather condition*, *flying altitude*, *camera view*, *vehicle category* and *levels of occlusion*). Recently, to facilitate research on face recognition from video/UAV-based data, the 'DroneSURF' dataset [15] was released. This dataset is composed of 200 videos from 58 subjects, captured across 411K frames, and includes over 786K face annotations.

B. Pedestrian Analysis Datasets

As summarized in Table I, there are various datasets for supporting pedestrian analysis research. The pioneer set

TABLE I
COMPARISON BETWEEN THE P-DESTRE AND THE EXISTING DATASETS THAT SUPPORT THE RESEARCH IN PEDESTRIAN DETECTION, TRACKING AND SHORT/LONG-TERM RE-IDENTIFICATION (APPEARING IN CHRONOLOGICAL ORDER)

Dataset	Camera	Format	Task					Identities	Bound. Box	Environment	Height (m)
			Detection	Tracking	ReID	Search	Action Rec.				
PRID-2011 [14]	UAV	Still	✗	✗	✓	✗	✗	1,581	40K	Surveillance	[20, 60]
CUHK03 [17]	CCTV	Still	✗	✗	✓	✗	✗	1,467	13K	Surveillance	-
iLIDS-VID [32]	CCTV	Video	✗	✗	✓	✗	✗	300	42K	Surveillance	-
MRP [16]	UAV	Video	✓	✓	✓	✗	✗	28	4K	Surveillance	< 10
PRAI-1581 [32]	UAV	Still	✗	✗	✓	✗	✗	1,581	39K	Surveillance	[20, 60]
CSM [1]	(Various)	Video	✗	✗	✗	✓	✗	1,218	11M	TV	-
Market1501 [37]	CCTV	Still	✓	✓	✓	✗	✗	1,501	32,668	Surveillance	< 10
Mini-drone [6]	UAV	Videos	✓	✓	✗	✗	✓	-	> 27K	Surveillance	< 10
Mars [39]	CCTV	Video	✗	✗	✓	✗	✗	1,261	20K	Surveillance	-
AVI [30]	UAV	Still	✗	✗	✗	✗	✓	5,124	10K	Surveillance	[2, 8]
DukeMTMC-VideoReID [34]	CCTV	Video	✗	✗	✓	✗	✗	1,812	815K	Surveillance	-
iQIYI-VID [20]	(Various)	Video	✗	✗	✗	✓	✗	5,000	600K	TV	-
DRone HIT [11]	UAV	Still	✗	✗	✓	✗	✗	101	40K	Surveillance	25
LTCC [26]	CCTV	Still	✓	✗	✓	✓	✗	152	17K	Surveillance	-
P-DESTRE	UAV	Video	✓	✓	✓	✓	✓	269	> 14.8M	Surveillance	[5.5, 6.7]

was the 'PRID-2011' [14], containing 400 image sequences of 200 pedestrians. Next, the 'CUHK03' [17] set aimed at providing enough data for deep learning-based solutions, and contains images collected from 5 cameras, comprising 1,467 identities and 13,164 bounding boxes. The 'iLIDS-VID' [32] set was the first to release video data, comprising 600 sequences of 300 individuals, with sequence lengths ranging from 23 to 192 frames. The 'MRP' [16] was the first UAV-based dataset specifically designed for the re-identification problem, containing a 28 identities and 4,000 bounding boxes. Roughly at the same time, the 'PRAI-1581' [32] data reproduces undoubtedly real surveillance conditions, but UAVs flew at too high altitude to enable re-identification experiments (up to 60 meters). This set has 39,461 images of 1,581 identities, and is mainly used for detection and tracking purposes. The 'Market-1501' [37] set was collected using 6 cameras in front of a supermarket, and contains 32,668 bounding boxes of 1,501 identities. Its extension ('MARS' [39]) was the first video-based set specifically devoted to pedestrian re-identification. Singularly, the 'Mini-drone' [6] set was created mostly to support abnormal event detection analysis, and has been also used for pedestrian detection, tracking and short-term re-identification purposes.

The 'DukeMTMC-VideoReID' [34] is a subset of the DukeMTMC [27] tracking dataset, used for pedestrian re-identification purposes. Authors also defined a performance evaluation protocol, enumerating the 702 identities used for training, the 702 testing identities, and the 408 distractor identities. Overall, this set comprises 369,656 frames of 2,196 sequences for training and 445,764 frames of 2,636 sequences for testing. The 'AVI' [30] set enables pose estimation/abnormal event detection experiments, with subjects in each frame annotated with 14 body keypoints. More recently, the 'DRoneHIT' [11] set supports image-based pedestrian re-identification experiments from aerial data, containing 101 identities, each one with about 459 images.

The 'CSM' [1] and 'iQIQI-VID' [20] sets were included in this summary because they previously released data for the long-term re-identification problem. However, their video sequences have notoriously different features from the acquired in surveillance environments: predominantly regard TV shows/movies. Similarly, the 'Long-Term Cloth-Changing (LTCC)' [26] set also supports long-term re-identification research and has 17,119 images from 152 identities, collected using CCTV footage and annotated across clothing-changes and different views.

Among the datasets analyzed, note that the Market1501, MARS, CUHK03, iLIDS-VID and DukeMTMC-VideoReID were collected using stationary cameras, and their data have notoriously different features of the resulting from UAV-based acquisition. Also, even though the PRAI-1581 and DRone HIT sets were collected using UAVs, they do not provide consistent identity information between acquisition sessions, and cannot be used in pedestrian search problem.

III. THE P-DESTRE DATASET

A. Data Acquisition Devices and Protocols

The P-DESTRE dataset is the result of a joint effort from researchers in two universities: the University of Beira Interior[4] (Portugal) and the JSS Science and Technology University[5] (India). In order to enable the research on pedestrian identification from UAV-based data, a set of *DJI® Phantom 4*[6] drones controlled by human operators flew over various scenes of both university campi, acquiring data that simulate the everyday conditions in outdoor urban environments.

All subjects in the dataset offered explicitly as volunteers and they were asked to completely ignore the UAVs (Fig. 2), that were flying at altitudes between 5.5 and 6.7 meters, with the camera pitch angles varying between 45° to 90°.

[4]http://www.ubi.pt
[5]https://jssstuniv.in
[6]https://www.dji.com/pt/phantom-4-pro-v2

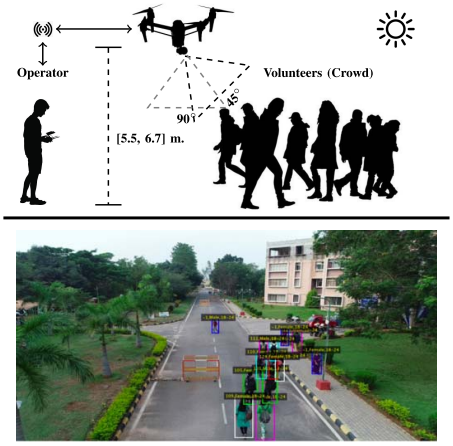


Fig. 2. At top: schema of the data acquisition protocol used. Human operators controlled DJI Phantom 4 aircrafts in various scenes of two university *campi*, flying at altitude between 5.5 and 6.7 meters, with gimbal pitch angles between 45° to 90°. The image at the bottom provides one example of a full scene of the P-DESTRE set.

TABLE II
THE P-DESTRE DATA ACQUISITION MAIN FEATURES

Image Acquisition Settings	
Camera Sensor: 1/2.3" CMOS, Effective pixels: 12.4 M	Frame Size: 3,840 × 2,160
Lens: FOV 94°, 20 mm (35 mm format equivalent) f/2.8 focus at ∞	ISO Range: 100-3200
Camera Pitch Angle: [45°, 90°]	Drone Altitude: [5.5, 6.7] meters
Format: MP4, 30 fps	Bit Depth: 24 bit
Volunteers	
Total IDs: 269	Gender: Male: 175 (65%); Female: 94 (35%)

Volunteers were students of both universities (mostly in the 18-24 age interval, $> 90\%$), $\approx$ 65/35% males/females, and of predominantly two ethnicities ('white' and 'indian'). About 28% of the volunteers were using glasses, 10% of them were using sunglasses. Data were recorded at 30fps, with 4K spatial resolution $(3,840 \times 2,160)$, and stored in "mp4" format, with H.264 compression. The key features of the data acquisition settings are summarized in Table II, and additional details can be found at the corresponding webpage.[7]

B. Annotation Data

The P-DESTRE set is fully annotated at the frame level, by human experts. For each video, we provide one text file with the same filename (plus the ".txt" extension), containing all the corresponding meta-information in comma-separated file format. In these files, each row provides the information for one bounding box in a frame (total of 25 numeric values). The annotation process was divided into four phases: 1) pedestrian detection; 2) tracking; 3) identification and soft biometrics characterisation; and 4) 3D head pose estimation.

[7]http://p-destre.di.ubi.pt/download.html

TABLE III
THE P-DESTRE DATASET ANNOTATION PROTOCOL. FOR EACH VIDEO, A TEXT FILE PROVIDES THE ANNOTATION AT FRAME LEVEL, WITH THE ROI OF EACH PEDESTRIAN IN THE SCENE, TOGETHER WITH THE ID INFORMATION AND 16 OTHER SOFT BIOMETRIC LABELS

Attributes	**Values**
Frame	**1**, **2**, …
ID	**-1**: 'Unknown', **1**, **2**, …
Bounding Box	$[x, y, h, w]$ (Top left column, top left row, height, width)
Head Pose	[**flag**, ***yaw***, ***pitch***, ***roll***] (flag: -1=not-available, 1=avaliable)
Age	**0**: 0-11, **1**: 12-17, **2**: 18-24, **3**: 25-34, **4**: 35-44, **5**: 45-54, **6**: 55-64, **7**: > 65, **8**: 'Unknown'
Height	**0**: 'Child', **1**: 'Short', **2**: 'Medium', **3**: 'Tall', **4**: 'Unknown'
Body Volume	**0**: 'Thin', **1**: 'Medium', **2**: 'Fat', **3**: 'Unknown'
Ethnicity	**0**: 'White', **1**: 'Black', **2**: 'Asian', **3**: 'Indian', **4**: 'Unknown'
Hair Color	**0**: 'Black', **1**: 'Brown', **2**: 'White', **3**: 'Red', **4**: Gray', **5**: 'Occluded', **6**: 'Unknown'
Hairstyle	**0**: 'Bald', **1**: 'Short', **2**: 'Medium', **3**: 'Long', **4**: Horse Tail,' **5**: 'Unknown'
Beard	**0**: 'Yes', **1**: 'No', **2**: 'Unknown'
Moustache	**0**: 'Yes', **1**: 'No', **2**: 'Unknown'
Glasses	**0**: 'Yes', **1**: 'Sunglass', **2**: 'No', **3**: 'Unknown'
Head Accessories	**0**: 'Hat', **1**: 'Scarf', **2**: 'Neckless', **3**: 'Occluded', **4**: 'Unknown'
Upper Body Clothing	**0**: 'T-shirt', **1**: 'Blouse', **2**: 'Sweater', **3**: 'Coat', **4**: 'Bikini', **5**: 'Naked', **6**: 'Dress', **7**: 'Uniform', **8**: 'Shirt', **9**: 'Suit', **10**: 'Hoodie', **11**: 'Cardigan'
Lower Body Clothing	**0**: 'Jeans', **1**: 'Leggins', **2**: 'Pants', **3**: 'Shorts', **4**: 'Skirt', **5**: 'Bikini', **6**: 'Dress', **7**: 'Uniform', **8**: 'Suit', **9**: 'Unknown '
Feet	**0**: 'Sport', **1**: 'Classic', **2**: 'High Heels', **3**: 'Boots', **4**: 'Sandals, **5**: 'Nothing', **6**: Unknown'
Accessories	**0**: 'Bag', **1**: 'Backpack', **2**: 'Rolling', **3**: 'Umbrella', **4**: 'Sportif', **5**: 'Market', **6**: 'Nothing', **7**: 'Unknown'
Action	**0**: 'Walk', **1**: 'Run', **2**: 'Stand', **3**: 'Sit', **4**: 'Cycle', **5**: 'Exercise', **6**: 'Pet', **7**: 'Phone', '**8**: 'Leave Bag', **9**: 'Fall', **10**: 'Fight', **11**: 'Date', **12**: 'Offend', **13**: 'Trade'

At first, the well-known Mask R-CNN [13] method was used to provide an initial estimate of the position of every pedestrian in the scene, with the resulting data subjected to human verification and correction. Next, the deep sort method [33] provided the preliminary tracking information, which again was corrected manually. As result of these two initial steps, we obtained the rectangular bounding boxes providing the regions-of-interest (ROI) of every pedestrian in each frame/video. The next phase of the annotation process was carried out manually, with human annotators that knew personally the volunteers of each university setting the ID information and characterising the samples according to the soft labels. Finally, we used the Deep Head Pose [29] method to obtain the 3D head pose angles for all elements (except backside views), expressed in terms of *yaw*, *pitch* and *roll* values.

Table III provides the details of the labels annotated for every instance (pedestrian/frame) in the dataset, along with the ID information, the bounding box that defines the ROI

Fig. 3. Examples of the six factors that - under visual inspection and in a qualitative analysis - constitute the major challenges to automated image analysis in video/UAV-based data. These are the predominant data degradation factors in the P-DESTRE set and the most important co-variates for the responses of automated systems.

and the frame information. For every label, we also provide a list of its possible values.

C. Typical Data Degradation Factors

As expected, the acquisition of video/UAV-based data in crowded outdoor environments, from at-a-distance and simulating covert protocols, has led to extremely heterogeneous samples, degraded in multiple perspectives. Under visual inspection, we identified the six major factors that the most frequently reduced the quality data, and augment the challenges of automated image analysis:

1) **Poor resolution/blur**. As illustrated in the top row of Fig. 3, some subjects were acquired from large distances (over 40 m.), with the corresponding ROIs having very poor resolution. Also, some parts of the scenes laid outside the cameras depth-of-field, in result of a large range in objects depth. This led to blurred samples. In both cases, the amount of information available per bounding box is reduced;
2) **Motion blur**. This factor yielded from the non-stationary nature of the cameras and the subjects' movements. In practice, for some bounding boxes, an apparent streaking of the body silhouettes is observed;
3) **Partial occlusions**. As a result of the scene dynamics and due to the multiple objects in the scenes, partial occlusions of subjects were particularly frequent. According to our perception, this might be the most concerning factor of UAV-based data, as illustrated in the third row of Fig. 3;
4) **Pose**. Under covert data acquisition protocols and without accounting for subjects cooperation, many samples regard profile and backside views, in which identification and soft biometric characterisation are particularly difficult;
5) **Lighting/shadows**. As a consequence of the outdoor conditions, many samples are over/under-illuminated, with shadowed regions due to the remaining objects in the scene (e.g., buildings, cars, trees, traffic signs…);
6) **UAV elevation angle**. When using gimbal pitch angles close to 90°, the longest axis of the subjects body is almost parallel to the camera axis. In such cases, images contain exclusively a top-view perspective of the subjects, with reduced amount of discriminating information (bottom row of Fig. 3).

When comparing the major features of CCTV and UAV-based data, the *pitch* factor of images is particularly evident. Due to the UAVs altitude, subjects appear almost invariably with negative pitch angles (over 95% of the P-DESTRE images have pitch angles between -10° and 50°), which - according to the results reported in Section IV - appears to be a relevant data degradation factor. Also, the non-stationary feature of UAVs increases the heterogeneity of the resulting data, which again augments the challenges in performing reliable automated image analysis.

D. P-DESTRE Statistical Significance

Let α be a confidence interval. Let p be the error rate of a classifier and $\hat{p}$ be the estimated error rate over a finite number of test patterns. At an α-confidence level, we want that the true error rate does not exceed $\hat{p}$ by an amount larger than $\varepsilon(n,\ \alpha)$. Guyon *et al.* [12] defined $\varepsilon(n,\ \alpha) = \beta p$ as a fraction of p. Assuming that recognition errors are Bernoulli trials, authors concluded that the number of required trials n to achieve (1-α) confidence in the error rate estimate is given by:

$$n = -ln(\alpha)/(\beta^2 p). \tag{1}$$

Using typical values $\alpha = 0.05$ and $\beta = 0.2$, authors recommend a simpler form, given by: $n \approx \frac{100}{p}$.

Considering the statistics of the P-DESTRE set (Fig. 4), in terms of the number of data acquisition sessions/days per volunteer and the number of bounding boxes per volunteer/session, it is possible to obtain the lower bounds for the statistical confidence in experiments related with identity verification at the frame level, assuming the 1) short-term re-identification; and 2) long-term re-identification problems.

In the short-term re-identification setting, considering that each frame (bounding box) with a valid ID ($\geq$ 1) generates a valid template, that all frames of the same ID acquired in different sessions of the same day can be used

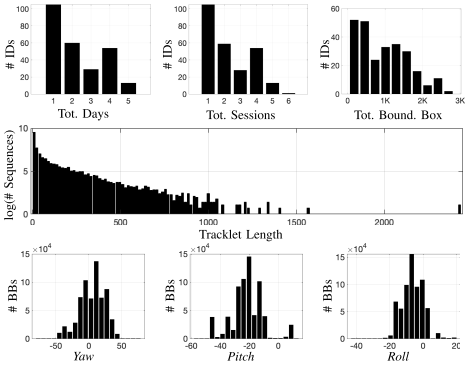


Fig. 4. P-DESTRE statistics. Top row: number of days with data per volunteer (at left), number of data acquisition sessions per volunteer (at center), and number of bounding boxes per volunteer (at right). The histogram at the middle row provides the summary statistics for the length of the tracklet sequences. Finally, the bottom row provides the total of bounding boxes (BBs) per 3D head pose angle, expressed in terms of *yaw*, *pitch* and *roll* values.

to generate *genuine* pairs and that frames with different IDs (including '*unknown*') compose the *impostors* set, the P-DESTRE dataset enables to perform 1,246,587,154 (genuine) + 605,599,676,264 (impostor) comparisons, leading to a $\hat{p}$ value with a lower bound of approximately 1.647×10^{-10}. Regarding the pedestrian long-term re-identification problem, where the *genuine* pairs must have been acquired in different days, the dataset enables to perform 2,160,586,581 (genuine) + 605,599,676,264 (impostor) comparisons, leading to a $\hat{p}$ value with a lower bound of approximately 1.645×10^{-10}. Note that these are lower bounds, that do not take into account the portions of data used for learning purposes. Also, these values will increase if we do not assume the independence between images and error correlations are taken into account.

IV. Experiments and Results

In this section we report the results obtained by methods that represent the state-of-the-art in four tasks: pedestrian 1) detection; 2) tracking; 3) short-term re-identification; and 4) long-term re-identification. For contextualisation, we report not only the performance obtained in the P-DESTRE set, but also provide baseline results attained by the same techniques in well-known datasets. Also, for each problem, we illustrate the typical failure cases that we have subjectively perceived during our experiments.

A. Pedestrian Detection

The RetinaNet [19], R-FCN [7] methods were initially considered to represent the state-of-the-art in pedestrian detection, as both outperformed in the PASCAL VOC 2007/2012 [10] challenge ('Person Detection' category). Then, the well-known SSD [21] method was also chosen as baseline, as it is the most widely detector reported in the literature, and its results can be easily contextualised. Accordingly, this section reports a comparison between the performance of the three object detectors in the P-DESTRE/PASCAL sets.

TABLE IV

Comparison Between the Average Precision (AP) Obtained by Three Methods Considered to Represent the State-of-the-Art in Pedestrian Detection, in the P-DESTRE and PASCAL VOC 2007/2012 Sets

Method	Backbone	PASCAL VOC	P-DESTRE
RetinaNet [19]	ResNet-50	86.44 ± 1.03	63.10 ± 1.64
R-FCN [7]	ResNet-101	84.43 ± 1.85	59.29 ± 1.31
SSD [21]	Inception-V2	74.70 ± 2.69	55.63 ± 2.93

In summary, RetinaNet is composed of a backbone network and two task specific subnetworks. It uses a feature pyramid network as backbone model, to obtain a convolutional feature map over the entire input image. Two sub-networks use this feature representation: the first one classifies the anchor boxes and the second model performs the bounding box regression, to refine the localization of the detected objects. R-FCN uses a fully convolutional architecture, where the translation invariance is obtained by position-sensitive score maps that use specialized convolutional layers to encode the deviations with respect to default positions. A position-sensitive ROI pooling layer is appended on top of the fully connected layers. The SSD model eliminates the proposal generation and feature resampling steps by encapsulating all the processing into a single network. It discretizes the output space of bounding boxes into a set of default boxes over different aspect ratios and scales per feature map location. In our experiments, in a data augmentation setting, the sizes of the learning patches were randomly sampled by [0.1, 1] factor, and horizontally flipped with probability 0.5.

For the PASCAL VOC 2007/2012 set, the official development kit[8] was used to evaluate the methods on the 'Person' category, using 10-fold cross validation. Regarding the P-DESTRE set, a 10-fold cross validation scheme was used, with the data in each split randomly divided into 60% for learning, 20% for validation and 20% for test, i.e., 45 videos were used for learning, 15 for validation and 15 videos for test purposes. The full specification of the samples used in each split and the scores returned by each method is provided in.[9]

The results are summarized in Table IV for all datasets/methods, in terms of the average precision obtained at intersection-of-union values equal to 0.5 (i.e., AP@IoU=0.5). Also, Fig. 5 provides the precision/recall curves for both data sets and all detection methods, with the P-DESTRE values being represented by red lines and the PASCAL VOC 2007/2012 results represented by green lines. The shadowed regions denote the standard deviation performance in the 10 splits, at each operating point. Overall, all methods decreased notoriously their effectiveness from the PASCAL VOC set to the P-DESTRE set, in some cases with error rates increasing over 160%. In the case of the R-FCN method, in a small region of the performance space (recall $\approx$ 0.2), the levels of performance for P-DESTRE and PASCAL VOC were approximately equal, yet the precision values then remain stable for much higher recall values in the PASCAL VOC set.

[8] http://host.robots.ox.ac.uk/pascal/VOC/voc2012/#devkit
[9] http://p-destre.di.ubi.pt/pedestrian_detection_splits.zip

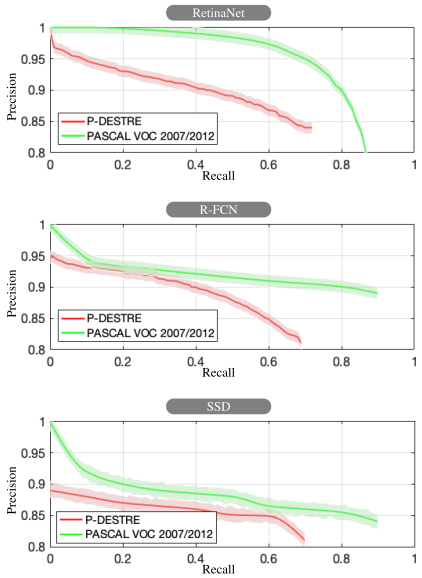


Fig. 5. Comparison between the precision/recall curves observed in the PASCAL VOC 2007/2012 (green lines) and P-DESTRE (red lines) sets. Results are given for the RetinaNet (top plot), R-FCN (middle plot) and SSD (bottom plot) object detection methods.

When comparing the performance of the three techniques tested, we observed that RetinaNet slightly outperformed the competitors in both datasets, in all cases with the R-FCN being the runner-up. The SSD algorithm not only got evidently the lowest average performance among all methods, but also its variance was the largest, which points for the lower robustness of this technique to most of the data co-variates in both the PASCAL VOC and P-DESTRE sets. The observed ranks among the three methods not only accord previous object evaluation initiatives [10], but also the substancial lower performance observed in P-DESTRE than in PASCAL VOC supports the hypothesis claimed in this paper: the P-DESTRE set has evidently different features with respect to previous similar sets.

In a qualitative perspective, we observed that all methods faced particular difficulties in crowded scenes, when only a small part of the subjects silhouette is unoccluded, as illustrated in Fig. 6. Considering that RetinaNet is anchor-based, and that the predefined anchor boxes have a set of handcrafted aspect ratios and scales that are data dependent, performance might have been seriously affected. Even though RetinaNet has clearly outperformed its competitors, the challenging conditions in the P-DESTRE set have still notoriously degraded its effectiveness, when compared to the PASCAL VOC baseline. By analysing the instances in both sets, we observed that the P-DESTRE set has notoriously more *hard* cases than PASCAL VOC, with a significant portion of severely degraded samples (i.e., with severe occlusions,

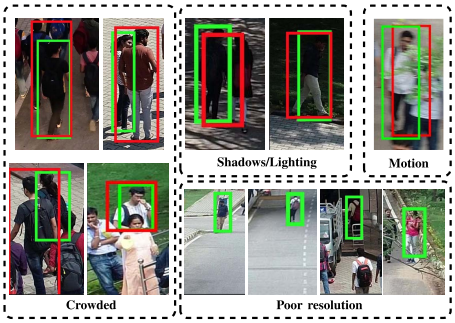


Fig. 6. Typical cases where the object detectors returned the worst scores, i.e., failed to appropriately detect the pedestrians. The green boxes represent the ground-truth, while the red colour denotes the detected boxes.

extreme poor resolution and strong local lighting variations/ shadows).

In summary, our experiments point for the requirement of novel strategies to handle the specific problems that yield from UAV-based data acquisition. Not only the state-of-the-art solutions provide levels of performance that are still far from the demanded to deploy this kind of solutions in real-environments, but most methods are also sensitive to particularly frequent co-variates in UAV-based imaging (e.g., motion-blur and shadows). Another concerning point is the density of subjects in the scenes, with crowded environments easily providing severe occlusions that constraint the effectiveness of the object detection phase.

B. Pedestrian Tracking

For the tracking task, the TracktorCV [2] and V-IOU [5] methods were initially selected to represent the state-of-the-art, according to: 1) their performance in the MOT challenge[10]; and 2) the fact that both provide freely available implementations, which is important to guarantee that we obtain a fair evaluation between datasets. Moreover, we considered additionally one method (IOU [4]) that is among the most widely reported in the literature. We compared the effectiveness attained by the three techniques in the P-DESTRE and MOT challenge sets, in order to perceive the relative hardness of tracking pedestrians in UAV-based data in comparison to a stationary camera setting. In terms of evaluation protocols, the rules provided for the MOT challenges were rigorously met for the MOT evaluation. For the P-DESTRE set, a 10-fold cross validation scheme was used, with the data in each split randomly divided into 60% for learning, 20% for validation and 20% for test, i.e., 45 videos were used for learning, 15 for validation and 15 videos for test purposes. The full details of each split are available at.[11]

The TracktorCV method comprises two steps: 1) a regression module, that uses the input of the object detection step to update the position of the bounding box at a subsequent frame; and 2) an object detector that provides the set of bounding

[10]https://motchallenge.net

[11]http://p-destre.di.ubi.pt/pedestrian_tracking_splits.zip

TABLE V
COMPARISON BETWEEN THE TRACKING PERFORMANCE ATTAINED BY THREE ALGORITHMS CONSIDERED TO REPRESENT THE STATE-OF-THE-ART IN THE P-DESTRE AND MOT-17 DATA SETS

Method	Dataset	MOTA	MOTP	F-1
TracktorCv [2]	**MOT-17**	65.20 ± 9.60	62.30 ± 11.00	89.60 ± 2.80
	P-DESTRE	56.00 ± 3.70	55.90 ± 2.60	87.40 ± 2.00
V-IOU [5]	**MOT-17**	52.50 ± 8.80	57.50 ± 9.50	86.50 ± 1.90
	P-DESTRE	47.90 ± 5.10	51.10 ± 5.80	83.30 ± 8.40
IOU [4]	**MOT-17**	45.51 ±13.61	46.02 ± 12.40	78.21 ± 3.12
	P-DESTRE	38.27 ± 8.42	39.68 ± 4.92	74.29 ± 6.87

boxes for the next frames. The IOU method was developed based on two assumptions: i) the detection step returns a detection per frame for every object to be tracked; and ii) the objects in consecutive frames have high overlap (according to an Intersection-over-Union perspective). Based on these two assumptions, IOU tracks objects without considering image information, which is a key point that contributes for its computational effectiveness. Further, the short tracks are eliminated according to an acceptance threshold. The V-IOU algorithm is an extension of the IOU algorithm that attenuates the problem of false negatives, by associating the detections in consecutive frames according to spatial overlap information. For all three methods, the hyper-parameters were tuned according to the way authors suggested, and are given in.[12]

In terms of performance measures, our analysis was based in the Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP) and F1 values, as described in [3]. The summary results attained by both algorithms and datasets are given in Table V. Once again, a consistent degradation in performance from the MOT-17 to the P-DESTRE set was observed, even though the deterioration was in absolute terms far less than the observed for the detection task (here, an decrease in the F1 values of around 10% was observed). It is interesting to observe the larger variance values obtained for tracking methods with respect to the values provided for the detection step. This was justified by the smaller number of learning/test instances available for tracking (working at *sequence*/video level) than for detection (that works at frame level).

When comparing the results of all methods, the Tracktor-Cv outperformed its competitors (V-IOU as runner-up) both in non-aerial and aerial data, decreasing the error rates around 9% with respect to the second best techniques. As expected, the IOU technique obtained invariably the worst performance among all methods tested, which also accords previous tracking performance evaluation initiatives carried out. In all cases, we observed a positive correlation between their typical failure cases, which were invariably related to crowded scenes, and two particularly concerning cases: 1) scenes where, due to extreme pedestrian density, subjects' trajectories cross

[12]http://p-destre.di.ubi.pt/parameters_tracking.zip

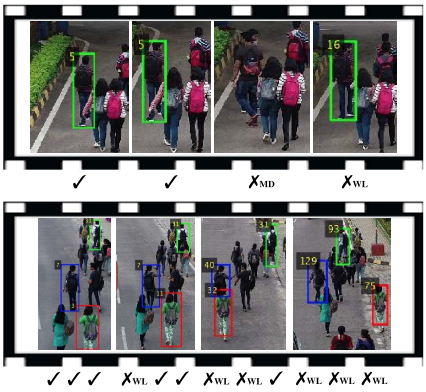


Fig. 7. Examples of sequences where the tracking methods faced difficulties, either missing the ground-truth targets at some point or producing a *fragmentation* that resulted in a wrong label assignment. *MD* stands for "missed detection" and *WL* represents "wrong label" assignment.

others at every moment; and 2) when severe occlusions of the body silhouettes occur. Both factors augment the likelihood of observing *fragmentations*, i.e., with the trackers erroneously switching identities of two trajectories in the scene, and wrong *merge* cases, with the trackers erroneously merging two ground truth identities into a single one.

When subjectively comparing the data in MOT-17 and P-DESTRE datasets, it is evident that P-DESTRE contains more complex scenarios, more cluttered backgrounds (e.g., many scenes have 'grass' grounds and tree branches) and more poor resolution subjects. Also, we noted that the trackability of pedestrians also depends on the tracklet length (i.e., the number of consecutive frames where an object appears), with the values in MOT-17 varying from 1 to 1,050 (average 304) and in P-DESTRE varying from 4 to 2,476 (average 63.7 ± 128.8), as illustrated in Fig. 4.

C. Pedestrian Short-Term Re-Identification

We selected three well known re-identification algorithms to represent the state-of-the-art and assessed their performance. The MARS [39] dataset was selected to represent the stationary datasets, as it is currently the largest video-based source that is freely available.

According to the results reported on a challenge [36], the GLTR [18], COSAM [31] and NVAN [22] methods were selected. The GLTR exploits multi-scale temporal cues in video sequences, by modelling separately short- and long-term features. Short-term components capture the appearance and motion of pedestrians, using parallel dilated convolutions with varying rates. Long-term information is extracted by a temporal self-attention model. The key in COSAM is to capture intra video attention using a co-segmentation module, extracting task-specific regions-of-interest that typically correspond to pedestrians and their accessories. This module is plugged between convolution blocks to induce the notion of co-segmentation, and enables to obtain representations of

TABLE VI
COMPARISON BETWEEN THE RE-IDENTIFICATION PERFORMANCE ATTAINED BY THREE STATE-OF-THE-ART METHODS IN THE P-DESTRE AND MARS DATA SETS

Method	Dataset	mAP	Rank-1	Rank-20
GLTR [18]	**MARS**	77.74 ± 1.07	84.72 ± 2.61	95.80 ± 2.34
	P-DESTRE	77.68 ± 9.46	75.96 ± 11.77	95.48 ± 3.17
COSAM [31]	**MARS**	78.35 ± 1.66	84.03 ± 0.91	96.97 ± 0.98
	P-DESTRE	80.64 ± 9.91	79.14 ± 12.43	97.10 ± 1.85
NVAN [22]	**MARS**	81.13 ± 1.35	85.94 ± 0.94	97.20 ± 0.97
	P-DESTRE	82.78 ± 10.35	80.42 ± 12.38	98.34 ± 1.93

both the spatial and temporal domains. Finally, the Non-local Video Attention Network (NVAN) exploits both spatial and temporal cues by introducing a non-local attention operation into the backbone CNN at multiple feature levels. Further, it reduces the computational complexity of the inference step by exploring the spatial and temporal redundancy that is observed in the learning data.

In a 5-fold setting, both datasets were divided into random splits, each one containing the learning, query and gallery sets, in proportions 50:10:40. For the MARS dataset, the evaluation protocol described in[13] was used. For the P-DESTRE dataset, we considered 1,894 tracklets of 608 IDs, with an average number of frames per tracklet of 67,4. The full specification of the samples used for learning/validation/test purposes in each split is given in.[14]

Regarding the GLTR method, the ResNet50 was used as backbone model, with the learning rate set to 0.01. In the COSAM method, the Se-ResNet50 architecture was used as backbone model. The COSAM layer was plugged between the forth and fifth convolution layers, with the learning rate set to 0.0001 and the reduction dimension size set to 256. For the NVAN method, we also used ResNet50 architecture as backbone network and plugged two non-local attention layers (after *Conv3_3* and *Conv3_4*) and three non-local layers (after *Conv4_4*, *Conv4_5*, and *Conv4_6*). The input frames were resized into 256 × 128. The model was trained using the Adam algorithm, with 300 epochs and learning rate set to 0.0001.

The summary results are provided in Table VI. In opposition to the detection and tracking problems, it is interesting to note that no significant decreases in performance were observed from the MARS to the P-DESTRE data, which points for the suitability of the existing short-term re-identification solutions for UAV-based data. Fig. 8 provides the cumulative rank-n curves for all algorithms/datasets. The red lines represent the P-DESTRE results and the green series denote the MARS values. Results are given in terms of the identification rate with respect to the proportion of gallery identities retrieved (i.e., hit/penetration plot). Apart the outperforming results of

[13]http://www.liangzheng.com.cn/Project/project_mars.html
[14]http://p-destre.di.ubi.pt/pedestrian_reid_splits.zip

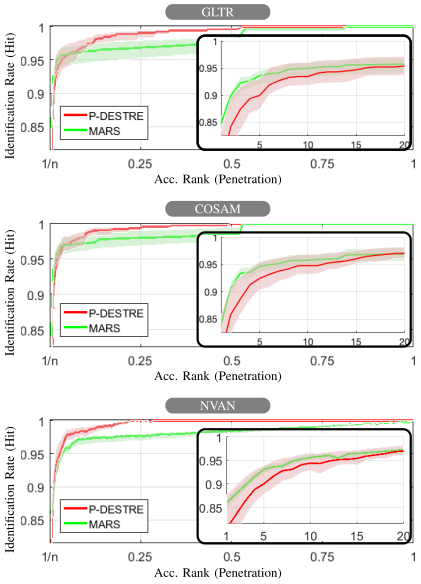


Fig. 8. Comparison between the closed-set identification (CMC) curves observed in the MARS (green lines) and P-DESTRE (red lines) sets for the GLTR, COSAM and NVAN re-identification techniques. Zoomed-in regions with the top-1 to 20 results are shown in the inner plots.

NVAN, it is particularly interesting to note the apparently contradictory results of the GLTR and COSAM algorithms in the MARS and P-DESTRE sets. In all cases, in terms of the top-20 performance, the P-DESTRE results were far worse than the corresponding MARS values. However, for larger ranks (starting at 5% of the enrolled identities), the P-DESTRE values were solidly better than the ranks observed for MARS. Also, in case of heavily degraded MARS instances, algorithms returned almost random results, which was not observed for the P-DESTRE. This might be justified by the fact that P-DESTRE contains more *poor quality* data than MARS, yet it does not provide *extremely degraded* (i.e., almost *impossible*) instances that turn the identification into a quasi-random process.

Based in these experiments, Fig. 9 highlights some notorious cases for re-identification purposes. The upper row represents the particularly hazardous cases in terms of *convenience*, where different IDs were erroneously perceived as the same. This was mostly due to similarities in clothing, together with shared soft biometric labels between different IDs. The bottom row provides the particularly dangerous cases for *security* purposes, where methods had difficulties in identifying a known ID. Here, errors often yielded from notorious differences in pose and scale between the query/gallery data. Along with the background clutter, these factors were observed to decrease the effectiveness of the feature representations, and were among the most concerning for re-identification performance.

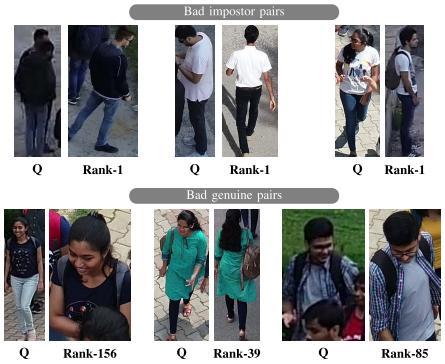


Fig. 9. Examples of the instances that got the worst re-identification performance. The upper row illustrates typical false matches, almost invariably related with clothing styles and colours. The bottom row provides some examples of cases where (due to differences in pose and scale), the true identities could not be retrieved among the top positions. "Q" represents the query image and "Rank-i" provides the rank of the corresponding gallery image.

D. Long-Term Pedestrian Re-Identification

As stated above, the pedestrian video-based long-term re-identification problem was the main motivation for the development of the P-DESTRE dataset. Here, there is not any guarantee about the clothing appearance of subjects, nor about the time elapsed between consecutive observations of one ID. In such circumstances, the analysis of alternative features should be considered (e.g., face, gait or soft-biometrics based).

Considering that there are not yet methods in the literature specifically designed for this kind of task, we have chosen a combination of two well-known re-identification techniques that combine face and body features. Similarly to the previous tasks, the goal was to obtain an approximation for the effectiveness attained by the existing solutions in UAV-based data. Such levels of performance constitute a baseline for this problem and can be used as basis for further developments.

The facial regions-of-interest were detected by the SSH method [25] (acceptance threshold=0.7), from where a feature representation was obtained using the ArcFace [8] model. For the body-based analysis, the COSAM [31] model provided the feature representation. Both models were trained *from scratch*. The data were sampled into 5 trials, each one containing learning/gallery/query instances in proportions 50:10:40. As for the previous tasks, the full specification of the samples used in each split is given in.[15]

For the ArcFace method, the MobileNetV2 was used as backbone model, and the learning rate set to 0.01. Regarding COSAM, the Se-ResNet50 was used as backbone model, and the COSAM layer was plugged into the forth and fifth convolutional layers, with learning rate equal to $1e^{-4}$ and dimension size equal to 256. Each model was trained reparately, and during the test phase, the mean value of the ArcFace facial features in the tracklet were appended to the body-based representation yielding from COSAM. The Euclidean norm was used as distance function between such concatenated representations.

[15]http://p-destre.di.ubi.pt/pedestrian_search_splits.zip

TABLE VII

BASELINE LONG-TERM PEDESTRIAN RE-IDENTIFICATION PERFORMANCE OBTAINED BY AN ENSEMBLE OF ARCFACE [8] + COSAM [31] IN THE P-DESTRE DATA SET

Method	mAP	Rank-1	Rank-20
ArcFace [8] + COSAM [31]	34.90 ± 6.43	49.88 ± 8.01	70.10 ± 11.25

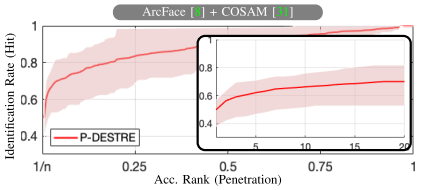


Fig. 10. Closed-set identification (CMC) curves obtained for the long-term re-identification problem in the P-DESTRE dataset. The inner plot provides the top-20 results as a zoomed-in region.

Fig. 10 provides the cumulative rank-n curves obtained, in terms of the successfull identification rates with respect to the proportion of gallery identities (i.e., hit/penetration plot). As expected, when compared to the short-term re-identification setting, performance was substantially lower (rank-1 $\approx$ 79.14% for re-identification $\rightarrow$ $\approx$ 49.88% for search), which accords the human perception for the additional difficulty of *search* with respect to *re-identify*.

Based in our qualitative analysis of the results, Fig. 11 provides three types of examples: the upper row shows some successful identification cases, in which the model retrieved the true identity in the first position. In most cases, we noted that subjects kept *some* piece of clothing/accessories between observations (e.g., glasses or backpack) and the same hairstyle. The remaining rows illustrate the failure cases: the second row provides examples of the hazardous cases for *convenience* purposes, in which due to similarities in pose, accessories and soft biometric labels between the query and gallery images, false matches have occurred. Finally, the bottom row provides examples of *security sensitive* cases, where the IDs of the queries were retrieved in high positions (ranks 56, 73 and 98), i.e., the system failed to detect a subject of interest in a crowd.

The challenges of long-term re-identification are illustrated in Fig. 12, providing the differences between the probabilities of obtaining a top-i correct identification (hit), $\forall i \in \{1, \ldots, n\}$, i.e., retrieve the identity corresponding to a query up to the i^{th} position, for the search and re-identification problems. Here, $P_s(i)$ and $P_r(i)$ denote the probabilities of observing a *hit* in the search P_s and re-identification P_r tasks, i.e., negative $\big(P_s(i)$ - $P_r(i)\big)$ denote higher probabilities for re-identification success than for search success. The zoomed-in region given at the right part of the Figure shows the additional difficulty (of almost 40 percentual points) in retrieving the true

Fig. 11. Examples of the instances where good/poor pedestrian search performance was observed. The upper row illustrates particularly successful cases, while the bottom rows show pairs of images where the used algorithm had notorious difficulties to retrieve the correct identity. "Q" represents the query image and "Rank-i" provides the rank of the retrieved gallery image.

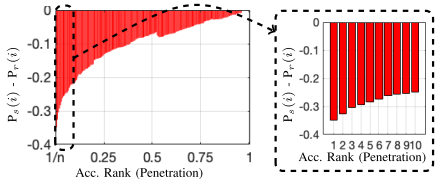


Fig. 12. Differences between the probability of retrieving the true identity of a query among the top-i positions, $\forall i \in \{1, \ldots, 100\}$, for the pedestrian long-term re-identification (P_s) and short-term re-identification (P_r) problems.

identity in a single shot (difference between top-1 values). Then, the gap between the accumulated values of P_s and P_r decreases in a monotonous way, and only approaches 0 near the full penetration rate, i.e., when all the known identities are retrieved for a query. In summary, it is much more difficult to identify pedestrians when no clothing information can be used, which paves the way for further developments in this kind of technology. According to our goals in developing this data source, the P-DESTRE set is a tool to support such advances in the state-of-the-art.

V. Conclusion

This paper announced the availability of the P-DESTRE dataset, which provides video sequences of pedestrians taken from UAVs in outdoor environments. The key point of the P-DESTRE set is to provide full annotations that enable the research on *long-term pedestrian re-identification*, where the time elapsed between consecutive observations of IDs forbids the use of clothing-based features. Apart this, the P-DESTRE set is also suitable for research on UAV/video-based pedestrian detection, tracking, short-term re-identification and soft biometrics analysis.

Additionally, as a secondary contribution, we offered the results of our own evaluation of the state-of-the-art in the pedestrian detection, tracking and short-term re-identification problems, comparing the performance attained in data acquired from stationary (CCTV) and from moving/UAV devices. Such results point for a particular hardness of the existing solutions to detect and track subjects UAV-based data. In opposition, the existing short-term re-identification techniques appear to be relatively robust to the features typical of UAV-based data.

Overall, the decreases in performance observed from CCTV to UAV-based data support the originality and usefulness of P-DESTRE. hence, potential directions for further developments of long-term UAV-based re-identification include the use of *attention-based* networks that disregard portions of the input data known to be ineffective for long-term re-identification (e.g., clothes or hairstyles). Another important field will be the development of *domain adaptation* techniques robust to changes in the UAV-acquisition settings and environments heterogeneity.

References

[1] M. Ahmed, M. Jahangir, H. Afzal, A. Majeed, and I. Siddiqi, "Using crowd-source based features from social media and conventional features to predict the movies popularity," in *Proc. IEEE Int. Conf. Smart City/SocialCom/SustainCom (SmartCity)*, Dec. 2015, pp. 273–278.

[2] P. Bergmann, T. Meinhardt, and L. Leal-Taixe, "Tracking without bells and whistles," 2019, *arXiv:1903.05625*. [Online]. Available: http://arxiv.org/abs/1903.05625

[3] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The CLEAR MOT metrics," *EURASIP J. Image Video Process.*, vol. 2008, pp. 1–10, Dec. 2008, doi: 10.1155/2008/246309.

[4] E. Bochinski, V. Eiselein, and T. Sikora, "High-speed tracking-by-detection without using image information," in *Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2017, pp. 1–6, doi: 10.1109/avss.2017.8078516.

[5] E. Bochinski, T. Senst, and T. Sikora, "Extending IOU based multi-object tracking by visual information," in *Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Nov. 2018, pp. 1–6, doi: 10.1109/avss.2018.8639144.

[6] M. Bonetto, P. Korshunov, G. Ramponi, and T. Ebrahimi, "Privacy in mini-drone based video surveillance," in *Proc. Workshop De-Identificat. Privacy Protection Multimedia*, 2015, pp. 1–3, doi: 10.13140/RG.2.1.4078.5445.

[7] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 379–387.

[8] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699, doi: 10.1109/cvpr.2019.00482.

[9] D. Du *et al.*, "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 370–386.

[10] M. Everingham, S. Eslami, L. Van Gool, C. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no,. 1, pp. 318–327, 2015.

[11] A. Grigorev, Z. Tian, S. Rho, J. Xiong, S. Liu, and F. Jiang, "Deep person re-identification in UAV images," *EURASIP J. Adv. Signal Process.*, vol. 2019, no. 1, p. 54, Dec. 2019, doi: 10.1186/s13634-019-0647-z.

[12] I. Guyon, J. Makhoul, R. Schwartz, and V. Vapnik, "What size test set gives good error rate estimates?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 1, pp. 52–64, Feb. 1998.

[13] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Person re-identification by descriptive and discriminative classification," 2018, *arXiv:1703.06870v3*. [Online]. Available: https://arxiv.org/abs/1703.06870v3

[14] M. Hirzer, C. Beleznai, P. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Proc. Scandin. Conf. Image Anal.*, 2011, pp. 91–102.

[15] I. Kalra, M. Singh, S. Nagpal, R. Singh, M. Vatsa, and P. B. Sujit, "DroneSURF: Benchmark dataset for drone-based face recognition," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–7.

[16] R. Layne, T. Hospedales, and S. Gong, "Investigating open-world person re-identification using a drone," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 225–240.

[17] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 152–159, doi: 10.1109/cvpr.2014.27.

[18] J. Li, J. Wang, Q. Tian, W. Gao, and S. Zhang, "Global-local temporal representations for video person re-identification," 2019, *arXiv:1908.10049*. [Online]. Available: http://arxiv.org/abs/1908.10049

[19] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[20] Y. Liu *et al.*, "IQIYI-VID: A large dataset for multi-modal person identification," 2018, *arXiv:1811.07548*. [Online]. Available: http://arxiv.org/abs/1811.07548

[21] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37, doi: 10.1007/978-3-319-46448-0_2.

[22] C. T. Liu, C. W. Wu, Y. C. F. Wang, and S. Y. Chien, "Spatially and temporally efficient non-local attention network for video-based person re-identification," in *Proc. Brit. Mach. Vis. Conf.*, 2019, pp. 1–13. [Online]. Available: https://arxiv.org/abs/1908.01683

[23] I. Mademlis *et al.*, "High-level multiple-UAV cinematography tools for covering outdoor events," *IEEE Trans. Broadcast.*, vol. 65, no. 3, pp. 627–635, Sep. 2019.

[24] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for UAV tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 445–461, doi: 10.1007/978-3-319-46448-0_27.

[25] M. Najibi, P. Samangouei, R. Chellappa, and L. S. Davis, "SSH: Single stage headless face detector," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4875–4884, doi: 10.1109/iccv.2017.522.

[26] X. Qian *et al.*, "Long-term cloth-changing person re-identification," 2020, *arXiv:2005.12633*. [Online]. Available: http://arxiv.org/abs/2005.12633

[27] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2016, pp. 17–35. [Online]. Available: http://arXiv:1609.01775v2, 2016.

[28] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory prediction in crowded scenes," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 549–565, doi: 10.1007/978-3-319-46484-8_33.

[29] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 2074–2083, doi: 10.1109/cvprw.2018.00281.

[30] A. Singh, D. Patil, and S. N. Omkar, "Eye in the sky: Real-time drone surveillance system (DSS) for violent individuals identification using ScatterNet hybrid deep learning network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 1629–1637, doi: 10.1109/cvprw.2018.00214.

[31] A. Subramaniam, A. Nambiar, and A. Mittal, "Co-segmentation inspired attention networks for video-based person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 562–572.

[32] X. Wang and R. Zhao, "Person re-identification: System design and evaluation overview," in *Person Re-Identification*. London, U.K.: Springer, 2014, doi: 10.1007/978-1-4471-6296-4_17.

[33] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3645–3649.

[34] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5177–5186, doi: 10.1109/cvpr.2018.00543.

[35] G.-S. Xia *et al.*, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3974–3983, doi: 10.1109/cvpr.2018.00418.

[36] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," 2020, *arXiv:2001.04193*. [Online]. Available: http://arxiv.org/abs/2001.04193

[37] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124, doi: 10.1109/iccv.2015.133.

[38] S. Zheng, J. Zhang, K. Huang, R. He, and T. Tan, "Robust view transformation model for gait recognition," in *Proc. 18th IEEE Int. Conf. Image Process.*, Sep. 2011, pp. 2073–2076, doi: 10.1109/icip.2011.6115889.

[39] L. Zheng *et al.*, "MARS: A video benchmark for large-scale person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, in Lecture Notes in Computer Science, vol. 9910. London, U.K.: Springer, 2016, pp. 868–884.

[40] P. Zhu, L. Wen, X. Bian, H. Ling, and Q. Hu, "Vision meets drones: A challenge," 2018, *arXiv:1804.07437*. [Online]. Available: http://arxiv.org/abs/1804.07437

S. V. Aruna Kumar received the Ph.D. degree in computer science and engineering from Visvesvaraya Technological University, India. He was a Post-Doctoral Researcher with the University of Beira Interior, Portugal. He is currently working as an Assistant Professor with the Department of Computer Science and Engineering, Ramaiah University of Applied Sciences, Bengaluru, India. His research interests include biometrics and medical image processing.

Ehsan Yaghoubi (Member, IEEE) received the B.Sc. degree from the Sadjad University of Technology in 2011 and the M.Sc. degree from the University of Birjand in 2016. He is currently pursuing the Ph.D. degree in biometrics with the University of Beira Interior, Portugal. His research interests broadly include computer vision and pattern recognition problems, with a particular focus on biometrics and surveillance.

Abhijit Das (Member, IEEE) received the Ph.D. degree from the School of Information and Communication Technology, Griffith University, Australia. He has worked as a Researcher with the University of Southern California, as a Post-Doctoral Researcher with the Inria Sophia Antipolis-Méditerranée, France, and as a Research Administrator with the University of Technology Sydney, Ultimo, NSW, Australia. He is currently a Visiting Scientist with the Indian Statistical Institute, Kolkata. During his research career, he has published several scientific articles in conferences, journals and a book chapter, having also received several awards. He is also involved in organizing scientific events.

B. S. Harish received the B.Eng. degree in electronics and communication and the master's degree in technology (networking and internet engineering) from Visvesvaraya Technological University, India, and the Ph.D. degree in computer science from the University of Mysore.

He was a Visiting Researcher with DIBRIS, Department of Informatics, Bio Engineering, Robotics and System Engineering, University of Genova, Italy. He is currently working as a Professor with the Department of Information Science and Engineering, JSS Science and Technology University, India. His research interests include machine learning, text mining, and computational intelligence. He is serving as a reviewer for many international journals and served as the Technical Program Chair for International Conferences. He successfully executed government funded projects sanctioned from the Government of India. He is a Member Life Member CSI (09872), a Life Member INSTICC (12844), a Life Member Institute of Engineers, and a Life Member ISTE.

Hugo Proença (Senior Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees from the University of Beira Interior in 2001, 2004, and 2007, respectively. He is currently an Associate Professor with the Department of Computer Science, University of Beira Interior. He has been researching mainly about biometrics and visual-surveillance. He was the Coordinating Editor of the IEEE Biometrics Council Newsletter and the Area Editor (ocular biometrics) of the IEEE BIOMETRICS COMPENDIUM JOURNAL. He is a member of the Editorial Boards of the *Image and Vision Computing*, IEEE ACCESS, and *International Journal of Biometrics*. Also, he served as a Guest Editor of Special Issues of the *Pattern Recognition Letters*, *Image and Vision Computing*, and *Signal, Image, and Video Processing* Journals.

A Quadruplet Loss for Enforcing Semantically Coherent Embeddings in Multi-Output Classification Problems

Hugo Proença, *Senior Member, IEEE*, Ehsan Yaghoubi, and Pendar Alirezazadeh

***Abstract*— This article describes one objective function for learning semantically coherent feature embeddings in multi-output classification problems, i.e., when the response variables have dimension higher than one. Such coherent embeddings can be used simultaneously for different tasks, such as identity retrieval and soft biometrics labelling. We propose a generalization of the triplet loss that: 1) defines a metric that considers the number of agreeing labels between pairs of elements; 2) introduces the concept of *similar* classes, according to the values provided by the metric; and 3) disregards the notion of *anchor*, sampling four arbitrary elements at each time, from where two pairs are defined. The distances between elements in each pair are imposed according to their *semantic similarity* (i.e., the number of agreeing labels). Likewise the triplet loss, our proposal also privileges small distances between *positive* pairs. However, the key novelty is to additionally enforce that the distance between elements of any other pair corresponds inversely to their semantic similarity. The proposed loss yields embeddings with a strong correspondence between the classes centroids and their semantic descriptions. In practice, it is a natural choice to jointly infer coarse (soft biometrics) + fine (ID) labels, using simple rules such as *k-neighbours*. Also, in opposition to its triplet counterpart, the proposed loss appears to be agnostic with regard to demanding criteria for mining learning instances (such as the *semi-hard* pairs). Our experiments were carried out in five different datasets (BIODI, LFW, IJB-A, Megaface and PETA) and validate our assumptions, showing results that are comparable to the state-of-the-art in both the identity retrieval and soft biometrics labelling tasks.**

***Index Terms*— Feature embedding, soft biometrics, identity retrieval, convolutional neural networks, triplet loss.**

I. Introduction

CHARACTERIZING pedestrians in crowds has been attracting growing attention, with soft biometrics (e.g., *gender*, *ethnicity* or *age*) being particularly important to determine the identities in a scene. This kind of labels is closely related to human perception and describes the visual appearance of subjects, with applications in identity retrieval [36], [40] and person re-identification [15], [27].

Manuscript received April 2, 2020; revised July 24, 2020 and August 26, 2020; accepted August 29, 2020. Date of publication September 10, 2020; date of current version September 30, 2020. This work was supported by the FCT/MCTES through National funds and co-funded by EU funds under the Project UIDB/50008/2020, and by the Fundo de Coesão and Fundo Social Europeu, (FEDER), PT2020 Program, under the Grant POCI-01-0247-FEDER-033395. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Andrew Beng Jin Teoh. *(Corresponding author: Hugo Proença.)*

The authors are with the Department of Computer Science, IT-Instituto de Telecomunicações, University of Beira Interior, 6200-001, Covilhã, Portugal (e-mail: hugomcp@di.ubi.pt; d2401@di.ubi.pt; d2389@di.ubi.pt).

Digital Object Identifier 10.1109/TIFS.2020.3023304

Deep learning frameworks have been repeatedly improving the state-of-the-art in many computer vision tasks, such as object detection and classification [25], [41], action recognition [6], [19], semantic segmentation [24], [44] and soft biometrics inference [32]. In this context, the triplet loss [34] is a popular concept, where three learning elements are considered at a time, two of them of the same class and a third one of a different class. By imposing larger distances between the elements of the *negative* than of the *positive* pair, the intra-class compactness and inter-class discrepancy in the destiny space are enforced. This strategy was successfully applied to various problems, upon the mining of the *semi-hard* negative input pairs, i.e., cases where the negative element is farther to the anchor than the positive, but still provides a positive loss due to an imposed margin.

This article describes one objective function that is a generalization of the triplet loss. Instead of dividing the learning pairs into *positive/negative*, we define a metric to perceive the semantic similarity between two classes (IDs). In learning time, four elements are considered at a time and the margins between the pairwise distances yield from the number of agreeing labels in each pair (Fig. 1). Under this formulation, elements of *similar* classes (e.g., two "*young, black, bald, male*" subjects) are projected into adjacent regions of the destiny space. Also, as we impose different margins between (almost) all *negative* pairs, we leverage the difficulties in mining appropriate learning instances, which is one of the main difficulties in the triplet loss formulation.

The proposed loss function is particularly suitable for *coarse-to-fine* classification problems, where some labels are easier to infer than others and the global problem can be decomposed into more tractable sub-components. This hierarchical paradigm is known to be an efficient way of organizing object recognition, not only to accommodate a large number of hypotheses, but also to systematically exploit the shared attributes. Under this paradigm, the identity retrieval problem is of particular interest, where the finest labels (IDs) are seen as the leaves of hierarchical structures with roots such as the *gender* or *ethnicity* features. However, note that the proposed formulation does not appropriately handle soft labels that vary among different images of a subject (e.g., *hairstyle*). Also, it does not take into account the varying difficulty of estimating the different labels, allowing further improvements based in metric learning concepts.

The remainder of this article is organized as follows: Section II summarizes the most relevant research in the scope

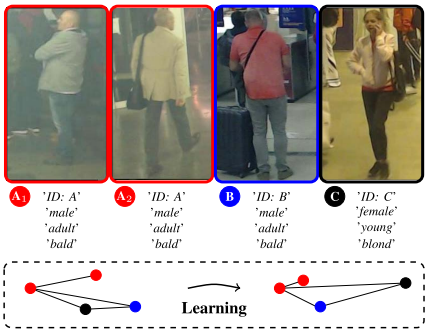


Fig. 1. Likewise the triplet loss [34], the proposed **quadruplet** formulation minimizes the distances between elements of *positive* pairs $\{A_1, A_2\}$. However, the key novelty is to additionally consider the semantic similarity between classes (A, B and C). In this example, assuming that A and B are semantically similar, our proposal privileges embeddings where the distances between (A , B) elements are smaller than the distances between (A , C) and between (B, C) elements.

of our work. Section III describes the proposed objective function. In Section IV we discuss the obtained results and the conclusions are given in Section V.

II. Related Work

Deep learning methods for biometrics can be roughly divided into two major groups: 1) methods that directly learn multi-class classifiers used in identity retrieval and soft biometrics inference; and 2) methods that learn low-dimensional feature embeddings, where inference yields from nearest neighbour search.

A. Soft Biometrics and Identity Retrieval

Bekele *et al*. [2] proposed a residual network for multi-output inference that handles classes-imbalance directly in the cost function, without depending of data augmentation techniques. Almudhahka *et al*. [1] explored the concept of comparative soft biometrics and assessed the impact of automatic estimations on face retrieval performance. Guo *et al*. [12] studied the influence of distance in the effectiveness of body and facial soft biometrics, introducing a joint density distribution based rank-score fusion strategy [13]. Vera-Rodriguez *et al*. [31] used hand-crafted features extracted from the distances between key points in body silhouettes. Martinho-Corbishley *et al*. [29] introduced the idea of *super-fine* soft attributes, describing multiple concepts of one trait as multi-dimensional perceptual coordinates. Also, using joint attribute regression and deep residual CNNs, they observed substantially better retrieval performance in comparison to conventional labels. Schumann and Specker used an ensemble of classifiers for robust attributes inference [35], extended to full body search by combining it with a human silhouette detector. He *et al*. [17] proposed a weighted multi-task CNN with a loss term that dynamically updates the weight for each task during the learning phase.

Several works regarded the semantic segmentation as a tool to support labels inference: Galiyawala *et al*. [10] described a deep learning framework for person retrieval using the height, clothes' color, and gender labels, with a segmentation module used to remove clutter. Similarly, Cipcigan and Nixon [3] obtained semantically segmented regions of the body that fed two CNN-based feature extraction and inference modules.

Finally, specifically designed for handheld devices, Samangouei and Chellappa [32] extracted various facial soft biometric features, while Neal and Woodard [26] developed a human retrieval scheme based on thirteen demographic and behavioural attributes from mobile phones data, such as calling, SMS and application data, having authors positively concluded about the feasibility of this kind of recognition.

A comprehensive summary of the most relevant research in soft biometrics is given in [38].

B. Feature Embeddings and Loss Functions

Triplet loss functions were motivated by the concept of *contrastive* loss [14], where the rationale is to penalize distances between *positive* pairs, while favouring distances between *negative* pairs. Kang *et al*. [21] used a deep ensemble of multi-scale CNNs, each one based on triplet loss functions. Song *et al*. [37] learned semantic feature embeddings that lift the vector of pairwise distances within the batch to the matrix of pairwise distances, and described a structured loss on the lifted problem. Liu and Huan [28] proposed a triplet loss learning architecture composed of four CNNs, each one learning features from different body parts that are fused at the score level.

A posterior concept was the *center* loss [42], which finds a center for each class and penalizes the distances between the projections and their corresponding class center. Jian *et al*. [20] combined additive margin *softmax* with center loss to increase the inter-classes distances and avoid over-confidence on classifications. Ranjan *et al*.'s *crystal* loss [30] restricts the features to lie on a hypersphere of a fixed radius, adding a constraint on the features projections such that their ℓ_2-norm is constant. Chen *et al*. [4] used deep representations to feed a Bayesian metrics learning module that maximizes the log-likelihood ratio between intra- and inter-classes distances. Deng *et al*.'s *Sphereface* [8] proposes an additive angular margin loss, with a clear geometric interpretation due to the correspondence to the geodesic distance on the hypersphere.

Observing that CNN-based methods tend to overfit in person re-identification tasks, Shi *et al*. [36] used siamese architectures to provide a joint description to a metric learning module, regularizing the learning process and improving the generalization ability. Also, to cope with large intra-class variations, they suggested the idea of *moderate positive mining*, again to prevent overfitting. Motivated by the difficulties in generate learning instances for triplet loss frameworks, Su *et al*. [39] performed adaptive CNN fine-tuning, along with an adaptive loss function that relates the maximum distance among the positive pairs to the margin demanded for separate *positive*

from *negative* pairs. Hu *et al*. [18] proposed an objective function that generalizes the Maximum Mean Discrepancy [33] metric, with a weighting scheme that favours good quality data. Duan *et al*. [9] proposed the *uniform* loss to learn deep equi-distributed representations for face recognition. Finally, observing the typical unbalance between positive and negative pairs, Wang *et al*. [41] described an adaptive margin list-wise loss, in which learning data are provided with a set of negative pairs divided into three classes (*easy*, *moderate*, and *hard*), depending of the distance rank with respect to the query.

Finally, we note the differences between our loss function and the (also *quadruplet*) loss described by Chen *et al*. [5]. These authors attempt to augment the inter-classes margins and the intra-class compactness without explicitly using any semantical constraint. As in the original triplet loss formulation, the concept of *similar* class doesn't exist in [5], and there is no rule to explicitly enforce the projection of identities that share most of the labels into neighbour regions of the latent space. In opposition, our method concerns essentially about such kind of semantical coherence, i.e., assures that similar classes are projected into adjacent regions of the embedding. Also, even the idea behind the loss formulation is radically different in both methods, in the sense that [5] still considers the concept of *anchor* (as the triplet-loss), which is also in opposition to our proposal.

III. Proposed Method

A. Quadruplet Loss: Definition

Consider a supervised classification problem, where t is the dimensionality of the response variable $\boldsymbol{y}_i$ associated to the input element $\boldsymbol{x}_i \in [0, 255]^n$. Let $f(.)$ be one embedding function that maps $\boldsymbol{x}_i$ into a d-dimensional space Ψ, with $\boldsymbol{f}_i = f(\boldsymbol{x}_i) \in \Psi$ being the projected vector. Let $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_b\}$ be a batch of b images from the learning set. We define $\phi(\boldsymbol{y}_i, \boldsymbol{y}_j) \in \mathbb{N}$, $\forall i, j \in \{1, \ldots, b\}$ as the function that measures the semantic similarity between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$:

$$\phi(\boldsymbol{y}_i, \boldsymbol{y}_j) = ||\boldsymbol{y}_i - \boldsymbol{y}_j||_0, \tag{1}$$

with $||.||_0$ being the ℓ_0-norm operator.

In practice, $\phi(.,.)$ counts the number of disagreeing labels between the $\{\boldsymbol{x}_i, \boldsymbol{x}_j\}$ pair, i.e., $\phi(\boldsymbol{y}_i, \boldsymbol{y}_j) = t$ when the i^{th} and j^{th} elements have fully disjoint classes membership (e.g., one "*black, adult, male*" and another "*white, young, female*" subjects), while $\phi(\boldsymbol{y}_1, \boldsymbol{y}_2) = 0$ when they have the exact same label (class) across all dimensions, i.e., when they constitute a *positive* pair.

Let $\{i, j, p, q\}$ be the indices of four images in the batch. The corresponding quadruplet loss value $\ell_{i,j,p,q}$ is given by:

$$\ell_{i,j,p,q} = sgn\Big(\phi(\boldsymbol{y}_i, \boldsymbol{y}_j) - \phi(\boldsymbol{y}_p, \boldsymbol{y}_q)\Big) \times \Big[\big(\|\boldsymbol{f}_p - \boldsymbol{f}_q)\|_2^2 - \|\boldsymbol{f}_i - \boldsymbol{f}_j\|_2^2\big) + \alpha\Big], \tag{2}$$

where $sgn()$ is the sign function, $||\boldsymbol{x}||_2^2$ denotes the square of the ℓ_2-norm of $\boldsymbol{x}$ $\big(||\boldsymbol{x}||_2 = (x_1^2 + \ldots x_n^2)^{\frac{1}{2}}$, i.e., $||\boldsymbol{x}||_2^2 = x_1^2 + \ldots x_n^2\big)$ and α is the desired margin ($\alpha = 0.1$ was used in our experiments). Evidently, the loss value will be zero when both image pairs have the same number of agreeing labels (as $sgn(0) = 0$ in these cases). In any other case, the sign function will determine the pair which distance in the embedding should be minimized. As an example, if the (p, q) elements are semantically closer to each other than the (i, j) elements $\big(\phi(\boldsymbol{y}_p, \boldsymbol{y}_q) < \phi(\boldsymbol{y}_i, \boldsymbol{y}_j)\big)$, we want to ensure that $\|\boldsymbol{f}_p - \boldsymbol{f}_q)\|_2^2 < \|\boldsymbol{f}_i - \boldsymbol{f}_j\|_2^2$.

The accumulated loss in the batch is given by the truncated mean of a sample (of size s) randomly taken from the subset of the $\binom{b}{4}$ individual loss values where $\phi(\boldsymbol{y}_i, \boldsymbol{y}_j) \neq \phi(\boldsymbol{y}_p, \boldsymbol{y}_q)$:

$$\mathcal{L} = \frac{1}{s} \sum_{z=1}^{s} \Big[\ell_z\Big]_+, \tag{3}$$

where $\boldsymbol{z} \in \{1, \ldots, s\}^4$ denotes the z^{th} composition of four elements in the batch and $[.]_+$ is the $\max(., 0)$ function. Even considering that a large fraction of the combinations in the batch will be invalid (i.e., with $\phi(.,.) = 0$), large values of b will result in an intractable number of combinations at each iteration. In practical terms, after filtering out those invalid combinations, we randomly sample a subset of the remaining instances, which is designated as the *mini-batch*.

B. Quadruplet Loss: Training

Consider four indices $\{i, j, p, q\}$ of elements in the mini-batch, with $\phi(\boldsymbol{y}_i, \boldsymbol{y}_j) > \phi(\boldsymbol{y}_p, \boldsymbol{y}_q)$. Let Δ_ϕ denote the difference between the number of disagreeing labels of the $\{i, j\}$ and $\{p, q\}$ pairs:

$$\Delta_\phi = \phi(\boldsymbol{y}_i, \boldsymbol{y}_j) - \phi(\boldsymbol{y}_p, \boldsymbol{y}_q). \tag{4}$$

Also, let Δ_f be the distance between the elements of the most alike pair minus the distance between the elements of the least alike pair in the destiny space (plus the margin):

$$\Delta_f = \|\boldsymbol{f}_p - \boldsymbol{f}_q)\|_2^2 - \|\boldsymbol{f}_i - \boldsymbol{f}_j\|_2^2 + \alpha. \tag{5}$$

Upon basic algebraic manipulation, the gradients of $\mathcal{L}$ with respect to the quadruplet terms are given by:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{f}_i} = \sum_z \begin{cases} 2(\boldsymbol{f}_j - \boldsymbol{f}_i), & \text{if } \Delta_\phi > 0 \ \wedge \Delta_f \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{f}_j} = \sum_z \begin{cases} 2(\boldsymbol{f}_i - \boldsymbol{f}_j), & \text{if } \Delta_\phi > 0 \ \wedge \Delta_f \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{f}_p} = \sum_z \begin{cases} 2(\boldsymbol{f}_p - \boldsymbol{f}_q), & \text{if } \Delta_\phi > 0 \ \wedge \Delta_f \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{f}_q} = \sum_z \begin{cases} 2(\boldsymbol{f}_q - \boldsymbol{f}_p), & \text{if } \Delta_\phi > 0 \ \wedge \Delta_f \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

In practice terms, the model weights are adjusted only when pairs have different number of agreeing labels (i.e., $\Delta_\phi > 0$) and when the distance in the destiny space between the elements of the most similar pair is higher than the distance between the elements of the least similar pair (plus the margin, $\Delta_f \geq 0$). According to this idea, using (6)-(9), the deep learning frameworks supervised by the proposed quadruplet loss are trainable in a way similar to its counterpart triplet

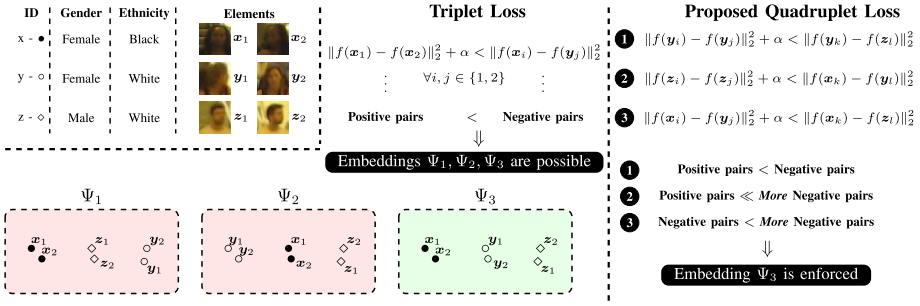


Fig. 2. Key difference between the triplet loss [34] formulation and the solution proposed in this article. Using a loss function that analyzes the semantic similarity (in terms of soft biometrics) between the different identities, we enforce embeddings (Ψ_3) that are semantically coherent, i.e., where: 1) elements of the same class appear near each other; but additionally 2) elements of similar classes appear closer to each other than elements with no labels in common. This is in opposition to the original formulation of the triplet loss, that relies mostly in image appearance to define the geometry of the destiny space, obtaining - in case of noisy image features - semantically incoherent embeddings (e.g., in Ψ_1 and Ψ_2, classes are compact and discriminative, but the $\boldsymbol{x}/z$ centroids are too close to each other).

loss and can be optimized according to the standard Stochastic Gradient Descend (SGD) algorithm, which was done in all our experiments.

For clarity purposes, Algorithm 1 gives a pseudocode description of the learning phase and of the batch/mini-batch definition processes.

Algorithm 1 Pseudocode Description of the Learning Phase and of the Batch/Mini-Batch Definition Processes

Precondition: $\boldsymbol{M}$: CNN, t_e: Tot. epochs, s: mini-batch size, b: batch size, $\boldsymbol{I}$: Learning set, n images
for 1 to t_e **do**
  for 1 to $\lfloor \frac{n}{s} \rceil$ **do**
    $\boldsymbol{b} \leftarrow$ randomly sample b out of n images from $\boldsymbol{I}$
    $\boldsymbol{c} \leftarrow$ create $\binom{b}{4}$ quadruplet combinations from $\boldsymbol{b}$
    $\boldsymbol{c}* \leftarrow$ filter out invalid elements from $\boldsymbol{c}$
    $\boldsymbol{s} \leftarrow$ randomly sample s elements from $\boldsymbol{c}^*$
    $\boldsymbol{M} \leftarrow$ update weights($\boldsymbol{M}, \boldsymbol{s}$) (eqs. (6-9))
  end for
end for
return $\boldsymbol{M}$

C. Quadruplet Loss: Insight and Example

Fig. 2 illustrates our rationale in the proposed loss. By defining a metric that analyses the similarity between two classes, we create the concept of *semantically similar* class. This enables to explicitly enforce that elements of the *least similar* classes (with no common labels) are at the farthest distances in the embedding. During the learning phase, we sample the image pairs in a stochastic way and enforce projections in a way that resembles the human perception of *semantic similarity*.

As an example, Fig. 3 compares the bidimensional embeddings resulting from the triplet and the quadruplet losses, for the LFW identities with more than 15 images in the dataset (using $t = 2$: {'ID', 'Gender'} labels). This plot yielded from the projection of a 128-dimensional embedding down to two dimensions, according to the Neighbourhood Component Analysis (NCA) [11] algorithm.

It can be seen that the triplet loss provided an embedding where the positions of elements are exclusively determined by their appearance, where 'females' appear nearby 'male tennis players' (upper left corner). In opposition, the quadruplet loss established a large margin between both genders, while keeping the compactness per ID. This kind of embedding is interesting: 1) for identity retrieval, to guarantee that all retrieved elements have soft labels equal to the query; 2) upon a semantic description of the query (e.g., "*find adult white males similar to this image*"), to guarantee that all retrieved elements meet the semantic criteria; and 3) to use the same embedding to directly infer fine (ID) + coarse (soft) labels, in a simple *k-neighbours* fashion.

IV. Results and Discussion

A. Experimental Setting and Preprocessing

Our empirical validation was conducted in one proprietary (BIODI) and four freely available datasets (LFW, PETA, IJB-A and Megaface) well known in the biometrics and re-identification literature.

The BIODI[1] dataset is proprietary of *Tomiworld*®,[2] being composed of 849,932 images from 13,876 subjects, taken from 216 indoor/outdoor video surveillance sequences. All images were manually annotated for 14 labels: gender, age, height, body volume, ethnicity, hair color and style, beard, moustache, glasses and clothing (x4). The Labeled Faces in the Wild (LFW) [16] dataset contains 13,233 images from 5,749 identities, collected from the web, with large variations in pose, expression and lighting conditions. PETA [7] is a combination of 10 pedestrian re-identification datasets, composed

[1] http://di.ubi.pt/~hugomcp/BIODI/
[2] https://tomiworld.com/

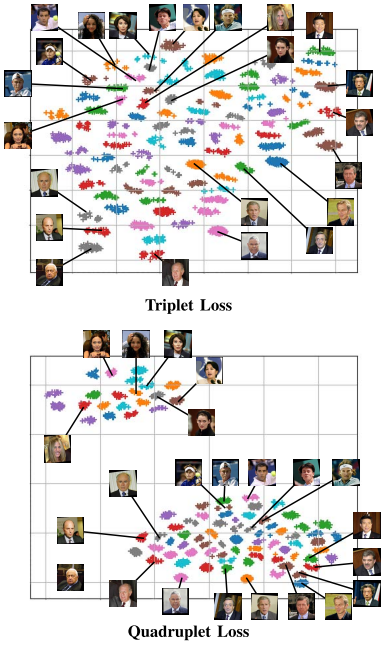


Fig. 3. Comparison between the 2D embeddings resulting from the triplet loss [34] (top plot), and from the proposed quadruplet loss (bottom plot). Results are given for $t = 2$ features {'ID', 'Gender'} for the LFW identities with at least 15 images (89 elements).

of 19,000 images from 8,705 subjects, each one annotated with 61 binary and 4 multi-output atributes. The IIJB-A [23] dataset contains 5,397 images plus 20,412 video frames from 500 individuals, with large variations in pose and illumination. Finally, the Megaface [22] set was released to evaluate face recognition performance at the million scale, and consists of a gallery set and a probe set. The gallery set is a subset of Flickr photos from Yahoo (more than 1,000,000 images from 690,000 subjects). The probe dataset includes FaceScrub and FGNet sets. FaceScrub has 100,000 images from 530 individuals and FGNet contains 1,002 images of 82 identities. Some examples of the images in each dataset are given in Fig. 4.

B. Convolutional Neural Networks

Two CNN architectures were considered: the *VGG* and *ResNet* models (Fig. 5). Here, the idea was not only to compare the performance of the quadruplet loss with respect to the baselines, but also to perceive the variations in performance with respect to different CNN architectures. A *TensorFlow* implementation of both architectures is available at.[3]

[3]https://github.com/hugomcp/quadruplets

Fig. 4. Datasets used in the empirical validation of the method proposed in this article. From top to bottom rows, images of the BIODI, PETA, LFW, Megaface and IJB-A sets are shown.

All the models were initialized with random weights, from zero-mean Gaussian distributions with standard deviation 0.01 and bias 0.5. Images were resized to 256×256, adding lateral white bands when needed to keep constant ratios. A batch size of 64 was defined, which results in too many combinations of pairs for the triplet/quadruplet losses. At each iteration, we filtered out the invalid triplets/quadruplets instances and randomly selected the mini-batch elements, composed of 64 instances in all cases. For every baseline, 64 pairs were also used as a batch. The learning rate started from 0.01, with momentum 0.9 and weight decay $5e^{-4}$. In the *learning-from-scratch* paradigm, we stopped the learning process when the validation loss didn't decrease for 10 iterations (i.e., *patience*=10).

We initially varied the dimensionality of the embedding (d) to perceive the sensitivity of the proposed method with respect to this parameter. Considering the LFW set, the average AUC values with respect to d are provided in Fig. 6 (the shadowed regions denote the $\pm$ standard deviation performance, after 10 trials). As expected, higher values for d were directly correlated to performance, even though results stabilised for dimensions higher than 128. In this regard, we assumed that using higher dimensions would require much more training data, having resorted from this moment to $d = 128$ in all subsequent experiments.

Interestingly, the absolute performance observed for very low d values was not too far of the obtained for much higher dimensions, which raises the possibility of using the position of the elements in the destiny space directly for classification and visualization, without the need of any dimensionality

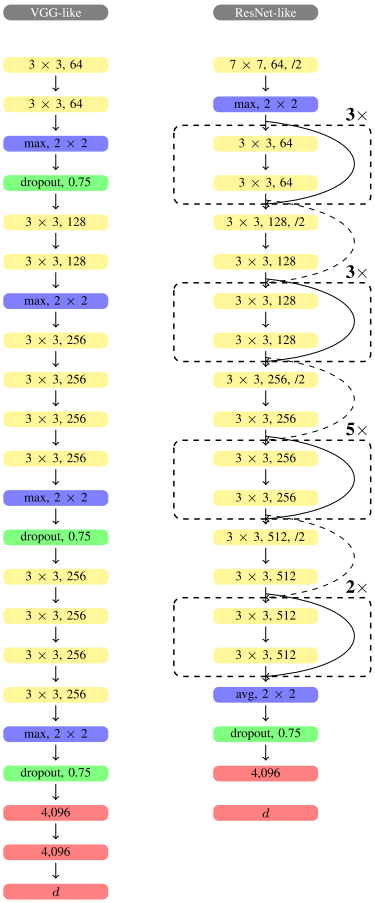


Fig. 5. Architectures of the CNNs used in the experiments. The yellow boxes represent convolutional layers, and the blue and green boxes represent pooling and dropout (keeping probability 0.75) layers. Finally, the red boxes denote fully connected layers. In the ResNet architecture, the dashed skip connections represent convolutions with stride 2×2, yielding outputs with half of the spatial input size. The '/2' symbol denotes stride 2×2 (the remaining layers use stride 1×1).

reduction algorithm (MDS, LLE or PCA algorithms are frequently seen in the literature for this purpose).

C. Single- vs. Multi-Output Embeddings Learning: Semantical Coherence

To compare the semantical coherence of the embeddings resulting from single-output (triplet and Chen *et al.*'s losses) and multi-output (ours) learning formulations, we measured the distances (ℓ_2-norm) between each element in an

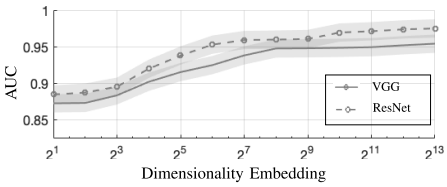


Fig. 6. Variations in the mean AUC values ($\pm$ the standard deviations after 10 trials, given as shadowed regions) with respect to the dimensionality of the embedding. Results are shown for the LFW validation set, when using the VGG-like (solid line) and ResNet-like (dashed line) CNN architectures.

embedding and all the others, grouping values into two sets: 1) *intra-label* observations, when two elements share a specific label (e.g., 'male'/'male' or 'asian'/'asian'); and 2) *inter-labels* observations, in case of different labels in the pair (e.g., 'male'/'female' or 'asian'/'black'). In practice, we measured the distances between elements of the same/different ID, gender, ethnicity and joint gender+ethnicity labels. Note that, in all cases, a unique embedding was obtained for each method, using the {ID} as feature for the triplet and Chen *et al.* methods, and the {ID, Gender, Ethnicity} ($t = 3$) for the proposed method, with the annotations for the IJB-A set provided by the Face++ algorithm and subjected to human validation. The VGG-like architecture was considered, as described in Section IV-B.

The results are given in Fig. 7 (LFW, Megaface and IJB-A sets). The green color represents the statistics of the *intra-label* values, while the red color represents the *inter-labels* values. Box plots show the median of the distance values (horizontal solid lines) and the first and third quartiles (top and bottom of the box marks). The upper and lower whiskers are denoted by the horizontal lines outside each box. All outliers are omitted, for visualisation purposes.

The leftmost group in each dataset is the root for the ID retrieval performance, and compares the distances in the embeddings between elements that have the same/different IDs. The remaining cases are the most important for our purposes, and provide the distances between elements that share (or not) some label: the second group compares the 'male'/'male and 'female'/'female' distances (green boxes) to 'male'/'female' values (red boxes). The third group provides the corresponding results for the *ethnicity* label, while the rightmost group provides the distances when jointly considering the *gender* and *ethnicity* features, i.e., when two elements constitute an *intra-label* pair *iff* they have the same gender *and* ethnicity labels.

These results turn evident the different properties of the embeddings yielding from the proposed loss with respect to the baselines. If we consider exclusively the ID to measure the distances between elements, the results almost do not vary among all methods. However, a different conclusion can be drawn when measuring the distances between the same/different *gender*, *ethnicity* and *gender/ethnicity* labels. Here, the proposed quadruplet loss was the unique method where the intra-label/inter-labels whiskers provided disjoint

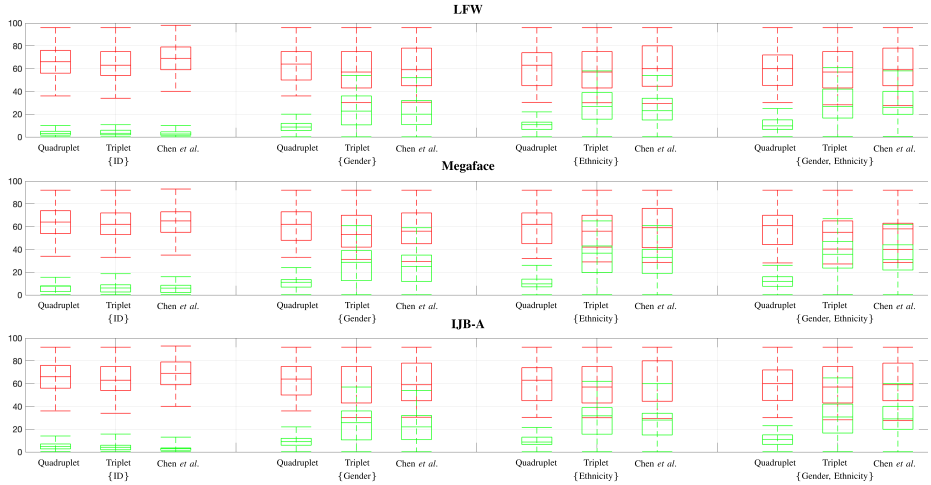


Fig. 7. Box plots of the distances between each element in the embedding with respect to others that share the same (green color) or different (red color) labels. We compare the multi-output learning solution proposed in this article (Quadruplet), with respect to the single-output learning methods (Triplet [34] and Chen *et al.* [5]). Values regard the LFW (top plot), Megaface (center plot) and IJB-A (bottom plot) sets, measuring the {ID}, {Gender}, {Ethnicity} and {Gender, Ethnicity} same/different label distances.

intersections, by a solid margin in all cases, i.e., the difference between the intra-label/inter-labels distances was far larger than in the remaining losses. Of course, such differences are due to the fact that the triplet and Chen *et al*. methods have not considered additional soft labels to define the topology of the embeddings, having exclusively resorted to the ID labels and images appearance for such purpose.

In practice, these experiments turn evident that single-label learning formulation yield embeddings that are semantically incoherent from other labels' perspectives, in the sense that 'males' are often nearby 'females', or 'white' nearby 'asian' elements. In this setting, using such embeddings for simultaneously ID retrieval and soft biometrics labelling is risky, and errors will often occur. In opposition, the proposed loss guarantees large margins between groups of intra-label/inter-labels observations, typically corresponding to *clusters* in the embeddings with respect to the set of learning labels considered.

D. Identity Retrieval

Even considering that the goals of our proposal are beyond the ID retrieval performance, it is important to compare the performance of the quadruplet loss with respect to the baselines in this task. As in the previous experiment, note that all the baselines (triplet loss, center loss, *softmax* and Chen *et al*. [5]) considered exclusively the ID to infer the embeddings, while the proposed loss used all the available labels for that purpose.

Fig. 8 provides the Cumulative Match curves (CMC, outer plots) and the Detection and Identification rates at rank-1 (DIR, inner plots). The results are also summarized in Table I, reporting the rank-1, top-10% values and the mean average precision (mAP) scores, given by:

$$\text{mAP} = \frac{\sum_{q=1}^{n} \bar{P}(q)}{n}, \tag{10}$$

where n is the number of queries, $\bar{P}(q) = \sum_{k=1}^{n} P(k)\Delta r(k)$, $P(k)$ is the precision at cut-off k and $\Delta r(k)$ is the change in recall from $k-1$ to k.

For the LFW set experiment, the BLUFR[4] evaluation protocol was chosen. In the verification (1:1) setting, the test set contained 9,708 face images of 4,249 subjects, which yielded over 47 million matching scores. For the open-set identification problem, the genuine probe set contained 4,350 face images of 1,000 subjects, the impostor probe set had 4,357 images of 3,249 subjects, and the gallery set had 1,000 images. This evaluation protocol was the basis to design, for the other sets, as close as possible experiments, in terms of the number of matching scores, gallery and probe sets.

Generally, we observed that the proposed quadruplet loss outperforms the other loss functions, which might be the result of having used additional information for learning. These improvements in performance were observed in most cases by a consistent margin for both the verification and identification tasks, not only for the VGG but also for the ResNet architecture.

In terms of the errors per CNN architecture, the ResNet-like error rates were roughly $0.9 \times (90\%)$ of the observed for the VGG-like networks (higher margins were observed for the *softmax* loss). Not surprisingly, the Chen *et al*. [5]'

[4]http://www.cbsr.ia.ac.cn/users/scliao/projects/blufr/

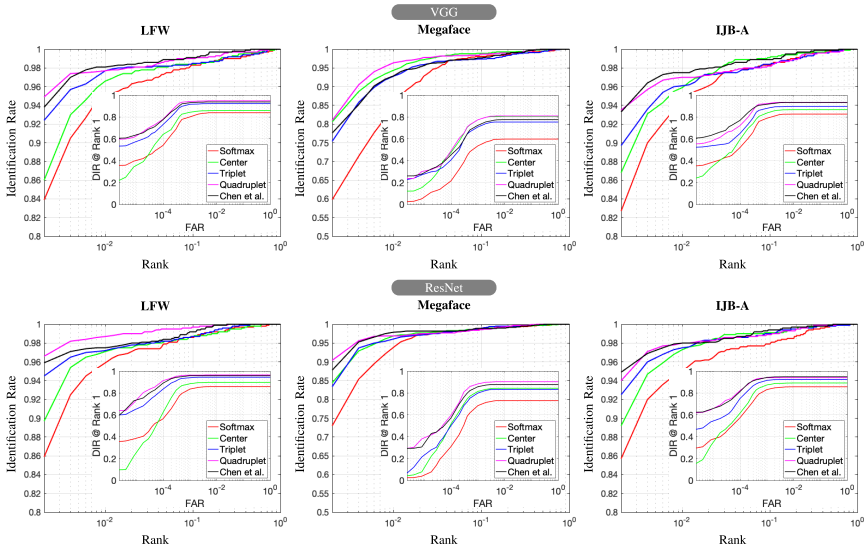


Fig. 8. Identity retrieval results. The outer plots provide the closed-set identification (CMC) curves for the LFW, Megaface and IJB-A sets, using the VGG and ResNet architectures. Inside each plot, the inner regions show the corresponding detection and identification rate (DIR) values at rank-1. Results are shown for the quadruplet loss function (purple color), and four baselines: the *softmax* (red color), center loss (green color), triplet loss (blue color) and Chen *et al.* [5]'s (black color) method.

method outperformed the remaining competitors, followed by the triplet loss function, which is consistent with most of the results reported in the literature. The *softmax* loss got repeatedly the worst performance among the five functions considered.

Regarding the performance per dataset, the values observed for Megaface were far worse for all objective functions than the values for LFW and IJB-A. In the Megaface set, we followed the protocol of the *small* training set, using 490,000 images from 17,189 subjects (images overlapping with Facescrub dataset were discarded). Also, note that the relative performance between the loss functions was roughly the same in all sets. Degradations in performance were slight from the LFW to the IJB-A set and much more visible in case of the Megaface set. In this context, the *softmax* loss produced the most evident degradations, followed by the center loss.

E. Soft Biometrics Inference

As stated above, the proposed loss can also be used for learning a soft biometrics estimator. In test time, the position to where one element is projected is used to infer the soft labels, in a simple nearest neighbour fashion. In these experiments, we considered only 1-NN, i.e., the label inferred for each query was given by the closest gallery element. Better results would be possibly attained if more neighbours had been considered, even though the computational cost of classification will also increase. All experiments were conducted according to a bootstrapping-like strategy: having n test images available, the bootstrap randomly selected (with replacement) $0.9 \times n$ images, obtaining samples composed of 90% of the whole data. Ten test samples were created and the experiments were conducted independently on each trail, which enabled to obtain the mean and the standard deviation at each performance value.

As baselines we used two commercial off-the-shelf (COTS) techniques, considered to represent the state-of-the-art [38]: the Matlab SDK for *Face++*[5] and the *Microsoft Cognitive Toolkit Commercial*.[6] Face++ is a commercial face recognition system, with good performance reported for the LFW face recognition competition (second best rate). Microsoft Cognitive Toolkit is a deep learning framework that provides useful information based on vision, speech and language. Also, in order to highlight the distinct properties of the embeddings generated by our proposal with respect to the state-of-the-art, we also measured the soft labelling effectiveness that can be attained by the Triplet loss [34] and Chen*et al*. [43] embeddings if a simple 1-NN rule is used to infer soft biometrics labels.

We considered exclusively the 'Gender', 'Ethnicity' and 'Age' labels ($t = 3$), quantised respectively into two classes for Gender ({'male', 'female'}), three classes for Age ({'young', 'adult', 'senior'}), and three classes for Ethnicity ({'white',

[5]http://www.faceplusplus.com/

[6]https://www.microsoft.com/cognitive-services/

TABLE I

IDENTITY RETRIEVAL PERFORMANCE OF THE PROPOSED LOSS WITH RESPECT TO THE BASELINES: *softmax*, CENTER AND TRIPLET LOSSES, AND CHEN *et al.* [5]'S METHOD. THE AVERAGE PERFORMANCE ± STANDARD DEVIATION VALUES ARE GIVEN, AFTER 10 TRIALS. INSIDE EACH CELL, VALUES REGARD (FROM TOP TO BOTTOM) THE LFW, MEGAFACE AND IJB-A DATASETS. THE BOLD FONT HIGHLIGHTS THE BEST RESULT PER DATASET AMONG ALL METHODS

Method	mAP	rank-1	top-10%
VGG			
Quadruplet loss	$0.958 \pm 3e^{-3}$	**0.951** $\pm$ 0.020	**0.979** $\pm 6e^{-3}$
	0.877 $\pm$ 0.011	**0.812** $\pm$ 0.053	**0.960** $\pm 9e^{-3}$
	0.953 $\pm 5e^{-3}$	**0.939** $\pm$ 0.037	$0.958 \pm 6e^{-3}$
Softmax loss	$0.897 \pm 4e^{-3}$	0.842 ± 0.034	0.953 ± 0.011
	0.727 ± 0.014	0.615 ± 0.060	0.863 ± 0.017
	0.849 ± 0.010	0.823 ± 0.039	0.941 ± 0.014
Triplet loss [34]	$0.934 \pm 4e^{-3}$	0.929 ± 0.033	$0.964 \pm 8e^{-3}$
	$0.854 \pm 9e^{-3}$	0.758 ± 0.059	0.946 ± 0.017
	$0.917 \pm 5e^{-3}$	0.901 ± 0.040	0.950 ± 0.011
Center loss [43]	$0.918 \pm 3e^{-3}$	0.863 ± 0.020	$0.962 \pm 6e^{-3}$
	0.850 ± 0.013	0.773 ± 0.052	0.939 ± 0.012
	0.862 ± 0.010	0.867 ± 0.041	0.944 ± 0.012
Chen *et al.* [5]	**0.961** $\pm 2e^{-3}$	0.945 ± 0.022	$0.976 \pm 6e^{-3}$
	0.864 ± 0.012	0.772 ± 0.061	$0.947 \pm 9e^{-3}$
	$0.948 \pm 6e^{-3}$	0.936 ± 0.055	**0.970** $\pm 4e^{-3}$
ResNet			
Quadruplet loss	**0.968** $\pm 2e^{-3}$	**0.966** $\pm$ 0.012	$0.981 \pm 4e^{-3}$
	$0.902 \pm 9e^{-3}$	**0.906** $\pm$ 0.048	**0.972** $\pm 8e^{-3}$
	0.959 $\pm 3e^{-3}$	0.947 ± 0.021	$0.980 \pm 4e^{-3}$
Softmax loss	$0.912 \pm 4e^{-3}$	0.861 ± 0.029	$0.960 \pm 8e^{-3}$
	0.730 ± 0.010	0.745 ± 0.051	0.899 ± 0.011
	$0.841 \pm 9e^{-3}$	0.860 ± 0.030	$0.958 \pm 8e^{-3}$
Triplet loss [34]	$0.947 \pm 4e^{-3}$	0.948 ± 0.026	$0.968 \pm 9e^{-3}$
	$0.872 \pm 8e^{-3}$	0.839 ± 0.052	$0.957 \pm 9e^{-3}$
	$0.919 \pm 5e^{-3}$	0.937 ± 0.031	0.961 ± 0.011
Center loss [43]	$0.939 \pm 3e^{-3}$	0.898 ± 0.016	$0.967 \pm 6e^{-3}$
	$0.847 \pm 9e^{-3}$	0.845 ± 0.048	$0.945 \pm 9e^{-3}$
	$0.877 \pm 7e^{-3}$	0.893 ± 0.035	$0.963 \pm 9e^{-3}$
Chen *et al.* [5]	$0.966 \pm 2e^{-3}$	0.959 ± 0.015	**0.983** $\pm 4e^{-3}$
	0.916 $\pm 8e^{-2}$	0.880 ± 0.050	$0.975 \pm 8e^{-3}$
	$0.952 \pm 4e^{-3}$	**0.960** $\pm$ 0.022	**0.986** $\pm 6e^{-3}$

TABLE II

SOFT BIOMETRICS LABELLING PERFORMANCE (MAP) ATTAINED BY THE PROPOSED METHOD, WITH RESPECT TO TWO COMMERCIAL-OFF-THE-SHELF SYSTEMS (FACE++ AND MICROSOFT COGNITIVE) AND TWO OTHER BASELINES. THE AVERAGE PERFORMANCE ± STANDARD DEVIATION VALUES ARE GIVEN, AFTER 10 TRIALS. INSIDE EACH CELL, THE TOP VALUE REGARDS THE VGG-LIKE PERFORMANCE, AND THE BOTTOM VALUE CORRESPONDS TO THE RESNET-LIKE VALUES

Method	Gender	Age	Ethnicity
BIODI			
Quadruplet loss	$0.816 \pm 6e^{-3}$	0.603 ± 0.014	0.777 ± 0.011
	0.834 $\pm 5e^{-3}$	**0.649** $\pm$ 0.011	$0.786 \pm 9e^{-3}$
Triplet loss [34]	0.684 ± 0.022	0.581 ± 0.034	0.599 ± 0.028
	0.690 ± 0.019	0.584 ± 0.025	0.600 ± 0.017
Chen *et al.* [43]	0.693 ± 0.020	0.602 ± 0.032	0.613 ± 0.019
	0.697 ± 0.015	0.604 ± 0.012	0.618 ± 0.018
Face++	$0.760 \pm 8e^{-3}$	0.588 ± 0.019	**0.788** $\pm$ 0.017
Microsoft Cognitive	$0.738 \pm 7e^{-3}$	0.552 ± 0.026	-
PETA			
Quadruplet loss	0.862 ± 0.024	0.649 ± 0.061	0.797 ± 0.053
	0.882 ± 0.018	0.658 ± 0.057	0.810 ± 0.036
Triplet loss [34]	0.720 ± 0.036	0.611 ± 0.038	0.612 ± 0.038
	0.722 ± 0.024	0.625 ± 0.022	0.628 ± 0.026
Chen *et al.* [43]	0.723 ± 0.034	0.613 ± 0.037	0.636 ± 0.025
	0.731 ± 0.027	0.630 ± 0.030	0.668 ± 0.021
Face++	0.870 ± 0.028	0.653 ± 0.062	**0.812** $\pm$ 0.054
Microsoft Cognitive	**0.885** $\pm$ 0.020	**0.660** $\pm$ 0.057	-
LFW			
Quadruplet loss	0.939 ± 0.021	0.702 ± 0.059	0.801 ± 0.044
	0.944 $\pm$ 0.017	0.709 ± 0.049	0.817 ± 0.041
Triplet loss [34]	0.794 ± 0.028	0.631 ± 0.032	0.652 ± 0.022
	0.799 ± 0.022	0.636 ± 0.020	0.670 ± 0.017
Chen *et al.* [43]	0.794 ± 0.030	0.639 ± 0.030	0.728 ± 0.027
	0.801 ± 0.021	0.659 ± 0.018	0.747 ± 0.022
Face++	0.928 ± 0.041	0.527 ± 0.063	**0.842** $\pm$ 0.061
Microsoft Cognitive	0.931 ± 0.037	**0.710** $\pm$ 0.051	-

'black', 'asian'}). The average and standard deviation performance values are reported in Table II for the BIODI, PETA and LFW sets.

Overall, the results achieved by the quadruplet loss can be favourably compared to the baseline techniques for most labels, particularly for the BIODI and LFW datasets. Regarding the PETA set, Face++ invariably outperformed the other techniques, even if at a reduced margin in most cases. This was justified by the extreme heterogeneity of image features in this set, in result of being the concatenation of different databases. This should had reduced the representativity of the learning data with respect the test set, being the Face++ model apparently the least sensitive to this covariate. Note that the 'Ethnicity' label is only provided by the Face++ framework. Regarding the Triplet [34] and Chen *et al.* [43] baselines, it is important to note that the reported values were obtained in embeddings that were inferred exclusively based in ID information. Under such circumstances, we confirmed that both solutions produce semantically inconsistent embeddings, in which elements with similar appearance but different soft labels are frequently projected to adjacent regions.

Globally, these experiments supported the possibility of using such the proposed method to estimate soft labels in a *single-shot* paradigm, which is interesting to reduce the computational cost of using specialized third-party solutions for soft labelling.

Finally, we analysed the variations in performance with respect to the number of labels considered, i.e., the value of the t parameter. At first, to perceive how the identity retrieval performance depends of the number of soft labels, we used the

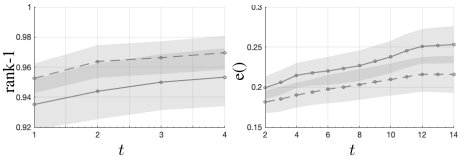


Fig. 9. At left: rank-1 identification accuracy in the LFW dataset, for $1 \leq t \leq 4$. At right: soft biometrics performance in the BIODI test set, for $2 \leq t \leq 14$, for the VGG (solid line) and ResNet (dashed line) architectures.

annotations provided by the ATVS group [38] for the LFW set, and measured the rank-1 variations for $1 \leq t \leq 4$, starting by the 'ID' label alone and then adding iteratively the 'Gender' $\rightarrow$ 'Ethnicity' $\rightarrow$ 'Age' labels. The results are shown in the left plot of Fig. 9. In a complementary way, to perceive the overall labelling effectiveness for large values of t, the BIODI dataset was used (the one with the largest number of annotated labels), and the values obtained for $t \in \{2, \ldots, 14\}$. In all cases, $d = 128$ was kept, with the average labelling error in the test set $\boldsymbol{X}$ given by:

$$e(\boldsymbol{X}) = \frac{1}{n.t} \sum_{i=1}^{n} ||\boldsymbol{p}_i - \boldsymbol{g}_i||_0, \tag{11}$$

with $\boldsymbol{p}_i$ denoting the t labels predicted for the i^{th} image and $\boldsymbol{g}_i$ being the ground-truth. $||\ ||_0$ denotes the ℓ_0-norm.

It is interesting to observe the apparently contradictory results in both plots: at first, a positive correlation between the labelling errors and the values of t is evident, which was justified by the difficulty of inferring some of the hardest labels in the BIODI set (e.g., the *type of shoes*). However, the average rank-1 identification accuracy also increased when more soft labels were used, even if the results were obtained only for small values of t (i.e., not considering the particularly hard labels, in result of no available ground truth). Overall, we concluded that the proposed loss obtain *acceptable* performance (i.e., close to the state-of-the-art) when a small number of soft labels is available (≥ 2), but also when a few more labels should be inferred (up to $t \approx 8$). In this regard, we presume that even higher values for t ($t \gg 8$) would require substantially more amounts of learning data and also higher values for d (dimension of the embedding).

F. Semantic Identity Retrieval

Finally, we considered the *semantic identity retrieval* problem, where - along with the query image - semantic criteria are used to filter the retrieved elements (i.e., "*Find this person*" $\rightarrow$ "*Find this female*", Fig. 10). In this setting, it is assumed that the ground-truth soft labels of the gallery IDs are known, even though the same does not apply for the queries.

We considered the hardest identity retrieval dataset (Megaface) and compared our results to Chen *et al.*'s (the most frequent runner-up in previous experiments). The soft label '*Gender*' (provided by the Microsoft Cognitive Toolkit for the queries) was used as additional semantic data, to filter the retrieved identities. The bottom plot in Fig. 10 provides the

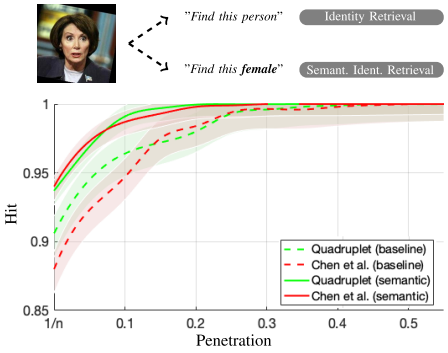


Fig. 10. Comparison between the hit/penetration rates of the proposed loss and Chen *et al.* [5]'s method, when disregarding (baseline) or considering semantic additional information to filter the retrieved results. Values are given for the *ResNet* architecture and Megaface dataset. The '*Gender*' was the semantic criterium in each query and "n" is the number of enrolled identities.

results in terms of the hit/penetration rates, being notorious the similar levels of performance of both methods in this setting ('semantic' data series), with Chen *et al.*'s method slightly outperforming up to the top-20 identities, and getting worse results than our solution for the remaining penetration values.

It can be concluded that - when coarse labels are available - our method and Chen *et al.*'s attain similar quality embeddings in terms of compactness and discriminability. However, the key point is that the baseline version of the proposed loss is a way to approximate the results attained by state-of-the-art methods when using semantic information to filter the retrieved identities.

V. Conclusion and Further Work

In this article we proposed a loss function for multi-output classification problems, where the response variables have dimension greater than one. Our function is a generalization of the well known triplet loss, replacing the *positive/negative* binary division of pairs and the notion of *anchor*, by: i) a metric that considers the *semantic similarity* between any two classes; and ii) a quadruplet term that imposes different distances between pairs of elements according to that similarity.

In particular, we considered the identity retrieval and soft biometrics problems, using the ID and three soft labels ('Gender', 'Age' and 'Ethnicity') to obtain semantically coherent embeddings. In such spaces, not only the intra-class compactness is guaranteed, but also the broad families of classes (e.g., "white young males" or "black senior females") appear in adjacent regions. This enables a direct correspondence between the ID centroids and their semantic descriptions, allowing that simple rules such as k-neighbours are used to jointly infer the identity/soft label information. The insight of the proposed loss is in opposition to single-label loss formulations, where elements are projected into the destiny

space based uniquely in ID information and image appearance, being assumed that semantical coherence yields naturally upon the similarity of image features.

As future directions for this work, we are exploring the possibility of fusing the concept described in this article to the original triplet and Chen *et al*. formulations. In this line of research, the concept of *anchor* will still be disregarded and all images in a triplet will regard different classes (IDs), with the margins imposed according to the soft biometrics similarity between pairs of elements. Also, two other possibilities are: 1) to differently weight the contribution of each soft label in defining the embedding topology; and 2) to consider the conceptual distance inside each label (e.g., 'young' is closer to 'adult' than to 'senior'). Both possibilities should also improve the overall ID+soft biometrics labelling performance.

Acknowledgement

The authors would like to thank support of NVIDIA Corporation®, with the donation of one *Titan X* GPU board.

References

[1] N. Y. Almudhahka, M. S. Nixon, and J. S. Hare, "Automatic semantic face recognition," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 180–185, doi: 10.1109/fg.2017.31.

[2] E. Bekele, C. Narber, and W. Lawson, "Multi-attribute residual network (MAResNet) for soft-biometrics recognition in surveillance scenarios," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 386–393, doi: 10.1109/fg.2017.55.

[3] E. B. Cipcigan and M. S. Nixon, "Feature selection for subject ranking using soft biometric queries," in *Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Nov. 2018, pp. 1–6, doi: 10.1109/avss.2018.8639319.

[4] J.-C. Chen, R. Ranjan, A. Kumar, C.-H. Chen, V. M. Patel, and R. Chellappa, "An end-to-end system for unconstrained face verification with deep convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 118–126, doi: 10.1109/iccvw.2015.55.

[5] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 403–412, doi: 10.1109/cvpr.2017.145.

[6] V. Choutas, P. Weinzaepfel, J. Revaud, and C. Schmid, "PoTion: Pose MoTion representation for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7024–7033, doi: 10.1109/cvpr.2018.00734.

[7] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," in *Proc. ACM Int. Conf. Multimedia MM*, 2014, pp. 789–792, doi: 10.1145/2647868.2654966.

[8] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," 2018, *arXiv:1801.07698*. [Online]. Available: http://arxiv.org/abs/1801.07698

[9] Y. Duan, J. Lu and J. Zhou, "UniformFace: Learning deep equidistributed representation for face recognition," 2019, *arXiv:1801.07698*. [Online]. Available: https://arxiv.org/abs/1801.07698

[10] H. Galiyawala, K. Shah, V. Gajjar, and M. S. Raval, "Person retrieval in surveillance video using height, color and gender," 2018, *arXiv:1810.05080*. [Online]. Available: http://arxiv.org/abs/1810.05080

[11] G. Jacob, R. Sam, H. Geoff, and S. Ruslan, "Neighbourhood Components Analysis," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 17, 2004, pp. 513–520, doi: 10.5555/2976040.2976105

[12] B. H. Guo, M. S. Nixon, and J. N. Carter, "Fusion analysis of soft biometrics for recognition at a distance," in *Proc. IEEE 4th Int. Conf. Identity, Secur., Behav. Anal. (ISBA)*, Jan. 2018, pp. 1–8, doi: 10.1109/isba.2018.8311457.

[13] B. H. Guo, M. S. Nixon, and J. N. Carter, "A joint density based rank-score fusion for soft biometric recognition at a distance," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, p. 3457, doi: 10.1109/icpr.2018.8546071.

[14] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 1735–1742., doi: 10.1109/cvpr.2006.100.

[15] M. Halstead, S. Denman, C. Fookes, Y. Tian, and M. S. Nixon, "Semantic person retrieval in surveillance using soft biometrics: AVSS 2018 challenge II," in *Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Nov. 2018, pp. 1–6, doi: 10.1109/avss.2018.8639379.

[16] E. Learned-Miller, G. Huang, A. RoyChowdhury, H. Li, and G. Hua, "Labeled faces in the wild: A survey," in *Proc. Adv. Face Detection Facial Image Anal.*, M. Kawulok, M. E. Celebi, and B. Smolka, Eds. New York, NY, USA: Springer, 2016, pp. 189–248, doi: 10.1007/978-3-319-25958-1_8.

[17] K. He, Z. Wang, Y. Fu, R. Feng, Y.-G. Jiang, and X. Xue, "Adaptively weighted multi-task deep network for person attribute classification," in *Proc. ACM Multimedia Conf. MM*, 2017, pp. 1636–1644, doi: 10.1145/3123266.3123424.

[18] Y. Hu, X. Wu, and R. He, "Attention-set based metric learning for video face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2827–2840, Nov. 2018.

[19] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[20] M. Jiang, Z. Yang, W. Liu, and X. Liu, "Additive margin softmax with center loss for face recognition," in *Proc. the 2nd Int. Conf. Video Image Process. ICVIP*, 2018, pp. 1–8, doi: 10.1145/3301506.3301511.

[21] B.-N. Kang, Y. Kim, and D. Kim, "Deep convolution neural network with stacks of multi-scale convolutional layer block using triplet of faces for face recognition in the wild," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2016, Art. no. 004460, doi: 10.1109/smc.2016.7844934.

[22] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The MegaFace benchmark: 1 million faces for recognition at scale," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4873–4882, doi: 10.1109/cvpr.2016.527.

[23] B. F. Klare *et al.*, "Pushing the frontiers of unconstrained face detection and recognition: IARPA janus benchmark a," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1931–1939, doi: 10.1109/cvpr.2015.7298803.

[24] F. Lateef and Y. Ruichek, "Survey on semantic segmentation using deep learning techniques," *Neurocomputing*, vol. 338, pp. 321–348, Apr. 2019.

[25] W. Liu *et al.*, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37, doi: 10.1007/978-3-319-46448-0_2

[26] T. J. Neal and D. L. Woodard, "You are not acting like yourself: A study on soft biometric classification, person identification, and mobile device use," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 1, no. 2, pp. 109–122, Apr. 2019.

[27] D. Li, Z. Zhang, X. Chen, and K. Huang, "A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1575–1590, Apr. 2019.

[28] H. Liu and W. Huang, "Body structure based triplet convolutional neural network for person re-identification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 1772–1776, doi: 10.1109/icassp.2017.7952461.

[29] D. Martinho-Corbishley, M. S. Nixon, and J. N. Carter, "Super-fine attributes with crowd prototyping," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 6, pp. 1486–1500, Jun. 2019.

[30] R. Ranjan *et al.*, "Crystal loss and quality pooling for unconstrained face verification and recognition," 2018, *arXiv:1804.01159*. [Online]. Available: http://arxiv.org/abs/1804.01159

[31] R. Vera-Rodriguez, P. Marin-Belinchon, E. Gonzalez-Sosa, P. Tome, and J. Ortega-Garcia, "Exploring automatic extraction of body-based soft biometrics," in *Proc. Int. Carnahan Conf. Secur. Technol. (ICCST)*, Oct. 2017, pp. 1–6, doi: 10.1109/ccst.2017.8167841.

[32] P. Samangouei and R. Chellappa, "Convolutional neural networks for attribute-based active authentication on mobile devices," in *Proc. IEEE 8th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Sep. 2016, pp. 1–8, doi: 10.1109/btas.2016.7791163.

[33] A. Gretton, K. Borgwardt, M. Rasch, B. Schlkopf, and J. Smola, "A kernel method for the two-sample-problem," in *Proc. Adv. Neural Inf. Process. Syst. Conf.*, Vancouver, BC, Canada, 2006, pp. 513–520.

[34] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823, doi: 10.1109/cvpr.2015.7298682.

[35] A. Schumann, A. Specker, and J. Beyerer, "Attribute-based person retrieval and search in video sequences," in *Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Nov. 2018, pp. 1–6, doi: 10.1109/avss.2018.8639114.

[36] H. Shi, X. Zhu, S. Liao, Z. Lei, Y. Yang, and S. Z. Li, "Constrained deep metric learning for person re-identification," 2015, *arXiv:1511.07545*. [Online]. Available: http://arxiv.org/abs/1511.07545

[37] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4004–4012, doi: 10.1109/cvpr.2016.434.

[38] E. Gonzalez-Sosa, J. Fierrez, R. Vera-Rodriguez, and F. Alonso-Fernandez, "Facial soft biometrics for recognition in the wild: Recent works, annotation, and COTS evaluation," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 8, pp. 2001–2014, Aug. 2018.

[39] C. Su, Y. Yan, S. Chen, and H. Wang, "An efficient deep neural networks training framework for robust face recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3800–3804, doi: 10.1109/icip.2017.8296993.

[40] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang, "Joint learning of single-image and cross-image representations for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1288–1296, doi: 10.1109/cvpr.2016.144.

[41] J. Wang, Z. Wang, C. Gao, N. Sang, and R. Huang, "DeepList: Learning deep features with adaptive listwise constraint for person reidentification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 513–524, Mar. 2017.

[42] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. 14th Eur. Conf. Comput. Vis.*, 2016, pp. 499–515, doi: 10.1007/978-3-319-46478-7_31.

[43] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A comprehensive study on center loss for deep face recognition," *Int. J. Comput. Vis.*, vol. 127, nos. 6–7, pp. 668–683, Jun. 2019.

[44] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. Adv. Neural Inf. Process. Syst. Conf.*, Montreal, QC, Canada, 2014, pp. 487–495.

Hugo Proença (Senior Member, IEEE). received the B.Sc., M.Sc., and Ph.D. degrees from the Department of Computer Science, University of Beira Interior, in 2001, 2004, and 2007, respectively. He is currently an Associate Professor with the Department of Computer Science, University of Beira Interior. He has been researching mainly about biometrics and visual-surveillance. He was the Coordinating Editor of the IEEE BIOMETRICS COUNCIL NEWSLETTER and the Area Editor (ocular biometrics) of the IEEE BIOMETRICS COMPENDIUM JOURNAL. He is a member of the Editorial Boards of the *Image and Vision Computing*, IEEE ACCESS, and the *International Journal of Biometrics*. He served as the Guest Editor of special issues of the *Pattern Recognition Letters*, *Image and Vision Computing and Signal*, and *Image and Video Processing Journals*.

Ehsan Yaghoubi, photograph and biography not available at the time of publication.

Pendar Alirezazadeh, photograph and biography not available at the time of publication.

All-in-one "*HairNet*": A Deep Neural Model for Joint Hair Segmentation and Characterization

Diana Borza
Babeş Boylai University,
Cluj-Napoca, Romania, 400000
dianaborza@cs.ubbcluj.ro

Ehsan Yaghoubi
Instituto de Telecomunicações
University of Beira Interior, 6201-001 Covilhã, Portugal
Ehsan.yaghoubi@ubi.pt

João Neves
TomiWorld
3500-106 Viseu, Portugal
JoaoNeves@tomiworld.com

Hugo Proença
Instituto de Telecomunicações
University of Beira Interior, 6201-001 Covilhã, Portugal
hugomcp@di.ubi.pt

Abstract

The hair appearance is among the most valuable soft biometric traits when performing human recognition at-a-distance. Even in degraded data, the hair's appearance is instinctively used by humans to distinguish between individuals. In this paper we propose a multi-task deep neural model capable of segmenting the hair region, while also inferring the hair color, shape and style, all from in-the-wild *images. Our main contributions are two-fold: 1) the design of an all-in-one neural network, based on depth-wise separable convolutions to extract the features; and 2) the use convolutional feature masking layer as an attention mechanism that enforces the analysis only within the 'hair' regions. In a conceptual perspective, the strength of our model is that the segmentation mask is used by the other tasks to perceive - at feature-map level - only the regions relevant to the attribute characterization task. This paradigm allows the network to analyze features from non-rectangular areas of the input data, which is particularly important, considering the irregularity of hair regions. Our experiments showed that the proposed approach reaches a hair segmentation performance comparable to the state-of-the-art, having as main advantage the fact of performing multiple levels of analysis in a* single-shot *paradigm.*

1. Introduction

Visual surveillance has grown astonishingly in the last decade at a worldwide level: more than 350 billion surveillance cameras were reported in 2016 [6]. Despite popular belief, a reliable, fully-automated visual surveillance system has not been yet developed, and state of the art artificial intelligence based models still struggle with high false-positive rates. Standard face biometric measures cannot be properly analyzed in surveillance systems, due to the poor image quality (low resolution, blurred, off-angle and occluded subjects), and soft biometric cues are often used to assist classical recognition systems. Despite this, research on external face features (hair, head and face shape) has been neglected in favor of other features, such as irises, eyes, mouth, etc.). [33] has shown that hair cues (namely, the hair length and color) are amongst the most discriminative soft biometric labels when dealing with person recognition at a distance. Moreover, neuroscience research studies confirm this hypothesis: the human visual system seems to perceive the face holistically [30], with emphasis on the head structure and hair feature, rather than internal cues [29, 31].

Automated hair analysis is undoubtedly a difficult task, as the hair structure, shape and visual appearance largely vary between individuals, as depicted in Fig. 1. Unlike internal face features (e.g., the eyes or mouth), it is hard to establish the appropriate region of interest for hair pixels. It is difficult to define a hair shape as a variety of hairstyles exist; defining hair texture and color is difficult too. Individuals naturally tend to have different hair colors and styles, but some tend to change their hair color and styles that affects the hair properties.

In this paper, we propose an all-in-one convolutional neural network (CNN) designed for complete hair analysis (segmentation, color, shape and hairstyle classification), which uses only depth wise separable convolutions [9], making it suitable for running on devices with limited

978-1-7281-9186-7/20/$31.00 @2020 European Union

Figure 1. Samples that illustrate the complexity of the *in-the-wild* hair analysis. Subjects have varying poses, with hair of varying shapes, densities and colors, often partially occluded and hard to distinguish from the background.

computational resources. The original architectural features of the network are: (1) the use of *convolutional feature masking* layers in order to keep the convolutions "focused" only on the hair pixels and (2) *convolutional feature selection* by using skip layers and feature-map masking. The network operates on images captured in uncontrolled, *in-the-wild* environments; the only constraint imposed on the input is that the head area is detectable by a state of the art face detector [22]. A cohesive perspective on the proposed solution is depicted in Fig. 2.

The remainder of this paper is organized as follows: in Section 2, we discuss the related work, and in Section 3 we detail the network architecture and the learning phase of the proposed method. Section 4 describes our experiments and, finally, the conclusions are given in Section 5.

2. Related Work

Early works on hair segmentation operated mainly on frontal images with relatively simple backgrounds. In [36], the positions of the face and eyes are used to establish the region of interest (ROI) for the hair analysis; next, based on spatial (anthropomorphic proportions) and color information a list of seeds is obtained, and region growing is performed to obtain the hair mask. Therefore, hair segmentation is problematic if the background has a similar texture to the hair area. The method also extracts several properties of the hair (volume, length, dominant color) using classical image processing techniques. The method described in [21] defines the hair ROI starting from the positions of the eyes and mouth. Next, the authors devised a region growing algorithm to distinguish the hair pixels from the skin and background pixels. The region growing algorithm operates on a set of 45 features, which includes color, gradient (Canny magnitude), and frequency descriptors. In [27] a raw localization of the hair area is obtained by fusing frequency and color information (YCrCb color space). [14] segments the hair based on the appearance of the upper hair region. First, this region is extracted using active shape and contour models. Based on the appearance parameters of this region, the entire hair region is extracted at pixel level using texture analysis. [17] operates on video sequences; the head area is computed using face detection and background subtraction. Within this region, a skin segmentation mask is obtained using flood-fill algorithm. Finally, the hair region is estimated as the difference between the head and skin pixels. Similarly, [35] extracts the head from video sequences, and the hair region is segmented through histogram analysis and k-means clustering. The hair length is determined through line scanning. [18] uses learned mixture models of color and location information to infer the hypothesis of the hair, face, and background regions. [34] relies on a *coarse hair probability map*, in which each pixel encodes the probability of belonging to the hair class. The hair segmentation map is inferred through regression techniques by finding pairs of isomorphic manifolds. In [28], the authors apply a shape detector to establish a ROI for the hair area; then, to extract the hair-pixels, they use graph-cuts based on solely color cues in the YCbCr color-space. Finally, k-means is applied as a post-processing step to ensure homogeneity between neighboring hair patches.

In [24], a two-layered hierarchical Markov Random Field (MRF) architecture is proposed for the segmentation and labeling of hair and facial hairstyles. The first layer operates at pixel level, modeling local interactions, while the latter extracts higher level, object information, providing coherent solutions for the segmentation problem. The method was tested on degraded images captured by an outdoor visual surveillance system.

Recently, the problem of hair segmentation was approached from a deep learning perspective ([19, 1, 13]); these methods achieve state of the art performance. The segmentation neural networks begin with "contracting" path, in which a sequence of convolutional layers extract meaningful features, but also reduce the spatial information. Next, a set of deconvolutional layers expands these condensed features into segmentation maps. To preserve high-resolution details, skip-connections are inserted to concatenate feature maps from the beginning of the network (higher level of details) with those from the expanding part of the network (higher semantic level). In [19], the loss function is tuned to preserve the high-frequency information of the hair by adding a term that penalizes the discrepancy between the gradients of the input image and those of the predicted hair mask.

2.1. Multi-task Convolutional Neural Networks

Multi-task learning (MTL) has been successfully used in machine learning as a strategy to improve generalization

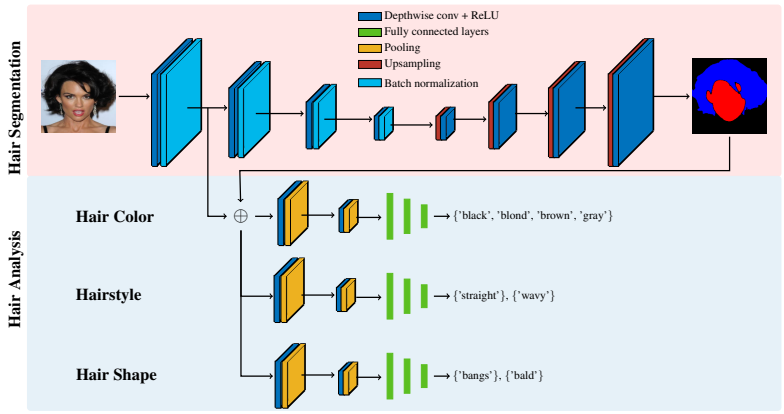


Figure 2. Solution outline: the network comprises several classification branches for the following hair attributes: hair-skin segmentation mask, hair color, shape and style. The segmentation output is used by the other classification branches to select, at feature map level, only the hair pixels via convolutional feature masking.

by learning several classification tasks at once while maintaining a shared representation of the data. A detailed description, including theoretical analysis and applications of multi-task learning, can be found in [2]. One of the pioneering works to performed multi-task facial attribute analysis using a single CNN in an end-to-end manner is [26]. The network simultaneously performs face detection and alignment, pose estimation, gender recognition, smile detection, age estimation, and face recognition. In this framework, the filters in the first convolutional layers of the network are shared between all the classification tasks, constraining a shared representation among the tasks, and reducing the risk of overfitting in these layers. Deep multi-task learning has also been applied for emotion analysis. In [3], the authors propose a deep learning framework for the tasks of facial attribute recognition, action unit detection, and valence-arousal estimation.

2.2. Feature Selection in Convolutional Neural Networks

Deep neural networks achieved state of the art performance on (almost) every field of computer vision and are often used as generic feature extractors. This adaptability of CNNs is also proved by transfer learning: features learned by the network on a (large) database can be successfully applied to other classification tasks. However, classical CNN architectures operate holistically, in the sense that the features are extracted globally, from the entire image, and thus capturing (potentially) irrelevant information. Therefore: How could the network be *guided* to *see* and extract features within some predefined ROI? This question arose for the problem of object detection, in which a bounding box and a class must be inferred for every object in an image. Clearly, the classification part should only analyze the region of interest of the localized object.

The R-CNN (Regions with CNN features) architecture solves this problem in a straightforward manner: a region proposal extraction step is first applied to extract potential objects, and each of these regions is fed to the CNN. Its successor, Fast R-CNN object detector [7], analyzes the entire image to extract a convolutional feature map. Next, each ROI is mapped to this feature map, warped into square regions of predefined size (ROI pooling), and fed to the object classification layer.

Similarly, the SPP-Net architecture [8] introduced the Spatial Pyramid Pooling layer (SPP layer), which masks convolutional feature maps by a rectangular region (i.e., zeros-out the features outside the ROI) and extracts a fixed-length feature vector out of each ROI. In [4], bounding boxes, which can be seen as coarse segmentation masks, are used to "supervise" the training of CNNs for semantic image segmentation. A step forward is taken by [5]; here, input masks of irregular shapes are used to eliminate irrelevant regions of the feature map. The input binary masks are projected into the domain of the convolutional feature maps: each activation is mapped to the input image domain as the center of its receptive field (similar to [8]), and each pixel of the input mask is assigned to the nearest projected receptive field. However, this approach requires an additional step to generate the input masks with the region proposals. In

[5], the proposal regions are extracted by grouping several super-pixels of the input image. In our approach, the masks are segmented directly by the neural network, therefore no pre-processing steps are required.

3. Proposed Method

3.1. Network Architecture

Formally, let X_i denote the feature vector (RGB-pixels) of the i^{th} sample, Y_i the corresponding annotations, and $\hat{Y}_i$ the network's prediction. The data associated to each image (Y_i or $\hat{Y}_i$) comprises the following attributes: $\{M_i, cl_i, st_i, wv_i, bg_i, bd_i\}$, where M_i is the face/hair segmentation mask, cl_i $\in$CL = {'black', 'blond', 'brown', 'gray'} denotes the hair color label, st_i and wv_i are binary values which indicate if the *hairstyle* is 'straight' or 'wavy' respectively. Finally, the values bg_i, bd_i compose the *hair shape* classification branch, indicating whether the person has bangs, or is bald respectively. The output of the network was chosen in accordance with the hair attribute information provided by CelebA database [23], which is, to the best of our knowledge, the largest image dataset providing multiple hair attributes annotations. We chose separate, binary attributes to describe the *hairstyle* (st_i and wv_i) and the *hair shape* (bg_i, bd_i), instead of a single multi-label classification layer, for two main reasons. First of all, not all the samples from the dataset are annotated with this information, or, on the other hand, some samples are annotated with multiple labels from the same logical group. The latter case results in a contradiction with the multi-label classification, which assumes that each example is appointed to *one and only one* label. Secondly, the annotations provided for these attributes are not exhaustive: for example, the *hair shape* analysis could also include one of the following: "long hair", "medium hair", "short hair", etc.

The backbone of the network is inspired by the lightweight MobileNet [9], on top of which we added several classification branches.

3.2. Hair Segmentation

The output of the hair segmentation branch $\hat{M}_i \in \mathbb{R}^{224\times224\times2}$ is a bi-dimensional, two-channel, mask of the same size as the input. The two channels (M_i^0, M_i^1) contain, for each pixel, the probability for belonging to the skin or hair class, respectively. The facial skin area is also segmented, as it provides essential information regarding the hair shape and length: one cannot make any inference about the shape of the hair without correlating its area to the face.

To obtain the segmentation mask, the feature map of the last convolutional layer in the network backbone is fed to a decoder. As suggested in [19], rather than using transposed convolutional layers, the upsampling is accomplished by a $2\times$ upsampling operation, followed by depth-wise and point-wise convolutions. Three such blocks are concatenated to obtain a mask of the same size as the input image. Similar to [19], skip connections to the corresponding, equal-sized layers in the network backbone are added such that the output includes information about the high resolution, but yet weak, features extracted by these layers.

Figure 3. Hair color perception is a contextual phenomenon and cannot be decoupled from the surrounding scene colors and light sources. Also, demographic attributes can influence the hair color estimation process.

Finally, the segmentation output is obtained by adding a 1×1 convolution with two filters (i.e., two output channels: one for the hair and one for the skin pixels) with *softmax* activation.

During training, we aim at minimizing the binary cross entropy loss (1) between the ground truth mask M_i and the predicted segmentation mask $\hat{M}_i$:

$$L_{seg}(M_i, \hat{M}_i) = -(M_i \cdot log(\hat{M}_i) + (1 - M_i) \cdot log(1 - \hat{M}_i)). \quad (1)$$

At test time, the single channel, output mask is obtained by assigning each pixel to the class (hair or skin) with the highest probability, given that it is larger than a threshold t, or to background otherwise:

$$M_{out} = \begin{cases} \arg\max(M^0, M^1) + 1, & \text{if } \max(M^0, M^1) > t \\ 0 \text{ (background)}, & \text{otherwise} \end{cases}, \quad (2)$$

where $t = 0.5$ was used in all our experiments.

3.3. Hair Color Inference

Color perception is a complex process, as the appearance of an object is highly dependent on the environmental context (both spatially and temporally) [12]. It is practically impossible to distinguish the apparent color of a patch, without having additional information regarding the surrounding colors and light sources. In the context of hair color estimation, demographic cues (such as gender and age) are also crucial in deciding the hair tone. An illustrative example is depicted in Figure 3.

Therefore, when deciding on the color tone, the network should use, not only information about the hair tone but also

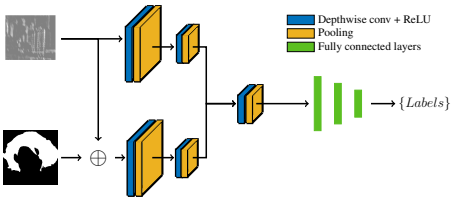


Figure 4. Hair color analysis module. Two separate convolutional branches analyze the image's feature map: the first captures information about the global scene lighting, while the second one focuses only on the hair region using convolutional feature masking.

some cues regarding the surrounding lighting conditions and light sources. With this in mind, the hair color classification task combines two convolutional branches (Fig. 4), which operate on the feature map extracted by the network backbone. The first analyzes the entire feature map, thus extracting information about the overall scene lighting conditions, while the latter masks this feature map using the output of the hair segmentation, to put emphasis solely on the hair features.

Finally, they are merged into a single feature vector FVC, which is flattened and passed to a fully-connected layer with *softmax* activation:

$$sm(FVC_i) = \frac{e^{FVC_i}}{\sum_{j=1}^{K} e^{FVC_j}}, \quad (3)$$

where FCV_i is the feature vector of the i-$^{\text{th}}$ sample.

As mentioned above, the hair color analysis module distinguishes the hair tone into one of the following classes CL = {'black', 'blond', 'brown', 'gray'}.

The loss function to be optimized in this case is the categorical cross-entropy loss:

$$L_{color} = \sum_{i} \sum_{j=1}^{|CL|} -cl_{ij} \cdot log(\widehat{cl_{ij}}), \quad (4)$$

where CL is the number of hair labels, cl_i is the one-hot encoding of the ground truth hair color, and $\hat{cl}_i$ are the predicted class probabilities.

3.4. Hairstyle Inference

The hairstyle analysis module comprises two separate binary classification layers, specialized for the 'wavy' or 'straight' structures respectively.

To decide on these tasks, the network should only analyze the hair pixels. Therefore, the input of each classification branch consists of a feature map extracted from the network backbone, masked with the hair segmentation mask, such that only the deemed hair regions are considered. Let FM be the feature map extracted from the network backbone and HS the binarized hair segmentation map. The input I of each of these branches is given by:

$$I = FM \ \Theta \ HS, \quad (5)$$

where Θ is the feature map masking operator as defined in Section 2.2. This input is passed to 2 convolutional layers, flattened and then fed to a fully convolutional classification layers. As we are dealing with binary attributes, the activation function for the output neurons O_b is the *sigmoid* function:

$$P(O_b) = \frac{1}{1 + e^{-O_b}}. \quad (6)$$

The loss function of these layers is the binary cross-entropy loss function:

$$L_a(a, \hat{a}) = -\frac{1}{N} \sum (a \cdot log(\hat{a}) + (1 - \hat{a}) \cdot log(1 - \hat{a})), \quad (7)$$

where $a = 1$ if the hair has the attribute and $a = 0$ otherwise; $\hat{a}$ is the predicted probability for the hair attribute.

3.5. Hair Shape Inference

The hair shape analysis task consists of two classification branches, each having a binary outcome: *Bangs* and *Bald*.

Intuitively, a piece of essential information in inferring these shape characteristics is the relationship between the face area and the hair area. Therefore, when applying the convolutional feature masking operation, we keep the hair pixels, as well as the facial skin pixels to better capture this relationship.

As the predictions are binary values, the activation and loss functions for the hair shape classification layers are identical to the ones used for hairstyle classification (Section 3.4).

4. Experiments and Discussion

4.1. Datasets and Experimental Setup

The main dataset used to train and validate the proposed model was CelebAMask-HQ [15], a subset of CelebA database [23]. CelebA [23] is suitable for training our model as it contains more than 200k images captured in real-world scenarios (blurred, occluded subjects and with large pose variations); in addition, each image is labeled with 40 binary attributes, including information about the hair color attributes {'black', 'blond', 'brown', 'gray'}, hairstyle attributes {'straight', 'wavy'} and shape attributes {'bangs', 'bald'}.

For the segmentation task, we used CelebAMask-HQ [15] which contains 30k images, selected from CelebA, together with manually annotated masks of face components

(skin, nose, eyes, eyebrows, ears, mouth, lip, hair, hat) and other accessories (eyeglass, earring, necklace, neck, and cloth).

In addition, to demonstrate the generalization ability of the proposed method, we also tested the segmentation module on three additional databases: (a) Labeled Parts in the Wild [11], (b) Figaro-1k [32] and (c) another subset of CelebA, independently annotated by [1]. Images from these datasets were not used at all in the training part. Labeled Parts in the Wild [10] (the *funnelled* version) is a subset of Label Faces in the Wild (LFW) [11] database; it contains 2927 face images segmented into hair/skin/background labels. The segmentation is performed at a coarse level: first the images are divided into super-pixels, and then each super-pixel is manually assigned to a label. Figaro-1k [32] contains 1050 images annotated with hair masks, gathered from the Internet, for the purpose of hair analysis in the wild.

4.2. Learning and Parameter Tuning

As annotated data (with hair masks and hair attributes) is limited, we used transfer learning to make sure that the network won't overfit the training data. So, instead of randomly initializing the weights of the neural network, the training starts from some weight values computed on a different task, for which larger datasets are available; this assumes that the low-level features extracted (edges, textures, gradients, etc.) are relevant across tasks. Therefore, the backbone of the network and the segmentation branch is first trained to segment objects from the COCO dataset [20]. COCO is a large scale image database, which comprises approximately 330K images, designed for object detection and segmentation. The dataset comprises more than 1.5 million object instances, captured in real-world scenarios, grouped into 80 object categories, thus providing enough generalization and data variance.

Next, we conduct the following training scheme:

1. Train the hair segmentation branch on CelebAMask-HQ dataset using the loss function described in L_{seg} (1). The segmentation branch is first trained, as the attribute classification problems use the segmentation mask to establish the ROIs (in the convolutional feature masking layers). Having a good estimate of the hair and face region would greatly speed-up the training process.

2. Freeze the shared layers of the network backbone and individually train all the hair analysis branches using their corresponding loss functions.

3. Finally, the neural network is trained on all the tasks, in an end-to-end manner, such that the common knowledge (filter values) is shared across all the classification problems. At this stage, the individual loss functions are combined into a weighted average as described in equation (8):

$$L = \sum_{i=0}^{T} \lambda_i \cdot L_i, \tag{8}$$

where T is the total number of tasks, $L_i \in \{L_{color}, L_{seg}, L_{straight}, L_{wavy}\}$ and λ_i are the loss value and weight for task i.

In all cases, the weights are optimized using Adam [16] optimizer. The initial learning rate α is set to $\alpha = 0.0001$ when training individually the classification branches, and decreased to $\alpha = 0.00001$ for the final, end-to-end training; in all cases, the exponential decay rate for the first moment estimates β_1 is set to 0.9, and the exponential decay rate for the second-moment estimates β_2 is fixed at 0.99.

4.3. Results

4.3.1 Hair Segmentation

Let n_{cl} be the number of segmentation classes ($n_{cl} = 2$ in our case), n_{ij} be the number of pixels belonging to class i but predicted to class j, and t_i the number of pixels in the ground truth annotation belonging to class i. For the numerical evaluation of the proposed method, we report the mean Intersection over Union ($mIOU$) and the mean pixel accuracy ($mAcc$), as defined in equations (9) and (10).

$$mIoU = \frac{1}{n_{cl}} \sum_i \frac{n_{ii}}{t_i + \sum_j n_{ij} - n_{ii}}. \tag{9}$$

The $mAcc$ metric defines the percentage of correctly classified pixels of a class, averaged over all the segmentation classes.

$$mAcc = \frac{1}{n_{cl}} \frac{\sum_i n_{ii}}{\sum_i t_i}. \tag{10}$$

A fraction of 3000 images (10%) of the CelebAHQ-Mask dataset, which were not used in the training process, are used to validate the proposed approach. The results of the proposed method compared to other state of the art works are reported in Table 1. The results are discussed in Section 4.3.1, with some of the predictions of the proposed method depicted in Fig. 5.

Baseline methods Table 1 displays the hair segmentation performance on CelebAHQ-Mask, Labeled Parts and Figaro-1k databases, compared to other state-of-the-art methods based on deep learning frameworks. In Table 1, CelebA* refers to the subset of CelebA dataset annotated by [1].

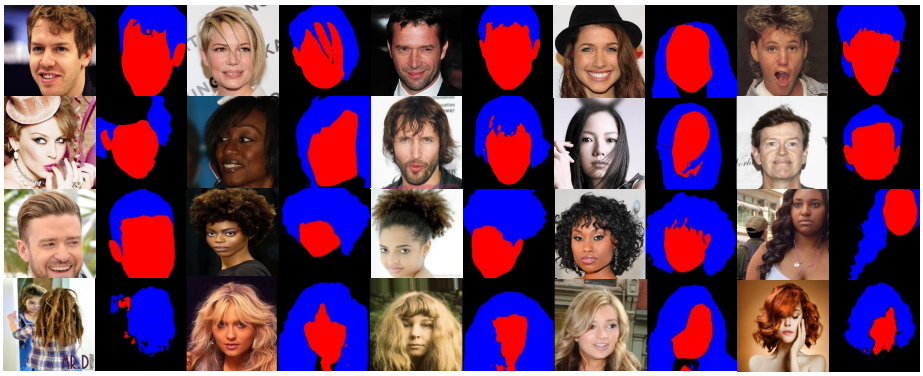


Figure 5. Segmentation masks obtained by the proposed solution on different datasets. The predicted hair pixels are depicted in blue, skin pixels appear in red and background pixels in black. Last row: some failure cases.

Table 1. Comparison of hair segmentation performance with respect to the state-of-the-art.

Method	Database	Pixel accuracy	IoU
[19]	LFW	**97.69**	NA
[1]	LFW	97.01	0.871
[25]	LFW	97.32	NA
Proposed	LFW	95.30	0.864
[1]	CelebA*	97.06	0.920
Proposed	CelebA*	**97.55**	0.881
Proposed	CelebA-MaskHQ	**98.79**	0.939
[1]	Figaro-1k	90.28	0.778
Proposed	Figaro-1k	**97.61**	0.903

Overall, the proposed method achieves high performance for the task of hair segmentation, even if it is surpassed by the other methods on the LFW dataset. In our view, this was due to the fact that most of these methods are intended for various fashion, visagisme or hair coloring applications, in which the hair shape needs to be accurately captured by the segmentation mask. [19] uses a secondary loss function besides binary cross-entropy to obtain accurate segmentation masks from coarse annotation data. This loss function enforces the consistency between the input image and the predicted mask edges. In [1] a more complex (VGG-16) fully convolutional neural networks, while [25] combines fully convolutional neural networks with conditional random fields to obtain an accurate hair matting result. Also, the lower performance in LFW might be due to the proposed method hasn't been trained on the LFW parts dataset and the segmentation masks provided by this database are quite different from the ones of CelebA-MaskHQ. First of all, they are provided at super-pixel level, so are not accurate enough

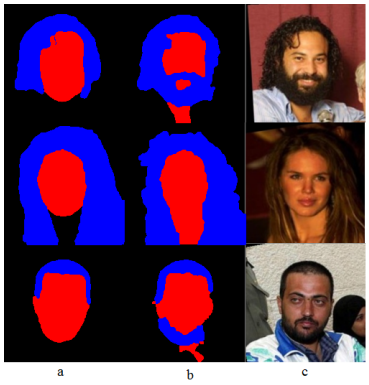


Figure 6. Examples of predicted segmentation masks (LFW dataset): **a)** predicted; **b)** ground truth; **c)** input image.

for high-accuracy evaluation. In addition, as opposed to CelebA-MaskHQ, the hair class also includes facial hair (moustache and beard), while the skin class comprises the neck area. Fig. 6 displays some ground truth segmentation masks versus predicted masks on the LFW dataset.

The proposed method is not intended for virtual try-on applications, where highly accurate hair segmentation masks are required, but for soft biometrics analysis in visual surveillance systems. Therefore, we are not interested in perfectly segmenting all the hair strands or contour details. Moreover, as discussed in the introductory section,

Table 2. Hair attributes classification performance of the proposed method

Metric	Feature masking	Hair color	Hairstyle		Hair shape	
			'wavy'	'straight'	'bangs'	'bald'
Accuracy	✗	88.16	93.20	92.10	92.71	**98.40**
Precision	✗	88.13	94.26	92.87	96.32	97.45
Recall	✗	88.16	92.00	91.2	88.82	99.40
F1 Score	✗	88.01	93.11	92.02	92.41	98.41
Accuracy	✓	**93.45**	**94.30**	**94.60**	**94.41**	98.10
Precision	✓	93.50	95.48	95.88	97.23	97.23
Recall	✓	93.45	93.00	93.20	91.41	99.00
F1 Score	✓	93.43	94.22	94.52	94.23	98.12

images captured by security cameras are often low resolution and blurred, and these hair details would be impossible to distinguish. Even so, from Figure 5 it can be observed that the proposed network is capable of capturing the overall hair shape by accurately segmenting larger strands of hair covering the face or bangs.

4.3.2 Hair Attributes Inference

To evaluate the classification branches, we randomly selected test images from the CelebA dataset (which are not a part of CelebA-MaskHQ) such that the number of samples in each class is the same. The standard metrics: acc - accuracy, pr - precision, rec -recall and F_1 - F1 score are used to numerically express the performance of the proposed solution. Table 2 summarizes the performance of our network in hair attribute characterization, with and without using convolutional feature masking (to prove the efficiency of the proposed convolutional feature masking layer). In the latter case, the network was trained as described in Section 4.2, but the input masks of the hair segmentation module are set to 1, such that the entire image is analyzed for classifying the hair shape.

For each hairstyle and shape classes we randomly selected 1,000 images from the CelebA dataset that are not part of CelebA-HQ. Our experiments showed that, except for the bald detection task, the convolutional feature masking resulted in an increase of the classification performance. For the bald attribute, the accuracy values between the masked and unmasked implementations are comparable (a difference of only 0.3%).

The hair color analysis branch was evaluated on 6,000 images (1,500 samples for each color class) randomly selected from the CelebA dataset. The proposed method uses *softmax* as a final classification layer for predicting the hair color, and considers the class with the highest probability as the hair color prediction. However, some images from the CelebA dataset are not labeled with any of the hair color attributes, or, on the other hand, are labeled with multiple colors (e.g., 'blond' and 'gray'). To be fair in the comparison, both for training and for testing, we randomly selected solely images that contain one and only one annotation of the hair color classes. Overall, the majority of confusions are between the 'brown'/'blonde', 'brown'/'gray' and 'blonde'/'gray' labels. In our view, this was mostly due to the subjective perception of hair color, with light brown/dark-blonde colors being easily mistaken with blond/light blonde color when performing the manual annotation of ground truth data.

The inference step (hair segmentation and hair attribute classification), takes, on average 350 milliseconds on an third generation iPad Pro device.

5. Conclusions

This paper described an *all-in-one* model for hair segmentation and attribute analysis, able to jointly extract the hair-facial skin segmentation mask while also inferring information about the hair color, shape and style. Also, as the proposed architecture uses only depth-wise separable convolutions, it is straightforward to running it in real time, even on devices with limited computational power (e.g., smartphones). To limit the influence of background and irrelevant features on the prediction of the network, an attention mechanism based on convolutional feature masking layers is proposed. Therefore, in our architecture, the inferred segmentation masks are used by the classification branches to determine, at the feature map level, any irregular shaped patches that might correspond to the hair pixels, which enables it to ignore the remaining regions that are deemed as irrelevant to the analysis problem. This feature masking strategy is preferred over traditional ROI-Pooling layers, as if we try to enclose the hair area into a rectangle, a large portion of that patch will be "filled" by the face area, which introduces irrelevant (but salient) features to the analysis problem.

Our experiments were performed in challenging *in the wild* datasets (CelebA, LFW and Fiagro-1k), obtaining high performance (similar or higher than the state of the art), at a lower computational cost.

References

[1] D. Borza, T. Ileni, and A. Darabant. A deep learning approach to hair segmentation and color extraction from facial images. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 438–449. Springer, 2018.

[2] R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, Jul 1997.

[3] W.-Y. Chang, S.-H. Hsu, and J.-H. Chien. Fatauva-net: An integrated deep learning framework for facial attribute recognition, action unit detection, and valence-arousal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 17–25, 2017.

[4] J. Dai, K. He, and J. Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1635–1643, 2015.

[5] J. Dai, K. He, and J. Sun. Convolutional feature masking for joint object and stuff segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3992–4000, 2015.

[6] S. Feldstein. The global expansion of ai surveillance. *Carnegie Endowment. https://carnegieendowment. org/2019/09/17/global-expansion-of-ai-surveillance-pub-79847*, 2019.

[7] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015.

[9] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[10] G. B. Huang, V. Jain, and E. Learned-Miller. Unsupervised joint alignment of complex images. In *ICCV*, 2007.

[11] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[12] A. Hurlbert and Y. Ling. Understanding colour perception and preference. In *Colour Design*, pages 169–192. Elsevier, 2017.

[13] T. Ileni, D. Borza, and A. Darabant. Fast in-the-wild hair segmentation and color classification. In *14th International Conference on Computer Vision Theory and Applications*, pages 59–66, 2019.

[14] P. Julian, C. Dehais, F. Lauze, V. Charvillat, A. Bartoli, and A. Choukroun. Automatic hair detection in the wild. In *2010 20th International Conference on Pattern Recognition*, pages 4617–4620. IEEE, 2010.

[15] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[16] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] A. Krupka, J. Prinosil, K. Riha, J. Minar, and M. Dutta. Hair segmentation for color estimation in surveillance systems. In *Proc. 6th Int. Conf. Adv. Multimedia*, pages 102–107, 2014.

[18] K.-c. Lee, D. Anguelov, B. Sumengen, and S. B. Gokturk. Markov random field models for hair and face segmentation. In *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–6. IEEE, 2008.

[19] A. Levinshtein, C. Chang, E. Phung, I. Kezele, W. Guo, and P. Aarabi. Real-time deep hair matting on mobile devices. In *2018 15th Conference on Computer and Robot Vision (CRV)*, pages 1–7. IEEE, 2018.

[20] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[21] U. Lipowezky, O. Mamo, and A. Cohen. Using integrated color and texture features for automatic hair detection. In *2008 IEEE 25th Convention of Electrical and Electronics Engineers in Israel*, pages 051–055. IEEE, 2008.

[22] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[23] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[24] H. Proenca and J. C. Neves. Soft biometrics: Globally coherent solutions for hair segmentation and style recognition based on hierarchical mrfs. *IEEE Transactions on Information Forensics and Security*, 12(7):1637–1645, 2017.

[25] S. Qin, S. Kim, and R. Manduchi. Automatic skin and hair masking using fully convolutional networks. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 103–108. IEEE, 2017.

[26] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa. An all-in-one convolutional neural network for face analysis. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 17–24. IEEE, 2017.

[27] C. Rousset and P.-Y. Coulon. Frequential and color analysis for hair mask segmentation. In *2008 15th IEEE International Conference on Image Processing*, pages 2276–2279. IEEE, 2008.

[28] Y. Shen, Z. Peng, and Y. Zhang. Image based hair segmentation algorithm for the application of automatic facial caricature synthesis. *The Scientific World Journal*, 2014, 2014.

[29] P. Sinha. Last but not least. *Perception*, 29(8):1005–1008, 2000.

[30] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russell. Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE*, 94(11):1948–1962, 2006.

[31] P. Sinha and T. Poggio. 'united'we stand. *Perception*, 31(1):133, 2002.

[32] M. Svanera, U. R. Muhammad, R. Leonardi, and S. Benini. Figaro, hair detection and segmentation in the wild. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 933–937. IEEE, 2016.

[33] P. Tome, J. Fierrez, R. Vera-Rodriguez, and M. S. Nixon. Soft biometrics and their application in person recognition at a distance. *IEEE Transactions on information forensics and security*, 9(3):464–475, 2014.

[34] D. Wang, S. Shan, H. Zhang, W. Zeng, and X. Chen. Isomorphic manifold inference for hair segmentation. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2013.

[35] Y. Wang, Z. Zhou, E. K. Teoh, and B. Su. Human hair segmentation and length detection for human appearance model. In *2014 22nd International Conference on Pattern Recognition*, pages 450–454. IEEE, 2014.

[36] Y. Yacoob and L. S. Davis. Detection and analysis of hair. *IEEE transactions on pattern analysis and machine intelligence*, 28(7):1164–1169, 2006.

Pose Switch-based Convolutional Neural Network for Clothing Analysis in Visual Surveillance Environment

Pendar Alirezazadeh[1], Ehsan Yaghoubi[2], Eduardo Assunção[3],João C. Neves[4], Hugo Proença[5]
IT-Instituto de Telecomunicações[1,2,3,5], Tomi World[4], Portugal
(Pendar.Alirezazadeh[1], Ehsan.Yaghoubi[2], Eduardo.Assuncao[3])@ubi.pt, Joaoneves@tomiworld.com[4], Hugomcp@di.ubi.pt[5]

***Abstract*—Recognizing pedestrian clothing types and styles in outdoor scenes and totally uncontrolled conditions is appealing to emerging applications such as security, intelligent customer profile analysis and computer-aided fashion design. Recognition of clothing categories from videos remains a challenge, mainly due to the poor data resolution and the data covariates that compromise the effectiveness of automated image analysis techniques (e.g., poses, shadows and partial occlusions). While state-of-the-art methods typically analyze clothing attributes without paying attention to variation of human poses, here we claim for the importance of a feature representation derived from human poses to improve classification rate. Estimating the pose of pedestrians is important to fed guided features into recognizing system. In this paper, we introduce pose switch-based convolutional neural network for recognizing the types of clothes of pedestrians, using data acquired in crowded urban environments. In particular, we compare the effectiveness attained when using CNNs without respect to *human poses* variant, and assess the improvements in performance attained by pose feature extraction. The observed results enable us to conclude that pose information can improve the performance of clothing recognition system. We focus on the key role of pose information in pedestrian clothing analysis, which can be employed as an interesting topic for further works.**

***Index Terms*—Soft biometrics, pedestrian clothing analysis, surveillance environment, human pose classification.**

I. Introduction

The analysis of the pedestrian appearance, and more specifically clothing analysis, has gained interest in machine learning technologies in order to increase accuracy of surveillance based recognition systems. Clothing is one of the most important soft biometrics to pedestrian analysis and has many different applications, such as clothing retrieval [1], [2], clothing recognition [3], [4], outfit recommendation [5] and visual search for matching fashion items [6]. Despite several works proposed in clothing analysis, clothing recognition can't be considered a solved task, especially for surveillance-based environment, that typically produce poor quality data. A good clothing recognition system is highly dependent on the training phase. If these systems are trained with images in controlled conditions, they will not achieve high performance in the real world with various clothing appearance, styles and poses.

One of the major problems in the analysis of clothing is the lack of comprehensive dataset with enough images. Recently two datasets have been published. The MVC Dataset [7] for view-invariant clothing retrieval with 161,638 images and the DeepFashion Dataset [4] with 800,000 annotated real-life images. Both datasets are image-based dataset. Nowadays with cities getting bigger and increasing the use of city-level scenes, researchers have shown an increased interest in clothing analysis of pedestrians which are captured by cameras in streets [8], [9].

To perform clothing analysis in surveillance environment with uncontrolled conditions, we collected a dataset composed of video-based images from outdoor and indoor advertisement panels in Portugal and Brazil. On the other hand, clothing attribute analysis is highly dependent on deformation and poses variation of the human body. By moving some parts of the body such as the knee, hip, neck, shoulder etc. in various gestures, different types of clothing may look like each other, which causes the similarity of the extracted feature vectors and decreases the classification rate. In order to have the ability to clothing recognition in the real application, in this paper, we consider switching CNN architecture that passes frames from a video within a surveillance environment on related Pose-CNN based on a pose-switch classifier. The related Pose-CNN is chosen based on pose information extracted from the video frames as in multi-column Pose-CNN networks to augment the ability to confront pose variations. A particular Pose-CNN is trained on a video frame if the performance of the network on the frame's pose is the best. Fig. 1 illustrates the architecture of our proposed approach.

II. Pose identification

The pose identification aims to explore the human pose group, to assist convolutional neural network for better pedestrian clothing recognition. The output of pose identification

This research is funded by the "FEDER, Fundo de Coesão e Fundo Social Europeu" under the "PT2020 - Portugal 2020" program, "IT: Instituto de Telecomunicações" and "TOMI: City's Best Friend". Also, the work is funded by FCT/MEC through national funds and when applicable co-funded by FEDER PT2020 partnership agreement under the project UID/EEA/50008/2019.

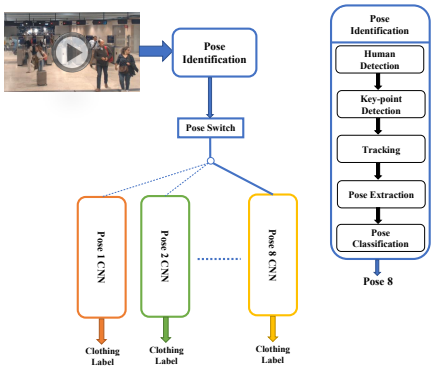


Fig. 1. Architecture of the proposed method, Pose Switch-CNN is shown. Video frames from the surveillance environment are relayed to one of the eight CNN networks based on the pose label inferred from pose identification.

is a pose number based on feature vector including a set of coordinates to describe the pose of the person. It consists of two main steps, including estimates human poses and classifies poses to select the appropriate network.

A. Human pose estimation

Human pose estimation also known as key-point detection, aims to detect the locations of *K* key-points or part of the body e.g. R-hip, L-hip, R-shoulder, L-shoulder etc. from bounding box images. So we have estimated *K* heatmaps where each heatmap indicates the location confidence of the defined key-point. In order to obtain pedestrian bounding boxes (BBs), we use the effective object detection technique VGG-based SSD 512 as pedestrian detector. Pedestrian BBs are fed into pose estimator and key points are generated automatically. In this paper we use CNN based Single Person Pose Estimator (SPPE) method to estimates poses. SPPE network is designed to train on single person images and it is very sensitive to localization errors [10]. On the other hand, pose information consists of a set of key points that each key point belongs to specific region. To select region of interests which have high quality for SPPE network, we use Spatial Transformer Networks (STN) [11]. The STN has shown excellent performance in modeling the variance of scale and pose for adaptively region localization [12]. The STN performs a 2D pointwise transformation with the affine parameters θ which can be expressed as:

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} \tag{1}$$

where (x_i^t, y_i^t) are the target coordinates of the regular grid in the output feature map and the (x_i^s, y_i^s) are the source coordinates in the input feature map that define the sample points. The output of the SPPE network is a set of 16 key points which are used to pose estimation. After human poses estimation for each BBs, we use pose similarity to track multi-person poses in videos to indicate the same person across different frames. Pose metric similarity is used to eliminate the poses which are too close and too similar to each others. We used intra-frame d_f and inter-frame d_c pose distance metrics to measure the pose similarity between two poses P_1 and P_2 in a frame and two sequential frames [13]:

$$\begin{aligned} &d_f\left(P_1, P_2|\sigma_1, \sigma_2, \lambda\right) = \\ &K_{\text{Sim}}\left(P_1, P_2|\sigma_1\right)^{-1} + \lambda H_{\text{Sim}}\left(P_1, P_2|\sigma_2\right)^{-1}, \\ &d_c\left(P_1, P_2\right) = \sum_{n=1}^{N} \frac{f_2^n}{f_1^n} \end{aligned} \tag{2}$$

where

$$K_{\text{Sim}}\left(P_1, P_2|\sigma_1\right) = \tag{3}$$
$$\begin{cases} \sum_{n=1}^{N} \tanh \frac{c_1^n}{\sigma_1} \cdot \tanh \frac{c_2^n}{\sigma_1} & \text{if } p_2^n \text{ is within } B\left(p_1^n\right) \\ 0; \textit{ otherwise} \end{cases}$$

$$H_{Sim}\left(P_1, P_2|\sigma_2\right) = \sum_{n=1}^{N} \exp\left[-\frac{\left(p_1^n - p_2^n\right)^2}{\sigma_2}\right] \tag{4}$$

where p_1^n and p_2^n are the n^{th} key points of pose P_1 and P_2 in $B\left(p_1^n\right)$ and $B\left(p_2^n\right)$ boxes respectively, *N*=16 is number of body keypoints, f_1^n and f_2^n are feature point extracted from boxes, and σ_1, σ_2 and λ can be determined in a data-driven manner. We have extracted coordinates(x,y) for 16 key-points of the full body for all the images. Then, these 16 coordinates points are concatenated to generate a 32 dimensional body coordinate-features vector for each human BBs.

B. Pose classification

Posed-based features may not necessarily be numerically similar for similar motions [14] and it is an important challenge in pose-based feature applications. One of the practical solutions is finding a suitable pattern that aims to grouping a set of pose-based feature in such a way that features in the

same group are more similar to each other than to those in other groups. For this purpose in this study we have used K-means classification algorithm. In order to raise the accuracy of the K-means, we use T-distributed Stochastic Neighbor Embedding (t-SNE) [15] method before classification. This method is known as a nonlinear dimensionality reduction technique for visualization high-dimensional data in a low-dimensional space of two or three dimensions that similar feature vectors are modeled by nearby points and dissimilar feature vectors are modeled by distant points with high probability. The t-SNE method aims to best capture neighborhood identity by considering the probability that one point is the neighbor of all other points. Conditional neighborhood probability of object x_i with object x_j is defined as:

$$p_{j|i} = \frac{\exp\left(-\left\|x_i - x_j\right\|^2 / 2\tau_i^2\right)}{\sum_{k \neq i} \exp\left(-\left\|x_i - x_k\right\|^2 / 2\tau_i^2\right)}, \tag{5}$$

where τ_i^2 is the variance for the Gaussian distribution centered around x_i. Since p_{ij} is not necessarily equal to p_{ji}, because τ_{ij} is not necessarily equal to τ_{ji}, so joint probabilities p_{ij} is defined by symmetrizing two conditional probabilities as:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}. \tag{6}$$

We have trained K-means with low dimension feature vectors resulted from t-SNE method and classified the body coordinate-features to *K* classes.

III. Results and Discussion

In this section, we briefly introduce the datasets, implementation details and results of the proposed method and comparison methods. The experimental results empirically validate the effectiveness of the proposed method.

A. Dataset

Due to the lack of comprehensive dataset for pedestrian clothing analysis in surveillance environment, we collected Biometria e Deteção de Incidentes (BIODI) dataset. The BIODI dataset collected from 216 videos recorded by 36 advertisement panels in Portugal and Brazil. These videos captured in various indoor and outdoor environments such as roads, beaches, airports, streets and metro stations at different hours of the day, lighting, pose, style and various weathers. In each panel, a camera is placed at a distance of 1.5 meters from the ground. All cameras have the same brand with different adjustments, which lead to videos with different qualities. There was no precondition and all of the videos were recorded in unconstraint environments. The statistics of BIODI dataset are summarized in the Table I. To recognize the enormous upper-body and lower-body clothing items, we have labeled the BBs manually. We generated category list bikini, blouse, coat, hoodie, shirt and t-shirt for the upper-body part and jean, legging, pant and short for the lower-body part. Each image received at most one category label for each part.

TABLE I
STATISTICS OF BIODI DATASET

Factors	Statistics
No. of Videos	216
Length of Videos	7 minutes
Frame rate extraction	7 frames/sec.
No. of Subjects	13876
No. of Bounding Boxes (BBs)	503433
Aspect ratio of BBs (Height/Width)	1.75

To further show the efficacy of our proposed methods, we conducted clothing recognition experiments on the RAP-2.0 [16] dataset and compared our results with the performance of their best method. RAP-2.0 comes from a realistic HighDefinition (1280 $\times$ 720) surveillance network at an indoor shopping mall and all images are captured by 25 cameras scenes. This dataset contains 84928 images (2589 subjects) with resolution ranging from 33 $\times$ 81 to 415 $\times$ 583.

B. Implementation Details

We have adopted *K*=8 typical poses, empirically. We consider a subset of 300,000 images of BIODI as training data and a subset of 100,000 images as validation data. Based on pose identification method, the training and validation data are divided to 8 typical pose groups. Clothes bounding boxes for upper-body and lower-body are detected by use of extracted key points. Time performance of pose identification algorithm for a frame including 20 people is about 0.3 second. To evaluate the performance of our proposed system after pose identification, we adopt end-to-end CNN approaches as clothing recognition. End-to-end deep learning methods have made jointly learn features and classifiers. We use CNNs with same architectures for each pose group. We fine-tune VGG-16 [17] and ResNet50 [18] on training and validation data with weights of ImageNet [19] dataset for each pose group. In testing, we employ the remain part of BIODI to test the fine-tuned models. We ensure that no subject BBs overlaps between fine-tuning and testing sets. The stochastic gradient descent (SGD) is adopted to optimize the networks. For both models, we use the initial learning rate 1×10^{-4} and weight decay with 1×10^{-6}. The models have implemented in Python 3.6.7 using the Keras 2.1.6 deep learning library on top of the Tensorflow 1.10 backend and trained for 100 epochs with one NVIDIA GeForce RTX 2080 Ti GPU.

C. Results

We present an extensive evaluation of our proposed method on upper-body and lower-body clothing recognition. We firstly compare our framework with two baseline models (i.e. VGG-16 and ResNet50 without pose information) on BIODI dataset to validate the effectiveness of Pose Switch-based CNN. Table II, III show the classification accuracy of baseline models on upper-body and lower-body BIODI clothing categories recognition, respectively. From the obtained results, the proposed technique increased the performance of clothing recognition rate on all pose groups of upper-body and lower-body parts.

TABLE II
THE PERFORMANCE OF THE PROPOSED METHOD (%) FOR BIODI UPPER-BODY

Network	Pose 1	Pose 2	Pose 3	Pose 4	Pose 5	Pose 6	Pose 7	Pose 8	Mean Accuracy	without Pose
VGG-16	88.94	88.98	89.52	88.79	88.93	89.59	88.54	88.20	88.93	87.41
ResNet50	88.02	87.43	87.54	87.44	88.13	88.14	87.37	87.14	87.65	86.23

TABLE III
THE PERFORMANCE OF THE PROPOSED METHOD (%) FOR BIODI LOWER-BODY

Network	Pose 1	Pose 2	Pose 3	Pose 4	Pose 5	Pose 6	Pose 7	Pose 8	Mean Accuracy	without Pose
VGG-16	87.5	85.21	87.34	88.13	87.97	86.62	85.42	87.62	86.98	86.15
ResNet50	86.89	85.66	87.03	88.44	88.04	85.68	85.43	87.54	86.83	85.17

In order to visualize performance of the proposed framework which has had better performance compared to the situation without pose information, we have drawn the receiver operating characteristic (ROC) curve per category for the VGG-16 network (Fig.2). We have drawn the ROC curve for coat and blouse classes from upper-body and pants and short classes for lower-body. As it derives from the ROC curves, performance of VGG-16 network is improved using pose information. Secondly, to further show the efficacy of our approach, we conducted clothing recognition experiments on RAP-2.0 dataset and compared our results with the performance of the baseline method which is achieved best recognition rate. Based on the full body's direction, RAP-2.0 images are annotated to four types of viewpoints, including facing front (F), facing back (B), facing left (L) and facing right (R). Due to this background, we have classified images to four typical pose groups. The results of employing VGG-16 network in each of 4 typical pose groups on RAP-2.0 upper-body and lower-body parts are shown in Table IV, V, respectively. It is clear that for all typical pose groups, the recognition rates are improved compared to without pose information, significantly.

TABLE IV
THE PERFORMANCE OF THE PROPOSED METHOD AND DEEPMAR-R (%) FOR RAP-2.0 UPPER-BODY

Network	Pose 1	Pose 2	Pose 3	Pose 4	Mean
VGG-16	82.66	82.89	83.44	83.84	83.20
DeepMAR-R [16]	-	-	-	-	76.68

TABLE V
THE PERFORMANCE OF THE PROPOSED METHOD AND DEEPMAR-R (%) FOR RAP-2.0 LOWER-BODY

Network	Pose 1	Pose 2	Pose 3	Pose 4	Mean
VGG-16	87.13	86.93	87.49	87.06	87.15
DeepMAR-R [16]	-	-	-	-	81.33

IV. CONCLUSION AND FUTURE WORKS

Since surveillance-based images are collected in unconstrained environment with various pose and styles, different types of clothing may look like each other which causes the similarity of the extracted feature vectors and decreases the classification rate. In this paper, we propose pose switch-based convolutional neural network that leverages pose variation to improve the accuracy of the pedestrian clothing recognition in crowded urban environments. The proposed method employs pose estimation techniques to key point detection for coordinate-features representation. We have classified all BBs to eight typical pose groups using these features. The convolutional neural networks are trained for each pose group and recognized upper-body and lower-body clothing images. Extensive experiments on RAP-2 datasets show that our method exhibits state-of-the art performance on major dataset in real surveillance scenarios. In the future, we plan to extend the proposed method to explore more efficient human semantic structure knowledge to assist pedestrian attribute recognition.

REFERENCES

[1] Z. Li, Y. Li, W. Tian, Y. Pang, and Y. Liu, "Cross-scenario clothing retrieval and fine-grained style recognition," in *2016 23rd International Conference on Pattern Recognition (ICPR)*, pp. 2912–2917, IEEE, 2016.

[2] X. Liang, L. Lin, W. Yang, P. Luo, J. Huang, and S. Yan, "Clothes co-parsing via joint image segmentation and labeling with application to clothing retrieval," *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 1175–1186, 2016.

[3] A. Y. Ivanov, G. I. Borzunov, and K. Kogos, "Recognition and identification of the clothes in the photo or video using neural networks," in *2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, pp. 1513–1516, IEEE, 2018.

[4] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1096–1104, 2016.

[5] P. Tangseng, K. Yamaguchi, and T. Okatani, "Recommending outfits from personal closet," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2275–2279, 2017.

[6] J. Lasserre, C. Bracher, and R. Vollgraf, "Street2fashion2shop: Enabling visual search in fashion e-commerce using studio images," in *International Conference on Pattern Recognition Applications and Methods*, pp. 3–26, Springer, 2018.

[7] K.-H. Liu, T.-Y. Chen, and C.-S. Chen, "Mvc: A dataset for view-invariant clothing retrieval and attribute prediction," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pp. 313–316, ACM, 2016.

[8] J. Huang, X. Wu, J. Zhu, and R. He, "Real-time clothing detection with convolutional neural network," in *Recent Developments in Intelligent Computing, Communication and Devices*, pp. 233–239, Springer, 2019.

[9] M. Yang and K. Yu, "Real-time clothing recognition in surveillance videos," in *2011 18th IEEE International Conference on Image Processing*, pp. 2937–2940, IEEE, 2011.

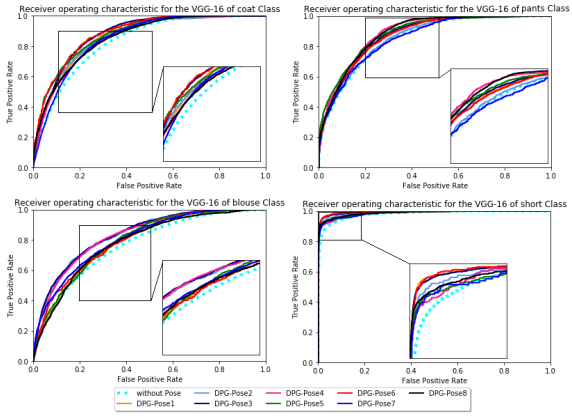


Fig. 2. ROC curves of the VGG-16 for different categories on BIODI upper-body and lower-body.

[10] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2334–2343, 2017.

[11] M. Jaderberg, K. Simonyan, A. Zisserman, *et al.*, "Spatial transformer networks," in *Advances in neural information processing systems*, pp. 2017–2025, 2015.

[12] D. Li, X. Chen, Z. Zhang, and K. Huang, "Pose guided deep model for pedestrian attribute recognition in surveillance scenarios," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, IEEE, 2018.

[13] Y. Xiu, J. Li, H. Wang, Y. Fang, and C. Lu, "Pose flow: Efficient online pose tracking," *arXiv preprint arXiv:1802.00977*, 2018.

[14] A. Yao, J. Gall, G. Fanelli, and L. Van Gool, "Does human action recognition benefit from pose estimation?," in *BMVC 2011-Proceedings of the British Machine Vision Conference 2011*, 2011.

[15] H. Zhou, F. Wang, and P. Tao, "t-distributed stochastic neighbor embedding method with the least information loss for macromolecular simulations," *Journal of chemical theory and computation*, vol. 14, no. 11, pp. 5499–5510, 2018.

[16] D. Li, Z. Zhang, X. Chen, and K. Huang, "A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios," *IEEE transactions on image processing*, vol. 28, no. 4, pp. 1575–1590, 2019.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.