ISTITUTO ITALIANO
DI TECNOLOGIA

UNIVERSITÀ DEGLI STUDI
DI GENOVA

DEPARTMENT OF PATTERN ANALYSIS AND COMPUTER VISION (PAVIS),
ISTITUTO ITALIANO DI TECNOLOGIA

DIPARTIMENTO DI INGEGNERIA NAVALE, ELETTRICA, ELETTRONICA E DELLE
TELECOMUNICAZIONI (DITEN), UNIVERSITÀ DEGLI STUDI DI GENOVA

PhD in Science and Technology for Electronic and Telecommunication Engineering

Curriculum: Computational Vision, Automatic Recognition and Learning

---

Learning Discriminative Features for Person Re-Identification

Xiangping Zhu

A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy

Supervisor: Prof. Dr. Vittorio Murino

Coordinator of the PhD Course: Prof. Dr. Mario Marchese

MARCH 2020 - XXXII CYCLE

Acknowledgments

Time flies and now in finishing my thesis, I am also moving to the end of my Ph.D. career in Pattern Analysis and Computer Vision (PAVIS), Istituto Italiano di Tecnologia (IIT). First of all, I should thank Prof. Vittorio Murino who is my supervisor in these three years. Many thanks to him for accepting me as a Ph.D. student in his PAVIS team and also for his patience and guidances. I should also thank Prof. Shaogang Gong who is my supervisor when I was an intern in his Vision Group, Queen Mary University of London. I experienced six great months in his lab.

I would like to express my thanks to Dr. Mohamed Lamine Mekhalfi, Dr. Paolo Rota, Dr. Pietro Morerio in PAVIS who directly supervised me and gave me detailed guidances on my research in the past three years. When I was an intern in Queen Mary University of London, I worked closely with Dr. Xiatian Zhu and Dr. Minxian Li. They gave me a lot of helps on my research, especially Dr. Xiatian Zhu who spent lots of time discussing with me about my work. In addition, thanks also go to the kind secretary Sara Curreli and the excellent technician Matteo Bustreo for their great supports.

Lastly, I want to express my thanks to my family members, in particular my parents, my aunt, my brothers and my sisters. Their enduring love and endless supports always encouraged me to move forward in this journey.

Abstract

For fulfilling the requirements of public safety in modern cities, more and more large-scale surveillance camera systems are deployed, resulting in an enormous amount of visual data. Automatically processing and interpreting these data promote the development and application of visual data analytic technologies. As one of the important research topics in surveillance systems, person re-identification (re-id) aims at retrieving the target person across non-overlapping camera-views that are implemented in a number of distributed space-time locations. It is a fundamental problem for many practical surveillance applications, *e.g.*, person search, cross-camera tracking, multi-camera human behavior analysis and prediction, and it received considerable attentions nowadays from both academic and industrial domains.

Learning discriminative feature representation is an essential task in person re-id. Although many methodologies have been proposed, discriminative re-id feature extraction is still a challenging problem due to: (1) Intra- and inter-personal variations. The intrinsic properties of the camera deployment in surveillance system lead to various changes in person poses, view-points, illumination conditions *etc*. This may result in the large intra-personal variations and/or small inter-personal variations, thus incurring problems in matching person images. (2) Domain variations. The domain variations between different datasets give rise to the problem of generalization capability of re-id model. Directly applying a re-id model trained on one dataset to another one usually causes a large performance degradation. (3) Difficulties in data creation and annotation. Existing person re-id methods, especially deep re-id methods, rely mostly on a large set of inter-camera identity labelled training data, requiring a tedious data collection and annotation process. This leads to poor scalability in practical person re-id applications.

Corresponding to the challenges in learning discriminative re-id features, this thesis contributes to the re-id domain by proposing three related methodologies and one new re-id setting:

(1) **Gaussian mixture importance estimation.** Handcrafted features are usually not discriminative enough for person re-id because of noisy information, such as background clutters. To precisely evaluate the similarities between person images, the main task of distance metric learning is to filter out the noisy information. Keep It Simple and Straightforward MEtric (KISSME) is an effective method in person re-id. However, it is sensitive to the feature dimensionality and cannot capture the multi-modes in dataset. To this end, a Gaussian Mixture Importance Estimation re-id approach is proposed, which exploits the Gaussian

Mixture Models for estimating the observed commonalities of similar and dissimilar person pairs in the feature space.

(2) **Unsupervised domain-adaptive person re-id based on pedestrian attributes.** In person re-id, person identities are usually not overlapped among different domains (or datasets) and this raises the difficulties in generalizing re-id models. Different from person identity, pedestrian attributes, *e.g.*, hair length, clothes type and color, are consistent across different domains (or datasets). However, most of re-id datasets lack attribute annotations. On the other hand, in the field of pedestrian attribute recognition, there is a number of datasets labeled with attributes. Exploiting such data for re-id purpose can alleviate the shortage of attribute annotations in re-id domain and improve the generalization capability of re-id model. To this end, an unsupervised domain-adaptive re-id feature learning framework is proposed to make full use of attribute annotations. Specifically, an existing unsupervised domain adaptation method has been extended to transfer attribute-based features from attribute recognition domain to the re-id domain. With the proposed re-id feature learning framework, the domain invariant feature representations can be effectively extracted.

(3) **Intra-camera supervised person re-id.** Annotating the large-scale re-id datasets requires a tedious data collection and annotation process and therefore leads to poor scalability in practical person re-id applications. To overcome this fundamental limitation, a new person re-id setting is considered without inter-camera identity association but only with identity labels independently annotated within each camera-view. This eliminates the most time-consuming and tedious inter-camera identity association annotating process and thus significantly reduces the amount of human efforts required during annotation. It hence gives rise to a more scalable and more feasible learning scenario, which is named as Intra-Camera Supervised (ICS) person re-id. Under this ICS setting, a new re-id method, *i.e.*, Multi-tAsk mulTi-labEl (MATE) learning method, is formulated. Given no inter-camera association, MATE is specially designed for self-discovering the inter-camera identity correspondence. This is achieved by inter-camera multi-label learning under a joint multi-task inference framework. In addition, MATE can also efficiently learn the discriminative re-id feature representations using the available identity labels within each camera-view.

Keywords: *Visual Surveillance, Metric Learning, Deep Learning, Unsupervised Domain Adaptation, Intra-Camera Supervised Person Re-Identification*

Publication List

The work of this thesis is based on the following papers, which have been published or submitted:

Chapter 3

1. Xiangping Zhu, Amran Bhuiyan, Mohamed Lamine Mekhalfi and Vittorio Murino, *Exploiting Gaussian Mixture Importance for Person Re-identification*, IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2017. (**Best Paper, Honorable Mention**)

Chapter 4

1. Xiangping Zhu, Pietro Morerio and Vittorio Murino, *Unsupervised Domain-Adaptive Person Re-identification Based on Attributes*, IEEE International Conference on Image Processing (ICIP), 2019.

Chapter 5

1. Xiangping Zhu, Xiatian Zhu, Minxian Li, Vittorio Murino and Shaogang Gong, *Intra-Camera Supervised Person Re-Identification: A New Benchmark*, International Conference on Computer Vision (ICCV) Workshop on Real-World Recognition from Low-Quality Images and Videos, 2019.

2. Xiangping Zhu, Xiatian Zhu, Minxian Li, Pietro Morerio, Vittorio Murino and Shaogang Gong, *Intra-Camera Supervised Person Re-Identification*, Submitted to International Journal of Computer Vision (IJCV).

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Visual Surveillance

More and more large-scale surveillance camera systems are being deployed for increasing the public safety in modern cities. Based on the statistics in [1], the number of closed-circuit television (CCTV) surveillance cameras in the world has been increased from less than 10 millions in 2006 to over 100 millions in 2016, and the number is still rapidly increasing in recent years. This huge amount of surveillance cameras are installed in different public spaces for various surveillance application purposes, for example monitoring traffic flow on highway and detecting abnormal activities in train station or shopping mall. Figure 1.1 gives examples of several application scenarios of CCTV cameras.

Visual surveillance refers to a visual monitoring process that aims at analyzing and interpreting visual data generated by CCTV cameras in order to understand the visual events of the scene [57]. The traditional surveillance visual data processing mostly depends on manual methods. The visual data from the cameras in a surveillance system are pooled in the control room [38]. Several or more operators are involved in data analyzing. Specifically, operators in the control room use a bank of wall monitors to get the quick snapshots of the scenes

(a) (b) (c)

Figure 1.1: Examples of three application scenarios of CCTV cameras. From (a) to (c), cameras are deployed in airport, highway and metro station[1].

[1]The images are from https://thecity.nyc/2019/10/subway-surveillance-cameras-turned-toward-the-homeless.html, http://www.cchargelondon.co.uk/operation.html and https://camscan.ca/airports.php.

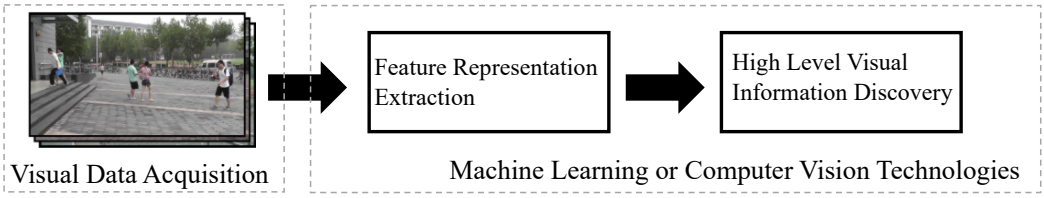


Figure 1.2: The pipeline of visual data processing in autonomous visual surveillance. The visual data is acquired from CCTV cameras. The machine learning or computer vision methods are used to extract feature representations and discover high level visual information, for example activity name in activity recognition and trajectory in pedestrian trajectory prediction.

in the area that is under surveillance. With the expansion of CCTV camera numbers, visual surveillance done solely by human is becoming unfeasible. As one example reported in [38], there is a control room in which one operator was responsible for 153 CCTV cameras, resulting in many cameras not being monitored for long periods of time. As unexpected events (involving specific individuals) often take place in a split second, real-time apprehension of the events may be missed. In addition, CCTV cameras keep recording videos consistently while for operators, continuously watching and analyzing videos is a very labor-intensive task.

Autonomous visual surveillance provides an alternative way to solve the challenges encountered in using manual methods, as mentioned above. It aims at automatically processing and interpreting visual data with the assistance of machine learning or computer vision technologies, for example object or human detection, tracking, action recognition and person re-identification (re-id). The visual data processing pipeline in autonomous visual surveillance is presented in Fig. 1.2. Compared with manual visual surveillance methods, autonomous visual surveillance is characterized with several advantages. For example, it can provide stable and real-time monitoring results. It is generally agreed that for human, long time of continuous watching videos requires a bigger level of visual attention than most every day tasks [47]. This can cause visual fatigue and thus result in errors in observing the incident happened in the monitored areas, which can further lead to unstable monitoring results. Autonomous visual surveillance methods can efficiently solve these problems using automatic image or video processing algorithms.

Due to large demands in practical applications, substantial efforts have been devoted into developing autonomous visual surveillance technologies [47, 145, 172, 192, 7]. Person re-id is one of the fundamental problems in autonomous visual surveillance and it attracts lots of attentions of both academy and industry [192, 7, 117, 132, 78, 175, 105, 186, 20]. In

(a) Camera layout

(b) Person re-id

Figure 1.3: (a) An example of camera layout that is considered in person re-id problem [7]. There are 15 person identities (*i.e.*, , No. 1-15) and they are color encoded. The dashed lines denote the person trajectories; (b) A toy example of person re-id.

the following section, detailed introductions will be sequentially presented about different aspects of person re-id, *i.e.*, problem formulation, applications, challenges, datasets and evaluation metrics.

1.2 Person Re-Identification

1.2.1 Problem Formulation

The origin of person re-id can be dated back to multi-camera pedestrian tracking [192, 145]. In multi-camera surveillance system, it commonly happens that pedestrians leave one camera-view and re-appear in another camera-view. A successful tracking algorithm should be able to associate the same pedestrian appeared in different camera views. This is a non-trivial problem due to the variations of person poses, illuminations, occlusions *etc*. Person re-id independently considers cross-camera person association procedure but under a more challenging scenario with no overlaps between camera views. Fig. 1.3(a) gives an example of the camera layout considered in person re-id. A general assumption in person re-id is that individuals keep the same clothing in different camera views. This is reasonable in most of practical cases. In a surveillance camera network, it only takes a short period of time for pedestrians walking from one camera-view to another one and most probably, pedestrians do not change their clothes. Based on this assumption, most of person re-id works mainly rely on pedestrian appearance as the cue for re-identification.

In practical person re-id, associating pedestrians across non-overlapping camera views is

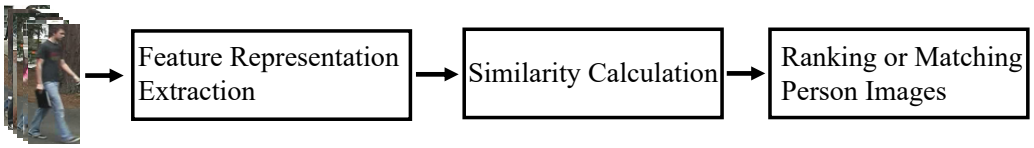


Figure 1.4: The general pipeline of person re-id. Input contains all of person images from both probe and gallery sets. Different methods can be applied in feature representation extraction part, *e.g.*, handcrafted features, metric learning and deep learning. A full version diagram of person re-id system can be found in [7].

converted into a cross-camera image retrieval problem. As in image retrieval, there are also probe and gallery sets in person re-id data. The probe set contains the images of target person while gallery set contains the images of candidates. Fig. 1.3(b) presents a toy example of person re-id. For simplifying the problem, only two cameras are illustrated here. Given a probe image, person re-id retrieves the target person images from the gallery set by ranking the candidate images according to their similarities to the probe. For a successful person re-id method, if there are images of the probe in the gallery set, theses images should be on the top positions in the ranking list.

Generally, most of person re-id methods are composed of: (1) feature representation extraction, (2) similarity calculation and (3) ranking or matching person images. Fig. 1.4 shows the general pipeline of person re-id. Learning discriminative feature representation is the essential task in person re-id. Various methods have been proposed for extracting re-id feature representations, including handcrafted features [78, 31, 25, 37], metric learning [123, 61, 78, 79, 44] and deep learning [186, 23, 200, 134, 181]. According to the existing works [192, 63, 205, 204, 200], after extracting the discriminative re-id features, simple distance functions, *e.g.*, euclidean or cosine distance, can be applied for calculating similarities between probe and gallery images. Although many methodologies have been proposed, discriminative re-id feature extraction is still a hard problem resulted from different challenges in person re-id, as detailed in Section 1.2.3. This thesis mainly focuses on designing discriminative feature extraction methods.

Although there are works *directly* applying distance functions on handcrafted features to calculate similarities without *learning or training* phase [192, 31, 25, 188], most of person re-id methods have two phases: (1) Training. With the given person re-id data, the main task of this phase is to train the re-id model to be capable of extracting discriminative features. There are two major types of training strategies, *i.e.*, verification and identification [197]. Based on person re-id purpose of matching person images, verification fulfills this by applying the idea that samples from the same identity are pulled together while samples from

different identities are pushed away. The methods based on this strategy include distance metric learning [173, 52, 123, 61, 78, 79, 44] and deep metric learning [51, 18, 192, 177]. Some existing works also treat verification strategy as a binary classification task which takes a pair of images as input and predicts if they are from the same person [72, 192]. On the other hand, identification treats person re-id as a multi-class classification task, in which person images are input into the re-id model, which provides predictions for their identities [192, 205, 165, 14, 109]. Recently, there are also works trying to benefit from both verification and identification strategies by combining them together in training re-id model [197, 142, 185, 134, 138]. (2) Test or evaluation. In this procedure, the images from gallery set are ranked according to their similarities to the probe, and evaluation metrics, which will be detailed in Section 1.2.5, are applied to evaluate the performances of the designed person re-id methods.

1.2.2 Applications

In addition to visual surveillance, person re-id also has a wide applications in other areas, for example robotics [39]. Several typical person re-id applications are introduced as follows:

(1) *Person Search*: Given a target person image and the whole scene images, person search tries to find the target person in the scene images [169, 166, 112, 171]. For person re-id, the dataset is created by detecting the pedestrians in the scene images and then the person image bounding boxes are cropped out. The re-id algorithm will be performed on these cropped person images. Person search unifies pedestrian detection and person re-id by directly performing person matching on scene images. In [69], Li *et al*. proposed another person search problem in which the natural language description is used for describing the target person, and this person search aims at searching a person in the database whose attributes is same as or close to the text description. Compared with person re-id, this person search problem has the additional task to encode the text description into feature vector that can be used for matching the person images in the database.

(2) *Pedestrian Tracking*: Visual tracking plays an important role in computer vision. The objects to be tracked can be pedestrians [172, 116], vehicles [60, 8], sport players [95], animals [96, 135] *etc*. Compared with other objects, pedestrian tracking is a more challenging problem both within and across camera views. In addition to the fact that human body is non-rigid that results in pose variations, the occlusions caused by other pedestrians or objects, especially in a crowded scenario, challenges pedestrian tracking even within camera-view. For tracking pedestrian across camera views, the illumination changes from one camera-view to another one, for example from indoor to outdoor environment, introduces new difficul-

ties in tracking. On the other hand, these challenges are fully considered in many existing person re-id algorithms. Based on this observation, several works have started to include person re-id technology in pedestrian tracking [145, 33]. In [145], a person re-id model is included for matching person hypotheses over longer temporal gaps in the proposed tracking framework. In [33], the pedestrian tracking is considered in multiple cameras but with overlapping Field of Views (FOVs). Person re-id is used for solving the problem of identity switches when pedestrians come close to each other. Both of these two introduced works show that pedestrian tracking can benefit from person re-id.

(3) *Vehicle Re-Identification*: Vehicle is a significant object class in urban video surveillance and vehicle detection, re-identification, tracking and classification are attracting more and more attentions. Vehicle re-identification is a relative new research topic in vehicle monitoring. Given a query vehicle image, vehicle re-id, similar with person re-id, is to search in a database for images contained the same vehicle captured by multiple cameras [87, 88, 183, 48, 146, 156]. Thus different from vehicle detection, tracking or classification, vehicle re-id can be regarded as an instance-level object search problem. It is a non-trivial problem due to the large intra-instance differences of the same vehicle in different cameras, and subtle inter-instance differences between different vehicles in the same views. Based on the observations on the similarities between person and vehicle re-id, the existing person re-id methods can be adapted for vehicle re-id, for example handcrafted feature design, siamese network and also performance evaluation metrics [88].

(4) *Human-Machine Interaction*: Robots are more and more involving in our human's daily life. One of the main tasks of mobile service robots is to accurately follow its master or other target person. For example, he/she can be a target customer for the shopping guide robot and a child for the nanny robot. Imagine a scenario where the robot and its master are in a crowded place, for example shopping mall. In this case, from the FOV of the robot, its master can disappear for several seconds due to the occlusions caused by other persons or objects. In order to continuously follow its master, the robot needs to re-identify its master. Re-identifying the target person is exactly the problem considered in person re-id. Thus the algorithms in person re-id can be adapted for service robot to re-identify its master. In addition, person re-id can be also applied for assisting industrial robot to re-identify its cooperator.

(5) *Human Behavior and Activity Analysis*: The task of human behavior and activity analysis is to automatically interpret behavior and activity patterns generated when humans interact with others or with machines. It has a wide applications including surveillance systems, patient monitoring systems, and a variety of systems that involve interactions between persons and electronic devices such as human-computer interfaces [2]. There are still many

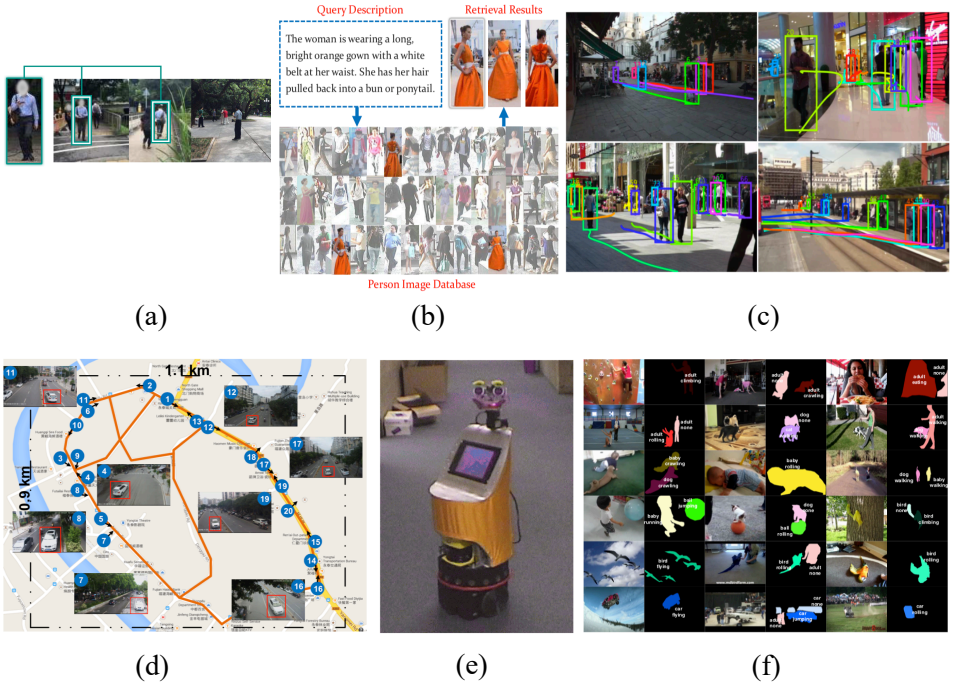


Figure 1.5: Examples of six applications that can benefit from person re-id: (a) Person search from the whole scene images [166], (b) Person search using natural language descriptions [69], (c) Pedestrian tracking [145], (d) Vehicle re-identification [87], (e) Mobile Service robot [6] and (f) Activity analysis [168].

open issues in this research topic, including the joint modeling of behavioral cues taking place at different time scales, the inherent uncertainty of machine detectable evidences of human behavior, the mutual influence of people involved in interactions, the presence of long term dependencies in observations extracted from human behavior, and the important role of dynamics in human behavior understanding [124]. Person re-id can be an assistance technology in human behavior and activity analysis. In work [94, 93], person re-id is applied for modelling correlations between multi-camera activities.

The examples of each described person re-id application are shown in Fig. 1.5.

1.2.3 Challenges

Person re-id is an inherently challenging problem due to the fact that it suffers from not only the challenges, for example domain variations, in generic computer vision and machine learning problems but also some specific challenges, for example intra- and inter-personal variations. In [7], the challenges in person re-id system are summarized in two categories, *i.e.*, system-level challenges and component-level challenges. The component-level chal-

lenges are further divided into descriptor issues and correspondence issues. In this thesis, learning discriminative features for person re-id is considered.

Learning discriminative features is an essential task in person re-id. With discriminative features, simple distance functions, *e.g.*, euclidean and cosine distance, can be applied for calculating the similarities between person images. Although many related methodologies have been proposed, extracting discriminative re-id features still remains a challenging problem because of intra- and inter-personal variations, domain variations and complexities in data creation and annotation.

1. Intra- and Inter-Personal Variations

Fig. 1.3 presents an illustrative example of the camera layout in a surveillance system. From the figure, it can be observed that cameras are dispersedly deployed in an area with no overlaps between camera FOVs. Due to the problems, such as the differences in camera locations and installation angles, the same person captured with different cameras can present with many variations:

(a) *Person poses*: The non-rigid property of human body results in the variations of person poses. This causes the problem in person re-id which is mostly based on the person appearance features. Pose variations destroy the body part alignment in the extracted feature vectors. As shown in Fig. 1.6(a), the same positions (as indicated in the red bounding boxes) between two images do not correspond to the same body part because of the changes in person pose, *i.e.*, the person is walking in one image while changes to ride the bike or motorbike in the other image. This can mislead the person image matching using appearance features. In order to mitigate this problem, the designed re-id algorithms should be capable of extracting feature vectors against pose variations.

(b) *Illuminations*: Different cameras in different locations can have different illumination conditions, results in the common illumination variation problem in person re-id data. For images from indoor camera, usually it is darker than the images from outdoor cameras. In addition, the images from same camera but different time, there may be also illumination variations. Two examples are presented in Fig. 1.6(b). From the figures, it can be found the appearance of the same person under different illuminations can be very different.

(c) *Viewpoints*: Due to different camera installation angles and locations, the same person presented in images from different cameras can have different viewpoints. Take person 2 who walks from camera 2 to 1 in Fig. 1.3 for example. Most probably, the front view of person 2 can be captured in camera 2 but after entering the FOV of camera 1, normally the camera will only capture the back or side view of the person. This can significantly increase

(a) Person pose variations (b) Illumination variations (c) Viewpoint variations

(d) Resolution variations (e) Occlusions (f) Clothing similarities

Figure 1.6: Examples of different kinds of intra- and inter-personal variations. In sub-figures (a)-(e), the images in the same bounding box contain the same person whilst in sub-figure (f), the images in the same bounding box refer to two different persons.

the difficulty in person re-id, especially when the target person is carrying a backpack as shown in Fig. 1.6(c). This causes the large difference between the appearances of the front and back view of the person.

(d) *Resolutions*: To reduce the number of cameras in the surveillance system, cameras are usually installed in high places to get the large FOVs. Thus, pedestrians are usually far away from cameras. Even for high resolution cameras, the person image can still be of relative low resolution. Due to the changes of the distance between pedestrians and cameras, resolution variation is also a very common problem both for the person images from one specific camera-view or different camera views, as illustrated in Fig. 1.6. This requires that the designed re-id algorithm should be able to match person images not only in low-resolution but also across different resolutions.

(e) *Occlusions*: In re-id data, pedestrians may be partially occluded by other pedestrians or objects. This can happen when the pedestrian is carrying the backpack. Fig. 1.6(e) gives the examples that the pedestrian is occluded by the motorbike or backpack. As shown in the figure, occlusions can result in large differences between the person images from the same person and finally leads to the misleading person re-id results.

(f) *Clothing similarities*: Two different pedestrians wearing very similar clothes can be also very challenging for person re-id. Fig. 1.6(f) gives two examples. The person images in the same bounding box are from two different pedestrians. It can be observed from the figure that the person images are very similar to each other and hard to be distinguished even for humans.

(a) CAVIAR4ReID (b) QMUL iLIDS

Figure 1.7: Samples from two person re-id datasets for illustrating dataset domain variations. The person images in the same column represent the same person. (a) CAVIAR4ReID dataset is collected from a shopping mall [25] while (b) QMUL iLIDS dataset is collected from an airport [161].

All of these different kinds of intra- and inter-personal variations can cause the problems in extracting discriminative re-id features. This further results in the problems of similarity comparison.

2. Domain Variations

Domain variation is a common problem in machine learning and computer vision. The data collected from different domains, *e.g.*, different time and locations, usually present different domain specific information. The conventional learning algorithms rely heavily on the assumption that data used for training and test are drawn from the same distribution (from same domain). If this kind of algorithms trained on one domain are directly applied on a different one, a large performance degradation will be observed. In order to solve this problem, many domain adaptation methodologies have been proposed [151, 152, 103, 35, 26, 141].

The domain variations in person re-id data can be roughly divided into two categories, *i.e.*, dataset domain variations [179, 155, 81, 180, 178, 200] and camera-view domain variations [199, 22, 77, 201]:

(a) *Dataset domain variations*. Different re-id datasets are usually collected from different domains and this leads to different domain information presented in datasets. Fig. 1.7 shows the samples from two different datasets. One is CAVIAR4ReID collected from a shopping mall [25] while the other one is QMUL iLIDS collected from an airport [161]. From the figure, it can be observed that pedestrians in airport are carrying luggages while this not happens for the pedestrians in the shopping mall. If the re-id model is trained

(a) Camera 1 (b) Camera 2

Figure 1.8: Camera-view domain variations in Market1501 dataset [191]. The persons in (a) camera 1 are the same as in (b) camera 2.

using QMUL iLIDS dataset, the model will not only focus on person appearance but also luggages in extracting re-id features since luggages can also provide discriminative features for re-identifying persons. If such re-id model is directly applied on CAVIAR4ReID dataset, it can be expected that the performance will largely degrade since most of pedestrians in CAVIAR4ReID dataset are not carrying luggages. In addition, it can be also observed that the resolutions and illuminations in these two datasets are also different.

(b) *Camera-view domain variations*. As aforementioned, person re-id is a problem of retrieving person images across non-overlapping camera views. Thus, re-id data are collected from different camera views. Because of the distributed deployment of cameras in surveillance system, person images from one camera-view usually contain domain specific information [201]. For example, for two cameras located in places with different illuminations, the person images from these two cameras will also present with different illuminations. Fig. 1.8 gives the examples of the camera-view domain variations. The person images from camera 2 are darker than the ones from camera 1. In addition to illumination, the viewpoint can be also one of the camera-view domain variations. For example, one camera mostly captures the front views of pedestrians while the other one camera mostly captures the back views of pedestrians. This can happen when two cameras are separately installed on the top of the entrance and exit gate.

3. Difficulties in Dataset Creation and Annotation

Existing person re-id methods depend mostly on a large set of cross-camera identity labelled training data, especially the deep learning-based methods. This requires a tedious data collection and annotation process.

Specifically, in order to label a fully supervised person re-id training dataset, a human annotator often needs to match manually a given person identity from one camera view with all the persons from the other camera views. This has a quadratic complexity with the number of both camera views and person identities. Assume an ideal case with M cameras and N

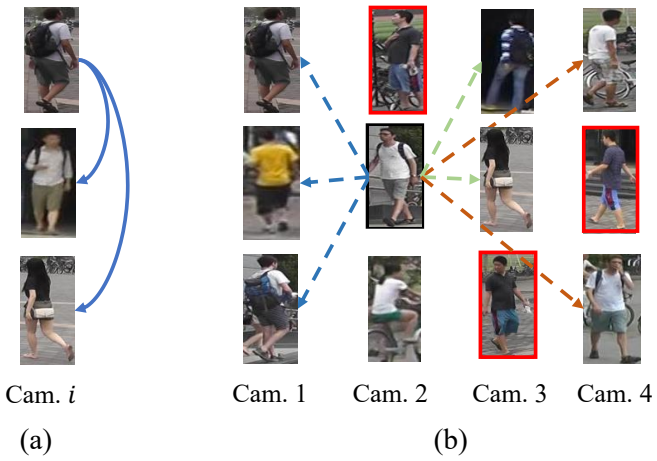


Figure 1.9: (a) Illustration of manually annotating identity labels for the samples in i-th camera view. (b) Illustration of manually associating identity labels across camera views (only four camera-views are illustrated here). For inter-camera identity association, the selected identity (e.g., the identity with black bounding box as in (b)) needs to be compared with the identities from all of the other camera views. The arrow line denotes the comparison made between two identities. The identities in red bounding boxes denote that they have already been associated and will not be compared for further association. In each column, the identities are from the same camera-view.

identities in each camera view, and each identity has one person image in each camera-view. The cost of annotating identity labels in one single camera view is $O(N)$. This is because for most people, re-appearing in a camera view is rare during a limited time period. However, the cross-camera identity association complexity is $O(M^2N^2)$. Fig. 1.9 gives the illustrative examples for manually annotate re-id data.

In addition, given the large size of most current machine learning and computer vision datasets, the data annotation process is almost impossible to be completed by one annotator. Thus, several or more annotators are usually involved in annotating one dataset. The whole dataset is divided into many parts and each part is assigned to one annotator. For fully supervised re-id dataset, in order to get a unified identity space, the annotators need to communicate to each other in the data annotation process to guarantee that the person images are assigned with right identity labels. This communication process adds the extra complexity in data annotation.

1.2.4 Datasets

As one of the important research topics in visual surveillance, person re-id is receiving a large amount of attentions nowadays and various new methodologies have been proposed in

the past one decade. In order to evaluate the designed methodologies, many datasets have been created.

Person re-id datasets can be categorized into two groups. One is image-based datasets and the other one is video-based datasets. Image based person re-id is a more generic problem compared with video-based person re-id. The methods proposed for image-based person re-id can be easily extended to video-based case by considering each frame in the video as one image. Image based person re-id can be further separated into single-shot re-id, which consists in matching pairs of images, a probe and a gallery image for each individual, and multi-shot re-id, in which each individual has multiple images, either in the gallery and/or the probe set. Compared to single shot re-id, multi-shot re-id can be exploited to accumulate more visual information and ensure higher re-id accuracy. In this thesis, the image-based person re-id is considered with emphasizing on multi-shot case.

A comprehensive list of person re-id datasets can be found in [42]. Here, a part of image-based person re-id datasets are summarized in Table 1.1. These datasets are collected from different scenarios, for example airport [161], underground station [85], shopping mall [25] and campus [71, 70, 72, 191]. Different person re-id challenges are included in these datasets. The re-id datasets that are used for evaluating the proposed methods in Chapters 3-5 are briefly introduced as follows:

VIPeR dataset [43] is widely used for evaluating the performance of re-id methods. It contains 632 person image pairs from two cameras. Large variations of viewpoint and illu-

Table 1.1: Person re-id datasets.

Dataset	Release Year	Identity No.	Camera No.	Image No.
VIPeR [43]	2007	632	2	1264
QMUL iLIDS [161]	2009	119	2	476
GRID [85]	2009	1025	8	1275
CAVIAR4ReID [25]	2011	72	2	1220
CUHK01 [71]	2012	971	2	3884
CUHK02 [70]	2013	1816	10 (5 pairs)	7264
CUHK03 [73]	2014	1467	10 (5 pairs)	13164
Market1501 [191]	2015	1501	6	32217
PKU-Reid [99]	2016	114	2	1824
PRW [193]	2016	932	6	34304
DukeMTMC-reID [196]	2017	1812	8	36441
MSMT17 [159]	2018	4101	15	126441
PKU Sketch-ReID [115]	2018	200	2	400

mination in images make it a challenging dataset. In compliance with common evaluation scenario, 316 person image pairs are randomly selected for training, and the rest of 316 image pairs are retained for test.

GRID dataset [85] is composed of 250 person image pairs. Each pair consists of two images of the same person but from different camera views. In addition, there are another 775 person images that do not belong to any of the 250 paired persons. In the experiment, 125 person image pairs are randomly selected for training. The remaining 125 person image pairs and 775 unpaired person images are used for test. As a result, in the test set, there are 125 probe images and 900 gallery images.

PRID 450S dataset [123] contains 450 single-shot image pairs that depict the walking persons captured in two spatially disjoint camera views. The dataset also provides the binary segmentation masks separating the foreground from background and the person part-level segmentation results are also contained.

Market-1501 Dataset [191] is a popular person re-id dataset in deep learning that is created using six cameras in front of a campus supermarket. It contains 1501 identities in which 751 identities for training and the other 750 identities for test. For each identity, multiple images are available and for training, it provide $12,936$ images. In test set, there are $19,732$ gallery images and $3,368$ probe images.

DukeMTMC-ReID Dataset [196] is created by selecting and annotating pedestrian images from a multi-target, multi-camera tracking dataset. There are $1,404$ identities in total from eight cameras. 702 identities with $16,522$ images are selected for training and the other 702 identities with $17,661$ images are for test. The query set is formulated by picking one person image for each identities in test set in each camera. Thus, there are $2,228$ query images in total.

MSMT17 Dataset [159] is collected from a 15-camera network in campus in which there are 12 outdoor and 3 indoor cameras. Faster RCNN [121] is utilized for pedestrian bounding box detection. The dataset is annotated with three annotators by checking all detected bounding boxes and annotating identity labels in 2 months. The final dataset contains $126,441$ bounding boxes of $4,101$ identities. The training set contains $32,621$ bounding boxes of 1,041 identities. The test set contains $93,820$ bounding boxes of $3,060$ identities, in which there are $11,659$ query images and $82,161$ gallery images.

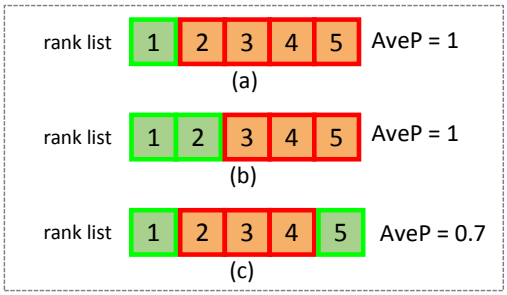


Figure 1.10: A toy example of the difference between CMC and average precision (AveP) measurements [191]. True matches and false matches are in green and red boxes, respectively. For all of three rank lists (a)-(c), the CMC curve remains 1, whilst for AveP, its value is 1, 1 and 0.7.

1.2.5 Evaluation Metrics

Evaluation metrics are important for evaluating the quality of the designed algorithm and also makes it possible for comparing performances of different related algorithms. In person re-identification, Cumulated Matching Characteristics (CMC) curve is the most popular metric for evaluating the performance of person re-id methodologies. However, as discussed in [191], CMC is valid only if there is only one ground truth match for a given query, *i.e.*, single-shot person re-id. For the multi-shot re-id scenario, CMC is not enough to provide a good evaluation about the person image retrieval quality since in CMC, the "recall" is not considered. Fig. 1.10 shows a toy example of the evaluation bias in CMC. For (b) and (c), it can be observed that the retrieval quality of the rank list (b) is better than the rank list (c), but CMC cannot reflect this and its value is 1 for both of them. In order to solve this problem, Zheng *et al.* proposed to use the mean average precision (mAP) for re-id algorithm evaluation [191]. Both CMC and mAP are related with two image retrieval metrics, *i.e.*, precision and recall, which are also used in Chapter 5. Thus in the following, precision and recall will be first described before the introduction of CMC and mAP.

1. Precision and Recall

In image (or information) retrieval, many metrics have been proposed and used for evaluating the retrieval quality of the designed algorithms [119, 89]. Precision and recall are two basic and widely used evaluation metrics. Usually these two metrics are used together to reflect different aspects of the algorithm. Precision is used to measure "how useful the search results are" while recall is to measure "how complete the results are". Given the top-k images retrieved using a query (or target) image, the precision $P(k)$ can be calculated as:

$$P(k) = \frac{|\{\text{relevant images}\} \cap \{\text{retrieved images}\}|}{|\{\text{retrieved images}\}|} \tag{1.1}$$

in which $|\{\cdot\}|$ denotes the number of items in the set and $\cap$ is the intersection operator between two sets. The corresponding recall $R(k)$ can be calculated as:

$$R(k) = \frac{|\{\text{relevant images}\} \cap \{\text{retrieved images}\}|}{|\{\text{relevant images}\}|} \tag{1.2}$$

From Eqs. (1.1) and (1.2), it can be observed that precision is the percentage of the top-k retrieved images that are relevant to the query and recall is the percentage of all the relevant images in the search database which are retrieved.

2. Cumulated Matching Characteristics

CMC curve is the most commonly used evaluation metrics in person re-id [192, 39, 73, 189, 191]. It represents results of an identification task by plotting the probability of correct identification against the number of candidates returned [114]. The faster the CMC curve approaches one, the better the person re-id algorithm. For the calculation of CMC curve, there is a difference between the single-shot and multi-shot person re-id case.

Given a query image in the single-shot case, the re-id model algorithm will rank all gallery images according to their similarities to the query image from large to small. The top-k CMC accuracy ACC_k for this query image is calculated as:

$$ACC_k = \begin{cases} 1 & \text{if top-}k\text{ ranked gallery images contain the query identity,} \\ 0 & \text{otherwise.} \end{cases} \tag{1.3}$$

The CMC accuracy will be calculated for all of query images and the final CMC curve is obtained with averaging every ACC_k over all query images. The k-th value in CMC curve is:

$$\text{CMC}(k) = \frac{\sum_{q=1}^{N_Q} ACC_k}{N_Q} \tag{1.4}$$

in which $k \in [1, 2, \ldots, N_G]$. N_G and N_Q denote the number of query and gallery images, respectively.

For the multi-shot person re-id case, CMC curve calculation is still not agreed in person re-id community. Take CUHK03 [73] and Market1501 [191] benchmarks for example, their calculations about CMC curves are different:

(a) In CUHK03 benchmark [73], the query and gallery images are from different camera views. In calculation of ACC_k, only one of gallery images is randomly sampled for each query identity. Then, the ACC_k and CMC curve are calculated as in single-shot person re-id case. This process is performed for N times ($N = 20$ in [73]) and the final CMC curve is obtained by averaging these repeatedly performed results.

(b) In Market1501 benchmark [191], the query and gallery images in Market1501 dataset can be from the same camera views. In the calculation of ACC_k for each individual query identity, the corresponding gallery images from the same camera-view are excluded. In addition, the random sampling is not performed on the considered gallery images for each query. Thus in the calculation of top-k CMC accuracy for each query, only the easiest positive gallery image (which shares the same identity as query) is considered, while for other positive gallery images, they are ignored.

3. Mean Average Precision

From the description of CMC curve, it can be observed that CMC curve is not comprehensive enough for evaluating the quality of the image rank list returned by the person re-id algorithm, especially for the multi-shot person re-id case. For the CMC curve calculation provided in CUHK03 benchmark [73], the randomly sampling process should be performed for N times in which N can be varied in different work and thus may cause the problem in algorithm comparisons. Although Market1501 benchmark provides a simple way to calculate CMC cuve as shown in Fig. 1.10, it does not consider the recall and thus gives a biased evaluation result. Zheng *et al*. introduced the mean average precision (mAP) metric for solving the problems in CMC curve [191].

mAP is a popularly used metric in information retrieval and it starts to be widely used in evaluating person re-id algorithms since the work [191]. As introduced before, precision and recall are two complementary metrics that are used for evaluating different aspects of the image retrieval algorithm and usually, they are used together. mAP can be regarded as a metric that makes the trade off between precision and recall. Given a query image and the corresponding rank list returned by one person re-id algorithm, the precision and recall can be calculated using Eqs. (1.1) and (1.2) at every position in the rank list. A example of

precision-recall curve is presented in Fig. 1.11. Suppose $p(r)$ denotes the precision at recall r. The average precision is the average value of $p(r)$ over the recall interval $[0, 1]$:

$$\text{AveP} = \int_0^1 p(r)dr. \tag{1.5}$$

AveP is exactly the area under the precision-recall curve.

In practical calculation, this integral is obtained with a finite sum over every position in the ranked list of gallery images:

$$\text{AveP} = \sum_{k=1}^{N_G} P(k)\Delta R(k) \tag{1.6}$$

$P(k)$ and $R(k)$ are the same as in Eqs. (1.1) and (1.2). N is the number of gallery images. $\Delta R(k)$ is the change of the recall from $k-1$ to k: $\Delta R(k) = R(k) - R(k-1)$.

Based on the average precision, *i.e.*, AveP, the mAP is the average of AveP over all query images and it can be formulated as:

$$\text{mAP} = \frac{\sum_{q=1}^{N_Q} \text{AveP}_q}{N_Q} \tag{1.7}$$

in which AveP_q denotes the AveP for the q-th query image.

Compared with CMC curve, mAP gives a more comprehensive evaluation of person image retrieval results by considering both precision and recall. Consider the rank lists (b) and (c) in Fig. 1.10, the true matches are all in top-2 in (b) and its retrieval result is better than rank list (c) in which only one true match is contained in top-2. The CMC curve of these two rank lists cannot reflect their retrieval qualities with the value 1 for both of them, while mAP gives its metric value 1 for rank list (b) and 0.7 for rank list (c). However, mAP is not as intuitive as CMC curve. In most of recent person re-id works [192, 73, 189, 39, 7], both of CMC curve and mAP are considered for algorithm performance evaluation.

1.3 Contributions of This Thesis

As discussed in Section 1.2.3, learning discriminative person re-id features has the challenges of (1) Intra- and inter-personal variations, (2) domain variations and (3) difficulties in data creation and annotation. Corresponding to these three challenges, this thesis contributes to the re-id domain by proposing three related methodologies and one new re-id setting:

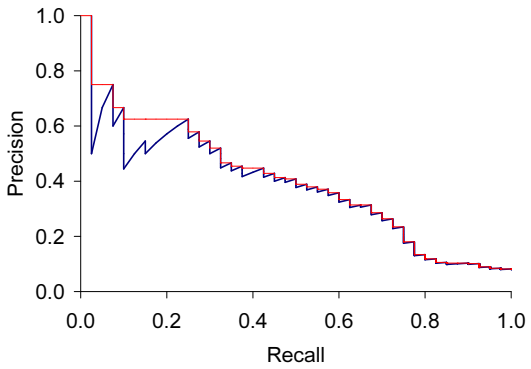


Figure 1.11: Precision-recall curve. The blue line denotes the interpolated precision at a certain recall level. Refer to [127] for more information

(1) **Gaussian mixture importance estimation.** Handcrafted features are usually not discriminative enough for person re-id because of noisy information, such as background clutters. To precisely evaluate the similarities between person images, the main task of distance metric learning is to learn the Mahalanobis matrix to filter out the noisy information. KISS metric learning is an effective method in person re-id. However, it is sensitive to the feature dimensionality and can not capture the multi-modes in dataset. To this end, a Gaussian Mixture Importance Estimation re-id approach is proposed, which exploits the Gaussian Mixture Models for estimating the observed commonalities of similar and dissimilar person pairs in the feature space. This work has been accepted in *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2017* [203] and it was recognized as the *Best Paper, Honorable Mention*.

(2) **Unsupervised domain-adaptive person re-id based on pedestrian attributes.** Different from person identity, pedestrian attributes, *e.g.*, hair length, clothes type and color, are consistent across different domains (or datasets). However, most of re-id datasets lack attribute annotations. On the other hand, in the field of pedestrian attribute recognition, there is a number of datasets labeled with attributes. Exploiting such data for re-id purpose can alleviate the shortage of attribute annotations in re-id domain and improve the generalization capability of re-id model. To this end, an unsupervised domain-adaptive re-id feature learning framework is proposed to make full use of attribute annotations. Specifically, an existing unsupervised domain adaptation method has been extended to transfer attribute-based features from attribute recognition domain to the re-id domain. With the proposed re-id feature learning framework, the domain invariant feature representations can be effectively extracted. This work has been accepted in the *IEEE International Conference on Image Processing (ICIP), 2019* [204].

(3) **Intra-camera supervised person re-id.** Annotating the large-scale re-id datasets re-

quires a tedious data collection and annotation process and therefore leads to poor scalability in practical person re-id applications. To overcome this fundamental limitation, a new person re-id setting is considered without inter-camera identity association but only with identity labels independently annotated within each camera-view. This eliminates the most time-consuming and tedious inter-camera identity association annotating process and thus significantly reduces the amount of human efforts required during annotation. It hence gives rise to a more scalable and more feasible learning scenario, which is named as Intra-Camera Supervised (ICS) person re-id. Under this ICS setting, a new re-id method, *i.e.*, Multi-tAsk mulTi-labEl (MATE) learning method, is formulated. Given no inter-camera association, MATE is specially designed for self-discovering the inter-camera identity correspondence. This is achieved by inter-camera multi-label learning under a joint multi-task inference framework. In addition, MATE can also efficiently learn the discriminative re-id feature representations using the available identity labels within each camera-view. This work has been accepted in a workshop of *IEEE International Conference on Computer Vision (ICCV), 2019* [205] and its journal extension has been submitted to *International Journal of Computer Vision (IJCV)*.

1.4 Thesis Outline

This thesis is organized as follows:

Chapter 2 presents a review on related works about different categories of person re-id problems and various related strategies and methodologies that proposed for solving these problems.

Chapter 3 describes a new metric learning methodology, *i.e.*, Gaussian Mixture Importance Estimation (GMIE), for person re-id. Different from the existing re-id methodologies, it shows that GMIE not only can model the multi-modes in re-id dataset but also it is robust to the increase of feature dimension.

Chapter 4 proposes a new person re-id framework which bridges the pedestrian attribute recognition and person re-id. Based on the observation that attributes are consistent across dataset while person identities are not, the proposed framework applies an extended unsupervised domain-adaptive method to adapt the trained attribute recognition model to the re-id domain for extracting attribute related features for person re-id.

Chapter 5 explains the intra-camera supervised (ICS) person re-id setting and a new methodology using this re-id setting. It shows that in ICS person re-id, the data is annotated with-

out inter-camera identity association but only with identity labels independently annotated within each camera-view and thus significantly reduces the data annotation efforts. Under this ICS setting, a new re-id method, *i.e.*, Multi-tAsk mulTi-labEl (MATE) learning method, is formulated and the experiments validates that the effectiveness of MATE in solving ICS person re-id problem.

Chapter 6 concludes this thesis and several potential re-id research directions are discussed as the future works.

The outline of this thesis is visualized in Fig. 1.12.

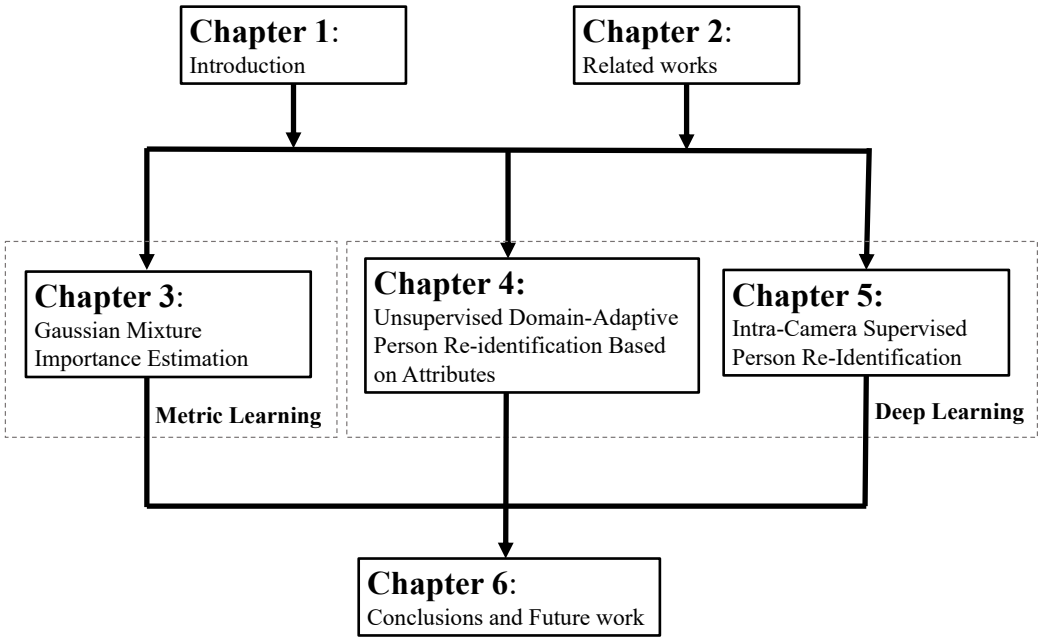


Figure 1.12: Visualization of thesis outline.

Chapter 2

Related Works

The related works of person re-id are roughly separated into (1) Handcrafted Feature Extraction and Metric Learning, (2) Pedestrian Attribute Based Methods and (3) Deep Learning Methods. Although the second category has overlaps with first and third categories, there are a large number of person re-id methods based on pedestrian attributes and these methods are grouped as a single category.

2.1 Handcrafted Feature Extraction and Metric Learning

Before deep learning, person re-id is mainly based on handcrafted feature extraction and metric learning. In this category, part of works propose to *directly* apply distance functions on handcrafted features for calculating similarities for person re-id [31, 25, 188, 192]. However, handcrafted features are usually not discriminative enough due to the noisy information, for example background clutters. The *learning-based* methods aims at applying metric learning to filter out noisy information and extract more discriminative features [123, 61, 78, 79, 44, 192].

2.1.1 Handcrafted Feature Extraction

Handcrafted feature extraction aims at manually designing feature extraction methods to extract appearance features, *e.g.*, color and texture, based on the experiences from designer. A common way in designing handcrafted feature is firstly dividing the person image in hand into horizontal stripes [10, 78, 175], triangular graph [37], regions clustered by color [157], symmetry and assymetrical parts [31], semantic or meaningful parts [25, 9], concentric rings [37] or grid of localized patches [5]. Then, different kinds of descriptors, *e.g.*, HSV histogram, Scale Invariant Local Ternary Pattern (SILTP) [80] and Local Binary Pattern (LBP) [113], are applied for extracting color and/or texture features from each divided image parts. In order to extract robust and discriminative person re-id feature representations, many handcrafted features have been proposed, *e.g.*, the ensemble of local features

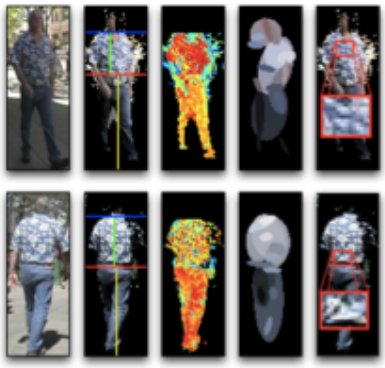


Figure 2.1: SDALF handcrafted feature extraction [31]. (a) two images of the same person; (b) x- and y-axes of asymmetry and symmetry, respectively; (c) weighted histogram back-projection (brighter pixels mean a more important color), (d) Maximally Stable Color Regions; (e) Recurrent Highly Structured Patches.

(ELF) [44], Symmetry-Driven Accumulation of Local Features (SDALF) [31], Custom Pictorial Structure (CPS) [25], kBiCov [98], fisher vectors (LDFV) [97], Local Maximal Occurrence (LOMO) [78], Gaussian Of Gaussian (GOG) [105, 106]. Several popular handcrafted feature extraction methods are presented in the following.

In [31], a handcrafted feature named SDALF has been proposed. Fig. 2.1 presents the SDALF handcrafted feature extraction method. The person images are pre-processed in SDALF by segmenting out the pedestrian foreground, and then salient parts of the body figure are selected by adopting perceptual principles of symmetry and asymmetry. Specifically, as shown in Fig. 2.1(b), two horizontal axes of asymmetry (blue and red lines) are firstly obtained in each person image which isolate the person body into three main regions, *i.e.*, head, torso and legs. Then, the vertical axes of appearance symmetry (green and yellow lines) are separately estimated for the torso and legs part. At last, three complementary aspects of the human body appearance are extracted from each part, including: (i) the general chromatic content via HSV histogram; (ii) the per-region color displacement, through Maximally Stable Colour Regions (MSCR) [34]; (iii) the presence of Recurrent Highly Structured Patches , estimated through a novel per-patch similarity analysis. In order to minimize the effects of pose variations, the extracted features are weighted by the distance with respect to the vertical axis. Matching these features gives the similarity measure between the candidates.

In [25], Pictorial Structures (PS) [32] is applied to localize human body parts. With fitting PS to person image in single-shot re-id, an ensemble of features are extracted from each localized body parts. Fig. 2.2(a) illustrates the PS fitting result. As in SDALF [31], different local descriptors are introduced to encoding complementary aspects, such as the chromatic

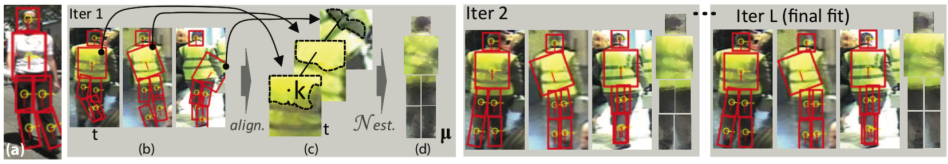
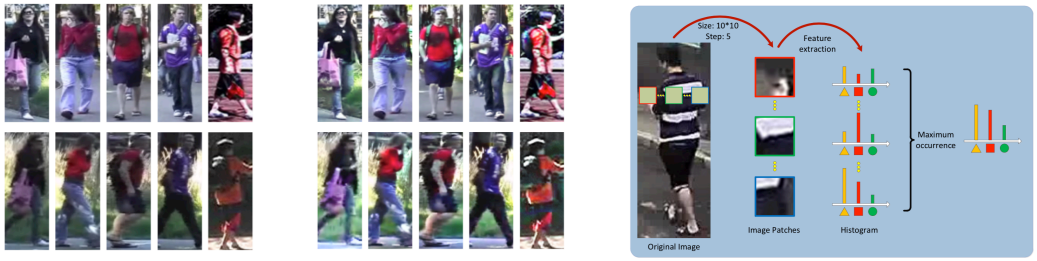


Figure 2.2: Illustration of localizing body parts [31]. Single-shot PS in (a). Multi-shot CPS at iteration 1: (b) initial PS fitting; (c) the parts are aligned and per-pixel statistics is collected employing spatio-temporal reasoning; (d) the ad-hoc part detectors are estimated, whose means $\hat{\mathrm{I}}$ij are shown. At every iteration until L, the fitting becomes more accurate due to the improving part detectors.



(a) (b) (c)

Figure 2.3: (a) Example pairs of images from the VIPeR database [43]. The images in the same column are from same person. (b) Processed images in (a) by Retinex. (c) Illustration of the LOMO feature extraction method.

content and the spatial arrangement of colors. The local descriptors include HSV histograms and MSCR [34, 31]. The features of each part are subsequently combined into a single ID signature. Matching between signatures is carried out by standard distance minimization strategies. For the multi-shot re-id, a model called Custom Pictorial Structure (CPS) has been proposed to get more accurate body parts. The main idea is to learn the local appearance of each part in a given subject so that ad- hoc appearance part detectors can provide more accurate PS fitting [25]. Localizing body parts using CPS is illustrated in Fig. 2.2. As in single-shot re-id case, after localizing body parts, the features of each part are extracted and combined to get the final features for each person image.

In [78], Liao *et al.* proposed the Local Maximal Occurrence (LOMO) handcrafted feature for person re-id. LOMO feature is designed by considering two main challenges in extracting person re-id features: (1) Illumination variations. Color is an important feature for describing person images. However, as shown in Fig. 2.3(a), the illumination conditions across cameras can be very different. In order to deal with this problem, the Retinex algorithm is applied to pre-process person images. It aims at producing a color image that is consistent to human observation of the scene. The restored image usually contains vivid

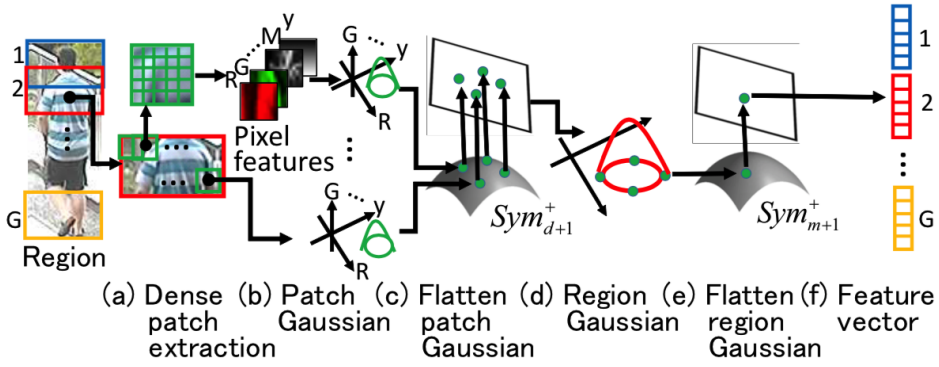


Figure 2.4: Illustration of GOG feature extraction process [105, 106].

color information, especially enhanced details in shadowed regions. Fig. 2.3(b) shows the images processed using Retinex algorithm. It can be observed that the illuminations variations have been reduced. In addition, SILTP descriptor, which is robust to image noise and has the invariant property under monotonic gray-scale transforms, is introduced for encoding person appearance features [78, 80]. (2) Viewpoint changes. As aforementioned in Chapter 1, pedestrians under different cameras usually present with different viewpoint. In LOMO, the sliding windows is used to describe local details of a person image. As shown in Fig. 2.3(c), a subwindow size of 10×10 is used with an overlapping step of 5 pixels to locate local patches in 128×48 images. Within each subwindow, the SILTP and HSV histograms are extracted. In order to deal with the viewpoint changes, all the subwindows at the same horizontal stripe are considered at one time and the maximal value of each patterns among these subwindows are selected and combined to formulated the final histogram of the horizontal stripe. In addition, to further consider the multi-scale information, a three-scale pyramid representation is built for each person image. The LOMO feature extraction process as shown in Fig. 2.3(c) is repeatedly performed on all of these scaled representations. The final LOMO feature is formulated by concatenating all of local maximal occurrences.

In [105], a Gaussian of Gaussian (GOG) handcrafted feature has been proposed. The covariance descriptor describes a region of interest based on the covariance of pixel features [149]. It can encode different modalities, *e.g.*, color and texture, of pixel features into a single meta-descriptor. According to the experimental results in [105], the mean of pixel features is also important in describing person appearance. Based on this observation, GOG feature introduces hierarchical Gaussian distribution of pixel features as region descriptor for person re-id. Fig. 2.4 sketches the process of GOG feature extraction. Specifically, regions are extracted from the original person image. For each region, it is densely divided into many patches as shown in Fig. 2.4(a). The pixel features in each patch is modeled by a Gaussian distribution. Thus for each region, it is represented with a set of Gaussian distri-

butions. This set of Gaussian distributions corresponding to one region are again modeled with one Gaussian distribution. The parameters of these Gaussian distributions are used as feature vectors to represent regions. The final GOG feature representation is formulated by concatenating all the feature vectors corresponding to different regions of the person image.

2.1.2 Metric Learning

After extracting handcrafted features from person images, there are usually two categories of post-processing methods. The first category is directly applying the distance function on handcrafted features for calculating similarities for person re-id [31, 25, 188]. For example, in work [31], the similarity is calculated as:

$$\begin{aligned} d(I_A, I_B) =& \beta_{WH} \cdot d_{WH}(\text{WH}(I_A), \text{WH}(I_B)) + \\ & \beta_{MSCR} \cdot d_{MSCR}(\text{MSCR}(I_A), \text{MSCR}(I_B)) + \\ & \beta_{RHSP} \cdot d_{RHSP}(\text{RHSP}(I_A), \text{RHSP}(I_B)), \end{aligned} \tag{2.1}$$

in which $\text{WH}(\cdot)$, $\text{MSCR}(\cdot)$ and $\text{RHSP}(\cdot)$ are the weighted histograms, MSCRs, and Recurrent High-Structured Patches, respectively, and β_{WH}, β_{MSCR} and β_{RHSP} are normalized weights. $d_{WH}(\cdot), d_{MSCR}(\cdot)$ and $d_{RHSP}(\cdot)$ are the distance functions.

However, handcrafted features are usually not discriminative enough for person re-id because of the included noisy information, for example the background information from person images. In order to precisely evaluate the similarities between person images, the second category is based on learning methods which is capable of two functions: (1) filtering out the noisy information contained in handcrafted features, and (2) calculating similarity. Take the commonly used distance metric learning [173] for example, its basic formulation is:

$$d^2_{\mathbf{M}}(\mathbf{z}_i, \mathbf{v}_j) = (\mathbf{z}_i - \mathbf{v}_j)^T \mathbf{M} (\mathbf{z}_i - \mathbf{v}_j), \tag{2.2}$$

in which $\mathbf{M} \succeq 0$ is the Mahalanobis matrix which is a positive semidefinite matrix. $\mathbf{z}_i$ and $\mathbf{v}_j$ are the handcrafted features corresponding to i-th and j-th samples from two different camera-views. Since $\mathbf{M}$ is a Hermitian and positive semi-definite matrix, $\mathbf{M}$ can be further decomposed into:

$$\mathbf{M} = \mathbf{L}^T \mathbf{L} \tag{2.3}$$

Thus, the distance $d^2_{\mathbf{M}}(\mathbf{z}_i, \mathbf{v}_j)$ can be reformulated as:

$$\begin{aligned} d^2_{\mathbf{M}}(\mathbf{z}_i, \mathbf{v}_j) =& (\mathbf{z}_i - \mathbf{v}_j)^T \mathbf{L}^T \mathbf{L} (\mathbf{z}_i - \mathbf{v}_j) \\ =& ||\mathbf{L}\mathbf{z}_i - \mathbf{L}\mathbf{v}_j||^2 \end{aligned} \tag{2.4}$$

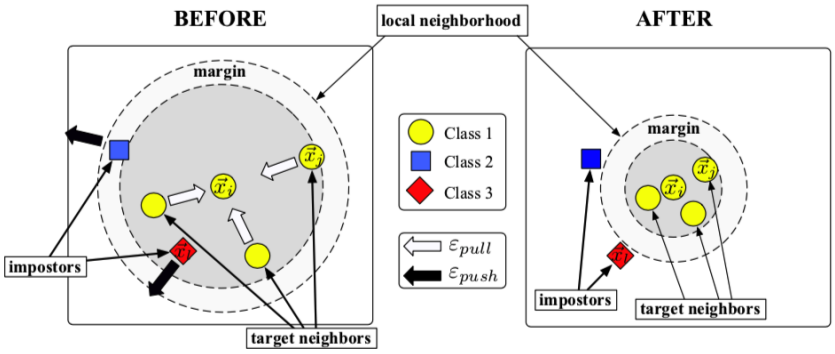


Figure 2.5: Illustration of training process in LMNN [163, 162]. Before training (*Left*), samples randomly locate in the space. During training, the target samples with same class (yellow circles) are gradually pulled into a smaller radius while differently labeled samples (squares) are pushed outside the smaller radius by some finite margin, as shown in the right diagram.

in which $||\cdot||^2$ is the Euclidean distance function. From Eq. (2.4), it can be found that the task of Mahalanobis matrix $\mathbf{M}$ functions is to filter out the noisy information in handcrafted features, *i.e.*, $\mathbf{z}_i$ and $\mathbf{v}_j$, and thus obtain more discriminative person re-id features, *i.e.*, $\mathbf{L}\mathbf{z}_i$ and $\mathbf{L}\mathbf{v}_j$. Finally, the Euclidean distance function is applied on these discriminative features. The main task of distance metric learning is to learn the Mahalanobis matrix $\mathbf{M}$.

Many learning based re-id methods have been proposed. As aforementioned, distance metric learning is popular in person re-id [173, 52, 123, 61, 78, 79, 44, 128, 110]. In [123], the existing metric learning methods in machine learning are customized for re-id. These methods include Large Margin Nearest Neighbor Learning (LMNN) [163, 162], Information Theoretic Metric Learning (ITML) [27] and Logistic Discriminant Metric Learning (LDML) [45]. LMNN aims at learning a Mahanalobis distance metric for k-nearest neighbor (kNN) classification [163, 162]. The metric is trained based on the idea that samples from the same class (positive pairs) are pulled together while samples from different classes (negative pairs) are separated by a large margin. The training process is illustrated in Fig. 2.5. However, LMNN is sometimes prone to the over-fitting problem due to the lack of regularization [61]. ITML mitigates overfitting by introducing a regularization step [27]. It is based on information-theoretic setting by leveraging the relationship between the multivariate Gaussian distribution and the set of Mahalanobis distances. For LDML [45], Guillaumin *et al.* introduce a probabilistic view on learning a Mahalanobis metric where the a posteriori class probabilities are treated as (dis)similarity measures. As in LMNN and ITML, the objective of LDML is also to learn a metric with which positive pairs have smaller distances than negative pairs. In addition to these existing metric learning methods in machine learning, there are also metric learning methods specifically designed for person re-id. As introduced

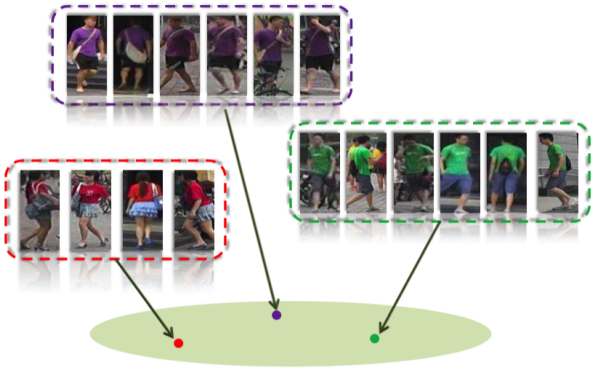


Figure 2.6: Illustration of the idea behind learning the discriminative null space [184]. Samples with the same identity are projected into a single point.

in Chapter 1, person re-id is a cross-camera image retrieval problem. Based on this observation, Hirzer *et al.* proposed a person re-id metric learning method, *i.e.*, Relaxed pairwise Metric Learning (RPML), with considering the transition of samples from one camera [52]. They found RPML can achieve state-of-the-art results even with less sophisticated features describing color and texture information. In [61], Köstinger *et al.* proposed a simple but efficient metric named KISSME (Keep It Simple and Straightforward Metric Learning), which measures the similarity and dissimilarity between samples based on the likelihood ratio test. Together with the LOMO handcrafted feature as introduced above, Liao *et al.* proposed a re-id metric learning method called Cross-view Quadratic Discriminant Analysis (XQDA) [78]. It learns a low dimensional subspace, which can be used for extracting the discriminative re-id features by cross-view quadratic discriminant analysis, and simultaneously, based on KISSME, XQDA also learns a metric for measuring the similarities between samples. One common problem existed in person re-id is that the numbers of positive and negative sample pairs are largely unbalanced. In order to deal with this problem, a logistic metric learning approach has been proposed for re-id in [79]. Logistic metric learning approach learns the metric using positive semidefinite (PSD) constraint based on the observation that PSD constraint provides a useful regularization to smooth the solution of the metric, and hence the learned metric is more robust than without the PSD constraint. One of the challenges in person re-id is dataset creation. As aforementioned, collecting training samples of matched person pairs across camera-views is labour intensive and tedious and this results in the small sample size (SSS) problem [17]. With a small sample size of the training dataset, the within-class scatter matrix becomes singular. Zhang *et al.* proposed to solve this small sample size problem by learning a discriminative null space of the training data [184]. Fig. 2.6 illustrates the idea behind learning discriminative null space, with which the samples of the same identity are projected into a single point and thus, the distances of positive pairs are extremely minimized and simultaneously, the distances of negative pairs are maximized.

Based on the fact that person re-id is a cross-camera image retrieval problem, the ranking methods as in image retrieval domain can be also used for training the re-id model to learn discriminative features. Several works started to exploit the ranking loss for re-id. In works [19, 114], the authors try to directly optimize the re-id evaluation metric CMC and mAP. Since these two list-wise based methods only use the binary similarity information, *i.e.*, relevant and irrelevant pair, they still cannot exploit the discriminative feature from negative pairs and only exploit the local discriminative features. In order to learn the discriminative features from negative image pairs, Chen *et al.* proposed a relevance metric learning method with list-wise constrains and the similarity of arbitrary image pairs can be learned from the algorithm [16].

2.2 Pedestrian Attribute Based Methods

Although person identity shows the good performance as the supervision to learn the feature representation for person re-id, many works have also demonstrated that re-id can also benefit a lot from person attribute supervision [64, 65, 136, 55, 107, 83]. Generally, person attribute and identity represent the features from different levels in person images. As shown in Fig. 2.7, person attribute represents the local part of a person [83], for example the hair length belonging to the attribute on the head part while the up-body clothing type mainly focusing on the torso part. For the person identity, it represents the global description of a whole person. Most of existing re-id algorithms are trained only considering the person identity [120, 143, 198, 192]. For these methods, if the training dataset is small or the person image variations are not sufficiently included in the training set, the algorithm may fail to learn discriminative local features and lead to inaccurate re-id results. Attribute is treated as a kind of mid-level feature that represents the local part of the person, it provides more details to describe the person. In addition, another one advantage of attribute over identity

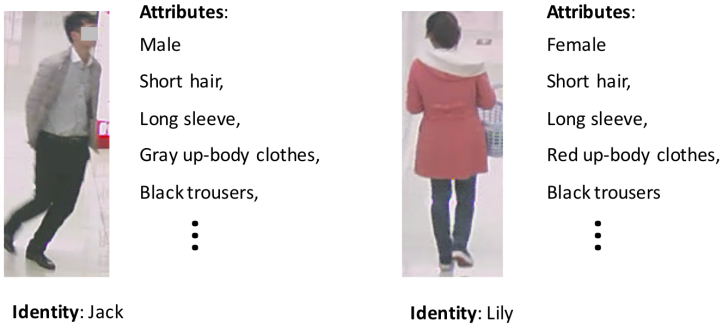


Figure 2.7: Two sample examples of identity and attribute labels.

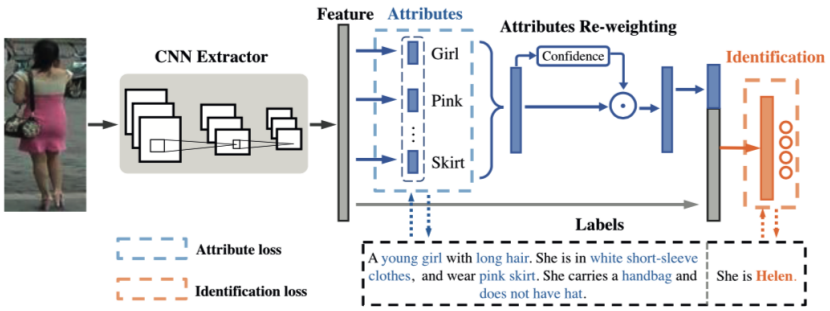


Figure 2.8: Overview of the learning strategy which combines identity and attribute as supervision for learning discriminative re-id features [83]. The model contains two classification parts, one for attribute recognition and the other for identification. The person feature representation of each image is extracted using the CNN extractor. The attribute classifiers predict attributes based on the image feature. For identity classification part, the attribute predictions are treated as additional cues. Specifically, the local attribute predictions are re-weighted by the Attribute Re-weighting Module and then concatenate them with the global image feature. The final identification is built upon the concatenated local-global feature.

label is that attribute is consistent between different domains [204] and thus the re-id model trained under the supervision of attributes can be easily generalized to other domains.

Many works have exploited as supervision for training person re-id model. In [64], a fusion strategy is designed to merge the discriminations of both person attributes and low level features for re-id. Considering the color and type of the clothes are the main cues for appearance based re-id, Li *et al.* proposed a latent Support Vector Machines framework to embed the clothing attributes into person re-id [65]. Su *et al.* proposed the low rank attribute embedding method as a preprocessing procedure to rectify the incorrect and incomplete attributes in the dataset, and with using a multi-task framework, a person identity and attribute based discriminative model is constructed [136]. For these works, the attributes are directly used as a feature vector and fused with low-level features for improving re-id performance. There are also part of works using attribute as the label to learn the attribute supervised discriminative features. In [55], the so-called attribute-consistent model is designed for leaning two projections to map the hand-crafted features into a joint subspace. These two projections respectively correspond to the person identity and attribute, and the experimental results show that, compared with the feature learned only considering person identity, the projected features in the joint subspace are more discriminative and efficient in re-id. Similarly, Lin *et al.* also jointly consider the identity and attribute as the supervisions but in a deep multi-task model to learn the discriminative re-id features [83], they found not only the attribute recognition can help person re-id, but re-id can also improve the attribute recognition performance. The learning strategy in [83] is illustrated in Fig. 2.8. Matsukawa *et al.* only consider the attribute as the supervision to learn the feature representation, and in addition

to the attribute classification loss, a combination attribute loss is designed to improve the feature discrimination [107]. The experiments show that the learned feature representation can be comparable with the hand-crafted features even in the small datasets.

For the works discussed above, one of the prerequisites is that the considered re-id dataset should contain the attribute labels. However, as aforementioned, most of the re-id datasets are still lack of attribute labels, while in attribute recognition domain, there are many datasets labeled with sufficient attributes. Based on this observation, several works started to exploit making use of the attribute recognition dataset for re-id. In [132], the so called Indian Buffet Process (IBP) is used to learn the mid-level representation based on the user defined attributes, and then transfer the learned attribute recognition knowledge from fashion domain to surveillance domain. With the learned attribute representation for each image in surveillance, the metric learning algorithm is used to measure the similarities between images. However, as stated in [117], the user defined semantic attributes may not enough to describe a person. Two different persons may share the same user defined attributes. In addition, there are some latent attributes that are not nameable/semantic but still useful for re-id. To overcome these problems, the work [117] introduces the dictionary learning model to jointly considers the semantic and latent attributes for both zero shot learning and person re-id. However, one limitation of this work lies in that the source and target domain should share user defined attributes if annotated and some latent attributes [117]. Although a same set of attributes across different domains can be extracted by removing the domain specific user defined attributes, the proposed algorithm cannot make full use of the user defined attributes. In [137], a deep attribute learning algorithm has been introduced, and the attribute recognition knowledge is transferred from one dataset to another one using a fine tuning strategy. However, it has been shown that fine tuning a network will degrade its performance on the original task (attribute recognition) [76], and thus the network may fail to extract discriminative features related to attributes. The proposed attribute adaption re-id framework learns to extract local discriminative features under the supervision of attribute labels, and it overcomes the limitations mentioned above inn domain adaptation. In addition, it can make full use of the label information for re-id in both the source and target domains.

2.3 Deep Learning Methods

In 2014, Yi *et al.* [177] and Li *et al.* [72] both proposed to use siamese neural network for person re-id by training the network to determine if a pair of input images belongs to the same identity or not. During the following years, many person re-id methods based on Convolutional Neural Networks (CNNs) have been proposed [192, 63]. In this section, these

CNN based re-id methods are grouped into several categories according to the label information used in training the re-id models, *i.e.*, fully supervised person re-id, unsupervised domain-adaptive person re-id, semi-supervised person re-id, weakly supervised person re-id, unsupervised person re-id, unsupervised tracklet learning person re-id. Except for fully supervised person re-id and unsupervised domain-adaptive person re-id, there are very few works in other groups and thus they are together grouped into the "others" category.

2.3.1 Fully Supervised Person Re-Identification

Most existing deep learning based person re-id models are created by supervised learning methods on a separate set of cross-camera identity labelled training data [73, 186, 23, 74, 133, 12, 130, 190, 192, 86]. Relying on the strong supervision of cross-camera identity labelled training data, they have achieved remarkable performance boost.

One of the challenges in person re-id is the pose variations. However, existing benchmarks, *e.g.*, Market1501 [191], DukeMTMC-reID [196], do not provide sufficient pose coverage to train a robust re-id model. In order to solve this problem, Liu *et al.* proposed a pose-transferrable person re-id framework [86]. Specifically, the framework is composed of two parts. One is the image generation. Based on GAN [40], person image is generated using the appearance from existing datasets and poses extracted from MARs dataset [189]. The second part is re-id model training. The generated images are combined with the realistic images to train the re-id model. In order to balance the contribution between real samples and generated samples during training, a label smoothness scheme is introduced for cross entropy training. The work [50] considers the occlusion problem in person re-id under deep neural network framework. The person re-id model can fail to re-identify a person when the person body is severely occluded. To solve this problem, several solutions have been

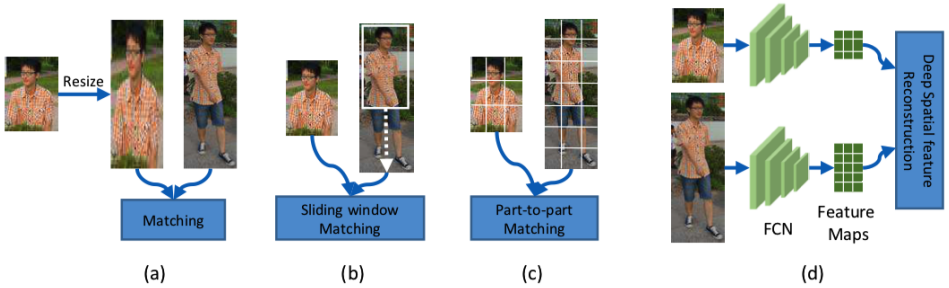


Figure 2.9: Different solutions for partial person re-id: (a) The probe person image and gallery person image are resized to fixed-size (Resizing model). (b) Sliding window matching. (c) Part-based model. (d) The proposed Deep Spatial feature Reconstruction [50].

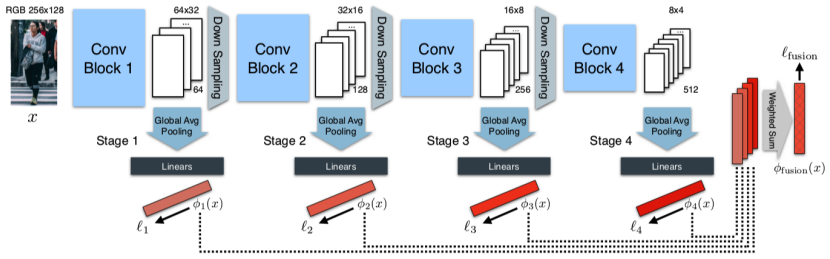


Figure 2.10: Illustration of the person re-id model proposed in [158]. Features from multiple convolutional network layers are fused together in the final loss function to train the model.

provided, *e.g.*, re-scaling an arbitrary patch of the person to a fixed-size image, Sliding Window Matching [194] and part-based model. These three solutions are illustrated in Fig. 2.9. However, all of them suffer from some drawbacks. For example re-scaling an arbitrary patch cannot deal with the case when the size of the probe person is bigger than the size of the gallery person and part-based model has a high computational cost. Deep Spatial feature Reconstruction (DSR) solves these problem by take advantage of both Fully Convolutional Network (FCN) and dictionary learning. FCN is utilized to generate spatial feature maps of certain sized. With the dictionary learning, each pixel in the probe spatial maps can be sparsely reconstructed on the basis of spatial maps of gallery images, and thus, the model is independent of the size of images and naturally avoids the time-consuming alignment step.

Multi-level features are considered in [13, 46, 158] for person re-id. Different-level features contains different discriminative information for re-identifying a person. In [13], the Multi-Level Factorisation Net (MLFN) is proposed in which a network architecture is designed to factorise the visual appearance of a person into latent discriminative factors at multiple semantic levels without manual annotation. Guo *et al.* propose an end-to-end fully convolutional Siamese network that computes the similarities at multiple levels [46]. According to their experimental results, the bottom convolutional layers contain low level visual information while the higher layers contain semantical information. The work [158] proposes to solve the resolution variations in person re-id by combining effective embeddings built on multiple convolutional network layers, trained with deep-supervision. Fig. 2.10 illustrates the person re-id model proposed in [158]. Li *et al.* designed a harmonious attention deep network for joint learning of soft pixel attention and hard regional attention along with simultaneous optimisation of feature representations [74]. Zheng *et al.* designed a pose fusion CNN architecture to reduce the impact of pose estimation errors and information loss in person re-id task [190].

As in handcrafted feature extraction, part-to-part matching is an intuitive idea for person re-id [31, 25]. This is also exploited under deep neural network framework [186, 187, 140, 142,

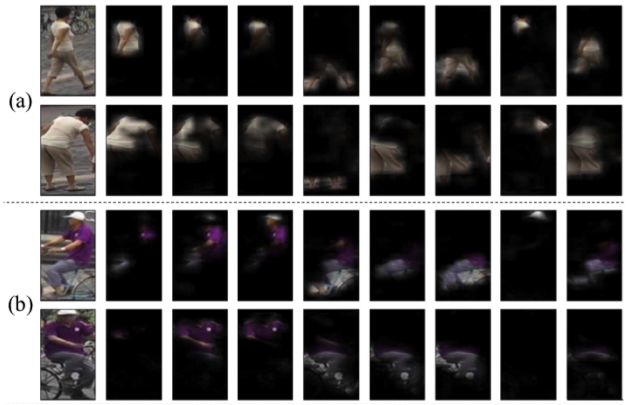


Figure 2.11: Examples of body regions extracted using the model in [187].

185, 174]. In [186], Zhao *et al*. proposed a novel CNN (Spindle Net) which is based on human body region guided multi-stage feature decomposition and tree-structured competitive feature fusion. Spindle Net features with (1) it separately captures semantic features from different body regions thus the macro- and micro-body features can be well aligned across images, and (2) the learned region features from different semantic regions are merged with a competitive scheme and discriminative features can be well preserved. The work [187] provides an approach to decompose the human body into regions (parts) which are discriminative for person matching. The proposed model can extract features from each regions and concatenate them together for formulating the final feature representation for person re-id. Several examples of the extracted body regions are presented in Fig. 2.11. Suh *et al*. proposed to align person body parts based on 2D pose estimation results [140]. The proposed model consists of a two-stream network, which generates appearance and body part feature maps respectively. A bilinear-pooling layer is added after two network streams for fusing two feature maps into one image descriptor. The network architecture is presented in Fig. 2.12.

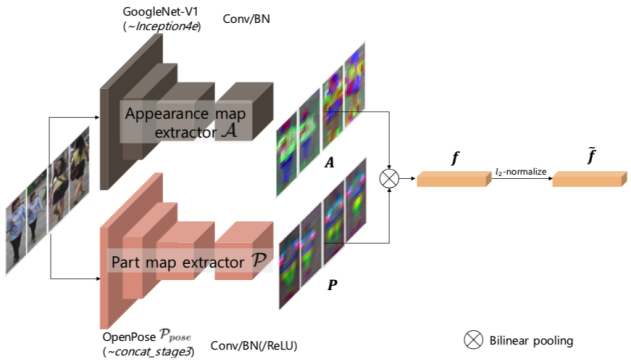


Figure 2.12: The network architecture proposed in [140].

2.3.2 Unsupervised domain-adaptive Person Re-Identification

Unsupervised domain adaptation is proposed to solve the domain shift problem when the label information is not provided in the target domain. The basic idea behind lots of domain adaptation works is to match the feature distributions in the source and target domains [35, 26, 152, 141]. Ganin *et al*. proposed an unsupervised domain adaptation method that can simultaneously train the network to learn discriminative and domain invariant feature representations by using the gradient reversal layer [35]. In works [141, 111], the correlation alignment (CORAL) is considered for unsupervised adaptation. CORAL minimizes the domain shift by aligning the second-order statistics of the source and target domains. In recent years, with the emergence of generative adversarial networks (GANs) [41], there are also some works trying to use adversarial adaptation strategy for adapting the network from source to target domain. Tzeng *et al*. proposed an adversarial adaptation framework in which the network is firstly trained in the source domain and then adversarially adapt to the target domain with the help of a discriminator as in GANs [41, 151]. The work [152] extends this adversarial adaptation framework with adding the step to train a feature generator and the domain invariant features can be learned.

In order to improve the generalization capability of re-id model and also reduce the human efforts consumed in annotating dataset, unsupervised domain-adaptive technologies are also exploited in person re-id [29, 199, 179, 155, 81, 180, 178, 200, 134, 164, 181]. The works [155, 81] try to adapt the re-id model by using both identity and attribute labels. Wang *et al*. developed a neural network method in which two network branches are included [155]. As shown in Fig. 2.13, one is for learning discriminative features under the supervision of identity labels and the other one is applied for learning the discriminative features under the supervision of attribute labels. Considering the fact that attribute label is consistent across domains while identity label is not, an encoder-decoder network is designed to bridge the gap

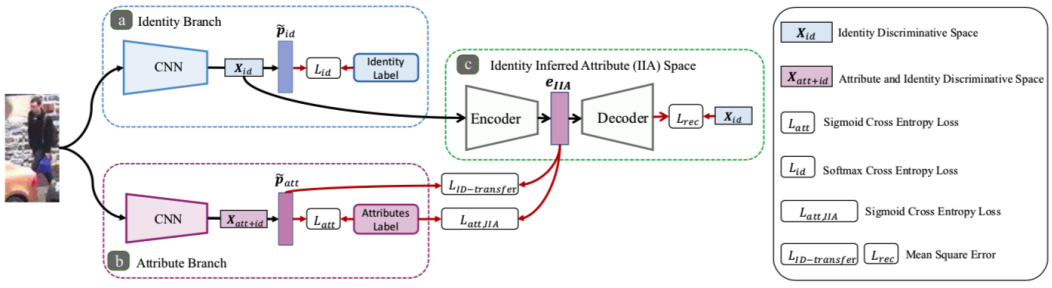


Figure 2.13: The re-id model proposed in [155].

between identity and attribute information, and then both identity and attribute knowledge are fused into one network branch which is further adapted to the target domain for person re-id. Lin *et al*. proposed to fuse attribute and identity knowledge by using a multi-task network framework and multi-level features are aligned for training a re-id model which works in both source and target domains [81]. Yu *et al*. proposed a deep model for unsupervised domain-adaptive person re-id using soft multilabel learning [179]. The soft multi-labels are obtained in the target domain by comparing the unlabeled person with a set of known reference persons from an auxiliary domain. Then the soft multilabel-guided hard negative mining is applied to learn a discriminative embedding for the unlabeled target domain by exploring the similarity consistency of the visual features and the soft multilabels of unlabeled target pairs.

The re-id works [155, 81, 179] try to adapt re-id model on the feature level while some other works are trying to adapt re-id model by transforming the image style between source and target domain [29, 4, 160, 22]. In [29], Deng *et al*. proposed a similarity preserving image-image translation model to transform the image style from the source domain to target domain. With the designed image-image translation model, two types of unsupervised similarities can be preserved: (1) self-similarity of an image before and after translation, and (2) domain-dissimilarity of a translated source image and a target image. The re-id model is trained based on the translated images. The diagram is illustrated in Fig. 2.14. Wei *et al*. also consider generative model for unsupervised domain adaptation [160]. A Person Transfer Generative Adversarial Network (PTGAN) has been proposed to transform the image style in the source domain to the target domain. The transferred persons from A can still keep their identities, meanwhile present similar styles, e.g., backgrounds, lightings, *etc*, with persons in B. In [164], Wu *et al*. proposed to solve unsupervised domain-adaptive person re-id problem by distilling the knowledges from teacher networks which have been trained on fully labeled re-id datasets. Although there are several large-scale person re-id datasets, it still cannot include enough variations for training a discriminative and robust re-id model. Based on this observation, the work [4] considers using generative model to synthesize virtual humans and then create samples from these virtual humans with different variations, *e.g*., different person poses and illuminations. These synthesized samples are

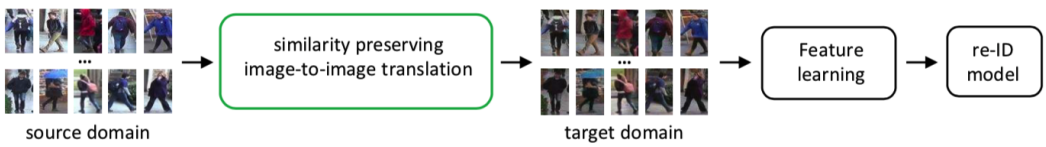


Figure 2.14: The pipeline of the re-id method proposed in [29].

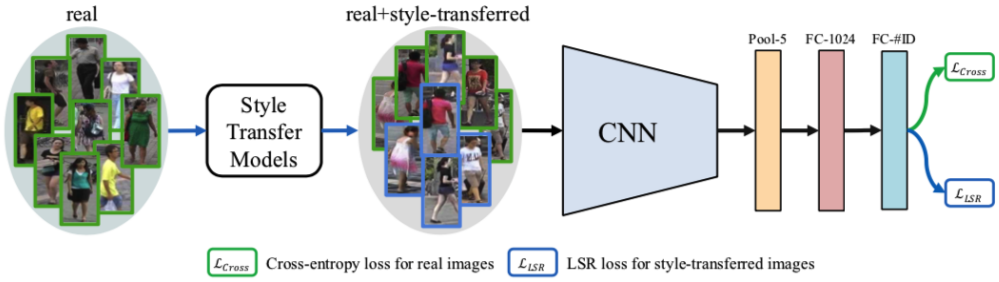


Figure 2.15: The pipeline of the re-id method proposed in [201].

transformed into the target re-id dataset by employing cycle-consistent adversarial networks. The translated images are then used to fine-tune the person re-id model.

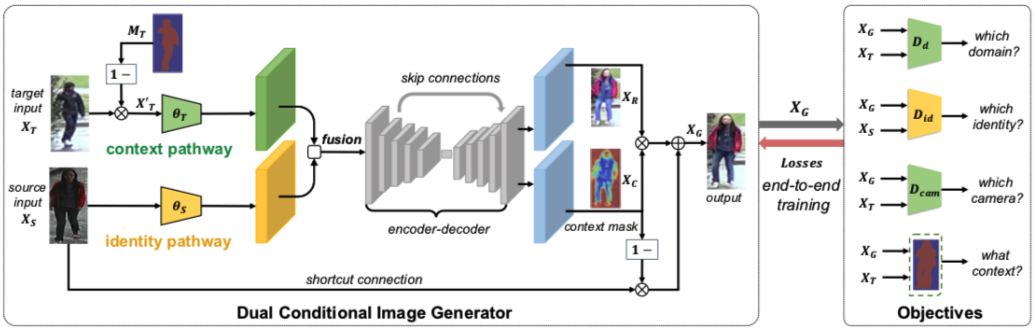


Figure 2.16: The re-id model proposed in [22].

In addition to dataset domain variations, several works also consider camera domain variations for improving re-id model generalization ability [199, 22, 77, 201]. Zhong *et al.* consider to take advantages camera-style variations to augment the re-id dataset and thus improve the discriminative and robust of the extracted re-id features [201]. The pipeline is illustrated in Fig. 2.15. Chen *et al.* proposed to generate person images with same person appearance but different contextual variations, *e.g.*, background and illuminations [22]. This is based on the observation that in open surveillance camera system, the contextual variations can be quite diverse, due to wide-of-the-field imagery and varying times of the day. Camera-view variation is also considered in the person image generation model. The designed model is illustrated in Fig. 2.16 In work [199], a Hetero-Homogeneous Learning (HHL) method is proposed. The method simultaneously can obtain two domain invariances: (1) camera invariance, learned via positive pairs formed by unlabeled target images and their camera style transferred counterparts; (2) domain connectedness, by regarding source/target images as negative matching pairs to the target / source images. The first property is obtained by homogeneous learning because training pairs are collected from the same domain

while the second property is achieved by heterogeneous learning because training pairs are sampled from both the source and target domains.

2.3.3 Other Methods

In addition to the fully supervised and unsupervised domain-adaptive person re-id works based on deep learning framework, there are also some other works trying to reduce the human efforts consumed in dataset annotation and improve the discriminative power of re-id features. These works include semi-supervised person re-id, weakly supervised person re-id, unsupervised person re-id, unsupervised tracklet learning person re-id.

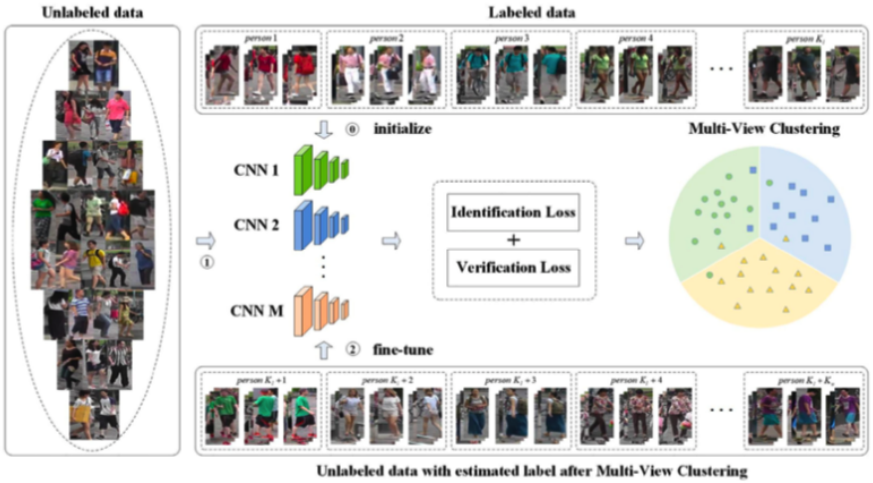


Figure 2.17: The semi-supervised person re-id method proposed in [167].

A typical strategy for reducing label supervision is by semi-supervised learning. The key idea is to self-mine supervision information from unlabelled training data based on the knowledge learned from a small proportion of labelled training data. The work [167] approaches the semi-supervised person re-id problem by constructing a set of heterogeneous CNNs fine-tuned using the labeled portion, and then propagating the labels to the unlabeled portion for further fine-tuning the overall system. A novel multi-view clustering method is proposed for estimating labels of the unlabeled samples. The method is presented in Fig. 2.17. Weakly supervised person re-id aims at training re-id model based on weakly labeled data. In [108], Meng *et al.* proposed a weakly supervised person re-id paradigm where the identity labels are annotated at the untrimmed video level. Unsupervised model learning is an intuitive solution to avoid the need of exhaustively collecting a large number of labelled training data for every application domain. Compared to the supervised learning methods, early hand-crafted feature based unsupervised learning methods

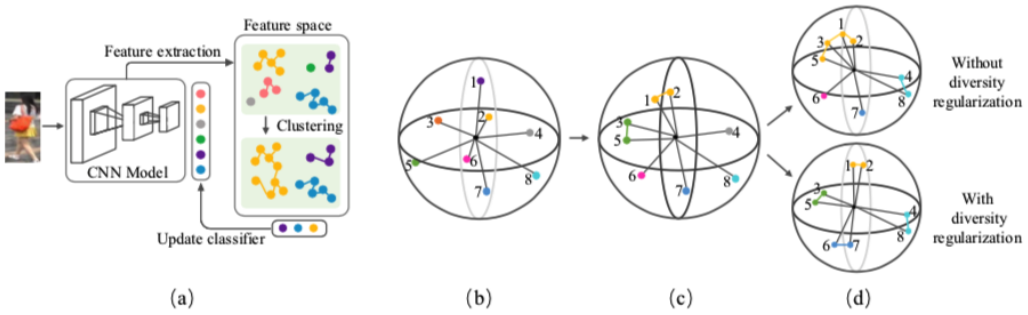


Figure 2.18: The unsupervised person re-id method proposed in [82]. (a) The unlabeled images are used as input to train the network and then features of all training images are extracted for clustering. Fig. (b)-(d) depict the cluster merging procedure.

[153, 59, 58, 56, 101, 176, 90] offer significantly inferior re-id matching performance. The deep learning based method [82] reduces this performance gap, in which the re-id model is first trained using the instance loss by treating each image belonging to one single unique identity. After pre-training the model, the clustering method is applied for further learning the discriminative features for person re-id. This unsupervised person re-id method is illustrated in Fig. 2.18. Instead of assuming transferable source domain training data, a small number of methods [67, 68, 21] leverage the auto-generated tracklet data with rich spatio-temporal information for unsupervised re-id model learning. In many cases this is a feasible solution as long as video data are available. However, it remains highly challenging to achieve good model performance due to noisy tracklets with unconstrained dynamics.

Chapter 3

Gaussian Mixture Importance Estimation

3.1 Introduction

Despite the devoted efforts, re-id remains a rather challenging task due to the nonrigid structure of the human body, the different perspectives in which a pedestrian can be observed, and the highly variable illumination conditions, which are also introduced in Chapter 1.

Most of the re-id works suggest that, normally, individuals do not change their clothings across a camera network. This assumption inspired the contributions to regard the visual aspect as a main cue in chacarterizing person images. In particular, the mainstream re-id literature can be broadly categorized in two classes, namely *direct* and *learning-based methods*. The former group tends to handcraft, robust features and potentially their combination thereof. In the latter group, i.e., *learning based methods*, a dataset of similar and dissimilar persons is used to 'learn' personalized features and/or a metric space where to match them. The underlying assumption is that the knowledge extracted from the training set generalizes to unseen samples.

In this chapter, a new re-id method has been proposed, *i.e.*, Gaussian mixture importance estimation (GMIE, for short), which is also based on likelihood ratio test, inspired by KISSME. However, GMIE offers several advantages with respect to KISSME. As described in [61], KISSME uses the Gaussian densities to separately approximate the distributions of intrapersonal and interpersonal variations. In this case, if the intrapersonal and/or interpersonal variations are characterized by a multi-modal distribution, which is commonly encountered in practice, the distribution approximation in KISSME would be inaccurate based only on Gaussian densities. Fig. 3.1 demonstrates a toy example of approximating a bimodal distribution with Gaussian density in a one projected dimension case. From the figure, it can be observed that the estimated (red) curve does not accurately fit the actual (blue) one. In addition, what is important in KISSME is to estimate the covariance matrices of intrapersonal and interpersonal variations. In practice, due to the small sample size problem in re-id dataset, it is always difficult to accurately estimate these two covariance matrices, especially in high-dimensional cases. Although the principal component analysis (PCA) is used for

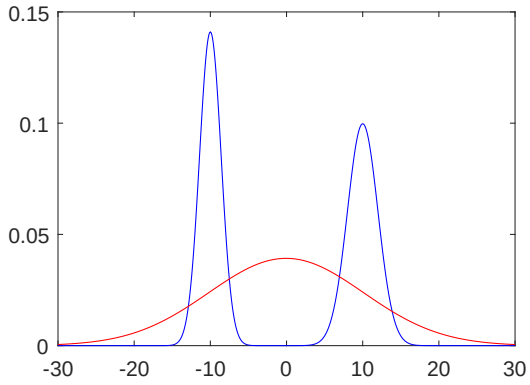


Figure 3.1: Illustration example of approximating a distribution with two modes using Gaussian density. Blue line is the true probability density and the red line is the approximated probability density.

dimension reduction, KISSME remains sensitive to the feature dimension as reported in the existing works [78, 105]. Considering these disadvantages in KISSME, we introduce a novel re-id method based on Gaussian mixture importance estimation, which is robust to feature dimension. Unlike KISSME, our method directly estimates the ratio of the aforementioned probability densities via GMMs, which maintains the performance as the feature dimension rises. Rigorous experiments are performed to validate the advantages of our approach over existing alternatives on multiple benchmark datasets.

3.2 Background

The proposed GMIE is related to the Mahalanobis metric, KISSME [61] and Gaussian mixture models. As in KISSME, GMIE is also based on the likelihood ratio between intrapersonal and interpersonal probability densities. However, in order to capture the multi-modes in the data structure, the Gaussian mixture models is used in GMIE. In the following, the brief descriptions about KISSME and Gaussian mixture models will be provided as the backgrounds of the proposed methodology.

3.2.1 KISS Metric Learning

Given a pair of labeled samples $\{\mathbf{z}_i, y_i\}$ and $\{\mathbf{v}_j, l_j\}$, in which y_i and l_j denote the label of person identities corresonding to the i-th and j-th person images, respectively. $\mathbf{z}_i$ and $\mathbf{v}_j$ are the feature vectors extracted from i-th and j-th person image in the corresponding camera-

view. The difference between these two feature vectors is calculated as $\mathbf{x}_{ij} = \mathbf{z}_i - \mathbf{v}_j$. If $y_i = l_j$, $\mathbf{x}_{ij}$ is called the intrapersonal difference, and if $y_i \neq l_j$, $\mathbf{x}_{ij}$ is called the interpersonal difference. In the following, the indexes i, j are omitted in $\mathbf{x}_{ij}$ for simplicity. Let Ω_I denotes the class that includes the intrapersonal differences, while Ω_E denotes the class that holds the interpersonal differences. In KISSME, the likelihood ratio test is used to determine if $\mathbf{x}$ belongs to Ω_E. The likelihood ratio is:

$$\delta(\mathbf{x}) = \log\left(\frac{p_E(\mathbf{x})}{p_I(\mathbf{x})}\right), \tag{3.1}$$

where $p_E(\mathbf{x})$ is the probability density for class Ω_E and $p_I(\mathbf{x})$ is for Ω_I. Since the set of sample difference $\mathbf{x}$ is zero mean, the approximations of $p_E(\mathbf{x})$ and $p_I(\mathbf{x})$ with Gaussian densities can be formulated as:

$$p_E(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}_E|^{1/2}} \exp(-\frac{1}{2}\mathbf{x}^T\mathbf{\Sigma}_E^{-1}\mathbf{x}), \tag{3.2}$$

$$p_I(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}_I|^{1/2}} \exp(-\frac{1}{2}\mathbf{x}^T\mathbf{\Sigma}_I^{-1}\mathbf{x}), \tag{3.3}$$

in which $\mathbf{\Sigma}_E$ and $\mathbf{\Sigma}_I$ represent the covariance matrices of Ω_E and Ω_I, respectively, and d is the dimension of $\mathbf{x}$.

Substitute Eqs. (3.2)-(3.3) into Eq. (3.1), the likelihood ratio can be reformulated as:

$$\delta(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T(\mathbf{\Sigma}_I^{-1} - \mathbf{\Sigma}_E^{-1})\mathbf{x} + \frac{1}{2}\log\left(\frac{|\mathbf{\Sigma}_I|}{|\mathbf{\Sigma}_E|}\right). \tag{3.4}$$

Removing the constant terms, we can get the simplified likelihood ratio:

$$\delta(\mathbf{x}) = \mathbf{x}^T(\mathbf{\Sigma}_I^{-1} - \mathbf{\Sigma}_E^{-1})\mathbf{x}. \tag{3.5}$$

To ensure the nonnegative value of $\delta(\mathbf{x})$, KISSME further re-projects $\mathbf{M} = \mathbf{\Sigma}_I^{-1} - \mathbf{\Sigma}_E^{-1}$ into the cone of positive semi-definite matrix.

With Eq. (3.5), the likelihood ratio $\delta(\mathbf{x})$ is converted into the calculations of the covariance matrices $\mathbf{\Sigma}_E$ and $\mathbf{\Sigma}_I$. However, there is a high computation requirement for calculating these two covariance matrices. Suppose there are two sets $\mathbf{Z}$ and $\mathbf{V}$ respectively containing the samples from two different camera views. Specifically, $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n) \in \mathbb{R}^{d\times n}$ and $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m) \in \mathbb{R}^{d\times m}$. The computations of $\mathbf{\Sigma}_E$ and $\mathbf{\Sigma}_I$ require $O(Nkd^2)$ and $O(nmd^2)$ multiplication operations, respectively, in which $N = \max(m, n)$, and k is the average number of images for each class (identity). To reduce the computation, the work

[78] proposed to derive $\mathbf{\Sigma}_E$ and $\mathbf{\Sigma}_I$ as:

$$n_I \mathbf{\Sigma}_I = \tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^T + \tilde{\mathbf{V}}\tilde{\mathbf{V}}^T - \mathbf{S}\mathbf{R}^T - \mathbf{R}\mathbf{S}^T \tag{3.6}$$

in which

$$\begin{cases} \tilde{\mathbf{Z}} = (\sqrt{m_1}\mathbf{z}_1, \sqrt{m_1}\mathbf{z}_2, \ldots, \sqrt{m_1}\mathbf{z}_{n_1}, \ldots, \sqrt{m_c}\mathbf{z}_n) \\ \tilde{\mathbf{V}} = (\sqrt{n_1}\mathbf{v}_1, \sqrt{n_1}\mathbf{v}_2, \ldots, \sqrt{n_1}\mathbf{z}_{m_1}, \ldots, \sqrt{n_c}\mathbf{v}_m) \\ \mathbf{S} = (\sum_{y_i=1}\mathbf{z}_i, \sum_{y_i=2}\mathbf{z}_i, \ldots, \sum_{y_i=c}\mathbf{z}_i) \\ \mathbf{R} = (\sum_{l_i=1}\mathbf{v}_i, \sum_{l_i=2}\mathbf{v}_i, \ldots, \sum_{l_i=c}\mathbf{v}_i) \end{cases} \tag{3.7}$$

in which n_k is the number of samples of class k in $\mathbf{Z}$ and similarly, m_k is the number of samples of class k in $\mathbf{V}$ and c is the number of classes.

Based on $\mathbf{\Sigma}_I$, $\mathbf{\Sigma}_E$ can be derived as:

$$n_E \mathbf{\Sigma}_E = m\mathbf{Z}\mathbf{Z}^T + n\mathbf{V}\mathbf{V}^T - \mathbf{s}\mathbf{r}^T - \mathbf{r}\mathbf{s}^T - n_I\mathbf{\Sigma}_E \tag{3.8}$$

in which

$$\begin{cases} \mathbf{s} = \sum_{i=1}^{n}\mathbf{z}_i \\ \mathbf{r} = \sum_{i=1}^{m}\mathbf{v}_i \end{cases} \tag{3.9}$$

With Eqs. (3.6) and (3.8), the computations of $\mathbf{\Sigma}_E$ and $\mathbf{\Sigma}_I$ are both reduced to $O(Nd^2)$. In addition, Eqs. (3.6) and (3.8) also show that both $\mathbf{\Sigma}_E$ and $\mathbf{\Sigma}_I$ can be computed from the sample mean and covariance of each class and all classes. Thus, there is no need to actually calculate the mn pairs of sample differences as in the original KISS metric learning [61].

3.2.2 Gaussian Mixture Models

The Gaussian Mixture Models (GMMs) is a parametric probability density function that is represented with a weighted sum of Gaussian components. It has the wide applications in different areas involving clustering and classification, for example pattern recognition, data mining, image analysis and machine learning *etc*.

Given the feature vectors $\mathbf{x}$ as in subsection 3.2.1, the GMMs can be formulated as:

$$f(\mathbf{x}) = \sum_{l=1}^{b}\pi_l N(\mathbf{x}|\mathbf{u}_l, \mathbf{\Sigma}_l), \tag{3.10}$$

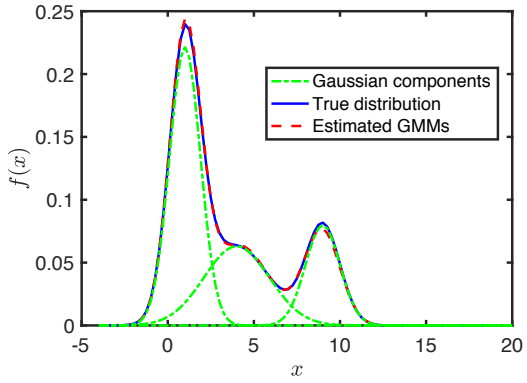


Figure 3.2: Illustration example of approximating a one dimensional distribution with GMMs (composed of three Guassian components). There are multi-modes existed in the true distribution denoted with blue line. The approximated distribution is denoted using dashed red line which is composed of three weighted Gaussian components which are denoted using dashed green lines.

where π_l, $l \in [1, 2, \ldots, b]$ are the weights, and b is the number of Gaussian mixture components. $\mathbf{u}_l$ and $\mathbf{\Sigma}_l$ respectively denote the mean vector and covariance matrix of the l-th Gaussian component which can be formulated as:

$$N(\mathbf{x}|\mathbf{u}_l, \mathbf{\Sigma}_l) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{u}_l)^T\mathbf{\Sigma}_l^{-1}(\mathbf{x}-\mathbf{u}_l)\right)}{(2\pi)^{d/2}|\mathbf{\Sigma}_l|^{1/2}}. \tag{3.11}$$

where d is the dimension of the feature vector $\mathbf{x}$.

The estimation of GMMs $f(\mathbf{x})$ relies on estimating the parameters π_l, $\mathbf{u}_l$ and $\mathbf{\Sigma}_l$ ($l \in [1, 2, \ldots, b]$) and usually they can obtained using Expectation-Maximization (EM) algorithm based on training data. Fig. 3.2 gives an illustration example of using GMMs to approximate a multi-mode distribution. From the figure, it can be observed that GMMs can effectively capture the multi-modes in the true data distribution and thus offers a more accurate distribution approximation compared with using Gaussian density (only with one Gaussian component).

3.3 Our Proposed Approach

Based on the aforementioned background, the proposed approach, *i.e.*, Gaussian mixture importance estimation (GMIE), aims at directly approximating the ratio of intrapersonal and interpersonal probability densities $p_E(\mathbf{x})$ and $p_I(\mathbf{x})$ using the GMMs. In this way, it avoids explicitly estimating $p_I(\mathbf{x})$ and $p_E(\mathbf{x})$ and effectively improves its robustness to

high dimensional features as demonstrated in the experiments. In addition, using GMMs as approximation function, GMIE can also capture the multi-modes existed in $p_I(\mathbf{x})$ and/or $p_E(\mathbf{x})$. The details of GMIE is introduced in the following.

3.3.1 Gaussian Mixture Importance Estimation

Different from KISSME, we here consider the likelihood of $\mathbf{x}$ belonging to Ω_I. Thus, the ratio used in GMIE between $p_E(\mathbf{x})$ and $p_I(\mathbf{x})$ is:

$$w(\mathbf{x}) = \frac{p_I(\mathbf{x})}{p_E(\mathbf{x})}. \tag{3.12}$$

As aforementioned, GMIE approximates the ratio $w(\mathbf{x})$ via the GMMs, which combines a number of Gaussian components. Hence, the approximation of $w(\mathbf{x})$ is:

$$\hat{w}(\mathbf{x}) = \sum_{l=1}^{b} \pi_l N(\mathbf{x}|\mathbf{u}_l, \mathbf{\Sigma}_l), \tag{3.13}$$

where as in Eq. (3.10), π_l ($l \in [1, 2, \ldots, b]$) are the weights, and b is the number of Gaussian components. $\mathbf{u}_l$ and $\mathbf{\Sigma}_l$ respectively denote the mean vector and covariance matrix of the l-th Gaussian component.

Based on Eq. (3.13), the likelihood ratio as in Eq. (3.5) can be reformulated as:

$$\delta(\mathbf{x}) = \log w(\mathbf{x}) \approx \log \sum_{l=1}^{b} \pi_l N(\mathbf{x}|\mathbf{u}_l, \mathbf{\Sigma}_l). \tag{3.14}$$

To estimate the parameters π_l, $\mathbf{u}_l$, $\mathbf{\Sigma}_l$ and b, the minimization of the Kullback-Leibler divergence from $p_I(\mathbf{x})$ to its approximation $\hat{p}_I(\mathbf{x})$ is used:

$$\mathrm{KL}[p_I(\mathbf{x})||\hat{p}_I(\mathbf{x})] = \int p_I(\mathbf{x}) \log \frac{p_I(\mathbf{x})}{\hat{p}_I(\mathbf{x})} d\mathbf{x}. \tag{3.15}$$

Take Eqs. (3.12) and (3.13) into consideration, the estimation of $\hat{p}_I(\mathbf{x})$ can be formulated as:

$$\hat{p}_I(\mathbf{x}) = \hat{w}(\mathbf{x}) p_E(\mathbf{x}). \tag{3.16}$$

Substitute Eq. (3.16) into Eq. (3.15), the Kullback-Leibler divergence can be reformulated

as:

$$\begin{aligned}\mathrm{KL}[p_I(\mathbf{x})||\hat{p}_I(\mathbf{x})] = &\int p_I(\mathbf{x})\log\frac{p_I(\mathbf{x})}{p_E(\mathbf{x})}d\mathbf{x} \\ &- \int p_I(\mathbf{x})\log\hat{w}(\mathbf{x})d\mathbf{x}.\end{aligned} \tag{3.17}$$

The unknown parameters are only contained in the second term of the equation. Thus, minimizing Kullback-Leibler divergence equals to maximizing the second term as:

$$J = \int p_I(\mathbf{x})\log\hat{w}(\mathbf{x})\, d\mathbf{x} = \frac{1}{n_I}\sum_{i=1}^{n_I}\log\hat{w}(\mathbf{x}_i^I), \tag{3.18}$$

where n_I is the number of samples in Ω_I, and $\mathbf{x}_i^I$ denotes the i^{th} sample from Ω_I.

Since $p_I(\mathbf{x})$ is a probability density, thus the following constraint should be held:

$$\begin{aligned}1 = &\int \hat{p}_I(\mathbf{x})\, d\mathbf{x} = \int \hat{w}(\mathbf{x})p_E(\mathbf{x})\, d\mathbf{x} \\ \approx &\frac{1}{n_E}\sum_{j=1}^{n_E}\hat{w}(\mathbf{x}_j^E),\end{aligned} \tag{3.19}$$

where n_E is the number of samples in Ω_E, and $\mathbf{x}_j^E$ denotes the the j^{th} sample from Ω_E.

Consider Eqs. (3.13), (3.18) and (3.19), the optimization problem of GMIE can be formulated as:

$$\begin{aligned}\max_{\{\pi_l,\mathbf{u}_l,\mathbf{\Sigma}_l\}_{l=1}^b} \quad & \sum_{i=1}^{n_I}\log\left(\sum_{l=1}^{b}\pi_l N(\mathbf{x}_j^I|\mathbf{u}_l,\mathbf{\Sigma}_l)\right) \\ \text{s.t.} \quad & \sum_{j=1}^{n_E}\sum_{l=1}^{b}\pi_l N(\mathbf{x}_j^E|\mathbf{u}_l,\mathbf{\Sigma}_l) = n_E \\ & \pi_1,...,\pi_b \geq 0.\end{aligned} \tag{3.20}$$

The parameters $\boldsymbol{\mu}_l$, $\mathbf{\Sigma}_l$ and π_l can be estimated by employing the Lagrangian multiplier method on this optimization problem. More detailed estimation procedure can be found in [170].

The parameter b can be determined through the likelihood cross validation method as introduced in [139, 170]. Considering the small simple size problem in re-id, we determine b by providing several candidates and select the one that maximize J in Eq. (3.18). It is worth to note that a large b will lead to the over fitting problem.

3.3.2 Person Re-Identification Process Using GMIE

With the estimated GMMs parameters, *i.e.*, $\boldsymbol{\mu}_l$, $\boldsymbol{\Sigma}_l$ and π_l ($l \in [1, 2, \ldots, b]$), the likelihood ratio of two given samples $\mathbf{z}_i$ and $\mathbf{v}_j$ can be calculated using:

$$\delta(\mathbf{x}_{ij}) \approx \log \sum_{l=1}^{b} \pi_l N(\mathbf{x}_{ij}|\mathbf{u}_l, \boldsymbol{\Sigma}_l). \tag{3.21}$$

in which $\mathbf{x}_{ij} = \mathbf{z}_i - \mathbf{v}_j$. $\mathbf{z}_i$ and $\mathbf{v}_j$ can be obtained using the handcrafted feature extraction methodology, for example LOMO [78] and GOG [105]. A high value of $\delta(\mathbf{x}_{ij})$ means that $\mathbf{z}_i$ and $\mathbf{v}_j$ are highly similar to each other and thus has the high probability to be the same person. The re-id procedure of GMIE is summarized in Algorithm 1.

Algorithm 1 Person re-id procedure of GMIE.

Training:
step 1: Extract handcrafted features from training dataset;
step 2: Calculate differences **x** between feature vectors;
step 3: Estimate GMMs parameters by solving the optimization problem in Eq. (3.20).

Evaluation:
step 1: Extract handcrafted features from test dataset;
step 2: Calculate likelihood ratios between query and gallery images using Eq. (3.21);
step 3: Rank the similarities between query and gallery images based on likelihood ratios.

3.4 Experiments

We assess our introduced method on three datasets, i.e., VIPeR dataset [43], GRID dataset [85] and PRID 450S dataset [123]. In the experiments, b is determined using five candidates, i.e., $b = [1, 2, 3, 4, 5]$, and the one that maximizes J is selected as mentioned earlier. The evaluation procedure in every dataset is repeatedly carried out 10 times. It is to note that PCA is employed for feature dimension reduction. We report the results in terms of re-id rate by means of Cumulative Matching Characteristic (CMC) curve, which accumulates the re-id accuracy as the rank index increases. For a through analysis, the performance of our method is evaluated from three different aspects, i.e. robustness to subspace dimensions, effect of different descriptors and re-id rate. The details and discussions on each dataset are presented in following three subsections.

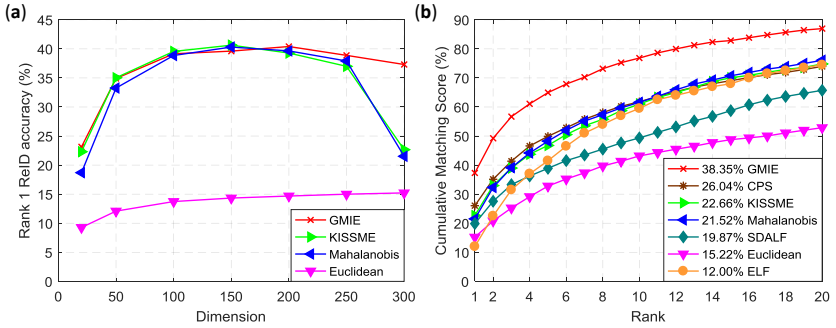


Figure 3.3: Experimental results on VIPeR dataset with GOG feature: (a) dimensionality influence; (b) CMC curves with 300 dimensions.

3.4.1 Experiments on VIPeR Dataset

Fig. 3.3 shows evaluation results on VIPeR dataset by adopting the Gaussian Of Gaussian (GOG) feature [105]. Its sub-figure (a) presents the trend of rank 1 scores with increasing the subspace dimension. For comparison, the KISSME, Mahalanobis distance trained with genuine pairs [61] and Euclidean distance are implemented. It can be found in the figure that KISSME and Mahalanobis distance work well only within the range from 100^{th} to 250^{th} dimensions. For the dimension larger than 250, their rank 1 scores decrease dramatically since the estimated covariance matrices become inaccurate in high dimensional subspace. However, the proposed GMIE method estimates the parameters based on the Kullback-Leibler divergence instead of a direct inference from samples. This renders it robust to the dimension increase which is also demonstrated in the Fig. 3.3 (b). Although the Euclidean distance is also robust to dimensionality, its re-id accuracy is relatively low. We further compare our method for a dimension of 300 with other methods, i.e., CPS [24], SDALF [31], and ELF [44]. The CMC curves for these considered methods are displayed in Fig. 3.3 (b). It can be seen that GMIE works better than the considered methods, often by a large margin.

To assess the performance of our method in other feature descriptors, we also repeat the experiments on VIPeR dataset using the LOMO feature [78]. The experimental results are depicted in Fig. 3.4. From the figure, the same behavior is observed. It is to note from these outcomes that GOG seems to incur more robustness on the KISSME and Mahalanobis distance.

3.4.2 Experiments on GRID Dataset

In the experiment, the KISSME, Mahalanobis and Euclidean distance are also included for comparison as in VIPeR dataset. To evaluate the effectiveness of GMIE on different feature

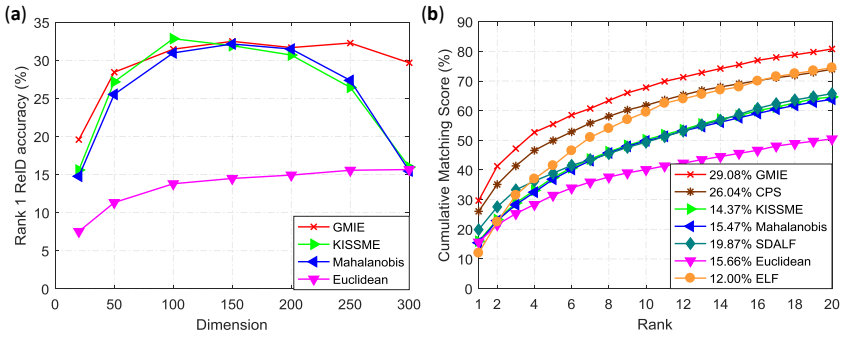


Figure 3.4: Experimental results on VIPeR dataset with LOMO feature: (a) dimensionality influence; (b) CMC curves with 300 dimensions.

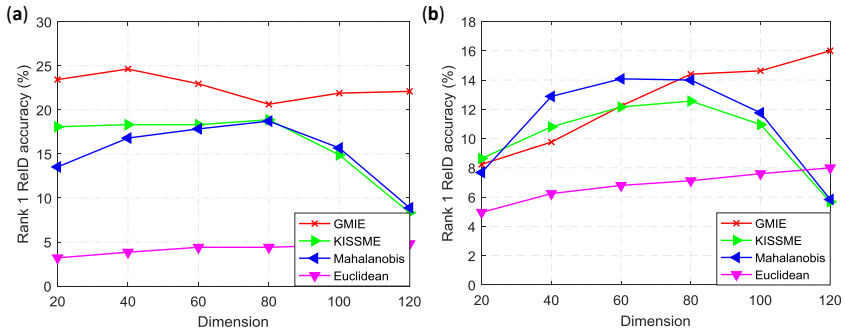


Figure 3.5: Experimental results of dimensionality influence on grid dataset with (a) GOG feature and (b) LOMO feature.

descriptors, the experiments are performed with both GOG and LOMO features. The results are presented in Fig. 3.5. It can be viewed from the figure that KISSME and Mahalanobis distance are not robust to the dimension increase. In the high dimensional subspace (the dimension greater than 80), their re-id accuracies decrease largely, especially at 120-th dimension of LOMO feature, the re-id accuracies of KISSME and Mahalanobis distance are even lower than the Eluclidean's. However, GMIE always works well in the high subspace dimension both in GOG and LOMO cases as demonstrated in Fig. 3.5.

To make a comparison with the state of the art results reported on GRID dataset, we calculate the re-id accuracy in the subspace dimension with which the related method has the highest re-id accuracy, and the final results are listed in Tab. 3.1. The table shows that for the rank 1, re-id accuracy of GMIE (24.64%) almost equals to the highest state of art result (24.70%) reported by XQDA with using GOG feature [105]. However, for the rank 5, 10, 15 and 20, GMIE has the best accuracies compared with other considered methods. The main reason of the high re-id accuracy of GMIE traces back to the fact that GMIE is based on GMMs, which makes it more accurate in capturing the multi-modal properties in the densities $p_I(\mathbf{x})$ and $p_E(\mathbf{x})$ in Equ. (3.12).

Table 3.1: Experimental results on GRID dataset of different methods comparing with our GMIE approach (%).

Methods	Rank1	Rank5	Rank10	Rank15	Rank20
MRank-RankSVM [92]	12.24	27.84	36.32	42.24	46.56
MtMCML [100]	14.08	34.64	45.84	52.88	59.84
PolyMap [15]	16.30	35.80	46.00	52.80	57.60
SSDAL+XQDA [137]	22.40	39.20	48.00	–	58.40
KEPLER [104]	18.40	39.12	50.24	57.04	61.44
NLML [54]	24.54	35.86	43.53	–	55.25
DR-KISS [147]	20.60	39.30	51.40	–	62.60
SCSP [15]	24.24	44.56	54.08	–	59.68
GOG+KISS	18.32	39.36	51.84	58.48	63.44
GOG+XQDA [105]	**24.70**	47.00	58.40	–	69.00
GOG+GMIE (ours)	24.64	**49.52**	**63.84**	**69.48**	**73.32**

3.4.3 Experiments on PRID 450S Dataset

The experimental results are illustrated in Fig. 3.6. From its sub-figure (a), it can be found that KISS and Mahalanobis distance perform well only in the certain subspace dimension range (between 50-th and 150-th dimension) as in VIPeR and GRID datasets. For the low or high dimensions, their re-id accuracies decrease. However, GMIE still works well in the high dimensional case. It can be seen from the figure that the GMIE's trend of rank 1 re-id accuracy consistently keeps stable from 50-th dimension. By setting the subspace dimension to 220, the CMC curves for these four methods are plotted in Fig. 3.6 (b). As shown in the figure, Euclidean distance works better than both KISS and Mahalanobis distance in high dimensional subspace. GMIE always has the highest re-id accuracy compared with other considered methods.

From the experimental results on these three datasets, it can be found that KISS and Mahalanobis distance are sensitive to the subspace dimension. In particular their performances degrade significantly in the high dimensional case. Thus, in order to use the KISS and Mahalanobis distance for re-id task, a cross validation procedure may be necessary in the training stage in order to find an appropriate subspace dimension which will come at the cost of more complexity. However, our approach GMIE always performs well in the high dimensional subspace on all three datasets. In addition, GMIE also has high re-id accuracy, especially on the GRID datasets.

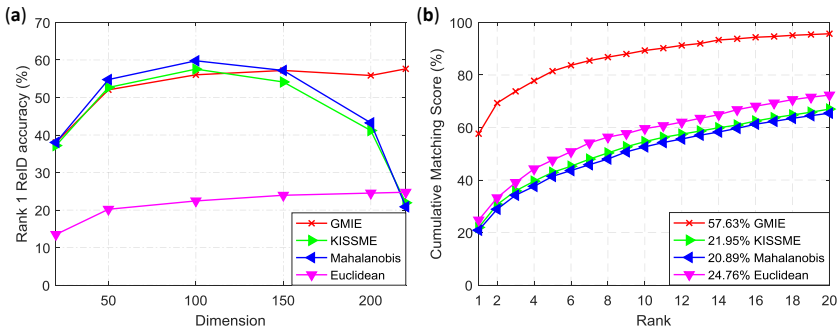


Figure 3.6: Experimental results on PRID 450S dataset with GOG feature: (a) dimensionality influence; (b) CMC curves with 220 dimensions.

3.5 Conclusions

In this chapter, a GMIE person re-id method has been proposed. Unlike KISSME, it directly approximates the density ratio between the intrapersonal and interpersonal variations. By adapting the Kullback-Leibler divergence technique, GMIE can maintain its re-id performance even in the high dimensional case, which is difficult for KISSME. In addition, thanks to the GMMs used for approximating the density ratio, GMIE is also capable of capturing the multi modal properties existed in the underlying densities of intrapersonal and interpersonal variations. These advantages of GMIE have been validated with detailed experiments on three datasets. The re-id accuracy of GMIE is also satisfactory compared with other works.

Chapter 4

Unsupervised Domain-Adaptive Person Re-identification Based on Attributes

4.1 Introduction

Pedestrian attributes, *e.g.*, hair length, clothes type and color, locally describe the appearance of a person. Training person re-id algorithms under the supervision of such attributes have proven to be effective in extracting the discriminative re-id features [64, 65, 136, 55, 83, 107]. Different from person identity, which denotes the global description of the person, attribute represents the local part of a person. For example, hair length refers to the head part while the up-body clothing type mainly focuses on the torso part. Thus, with the supervision of pedestrian attributes, the re-id model can learn the local semantic person appearance features.

However, most of datasets in re-id domain come without pedestrian attributes [192, 73, 189, 159]. In addition, attribute annotation is also a complex and time consuming task. Part of attributes, for example wearing sunglass or not and with or without backpack, are frequently not coherent along time since for the changes of illumination, person pose etc. Despite the effort of Lin *et al.* [83] in annotating two re-id datasets, *i.e.*, Market1501 [191] and DukeMTMC-reID [196], with attributes labels, there is still a big shortage of attribute-annotated large-scale re-id datasets. On the other hand, there is a number of attribute-labeled datasets in the domain of pedestrian attribute recognition [66, 75, 129]. Unfortunately, this kind of datasets cannot be directly exploited for re-id, since no identity label is provided and usually there is only one image for each identity.

As shown in Fig. 4.1, our work bridges the gap between attribute recognition and person re-id. In the proposed framework, an attribute recognition model is first trained on the attribute recognition dataset. Considering there is no attribute labels in the re-id dataset, unsupervised domain adaptation is applied for model adaptation. To learn the domain-invariant feature, a modified adversarial discriminative domain adaptation method is introduced based on [151, 152]. Compared with the original adversarial discriminative domain-adaptive (ADDA)

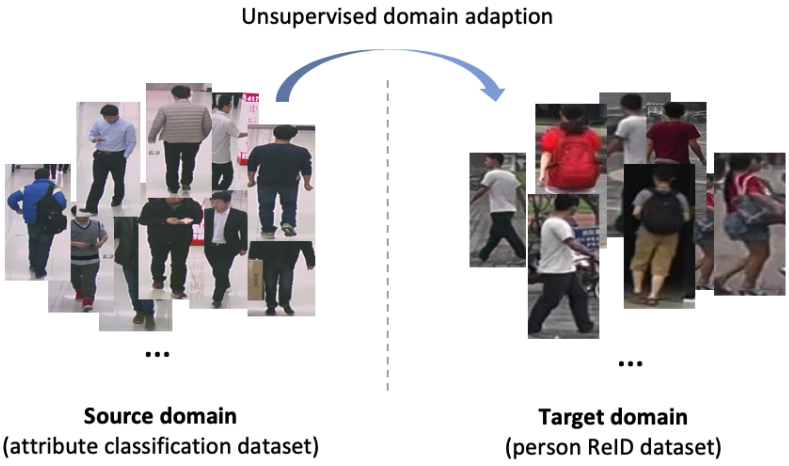


Figure 4.1: Overview of the considered re-id problem based on the annotated attribute recognition dataset. Source domain only labeled with attributes and no labels are needed in target domain. The images are picked from RAP (attribute classification dataset) [66] and Market1501 (re-id dataset) [191].

method in [151], both the source and target images are fed into the target model in the adversarial adaptation procedure of the proposed adaptation method. There are at least two advantages to do this. First, the domain-invariant capability of the model has been improved with that the adapted model is invariant to both source and target domains, while in ADDA, the adapted model (target CNN) can only be applied in the target domain [151]. The second advantage is that the modified ADDA is easier to converge compared with the original one according to our experimental results. The training of ADDA is separated into two steps, the model is first trained in the source domain and then it is adapted to the target domain. This stepwise training procedure can degrade the attribute recognition performance during domain adaptation and thus result in the decrease of re-id performance. In order to maintain the attribute recognition performance in the adversarial adaptation procedure, an additional classifier is added along with the discriminator. To evaluate the effectiveness of the proposed unsupervised adaptive re-id framework, three datasets have been used, namely the attribute recognition dataset RAP [66] and two re-id datasets, *i.e.*, Market-1501 [191] and DukeMTMC-reID [196].

4.2 Background

As aforementioned, in order to bridge the gap between attribute recognition and person re-id, unsupervised domain adaptation is applied in the proposed person re-id framework by extending an existing method, *i.e.*, ADDA [151]. Thus, before introducing the proposed

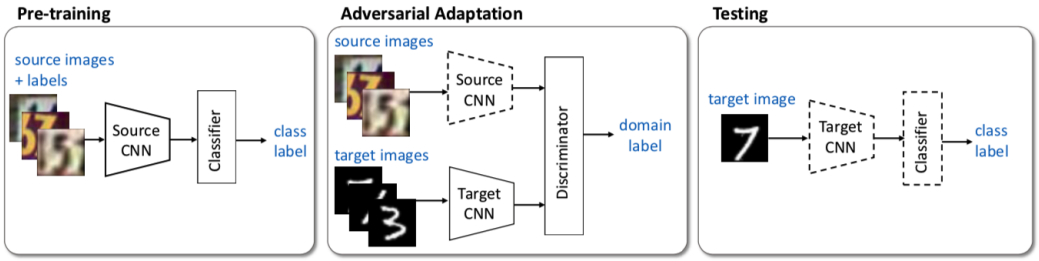


Figure 4.2: Overview of the ADDA method. The training and evaluation is separated into three steps, *i.e.*, pre-training, adversarial adaptation and testing. The figure is from [151].

re-id framework, the brief introductions about adversarial process [40] and ADDA will be made.

4.2.1 Adversarial Process

The adversarial training strategy has gained considerable attentions since the work [40] in which the effectiveness of adversarial process is validated to train the image generation nets. After this work, many efforts have been devoted in studying adversarial process from different machine learning and computer vision topics, *e.g.*, image generation [40], domain adaptation [151], feature learning, person re-id [204], face recognition *etc*.

Here the adversarial process is briefly introduced considering the image generation net in [40]. Suppose there is a image dataset $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$ with n samples. The generator is denoted as $G(\mathbf{z})$ which is used to generate images with inputing the noise that is sampled from a noise distribution $p(\mathbf{z})$. $\mathbf{Z}$ is used to denote the set containing the noise sampled from $p(\mathbf{z})$. The discriminator is represented as $D(\mathbf{x})$. In the adversarial training process, the discriminator $D(\mathbf{x})$ is trained to accurately separate the training samples (or true samples) and the generated samples from $G(\mathbf{z})$. At the same time, $G(\mathbf{z})$ is trained to generate the samples as similar as the samples in the dataset $\mathbf{X}$ so that the discriminator cannot correctly separate them from the true samples. This adversarial process can be converted to play a minimax game with value function L_{adv}:

$$\begin{aligned} \min_G \max_D L_{adv} = \mathbb{E}_{\mathbf{x}_i \sim \mathbf{X}} [\log D(\mathbf{x}_i)] + \\ \mathbb{E}_{\mathbf{z}_i \sim \mathbf{Z}} [\log(1 - D(G(\mathbf{z}_i))]. \end{aligned} \tag{4.1}$$

With a well trained generator $G(\mathbf{z})$, its generated samples will be similar to the true samples, *i.e.*, accurately approximating the true data distribution.

4.2.2 Adversarial Discriminative Domain Adaptation

ADDA is an unsupervised domain adaptation method which is based on adversarial process to adapt a trained network from the source domain to the target domain [151]. Different from the existing works [36, 150, 91, 11], ADDA considers independently the source and target mappings, *i.e.*, unshared weights between the two network streams, allowing domain specific feature extraction to be learned, where the target weights are initialized by the network pretrained on the source.

Suppose the samples from source and target domain are denoted as $\mathbf{X}^s = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{n_s}$ and $\mathbf{X}^t = \{(\mathbf{x}_i^t, y_i^t)\}_{i=1}^{n_t}$, respectively. n_s and n_t denote the number of samples in source and target domain. M_s is used to denote the network stream in source domain while M_t denotes the network stream in target domain. The discriminator is denoted as D. ADDA is trained and evaluated in three steps as shown in Fig. 4.2 and they can be formulated as follows:

Step one: Pre-train the classification network in the source domain. The classifier is denoted as C and the classification loss is formulated as:

$$\min_{M_s,\, C} \; L_{cls} = -\mathbb{E}_{\mathbf{x}_i^s \sim \mathbf{X}^s} \sum_{k=1}^{K} \mathbb{1}_{[k=y_i^s]} \log C(M_s(\mathbf{x}_i^s)) \tag{4.2}$$

Step two: For the domain variations between the source and target domain, a large classification performance drop will be observed if the network trained on source domain is directly applied to the target domain. The adversarial process is used in ADDA for adapting the trained network to the target domain. The adversarial adaptation loss is:

$$\begin{aligned} \max_{M_t} \min_{D} L_{adv} = &- \mathbb{E}_{\mathbf{x}_i^s \sim \mathbf{X}^s} [\log D(M_s(\mathbf{x}_i^s))] \\ &- \mathbb{E}_{\mathbf{x}_i^t \sim \mathbf{X}^t} [\log(1 - D(M_t(\mathbf{x}_i^t)))] \end{aligned} \tag{4.3}$$

Step three: As shown in Fig. 4.2, the trained target mapping M_t and the classifier C are used for evaluation purpose in the target domain.

4.3 Our Proposed Methodology

As shown in Fig. 4.1, the problem considered is to learn the attribute-related person re-id features with the assistance of the labeled attribute recognition dataset. Specifically, suppose there are n_a samples from the attribute recognition (source) domain $\mathcal{D}_a = \{(\mathbf{x}_i^a, \mathbf{a}_i)\}_{i=1}^{n_a}$, in

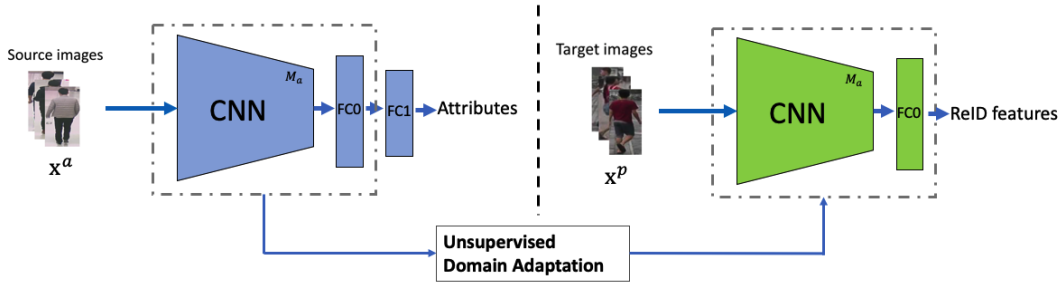


Figure 4.3: The proposed attribute related re-id feature learning framework.

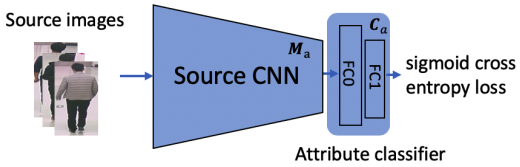


Figure 4.4: The attribute recognition network architecture.

which $\mathbf{x}_i^a$ represents the i-th sample, featuring m attributes $\mathbf{a}_i \in \mathbb{R}^m$. For the person re-id (target) domain $\mathcal{D}_p = \{\mathbf{x}_i^p\}_{i=1}^{n_p}$, there are n_p samples $\mathbf{x}_i^p$, but, on the contrary, no annotation is available. Based on the data from these two domains $\mathcal{D}_a$ and $\mathcal{D}_p$, we want to learn a domain-invariant mapping $\mathbf{M}$, which can be used to extract attribute-related features in re-id domain.

In the proposed framework, depicted in Fig. 4.3, the convolutional neural network (CNN) is used to learn the mapping $\mathbf{M}$. The network is first trained in the attribute recognition domain $\mathcal{D}_a$ and then adapted to the person re-id domain $\mathcal{D}_p$. Since the label information is not available in re-id domain, an unsupervised domain adaptation method, based on [151], is applied for adapting the trained attribute recognition network from attribute recognition domain $\mathcal{D}_a$ to person re-id domain $\mathcal{D}_p$. In the following, this framework will be introduced in detail from three aspects, *i.e.*, attribute recognition, unsupervised domain adaptation and feature extraction for re-id.

4.3.1 Attribute Recognition

As shown in Fig. 4.4, the attribute recognition is treated as a multi-label classification task. Given one sample $\mathbf{x}_i^a \in \mathcal{D}_a$, there are m attributes $\mathbf{a}_i = [a_{i,1}, \ldots, a_{i,m}] \in \mathbb{R}^m$. Thus, instead of using softmax cross entropy loss, the sigmoid cross entropy loss is used to train

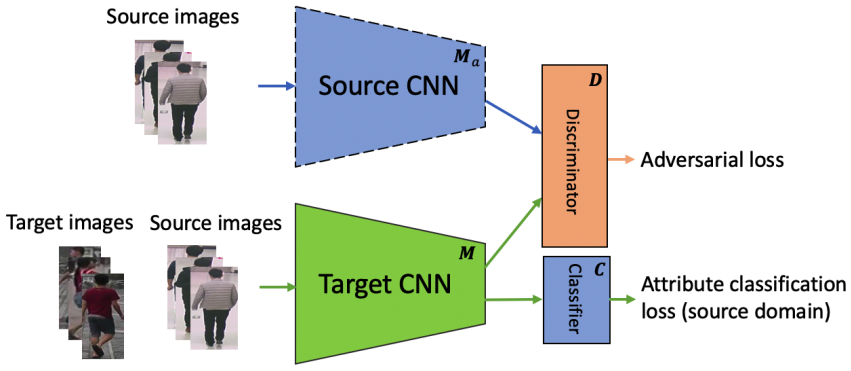


Figure 4.5: The proposed unsupervised adversarial adaptation. The source CNN is pre-trained on attribute recognition dataset. The classifier is only used for attribute classification on the source images (from attribute recognition domain $\mathcal{D}_a$). Dashed bounding box indicates fixed weights, while solid bound box indicates the weight that can be learned.

the network:

$$L_{attr} = -\mathbb{E}_{\mathbf{x}_i \sim \mathbf{X}^a} \sum_{j=1}^{m} \Big(a_{i,j} \log(\mathbf{C}_a(\mathbf{M}_a(\mathbf{x}_i))) + (1 - a_{i,j}) \log(1 - \mathbf{C}_a(\mathbf{M}_a(\mathbf{x}_i))) \Big), \tag{4.4}$$

in which $\mathbf{C}_a$ is the attribute classifier, as in Fig. 4.4, and $\mathbf{M}_a$ denotes the feature mapping in the attribute recognition domain. The attribute recognition network is trained by minimizing the loss L_{attr}.

4.3.2 Unsupervised Domain Adaptation

Given the learned mapping $\mathbf{M}_a$ in the attribute recognition domain $\mathcal{D}_a$, the objective of the domain adaptation step is to learn a domain-invariant mapping $\mathbf{M}$ for attribute discriminative feature extraction in the re-id domain $\mathcal{D}_p$. Since the attribute label is unavailable in $\mathcal{D}_p$, an unsupervised domain adaptation method is used to adapt the network trained in $\mathcal{D}_a$ to the person re-id domain $\mathcal{D}_p$. To this end, the ADDA proposed in work [151] is modified to make it more suitable for our purpose in network adaptation.

As aforementioned, the classification network in ADDA is first trained in source domain with the supervision of labels, and in order to adapt the trained source network to the target domain, the adversarial adaptation process is used to learn the target mapping. A good adapted target mapping means the discriminator network $\mathbf{D}$ cannot reliably predict the domain label of the feature vector from source or target mapping.

Since in ADDA only samples from the target domain are fed into the target network, the adapted target network is not domain-invariant and it can be only used in the target domain. To learn a domain-invariant mapping, samples from both the target (person re-id) and source (attribute recognition) domain are used as in [152]. There are at least two advantages to do this. Firstly, the target mapping is invariant to both the attribute recognition and re-id domains and thus the extracted features are more robust to the domain shift compared to the original ADDA. The second one is that more samples are used to train the target encoder. In our experiments, we found that this makes the modified adversarial adaptation easier to converge than the original one. The least square GANs, which uses the least square loss function for the discriminator, has been reported as a stable variation of GANs [152, 103]. The feature mapping in the target domain is denoted as $\mathbf{M}_t$ and $\mathbf{M}_t$ is usually initialized with the weights from $\mathbf{M}_s$ which is trained in the source domain. In the modified adversarial adaptation, the least square loss function is also used in the adversarial loss L_{adv} as:

$$\begin{aligned}\min_{\mathbf{M}}\max_{\mathbf{D}} L_{adv} =& \mathbb{E}_{\mathbf{x}_i\sim\mathbf{X}^a\cup\mathbf{X}^p}\|\mathbf{D}(\mathbf{M}(\mathbf{x}_i))-1\|^2 \\ &+\mathbb{E}_{\mathbf{x}_i\sim\mathbf{X}^a}\|\mathbf{D}(\mathbf{M}_a(\mathbf{x}_i))\|^2,\end{aligned} \tag{4.5}$$

in which $\mathbf{X}^p=[\mathbf{x}_1^p,\ldots,\mathbf{x}_{n_p}^p]$ and $\mathbf{M}$ is initialized with the weights from $\mathbf{M}_a$, which is fixed during domain adaptation.

With the domain-invariant $\mathbf{M}$, we can now extract attribute-related discriminative features in $\mathcal{D}_p$. In our experiments, we also found that the attribute recognition performance of the original ADDA decreases during the domain adaptation procedure and this influences the attribute-related feature extraction. In order to cope with this drop of attribute recognition performance, an additional attribute classifier is added for source samples, as shown in Fig. 4.5. Thus, the final optimization problem in unsupervised domain adaptation is:

$$\min_{\{\mathbf{M},\mathbf{C}\}}\max_{\mathbf{D}} L_{adv}+\alpha L_{attr}, \tag{4.6}$$

where α is a hyper-parameter. L_{attr} is the same as in Eq. (4.4) with $\mathbf{C},\mathbf{M}$ in place of $\mathbf{C}_a,\mathbf{M}_a$. The weights of $\mathbf{C},\mathbf{M}$ are initialized with $\mathbf{C}_a,\mathbf{M}_a$, as usually done in domain adaptation. Only the samples from source domain (attribute recognition domain) are used to calculate the loss L_{attr}. With the additional attribute classifier, the attribute recognition performance can be maintained during domain adaptation.

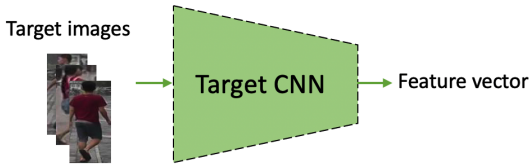


Figure 4.6: Person re-id feature extraction network. The target images are from re-id domain.

4.3.3 Feature Extraction

After domain adaptation, the target CNN (or feature mapping $\mathbf{M}$) is used for extracting attribute-related features of person images in the re-id domain $\mathcal{D}_p$. The feature extraction network is presented in Fig. 4.6. The similarities between these extracted features are calculated simply using the Euclidean distance as in deep re-id works [192, 189, 155]. Finally, matching or ranking person images can be performed based on their similarities.

4.4 Experiments

In this section, experimental results are presented to demonstrate the effectiveness of the proposed re-id feature learning framework. To measure the performances of the proposed method, the Cumulative Matching Characteristic (CMC) and mean average precision (mAP) are used.

4.4.1 Dataset

Three datasets, *i.e.*, RAP, Market-1501 and Duke-MTMC-reID, are included to evaluate the proposed re-id framework [192, 66]. RAP is an attribute recognition dataset that includes $41,585$ images in total. For each image 91 pedestrian attributes have been annotated. In our experiments we removed some extremely unbalanced attributes, selecting only 70 attributes. Market-1501 and DukeMTMC-reID are two popular large-scale re-id datasets. Thanks for the work [83], they are also labeled with pedestrian attributes *i.e.*, 27 attributes in Market-1501 and 22 in DukeMTMC-reID. However, in our method, we never use attribute annotations in the target domain.

Table 4.1: Performance comparisons with existing attribute based unsupervised adaptive person re-id methods.

Source → Target	Market-1501 → DukeMTMC-reID			
Metric	Rank1	Rank5	Rank10	mAP
TJ-AIDL	24.3%	38.3%	45.7%	10.0%
MMFA	15.8%	26.0%	48.2%	5.7%
Ours	**28.6%**	**44.2%**	**51.7%**	**13.1%**
Source → Target	DukeMTMC-reID → Market-1501			
Metric	Rank1	Rank5	Rank10	mAP
TJ-AIDL	38.0%	59.2%	67.6%	13.6%
MMFA	35.5%	55.3%	64.0%	12.7%
Ours	**43.0%**	**63.3%**	**70.6%**	**17.1%**

4.4.2 Implementation Details

In all experiments, the input images are resized to $(224, 224, 3)$. The hyper-parameter α is fixed to 0.1. MobileNet is selected as the backbone network and it is pretrained on ImageNet. Adam optimizer with a learning rate of 0.0001 is used in all experiments. Two fully connected (FC) layers are included in the attribute classifier ($1024 \rightarrow 512 \rightarrow m$, in which m is the number of pedestrian attributes as aforementioned). The discriminator is also composed of two FC layers ($1024 \rightarrow 384 \rightarrow 1$).

4.4.3 Comparisons with State-Of-The-Art Results

In order to compare with state-of-the-arts methods, the experiments using Market-1501 and DukeMTMC-reID datasets are performed. Compared with the re-id works using person identity supervision, pedestrian attribute supervised re-id works are much less. In the work [155], a Transferable Joint Attribute-Identity Deep Learning (TJ-AIDL) is proposed to simultaneously learn an attribute-semantic and identity-discriminative feature representation space transferrable to the new (unseen) target domain for re-id tasks. Lin *et al*. proposed the multi-task mid-level feature alignment (MMFA) network for the unsupervised cross-dataset person re-id task. Different from TJ-AIDL, the Maximum Mean Discrepancy (MMD) is used to transfer the learned label information from one fully labeled domain to another domain without label information. The comparisons are made using Market-1501 and DukeMTMC-reID datasets and as in TJ-AIDL [155] and MMFA [81], the experiments are one of the datasets is used as the source domain and the other one is for target domain. In our proposed unsupervised domain adaptation re-id method, only the attribute labels are used as the supervision information for training the model. Thus, for fair comparison, the

Table 4.2: Comparing experimental results before and after adaptation. The network is first trained on the attribute recognition dataset RAP.

	Market-1501		DukeMTMC-reID	
Metric	Rank1	mAP	Rank1	mAP
w/o adaptation	28.2%	8.7%	15.6%	4.9%
w adaptation	**32.1%**	**10.6%**	**18.7%**	**6.5%**

experimental results of TJ-AIDL and MMFA with only using attribute label are selected and presented in Table 4.1.

From Table 4.1, it shows our proposed method outperforms the state-of-arts. For example, in the case of Market-1501 $\rightarrow$ DukeMTMC-reID, there are 4.3% improvements in Rank1 compared with TJ-AIDL and 12.8% improvements compared with MMFA. This improvement mainly due to the proposed unsupervised adversarial adaptation which can effectively transfer the attribute recognition capability from source domain (with attribute labels) to target domain (not labelled), and with the added attribute classifier, it can also maintain the attribute recognition performance during domain adaptation. The table also shows that the re-id performance in the case of DukeMTMC-reID $\rightarrow$ Market-1501 is better than Market-1501 $\rightarrow$ DukeMTMC-reID (from 28.6% to 43.0% for Rank1 in our proposed method). This results from that DukeMTMC-reID contains more variations in samples and thus it makes the network trained on it has better generalization ability.

4.4.4 Domain Adaptation Influences

As aforementioned, since the label information is not available in the target domain, unsupervised domain adaptation is used for adapting the trained model from the source domain to the target domain. To evaluate the influence of the adaptation procedure on the proposed unsupervised adaptive re-id framework, Tab. 4.2 shows the experimental results before and after domain adaptation. The network is first trained on the attribute recognition dataset RAP. For the experiments without using adaptation, the trained attribute recognition network is directly used on re-id dataset, *i.e.*, Market-1501 or DukeMTMC-reID, to extract re-id feature representations. For the experiments with adaptation, the proposed unsupervised adversarial adaptation, as shown in Fig. 4.5, is used to adapt the trained model from attribute recognition domain to re-id domain before re-id feature extraction.

From the table, it can be found that after adaptation, the re-id performance of the network has been improved. There are 3.9% Rank1 improvements on Market-1501 and 3.1% for DukeMTMC-reID. This validates the positive influence of the proposed unsupervised ad-

versarial adaptation on transferring attribute related features across domains. An interesting observation from the table is that both the rank1 and mAP metrics of Market-1501 are much higher than DukeMTMC-reID. This is mainly because there are large domain differences between RAP and DukeMTMC-reID dataset while it is small between RAP and Market-1501 dataset. Fig. 4.7 gives the person image examples from RAP, Market1501 and DukeMTMC-reID dataset. From the figure, it can be observed that the domain variation between RAP and Market1501 dataset is smaller than the domain variation between RAP and DukeMTMC-reID dataset.

4.4.5 Attribute Classifier Influences

In the original adversarial discriminative domain adaptation (ADDA) [151], the classification task and unsupervised domain adaptation are separated into two steps. In our experiments, we found that the attribute recognition performance degrades in the domain adaptation procedure. In order to guarantee the attribute recognition performance, an attribute classifier is added based on ADDA as shown in Fig. 4.5. To evaluate the influence of this added attribute classifier on re-id performance, the comparison experiments are performed using the proposed unsupervised domain-adaptive re-id framework with and without the attribute classifier.

Fig. 3 shows the experimental results. The experiments are performed on RAP and Market-1501 datasets. The network is first trained on RAP dataset and then adapted to Market-1501 dataset. For the re-id framework without classifier, the classifier, as shown in Fig. 4.1, has been deleted in the adaptation step. In the experiments, we found the attribute recognition performance degrades during adapting the network. As observed in Fig. 4.8, this results in the decrease of the re-id performance during domain adaptation. With the additional classifier, the re-id performance, as the red curve in the figure, are more stable.

(a) RAP dataset (b) Market1501 dataset (b) DukeMTMC-reID dataset

Figure 4.7: Person image samples from (**a**) RAP dataset, (**b**) Market1501 dataset and (**c**) DukeMTMC-reID dataset.

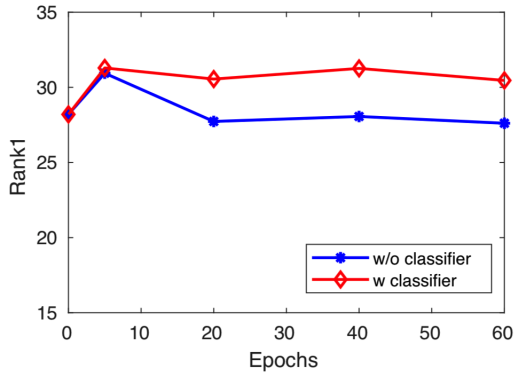


Figure 4.8: Adaptation performance comparisons between the unsupervised domain-adaptive re-id frameworks with and without the additional classifier. 0 epoch means there is no adaptation in the proposed re-id framework.

4.4.6 Sample Feeding Influences

In the adaptation procedure of ADDA, only samples from source domain are fed into the target network. However, in the proposed method, the samples from both source and target domains are fed into the target network as shown in Fig. 4.5. This is based on the observations that: (1) after domain adaptation with feeding samples from both source and target domains, the target network will be invariant to both these two domains; (2) with more samples for training, the adversarial adaptation becomes easier to converge as shown in our experiments.

In order to verify the advantages of the applied sample feeding strategy, three experiments are designed for ablation study: (1) the network is trained on Market-1501 dataset without adaptation and then it is used for re-id feature extraction on Market-1501 and DukeMTMC-reID dataset; (2) the network is first trained on Market-1501 dataset and then it is adapted to DukeMTMC-reID dataset. During adaptation, only the samples from source domain are fed into the target network. After the adaptation, the adapted network is used for extracting re-id features on Market-1501 and DukeMTMC-reID dataset; (3) as in the second experiment, the network is pre-trained on Market-1501 dataset but in adaptation, the samples from both domains are fed into the target network. The adapted network is used for extracting re-id feature on Market-1501 and DukeMTMC-reID dataset.

The experimental results are presented in Table 4.3. From the table, it can be observed that after the adaptation in the second experiment, the adapted network can get higher re-id performance on the target domain which verifies the effectiveness of adaptation. However, it can be also found that the adapted network has a degradation of the re-id performance on

Table 4.3: Ablation study results for evaluating the sample feeding influences. In all of three experiments, the networks are first trained on Market-1501 dataset.

	Market-1501		DukeMTMC-reID	
Metric	Rank1	mAP	Rank1	mAP
w/o adaptation	66.8%	40.9%	22.4%	9.54%
Target Samples	63.8%	37.8%	23.6%	10.9%
Target+Source Samples	67.7%	42.7%	25.4%	12.1%

source domain. This results from the fact that the target network has been adapted to target domain and it is not optimal for the source domain due to the domain variations between source and target domain. With feeding the samples from both source and target domains into the target network in adaptation, the results of third experiment show that the adapted network can maintain its re-id performance on source domain, which verifies that the adapted network is invariant to both source and target domains. It can be also found that the adapted network largely improves the re-id performance on target domain. The reason can lie in that compared with the second experiment, more samples are involved in the adaptation procedure for the third experiment. In the experiments, we also observed that the converging epoch number of the third experiment is generally less than that of the second experiment.

4.5 Conclusions

In this chapter, we presented an unsupervised domain-adaptive re-id framework for extracting attribute-related features. Considering that most re-id datasets are not labeled with pedestrian attributes, a modified domain adaptation method has been proposed for adapting the attribute recognition network. Based on the observation that the attribute recognition performance degrades during domain adaptation procedure, an additional classifier has been added. The experimental results using three large-scale datasets proved the effectiveness of the proposed unsupervised adaptive re-id framework.

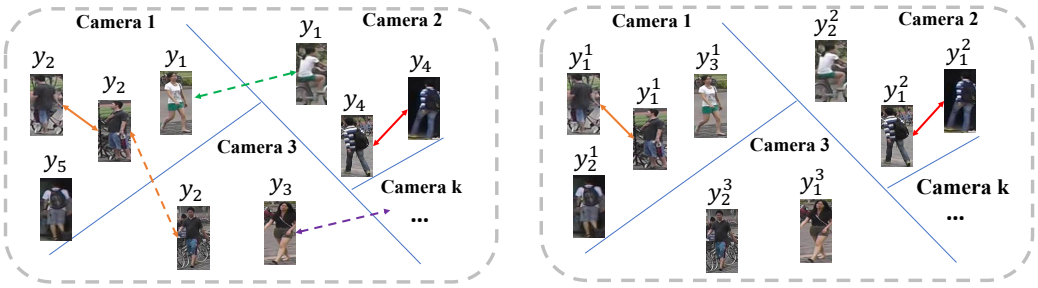
Chapter 5

Intra-Camera Supervised Person Re-Identification

5.1 Introduction

Although deep learning based person re-id methods [20, 74, 144, 53, 195, 202] have demonstrated remarkable performance advances, they rely on supervised model learning using a large set of cross-camera identity labelled training samples. This paradigm needs an exhaustive and expensive training data annotation process. As introduced in Chapter 1, for an ideal case with M cameras and N identities in each camera view, the annotation complexity can reach $O(MN + M^2N^2)$. This expensive data annotation process dramatically degrades the usability and scalability of re-id methods for large scale deployment in real-world application.

This problem has received significant attentions. Representative attempts for minimising the annotation cost include: **(1)** Domain generic feature design [44, 31, 78, 105, 191], **(2)** Unsupervised domain adaptation [118, 29, 155, 81, 199, 179, 22], **(3)** Unsupervised model learning [154, 21, 82, 68], and **(4)** Weakly supervised learning [108]. By hand-crafting generic appearance features with prior knowledge, the *first* paradigm of methods can perform re-id matching universally. However, their performances are often inferior due to limited knowledge encoded in such image representations. This can be addressed by transferring the labelled training data of a source dataset (domain), as demonstrated in the *second* paradigm of methods. Implicitly, these methods assume that the source and target domains share reasonably similar camera viewing conditions for ensuring sufficient transferable knowledge. The heavy reliance on the relevance and quality of source datasets [204] renders this approach less practically useful, since this assumption is often invalid. The *third* paradigm of methods is more scalable, as they need only unlabelled target domain data. While having high potential, unsupervised re-id methods usually yield the weakest performance, making them fail to meet the deployment requirements. In contrast, the *fourth* paradigm of methods considers a weakly supervised learning setting, where the person identity labels are annotated at the video level without fine-grained bounding boxes. Apart from insufficient re-id accuracy,



(a) Fully supervised person re-id data

(b) Intra-camera supervised person re-id data

Figure 5.1: Labels in person re-id data. (**a**) Fully supervised training data needs both *per-camera* and *cross-camera* identity labelling in a unified class space. (**b**) Intra-camera supervised (ICS) training data annotation only needs *per-camera* identity labelling *independently*, each camera view with a separate class space. Camera-view index is encoded as superscript of identity label in ICS person re-id data. Solid and dashed arrows denote intra-camera and inter-camera association, respectively.

this paradigm is mostly sensible only when such weak labels can be cheaply obtained from certain domain knowledge, which however is not generically accessible.

In this work, we suggest another novel person re-identification paradigm for scaling-up the model training process, called ***Intra-Camera Supervised*** (ICS) person re-id (Fig. 5.1(b)). As the name indicates, ICS eliminates the sub-process of cross-camera identity association during annotation, therefore its corresponding complexity $O(M^2N^2)$ has been eliminated, which is the majority component of the standard annotation cost, as discussed above. Under the ICS paradigm the training data involves only the intra-camera annotated identity labels with each camera view labelled *independently*. This labelling complexity is hence only $O(MN)$ with M the camera view number and N the average per-camera person identity number, therefore being significantly more affordable. Importantly, ICS naturally enables a parallel annotation process by camera views without labelling conflict due to no cross-camera identity association (Fig. 5.2(b)). This desirable merit is lacking in the conventional training data labelling due to the difficulty of obtaining disjoint labelling tasks, e.g. subsets of person identity classes without overlap (Fig. 5.2(a)). While being similar to the concurrent work [108] since they both consider explicitly the training data labelling process, the proposed ICS paradigm however does *not* assume specific domain knowledge therefore it is more generally applicable.

To solve the ICS re-id problem, we propose a ***Multi-tAsk mulTi-labEl*** (MATE) deep learning model. Unlike the conventional fully supervised re-id methods using inter-camera identity labels, MATE is designed specially for overcoming two ICS challenges: (1) how to learn effectively from per-camera independently labelled training data, and (2) how to discover re-

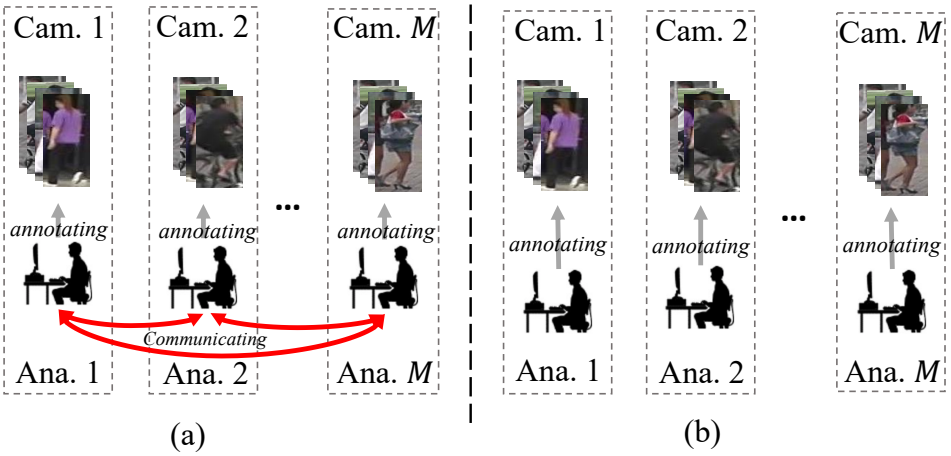


Figure 5.2: Illustrations of data annotation process. (**a**) Conventional fully supervised person re-id vs. (**b**) ICS person re-id in the process of *training data collection*. Suppose each annotator needs to label the training data from a different camera view. In order to minimise the labelling conflict, an annotator may have to check if a person has been labelled or not by others. This gives rise to expensive communication costs, which is totally eliminated in the proposed ICS re-id paradigm, due to the independence nature between camera views.

liably the missing identity association across camera views. Specifically, MATE integrates two complementary learning components into a unified model: (a) *Per-camera multi-task learning* that separately learn individual camera views for modelling their specificity and the implicit shared information in a multi-task learning manner (Sec. 5.3.1). This assigns a specific network branch (i.e. a learning task) for modelling each camera view while constraining all the per-camera tasks to share a feature representation space. (b) *Cross-camera multi-label learning* that associates the identity labels across camera views in a multi-label learning strategy (Sec. 5.3.2). This is based on an idea of curriculum cyclic association that can associate reliably multiple cross-camera identity classes from self-discovered identity matches for multi-label model optimisation.

The **contributions** of chapter are: **(1)** We present a novel person re-identification paradigm for scaling up the model training process, dubbed as *Intra-Camera Supervised* (ICS) person re-id. ICS is characterised by *no* need for exhaustive cross-camera identity matching during training data annotation, whilst allowing naturally parallel labelling by camera views without conflict. Consequently, it makes the training data collection substantially cheaper and faster than the standard cross-camera identity labelling, therefore offering a more scalable mechanism to large re-id deployments. **(2)** We formulate a *Multi-tAsk mulTi-labEl* (MATE) deep learning method for solving the proposed ICS person re-id problem. In particular, MATE combines the strengths of multi-task learning and multi-labelling learning in a unified framework to account for independent camera-specific identity label information and

self-discovering their cross-camera association relationships concurrently. This represents a natural strategy for fully leveraging the ICS supervision with per-camera independent identity label spaces. **(3)** Through extensive benchmarking and comparisons on the ICS variant of three large re-id datasets (Market-1501 [191], DukeMTMC-reID [196, 122], and MSMT17 [159]), we demonstrate the cost-effectiveness advantages of the ICS re-id paradigm using our MATE model over the existing representative solutions including supervised learning, semi-supervised learning, unsupervised learning, unsupervised domain adaptation, and tracklet learning.

5.2 Problem Formulation

We formulate the *Intra-Camera Supervised* (ICS) person re-identification problem. As illustrated in Fig. 5.1(b), ICS only needs to annotate intra-camera person identity labels *independently*, whilst eliminating the most-expensive inter-camera identity association as required in the conventional fully supervised re-id setting.

Suppose there are M camera views in a surveillance camera network. For each camera view $p \in \{1, 2, \cdots, M\}$, we *independently* annotate a set of training images $\mathcal{D}^p = \{(\mathbf{x}_i^p, y_k^p)\}$ where each person image $\mathbf{x}_i^p$ is associated with an identity label $y_k^p \in \{y_1^p, y_2^p, \cdots, y_{N^p}^p\}$, and N^p is the total number of unique person identities in $\mathcal{D}^p$ [1]. For clarity, we express the camera view index in the superscript due to the per-camera independent labelling nature in the ICS setting. By combining all the camera-specific labelled data $\mathcal{D}^p$, we obtain the entire training set as $\mathcal{D} = \{\mathcal{D}^1, \mathcal{D}^2, \ldots, \mathcal{D}^M\}$. For any two camera views p and q, their k-th person identities y_k^p and y_k^q usually describe two different people, i.e. they are two independent identity label spaces (Fig. 5.1(b)). This means exactly that the cross-camera identity association is *not* available, in contrast to the fully supervised re-id data annotation (Fig. 5.1(a)).

The ICS re-id problem presents a couple of new modelling challenges: (1) how to effectively exploit the per-camera person identity labels, and (2) how to automatically and reliably associate independent identity label spaces across camera views. The existing fully supervised re-id methods do not apply due to the need for identity annotation in a *single* label space across camera views. A new learning method tailored for the ICS setting is required to be developed.

[1] We use i, j to denote image indexes, k, l, t to denote identity indexes and p, q to denote camera indexes.

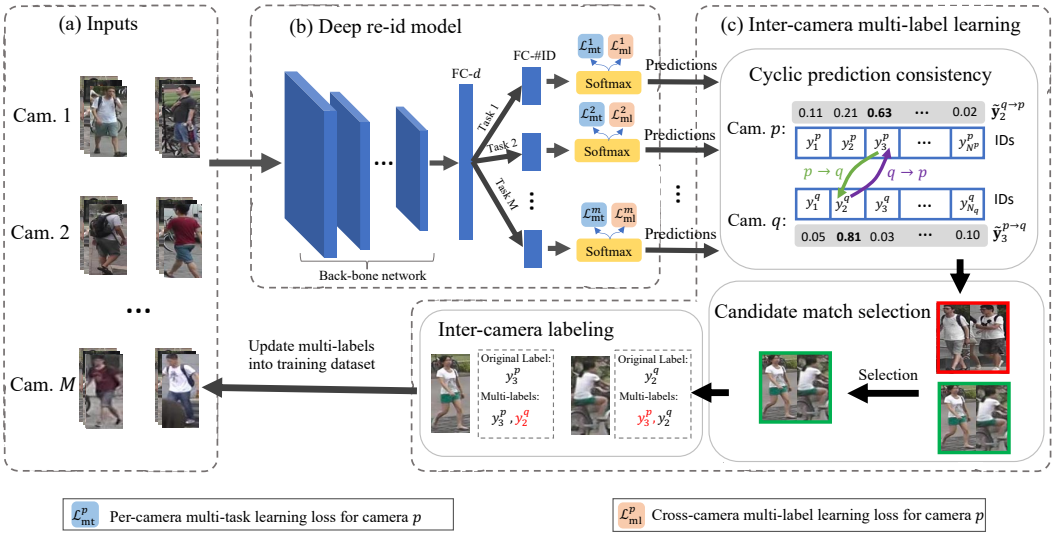


Figure 5.3: Overview of the proposed Multi-tAsk mulTi-labEl (**MATE**) deep learning method. **(a)** Given per-camera independently labelled training images, MATE aims to learn an identity discriminative feature representation model. This is achieved by designing two learning components: **(b)** *Per-camera multi-task* learning where we consider each individual camera view as a separate learning task with its own identity class space and optimise these camera-specific tasks on a common feature representation (Sec. 5.3.1), and **(c)** *Cross-camera multi-task* learning where we self-discover the underlying identity matching relationships across camera views via curriculum cyclic association and design a multi-label optimisation algorithm to exploit these discovered cross-camera association information during model training. The two components are integrated together in a single MATE formulation, resulting in an end-to-end trainable model.

5.3 Method

We introduce a novel ICS deep learning method, capable of conducting **M**ulti-t**A**sk mul**T**i-lab**E**l (**MATE**) model learning to fully exploit the independent per-camera person identity label spaces. In particular, MATE solves the aforementioned two challenges by integrating two complementary learning components into a unified solution: **(i)** *Per-camera multi-task learning* that assigns a separate learning task to each individual camera view for dedicatedly modelling the respective identity space (Sec. 5.3.1), **(ii)** *Cross-camera multi-label learning* that associates the independent identity label spaces across camera views in a multi-label strategy (Sec. 5.3.2). Combining the two capabilities with a unified objective function, MATE explicitly optimises their mutual compatibility and complementary benefits via end-to-end training. An overview of MATE is depicted in Fig. 5.3.

5.3.1 Per-Camera Multi-Task Learning

To maximise the use of *multiple* camera-specific identity label spaces with some underlying correlation (e.g. partial identity overlap) in the ICS setting, multi-task learning is a natural choice for model design [3]. This allows to not only mine the common knowledge among all the camera views, but also to improve per-camera model learning concurrently given augmented (aggregated) training data.

Specifically, given the nature of independent label spaces we consider each camera view as a separated learning task, all of which share a feature representation network for extracting the common knowledge in a multi-branch architecture design. One branch is in charge of a specific camera view. This forms *per-camera multi-task learning* in the ICS context. By such multi-task learning, our method can favourably derive a person re-id representation with *implicit* cross-camera identity discriminative capability, facilitating cross-camera identity association [68]. This is because during training, all the branches *concurrently* propagate the respective camera-specific identity label information through the shared representation network f_θ (Fig. 5.3(b)), leading to a *camera-generic* representation. This process is done by minimising the softmax cross-entropy loss.

Formally, for a training image $(\mathbf{x}_i^p, y_k^p) \in \mathcal{D}^p$ from camera view p, the softmax cross-entropy loss is used for formulating the training loss:

$$\mathcal{L}_{\text{mt}}^p(i) = -\mathbb{1}(y_k^p)\log\Big(g^p\big(f_\theta(\mathbf{x}_i^p)\big)\Big) \tag{5.1}$$

where given the *camera-shared* feature vector $f_\theta(\mathbf{x}_i^p) \in \mathbb{R}^{d\times 1}$, the classifier $g^p(\cdot)$ for the camera view p predicts an identity class distribution in its own label space with N_p classes: $\mathbb{R}^{d\times 1} \rightarrow \mathbb{R}^{N_p\times 1}$. The Dirac delta function $\mathbb{1}(\cdot) : \mathbb{R} \rightarrow \mathbb{R}^{1\times N_p}$ returns a one-hot vector with "1" at the specified index.

By aggregating the loss of training samples from all the camera views, we formulate the *per-camera multi-task learning* objective function as:

$$\mathcal{L}_{\text{mt}} = \frac{1}{M}\sum_{p=1}^{M}\Big(\frac{1}{B^p}\sum_{i=1}^{B^p}\mathcal{L}_{\text{mt}}^p(i)\Big) \tag{5.2}$$

where B^p denotes the number of training images from the camera view p in a mini-batch.

5.3.2 Cross-Camera Multi-Label Learning

Cross-camera person appearance variation is a key challenge for re-id. Whilst this is implicitly modelled by the proposed multi-task learning as detailed above, the per-camera multi-task learning is still insufficient to fully capture the underlying identity correspondence relationships across camera-specific label spaces.

However, it is non-trivial to associate identity classes across camera views. One major reason is that a different set of persons may appear in a specific camera view, leading to *no* one-to-one identity matching between camera views. Conceptually, this gives rise to a very challenging open-set recognition problem where a rejection strategy is often additionally required [125, 126]. Compared to generic object recognition in natural images, open-set modelling in re-id is more difficult due to small training data, large intra-class variation, subtle inter-class difference, and ambiguous visual observations of surveillance person imagery. Besides, existing open-set methods often assume accurately and completely labelled training data, and the unseen classes only in model test. In contrast, we need to discover cross-camera identity correspondences during training with small (unknown) overlap across different spaces.

This is hence a harder learning scenario with a higher risk of error propagation from noisy cross-camera association. An intuitive solution for open-set recognition is to find an operating threshold, e.g. by Extreme Value Theory [28] based statistical analysis. This relies on optimal *supervised* model learning from a sufficiently large training dataset, which however is unavailable in the ICS setting.

To circumvent the above problems, we design a ***cross-camera multi-label learning*** strategy for robust cross-camera identity association. This is realised by (i) designing a *curriculum cyclic association constraint* to find reliable cross-camera identity association, and (ii) forming a *multi-label learning algorithm* to incorporate the self-discovered cross-camera identity association into discriminative model learning (Fig. 5.3(c)).

1. Curriculum Cyclic Association

For more reliable identity association across camera views, we form a *cyclic prediction consistency* constraint. Specifically, given an identity class $y_k^p \in \{y_1^p, y_2^p, \cdots, y_{N^p}^p\}$ from a camera view $p \in \{1, 2, \ldots, M\}$, we need to find if a true matching identity (i.e. the same identity) exists in another camera view q. We achieve this in the following process.

(i) We first project all the images of each person identity y_k^p from camera view p to the classifier branch of camera view q to obtain a *cross-camera prediction* $\tilde{\mathbf{y}}_k^{p \to q}$ via averaging

as:

$$\tilde{\mathbf{y}}_k^{p\to q} = \frac{1}{S_k^p}\sum_{i=1}^{S_k^p} g^q\big(f(\mathbf{x}_i^p)\big) \in \mathbb{R}^{N_q\times 1}, \tag{5.3}$$

where S_k^p is the number of images from identity y_k^p. Each element of $\tilde{\mathbf{y}}_k^{p\to q}$, denoted as $\tilde{\mathbf{y}}_k^{p\to q}(l)$, means the identity class matching probability at which y_k^p (an identity from camera view p) matches y_l^q (an identity from camera view q) in a cross-camera sense.

(ii) We then nominate the person identity $y_{l^*}^q$ from camera view q with the maximum likelihood probability as the candidate matching identity:

$$l^* = \arg\max_l \tilde{\mathbf{y}}_i^{p\to q}(l),\ l \in \{1, 2, \ldots, N_q\}. \tag{5.4}$$

With such one-way ($p \to q$) association alone, the matching accuracy should be not satisfactory since it cannot handle the cases of *no-true-match* as typical in the ICS setting. To boost the matching robustness and correctness, we further design a *curriculum cyclic association* constraint.

(iii) Specifically, in an opposite direction of the above steps, we project all the images of identity $y_{l^*}^q$ from camera view q to the classifier branch of camera view p in a similar way as Eq. (5.3), and obtain the best candidate matching identity $y_{t^*}^p$ with Eq. (5.4). Given this back-and-forth matching between camera view p and q, we subsequently filter the above candidate pair $(y_k^p, y_{l^*}^q)$ by a cyclic constraint as:

$$(y_k^p, y_{l^*}^q)\begin{cases} \text{is a candidate match,} & \text{if } y_{t^*}^p = y_k^p, \\ \text{is not a candidate match,} & \text{otherwise.} \end{cases} \tag{5.5}$$

This removes non-cyclic association pairs. While being more reliable, it is observed that only the cyclic association in Eq. (5.5) is not sufficiently strong for *hard* cases (e.g. different people with very similar clothing appearance), leading to false association.

(iv) To overcome this problem, inspired by the findings of cognitive study which suggest a better learning strategy is to start *small* [30, 62], we design a curriculum association constraint. It is based on the cross-camera identity matching probability. Formally, we define a cyclic association degree as:

$$\psi_{k\Leftrightarrow l^*}^{p\Leftrightarrow q} = \tilde{\mathbf{y}}_k^{p\to q}(l^*)\cdot\tilde{\mathbf{y}}_{l^*}^{q\to p}(k) \tag{5.6}$$

which measures the joint probability of a cyclic association between two identities y_k^p and $y_{l^*}^q$. Given this unary measurement, we can deploy a *curriculum threshold* $\tau \in [0, 1]$ for

selecting candidate matching pairs via:

$$\text{Cyclic } (y_k^p, y_{l^*}^q) \begin{cases} \text{is a match,} & \text{if } \psi_{k \Leftrightarrow l^*}^{p \Leftrightarrow q} > \tau, \\ \text{is not a match,} & \text{otherwise.} \end{cases} \tag{5.7}$$

This filtering determines if a cyclically associated identity pair $(y_i^p, y_{k^*}^q)$ will be considered as a match.

Curriculum threshold. The design of the curriculum threshold τ has a crucial influence on the quality of cross-camera identity association. In the spirit of curriculum learning, we consider τ as an annealing function of the model training time to enable a progressive selection. Meanwhile, we need to take into account that the magnitude of maximum prediction usually increases along the training process as the model gets more mature. Taking these into consideration, we formulate the curriculum threshold as:

$$\tau^r = \min\left(\tau^u, \ \ \tau^l + \frac{r}{R-1}(1-\tau^l)\right) \tag{5.8}$$

where r specifies the current training round, with a total of R rounds. We maintain two thresholds: τ^u, which denotes the upper bound, and τ^l, which denotes the lower bound. Both of these two thresholds can be estimated by cross-validation.

Summary. We perform the above curriculum cyclic association process for every camera view pairs, which outputs a set of associated identity pairs across camera views. This self-discovered pairwise information will be used to improve model training as detailed in the following.

2. Multi-Label Learning

To leverage the above identity association results for improving model discriminative learning, we introduce a multi-label learning scheme in a cross-camera perspective. It consists of (i) multi-label annotation and (ii) multi-label training.

(i) Multi-label annotation. For easing presentation and understanding, we assume two camera views, and it is straightforward to extend to more camera views. Given an associated identity pair (y_k^p, y_{l*}^q) obtained as above, we annotate all the images X_i^p of y_i^p from camera view p with an extra label y_{l*}^q of camera view q. We do the same for all the images X_{l*}^q of y_{l*}^q in an inverse direction. Both image sets are therefore annotated with the same two identity labels, i.e. these images are associated. See an illustration example in Fig. 5.3(c). Given M camera views, for each identity y_k^p we perform at most $M-1$ times such annotation whenever a cross-camera association is found, resulting in a multi-label set $Y_i^p = \{y_k^p, y_{l*}^q, \cdots\}$ for X_i^p, with the cardinality $1 \leq |Y_i^p| \leq M$. When $|Y_i^p| = 1$, it means no cross-camera association is obtained. When $|Y_i^p| = M$, it means an identity association is found in every other camera view.

(ii) Multi-label training. Given such cross-camera multi-label annotation, we then formulate a multi-label training objective for an image $\mathbf{x}_i^p$ as

$$\mathcal{L}_{\text{ml}}^p(i) = \frac{1}{|Y_i^p|} \sum_{y^c \in Y_i^p} -\mathbb{1}(y^c) \log\Big(g^c\big(f_\theta(\mathbf{x}_i^p)\big)\Big) \tag{5.9}$$

where c indices the camera view of Y_i^p with the corresponding identity label simplified as y^c. For mini-batch training, we design the cross-camera multi-label learning objective as:

$$\mathcal{L}_{\text{ml}} = \frac{1}{B} \sum_{i,p} \mathcal{L}_{\text{ml}}^p(i) \tag{5.10}$$

which averages the multi-label training loss of all the B number of training images in a mini-batch.

Remarks. It is noteworthy to point out that, in contrast to the conventional single-task multi-label learning [148], we jointly form multi-label learning and multi-task learning in a unified framework, with a unique objective of associating different label spaces and merging the independently annotated labels with the same semantics.

5.3.3 Final Objective Loss Function

By combining per-camera multi-task (Eq. (5.2)) and cross-camera multi-label (Eq. (5.10)) learning objectives, we obtain the final model loss function as:

$$\mathcal{L} = \mathcal{L}_{\text{mt}} + \lambda \mathcal{L}_{\text{ml}}, \tag{5.11}$$

where the weight parameter $\lambda \in [0, 1]$ is to trade-off the two loss terms. With this formula as model training supervision, our method can effectively learn discriminative re-id model using both camera-specific identity label spaces available under the ICS setting ($\mathcal{L}_{\text{mt}}$) and cross-camera identity association self-discovered by MATE itself ($\mathcal{L}_{\text{ml}}$) concurrently. The MATE model training process is summarised in Algorithm 2.

Algorithm 2 The MATE model training procedure.

Input: Intra-camera independently labelled training data;
Output: A trained person re-id model;
Model training:
for $r = 1$ **to** R **do**:
  Calculate the curriculum threshold τ^r;
  Cross-camera identity association as in Eqs. (5.3)-(5.7);
  for $e = 1$ **to** *epoch_number* **do**:
    for $t = 1$ **to** *per-epoch mini-batch number* **do**:
      Feed forward a mini-batch of training images;
      Compute learning loss using Eq. (5.11);
      Update the network model by back-propagation;
    end for
  end for
end for

5.4 Experiments

Datasets. Due to *no* existing re-id datasets for the proposed scenario, we introduced three ICS re-id benchmarks. We simulated the ICS identity annotation process on three existing large person re-id datasets, Market-1501 [191], DukeMTMC-reID [122, 196] and MSMT17 [159]. Specifically, for the training data of each dataset, we *independently* perturbed the original identity labels for every individual camera view, and ensured that the same class labels of any pair of different camera views correspond to two unique persons (i.e. no labelled cross-camera association). We used the same original test data of each dataset for model performance evaluation.

Table 5.1: Benchmarking the ICS person re-id performance.

Dataset	Market-1501			
Metric (%)	R1	R10	R20	mAP
MCST	34.9	60.1	69.3	16.7
EPCS	42.6	64.6	71.2	19.6
PCMT	78.4	93.1	95.7	52.1
MATE (Ours)	**88.7**	**97.1**	**98.2**	**71.1**
Dataset	DukeMTMC-reID			
Metric (%)	R1	R10	R20	mAP
MCST	25.0	50.1	58.8	16.3
EPCS	38.8	58.9	64.6	22.1
PCMT	65.2	81.1	85.6	44.7
MATE (Ours)	**76.9**	**89.6**	**92.3**	**56.6**
Dataset	MSMT17			
Metric (%)	R1	R10	R20	mAP
MCST	12.1	26.3	33.0	4.8
EPCS	16.8	31.5	37.4	5.4
PCMT	39.6	59.6	65.7	15.9
MATE (Ours)	**46.0**	**65.3**	**71.1**	**19.1**

Performance metrics Following the common person re-id works, the Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP) metrics were used for model performance measurement.

Implementation details The ImageNet pre-trained ResNet-50 [49] was selected as the backbone network of our MATE model. As shown in Fig. 5.3, each branch in MATE was formed by a fully connected (FC) classification layer. We set the dimension of the re-id feature representation to 512. The person images were resized to 256×128 in pixel. The standard stochastic gradient descent (SGD) optimiser was adopted. The initial learning rate of the backbone network and classifiers were set to 0.005 and 0.05, respectively. We set a total of 10 rounds to anneal the curriculum threshold τ (Eq. (5.7)), with each round covering 20 epochs (except the last round where we trained 50 epochs to guarantee the convergence). We empirically estimated $\tau^l = 0.5$ (the lower bound of τ) and $\tau^u = 0.95$ (the upper bound of τ) for Eq. (5.8). In order to balance the model training across camera views, we randomly selected from each camera the same number of images, i.e. 4 images, per identity and the same number of identities, i.e. 2 identities, to construct a mini-batch. Unless stated otherwise, we set the loss weight $\lambda = 0.5$ for Eq. (5.11). In test, the Euclidean distance was applied to the camera-generic feature representations for re-id matching.

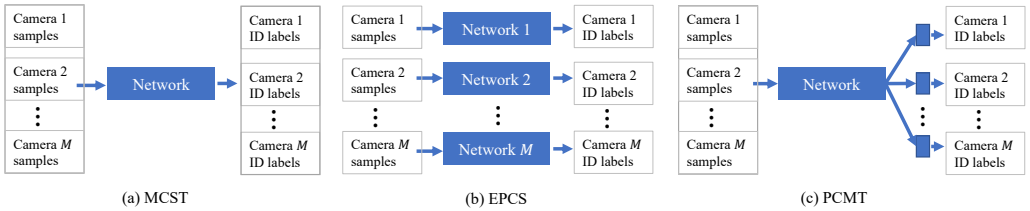


Figure 5.4: Three baseline learning methods for ICS person re-id: (**a**) Multi-Camera Single-Task (MCST) learning. (**b**) Ensemble of Per-Camera Supervised (EPCS) Learning. (**c**) Per-Camera Multi-Task (PCMT) learning.

5.4.1 Benchmarking the ICS Person Re-ID

While there has been no dedicated methods for solving the proposed ICS person re-id problem, we formulated and benchmarked three baseline methods based on the generic learning algorithms:

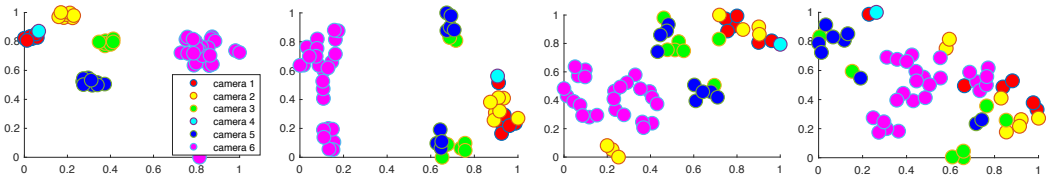
1. *Multi-Camera Single-Task* (MCST) learning (Fig. 5.4(a)): Given no identity association across camera views, we simply consider any identity classes from different camera views are distinct people and merge all the per-camera label spaces into a joint space cumulatively. This enables the conventional supervised model learning based on identity classification. We therefore train a single re-id model, as in the common supervised learning paradigm. At test time, we extract the re-id feature vectors and apply the Euclidean distance as the metrics for re-id matching.

2. *Ensemble of Per-Camera Supervised* (EPCS) learning (Fig. 5.4(b)): Without inter-camera identity labels, for each camera view we train a separate re-id model with its own single-camera training data. During deployment, given a test image we extract the feature vectors of all the per-camera models, concatenate them into a single representation vector, and utilise the Euclidean distance as the matching metrics for re-id.

3. *Per-Camera Multi-Task* (PCMT) learning (Fig. 5.4(c)): While being a variant of our MATE model *without* the cross-camera multi-label learning component, we simultaneously consider it as a baseline due to using the multi-task learning strategy.

To implement fairly the baseline learning methods, we used the same backbone ResNet50 as our method, a widely used architecture in the re-id literature. We trained each of these models with the softmax cross-entropy loss function in their respective designs.

Results We compared our MATE model with the three baseline methods in Table 5.1. Sev-



(a) MCST (b) EPCS (c) PCMT (d) MATE

Figure 5.5: Feature distribution visualization of a randomly selected person identity appearing under all the six camera views of the Market-1501 dataset. This is made by t-SNE [102]. Camera views are colour-coded. Best viewed in colour.

eral observations can be derived:

1. Concatenating simply the per-camera identity label spaces, MCST yields the weakest re-id performance. This is not surprised because there is a large (unknown) proportion of duplicated identities but mistakenly labelled with different classes, misleading the model training process.

2. The above problem can be addressed by independently exploiting camera-specific identity class annotations, as what EPCS does. This method does produce better re-id model generalisation consistently. However, the over accuracy is still rather low, due to the incapability of leveraging the shared knowledge between camera views and mining the inter-camera identity matching information.

3. To address this cross-camera association issue, PCMT provides an implicit solution and significantly improves the model performance.

4. Moreover, the proposed MATE model further boosts the re-id matching accuracy by explicitly associating the identity classes across camera views in a reliable formulation. This verifies the efficacy of our model in capitalising such cheaper and more scalable per-camera identity labelling.

To further examine the model performance, in Fig. 5.5 we visualised the feature distributions of a randomly selected person identity with images captured from all the camera views of Market-1501. It is shown that the feature points of our model present the best camera-invariance property, qualitatively validating the superior re-id performance over other competitors.

Table 5.2: Comparative evaluation of representative person re-id paradigms in the model training *supervision* perspective. '†': Results from [180]; '*': Results from [167].

Supervision	Method	Market-1501				DukeMTMC-reID				MSMT17			
		R1	R10	R20	mAP	R1	R10	R20	mAP	R1	R10	R20	mAP
None	RKSL†	34.0	-	-	11.0	-	-	-	-	15.4	-	-	4.3
	ISR†	40.3	-	-	14.3	-	-	-	-	21.5	-	-	6.1
	DIC†	50.2	-	-	22.7	-	-	-	-	22.8	-	-	7.0
	BUC	66.2	84.5	-	38.3	47.4	68.4	-	27.5	-	-	-	-
Tracking	TAUDL	63.7	-	-	41.2	61.7	-	-	43.5	-	-	-	-
	UTAL	69.2	85.5	89.7	46.2	62.3	80.7	84.4	44.6	31.4	51.0	58.1	13.1
Source Domain	CAMEL	54.5	-	-	26.3	-	-	-	-	-	-	-	-
	TJ-AIDL	58.2	-	-	26.5	44.3	-	-	23.0	-	-	-	-
	CR-GAN	59.6	-	-	29.6	52.2	-	-	30.0	-	-	-	-
	MAR	67.7	-	-	40.0	67.1	-	-	48.0	-	-	-	-
	ECN	75.1	91.6	-	43.0	63.3	80.4	-	40.4	30.2	46.8	-	10.2
Intra-Camera	**MATE (Ours)**	88.7	97.1	98.2	71.1	76.9	89.6	92.3	56.6	46.0	65.3	71.1	19.1
Cross-Camera (Semi)	ResNet50*	66.1	-	-	42.1	50.0	-	-	30.3	-	-	-	-
	WRN50*	65.8	-	-	42.2	49.4	-	-	30.9	-	-	-	-
	MVC	72.2	-	-	49.6	52.9	-	-	33.6	-	-	-	-
Cross-Camera	HA-CNN	91.2	-	-	75.7	80.5	-	-	63.8	-	-	-	-
	SGGNN	92.3	-	-	82.8	81.1	-	-	68.2	-	-	-	-
	PCB	93.8	-	-	81.6	83.3	-	-	69.2	68.2	-	-	40.4
	JDGL	94.8	-	-	86.0	86.6	-	-	74.8	77.2	-	-	52.3
	OSNet	94.8	-	-	84.9	88.6	-	-	73.5	78.7	-	-	52.9

5.4.2 Comparing Different Person Re-ID Paradigms

As a novel re-id person scenario, it is informative and necessary to compare with existing other scenarios in the problem-solving and supervision cost perspectives. To that end, we compared ICS with existing representative re-id paradigms in an increasing order of training supervision cost:

1. *Unsupervised learning* (no supervision): RKSL [154], ISR [84], DIC [59], and BUC [82];

2. *Tracking data modelling*: TAUDL [67] and UTAL [68];

3. *Unsupervised domain adaptation* (source domain supervision): CAMEL [178], TJ-AIDL [155], CR-GAN [22], MAR [180], and ECN [200];

4. *Semi-supervised learning* (cross-camera supervision at small size): ResNet50 [49], WRN50 [182], and MVC [167];

5. *Supervised learning* (cross-camera supervision): HA-CNN [74], SGGNN [131], PCB [144], JDGL [195], and OSNet [202].

Table 5.2 presents a holistic comparative evaluation of different person re-id paradigms in terms of the model performance *versus* supervision requirement. We have the following observations:

1. Early unsupervised learning re-id models (RKSL, ISR, DIC), which rely on hand-crafted visual feature representations, often yield very limited re-id matching accuracy. While deep learning clearly improves the performance as shown in BUC, the results are still largely unsatisfied.

2. By exploiting tracking information including spatio-temporal object appearance continuity, TAUDL and UTAL further improve the model generalisation.

3. Unsupervised domain adaptation is another classical approach to eliminating the tedious collection of labelled training data per domain. The key idea is knowledge transfer from a source dataset (domain) with cross-camera labelled training samples. This strategy continuously pushes up the matching accuracy. It has a clear limitation that a *relevant* labelled source domain is assumed which however is not always guaranteed in practice.

4. While semi-supervised learning enables label reduction, the model performance remains unsatisfactory and is relatively inferior to unsupervised domain adaptation. This

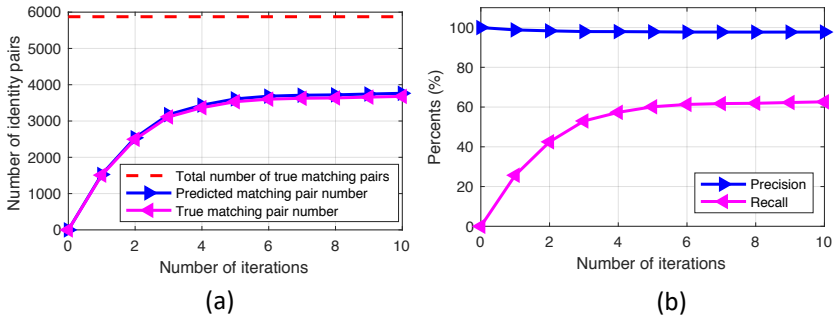


Figure 5.6: Dynamic statistics of cross-camera identity association over the training rounds. Dataset: Market-1501.

paradigm relies on expensive cross-camera identity annotation despite at smaller sizes.

5. With full cross-camera identity label supervision, supervised learning methods produce the best re-id performance among all the paradigms. However, the need for cross-camera identity association leads to very high labelling cost per domain, restricting significantly its scalability in realistic large scale applications typically with limited annotation budgets.

6. The ICS re-id is proposed exactly for solving this low cost-effectiveness limitation of the conventional supervised learning re-id paradigm, without the expensive cross-camera identity association labelling. Despite much weaker supervision, MATE can approach the performance of the latest supervised learning re-id methods on Market-1501. However, the performance gap on the largest dataset MSMT17 is still clearly bigger, suggesting a large room for further ICS re-id algorithm innovation.

5.4.3 Further Evaluations

We conducted a sequence of in-depth component evaluations for the MATE model on the Market-1501 dataset.

5.4.3.1 Component Analysis

We started by evaluating the three components of our MATE model: *Per-Camera Multi-Task* (PCMT) learning, *Cross-Camera Multi-Label* (CCML) learning, and *Curriculum Thresholding* (CT). The results in Table 5.3 show that: (1) Using the PCMT component alone, the model can already achieve fairly strong re-id matching performance, thanks to the ability of learning implicitly cross-camera feature representation via a specially designed multi-task

Table 5.3: Evaluating the model components of MATE: Per-Camera Multi-Task (PCMT) learning, Cross-Camera Multi-Label (CCML) learning, and Curriculum Thresholding (CT). Dataset: Market-1501.

Component	R1	R10	R20	mAP
PCMT	78.4	93.1	95.7	52.1
PCMT+CCML	85.3	96.2	97.6	65.2
PCMT+CCML+CT (**full**)	**88.7**	**97.1**	**98.2**	**71.1**

inference structure. (2) Adding the CCML component significantly boosts the accuracy, verifying the capability of our cross-camera identity matching strategy in discovering the underlying image pairs. (3) With the help of CT, a further performance gain is realised, validating the idea of exploiting curriculum learning and the design of our curriculum threshold.

As a key performance contributor, we further examined CCML by evaluating its essential part – cross-camera identity association. To that end, we tracked the statistics of self-discovered identity pairs across camera views over the training rounds, including the prevision and recall measurements. It is shown in Fig. 5.6 that our model can mine an increasing number of identity association pairs whilst maintaining very high precision which therefore well limits the risk of error propagation and its disaster consequence. This explains the efficacy of our cross-camera multi-label learning. On the other hand, while failing to identify around 40% identity pairs with further improvement to be made, our model can still achieve very competitive performance as compared to fully supervised learning models. This suggests that our method has already discovered the majority of re-id discrimination information from the associated identity pairs, missing only a small fraction embedded in those hard-to-match pairs. In this regard, we consider the proposed model is making a satisfactory trade-off between error association and knowledge mining. To check the impact of cross-camera identity association together with per-camera learning, we visualised the feature distribution change for a set of multi-camera images from a single person. It is observed from Fig. 5.7 that the same-person images are associated gradually in the re-id feature space, reaching a similar distribution as in the supervised learning case. This is consistent with the numerical performance evaluation above.

5.4.3.2 Hyper-Parameter Analysis

We examined the performance sensitivity of three parameters of MATE: the loss weight λ (default value 0.5) in Eq. (5.11), the lower (default value 0.5) and upper (default value 0.95) bound of curriculum threshold in Eq. (5.8). The evaluation in Fig. 5.8 shows that all these parameters have a wide range of satisfactory values in terms of performance. This suggests the convenience of setting up model training and good accuracy stability of our method.

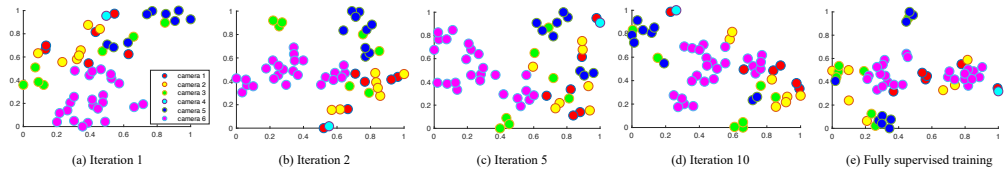


Figure 5.7: **(a-d)** The feature distribution evolution of a set of multi-camera images from a single random person over the training rounds, in comparison to **(e)** the feature distribution by supervised learning. Dataset: Market-1501. Best viewed in colour.

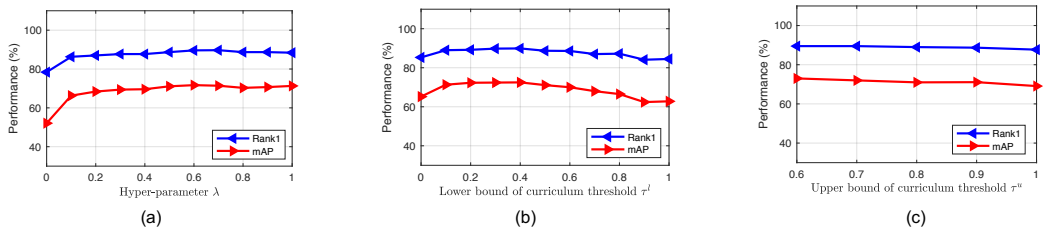


Figure 5.8: Hyper-parameter analysis: (**a**) the loss weight λ in Eq. (5.11), the (**b**) lower and (**c**) upper bound of curriculum threshold in Eq. (5.8). Dataset: Market-1501.

5.5 Conclusions

In this chapter, a novel person re-id paradigm, *i.e.*, intra-camera supervised (ICS) person re-id, is presented which is characterised by training re-id models with only per-camera independent person identity labels but no the conventional cross-camera identity labels. The key motivation is for eliminating the tedious and expensive process of manually associating identity classes across every pair of camera views in a surveillance network, which makes the training data collection too costly to be affordable in large-scale real-world application. To address the ICS re-id problem, a Multi-Task Multi-Label (MATE) learning model is formulated which is capable of fully exploiting the per-camera re-id supervision whilst simultaneously self-mining cross-camera identity association. Extensive evaluations were conducted on three re-id benchmarks to validate the advantages of the proposed MATE model over the state-of-the-art alternative methods in the proposed ICS learning setting. The detailed component analysis is also provided for giving insights on our model design. Extensive comparative evaluations have been conducted to demonstrate the cost-effectiveness advantages of the ICS re-id paradigm over existing representative re-id settings and the performance superiority of our MATE model over alternative learning methods. In addition, in-depth model component analysis is also performed to give insights on the MATE model formulation.

Chapter 6

Conclusions and Future Work

Person re-id is one of the fundamental problems in visual surveillance. Due to wide practical applications, it is attracting more and more attentions and substantial efforts have been made towards developing new technologies. This thesis focused on designing methodologies for learning discriminative features for person re-id. The first two chapters respectively presented an overview of person re-id and a summarization of existing related works. Extracting discriminative re-id features is inherently challenging due to (1) intra- and inter-personal variations, (2) domain variations and (3) difficulties in data creation and annotation. This thesis sequentially considered these three aspects with proposing three methodologies and one new person re-id setting. Specifically,

In Chapter 3, a robust metric learning, *i.e.*, Gaussian Mixture Importance Estimation (GMIE), has been proposed. Unlike KISSME, one of the popular metric learning method in person re-id, GMIE directly approximates the density ratio between the intra- and inter-personal variations. By adapting the Kullback-Leibler divergence technique, GMIE can maintain its re-id performance even in the high dimensional case, which is difficult for KISSME. In addition, thanks to the Gaussian Mixture Models used for approximating the density ratio, GMIE is also capable of capturing the multi modal properties existed in the underlying densities of intra- and inter-personal variations.

In Chapter 4, an unsupervised domain adaptive re-id framework has been proposed for extracting attribute-related features. Considering that most re-id datasets are not labeled with pedestrian attributes, a modified domain adaptation method has been proposed for adapting the attribute recognition model. Based on the observation that the attribute recognition performance degrades during domain adaptation procedure, an additional classifier has been added along with the discriminator.

In Chapter 5, a novel person re-id paradigm, *i.e.*, intra-camera supervised (ICS) person re-id, is presented which is characterized by training re-id models with only per-camera independent person identity labels but no the conventional cross-camera identity labels. The key motivation is for eliminating the tedious and expensive process of manually associating identity classes across every pair of camera views in a surveillance network, which

makes the training data collection too costly to be affordable in large-scale real-world application. To address the ICS re-id problem, a Multi-Task Multi-Label (MATE) learning model is formulated which is capable of fully exploiting the per-camera re-id supervision whilst simultaneously self-mining cross-camera identity association. Extensive evaluations were conducted on three re-id benchmarks to validate the advantages of the proposed MATE model over the state-of-the-art alternative methods in the proposed ICS learning setting. The detailed component analysis is also provided for giving insights on our model design. Extensive comparative evaluations have been conducted to demonstrate the cost-effectiveness advantages of the ICS re-id paradigm over existing representative re-id settings and the performance superiority of our MATE model over alternative learning methods. In addition, in-depth model component analysis is also performed to give insights on the MATE model formulation.

The potential research direction of future work can be on minimizing the supervision information required in training person re-id model. Annotating person re-id data requires an expensive and tedious data annotation process. This dramatically degrades the usability and scalability of re-id methods for large scale deployment in real-world application. Most of current works are trying to solve this problem based on unsupervised domain adaptation methods. However, this kind of methods need an auxiliary dataset that is fully annotated with identity labels. In addition, these methods have heavy reliance on the relevance and quality of source datasets which renders them less practically useful, since this assumption is not always valid. The proposed ICS person re-id setting in this thesis provides a way to significantly reduce the annotation efforts in creating dataset. The method MATE designed under this setting can automatically discover cross-camera identity associations. Although it gets satisfactory results on Market1501 dataset, its performance decreases with the increase of camera numbers in datasets. New methods or improvements on MATE are still needed under ICS person re-id setting. Fully unsupervised person re-id can be an alternative way to totally eliminate the data annotation process by training re-id model based on unlabeled data. Although fully unsupervised learning methods have been widely studied in other machine learning and computer vision topics, there are very few related works in person re-id domain, especially under deep learning framework.

Bibliography

[1] —. Security technologies top trends for 2019. IHS Market, 2019.

[2] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *ACM Computing Surveys*, 43(3):16, 2011.

[3] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 41–48, 2007.

[4] S. Bak, P. Carr, and J.-F. Lalonde. Domain adaptation through synthesis for unsupervised person re-identification. In *Proceedings of the European Conference on Computer Vision*, pages 189–205, 2018.

[5] S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Multiple-shot human re-identification by mean riemannian covariance grid. In *Proceedings of the IEEE International Conference on Advanced Video Signal-based Surveillance*, 2011.

[6] G. Baltus, D. Fox, F. Gemperle, J. Goetz, T. Hirsch, D. Magaritis, M. Montemerlo, J. Pineau, N. Roy, J. Schulte, et al. Towards personal service robots for the elderly. In *Proceedings of the Workshop on Interactive Robotics and Entertainment*, 2000.

[7] A. Bedagkar-Gala and S. K. Shah. A survey of approaches and trends in person re-identification. *Image and Vision Computing*, 32(4):270–286, 2014.

[8] M. Betke, E. Haritaoglu, and L. S. Davis. Real-time multiple vehicle detection and tracking from a moving vehicle. *Machine Vision and Applications*, 12(2):69–83, 2000.

[9] A. Bhuiyan, A. Perina, and V. Murino. Person re-identification by discriminatively selecting parts and features. In *Proceedings of the European Conference on Computer Vision Workshop.*, 2014.

[10] N. D. Bird, O. Masoud, N. P. Papanikolopoulos, and A. Isaacs. Detection of loitering individuals in public transportation areas. *IEEE Transactions on Intelligent Transportation Systems*, 2005.

[11] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan. Domain separation networks. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 343–351, 2016.

[12] X. Chang, T. M. Hospedales, and T. Xiang. Multi-level factorisation net for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[13] X. Chang, T. M. Hospedales, and T. Xiang. Multi-level factorisation net for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and*

Pattern Recognition, pages 2109–2118, 2018.

[14] B. Chen, W. Deng, and J. Hu. Mixed high-order attention network for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 371–381, 2019.

[15] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang. Similarity learning on an explicit polynomial kernel feature map for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1565–1573, 2015.

[16] J. Chen, Z. Zhang, and Y. Wang. Relevance metric learning for person re-identification by exploiting listwise similarities. *IEEE Transactions Image Processing*, 24(12):4741–4755, 2015.

[17] L.-F. Chen, H.-Y. M. Liao, M.-T. Ko, J.-C. Lin, and G.-J. Yu. A new lda-based face recognition system which can solve the small sample size problem. *Pattern recognition*, 33(10):1713–1726, 2000.

[18] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 403–412, 2017.

[19] Y. Chen, S. Duffner, A. Baskurt, A. Stoian, and J.-Y. Dufour. Similarity learning with listwise ranking for person re-identification. In *Proceedings of the IEEE International Conference on Image Processing*, pages 843–847. IEEE, 2018.

[20] Y. Chen, X. Zhu, and S. Gong. Person re-identification by deep learning multi-scale representations. In *Workshop of the IEEE international conference on computer vision*, pages 2590–2600, 2017.

[21] Y. Chen, X. Zhu, and S. Gong. Deep association learning for unsupervised video person re-identification. In *Proceedings of the British Machine Vision Conference*, 2018.

[22] Y. Chen, X. Zhu, and S. Gong. Instance-guided context rendering for cross-domain person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 232–242, 2019.

[23] Y. Chen, X. Zhu, W.-S. Zheng, and J.-H. Lai. Person re-identification by camera correlation aware feature augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(2):392–408, 2018.

[24] D. S. Cheng and M. Cristani. Person re-identification by articulated appearance matching. In *Person Re-Identification*. Springer, 2014.

[25] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *Proceedings of the British Machine Vision Conference*, volume 1, page 6, 2011.

[26] G. Csurka. Domain adaptation for visual applications: A comprehensive survey. *arXiv preprint arXiv:1702.05374*, 2017.

[27] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *Proceedings of International Conference on Machine Learning*, 2007.

[28] L. De Haan and A. Ferreira. *Extreme value theory: an introduction*. Springer Science & Business Media, 2007.

[29] W. Deng, L. Zheng, G. Kang, Y. Yang, Q. Ye, and J. Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[30] J. L. Elman. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99, 1993.

[31] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[32] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005.

[33] F. Fleuret, H. B. Shitrit, and P. Fua. Re-identification for improved people tracking. In *Person Re-Identification*, pages 309–330. Springer, 2014.

[34] P.-E. Forssén. Maximally stable colour regions for recognition and matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[35] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the International Conference on Machine Learning*, pages 1180–1189, 2015.

[36] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[37] N. Gheissari, T. B. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[38] M. Gill, A. Spriggs, J. Allen, M. Hemming, P. Jessiman, D. Kara, J. Kilworth, R. Little, and D. Swain. *Control room operation: findings from control room observations*. 2015.

[39] S. Gong, M. Cristani, S. Yan, and C. C. Loy. *Person re-identification*, volume 1. Springer, 2014.

[40] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[41] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Proceedings of the*

Advances in Neural Information Processing Systems, pages 2672–2680, 2014.

[42] M. Gou. http://robustsystems.coe.neu.edu.

[43] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proceedings of the IEEE international workshop on performance evaluation for tracking and surveillance*, 2007.

[44] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proceedings of the European Conference on Computer Vision*, pages 262–275. Springer, 2008.

[45] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *Proceedings of IEEE International Conference on Computer Vision*, 2009.

[46] Y. Guo and N.-M. Cheung. Efficient and deep person re-identification using multi-level similarity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2335–2344, 2018.

[47] A. Hampapur, L. Brown, J. Connell, S. Pankanti, A. Senior, and Y. Tian. Smart surveillance: applications, technologies and implications. In *Proceedings of the IEEE Pacific-Rim Conference On Multimedia*, pages 1133–1138, 2003.

[48] B. He, J. Li, Y. Zhao, and Y. Tian. Part-regularized near-duplicate vehicle re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3997–4005, 2019.

[49] 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*, pages 770–778, 2016.

[50] L. He, J. Liang, H. Li, and Z. Sun. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7073–7082, 2018.

[51] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[52] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *Proceedings of the European Conference on Computer Vision*, 2012.

[53] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen. Interaction-and-aggregation network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2019.

[54] S. Huang, J. Lu, J. Zhou, and A. K. Jain. Nonlinear local metric learning for person re-identification. *arXiv preprint arXiv:1511.05169*, 2015.

[55] S. Khamis, C.-H. Kuo, V. K. Singh, V. D. Shet, and L. S. Davis. Joint learning for attribute-consistent person re-identification. In *Proceedings of the European Conference on Computer Vision*, pages 134–146, 2014.

[56] F. M. Khan and F. Bremond. Unsupervised data association for metric learning in the context of multi-shot person re-identification. In *Proceedings of the IEEE International Conference on Advanced Video Signal-based Surveillance*, pages 256–262, 2016.

[57] I. S. Kim, H. S. Choi, K. M. Yi, J. Y. Choi, and S. G. Kong. Intelligent visual surveillance-a survey. *International Journal of Control, Automation and Systems*, 8(5):926–939, 2010.

[58] E. Kodirov, T. Xiang, Z. Fu, and S. Gong. Person re-identification by unsupervised l_1 graph learning. In *Proceedings of the European Conference on Computer Vision*, pages 178–195, 2016.

[59] E. Kodirov, T. Xiang, and S. Gong. Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification. In *Proceedings of the British Machine Vision Conference*, page 8, 2015.

[60] D. Koller, J. Weber, and J. Malik. Robust multiple car tracking with occlusion reasoning. In *Proceedings of the European Conference on Computer Vision*, pages 189–196. Springer, 1994.

[61] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[62] K. A. Krueger and P. Dayan. Flexible shaping: How learning in small steps helps. *Cognition*, 110(3):380–394, 2009.

[63] B. Lavi, M. F. Serj, and I. Ullah. Survey on deep learning techniques for person re-identification task. *arXiv preprint arXiv:1807.05284*, 2018.

[64] R. Layne, T. M. Hospedales, S. Gong, and Q. Mary. Person re-identification by attributes. In *Proceedings of the British Machine Vision Conference*, volume 2, page 8, 2012.

[65] A. Li, L. Liu, and S. Yan. Person re-identification by attribute-assisted clothes appearance. In *Person Re-Identification*, pages 119–138. Springer, 2014.

[66] 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.

[67] M. Li, X. Zhu, and S. Gong. Unsupervised person re-identification by deep learning tracklet association. In *Proceedings of the European Conference on Computer Vision*, pages 737–753, 2018.

[68] M. Li, X. Zhu, and S. Gong. Unsupervised tracklet person re-identification. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[69] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang. Person search with natural language description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1970–1979, 2017.

[70] W. Li and X. Wang. Locally aligned feature transforms across views. In *Proceedings*

of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3594–3601, 2013.

[71] W. Li, R. Zhao, and X. Wang. Human re-identification with transferred metric learning. In *Proceedings of the Asian conference on computer vision*, pages 31–44. Springer, 2012.

[72] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159, 2014.

[73] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014.

[74] W. Li, X. Zhu, and S. Gong. Harmonious attention network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2285–2294, 2018.

[75] Y. Li, C. Huang, C. C. Loy, and X. Tang. Human attribute recognition by deep hierarchical contexts. In *Proceedings of the European Conference on Computer Vision*, pages 684–700. Springer, 2016.

[76] Z. Li and D. Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[77] W. Liang, G. Wang, J. Lai, and J. Zhu. M2m-gan: Many-to-many generative adversarial transfer learning for person re-identification. *Proceedings of AAAI Conference on Artificial Intelligence*, 2018.

[78] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[79] S. Liao and S. Z. Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3685–3693, 2015.

[80] S. Liao, G. Zhao, V. Kellokumpu, M. Pietikäinen, and S. Z. Li. Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1301–1306, 2010.

[81] S. Lin, H. Li, C.-T. Li, and A. C. Kot. Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. 2018.

[82] Y. Lin, a. Dong, L. Zheng, Y. Yan, and Y. Yang. A bottom-up clustering approach to unsupervised person re-identification. In *Proceedings of AAAI Conference on Artificial Intelligence*, volume 2, 2019.

[83] 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*, 2019.

[84] G. Lisanti, I. Masi, A. D. Bagdanov, and A. Del Bimbo. Person re-identification by iterative re-weighted sparse ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(8):1629–1642, 2014.

[85] C. Liu, S. Gong, and C. C. Loy. On-the-fly feature importance mining for person re-identification. *Pattern Recognition*, 2014.

[86] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu. Pose transferrable person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4099–4108, 2018.

[87] X. Liu, W. Liu, H. Ma, and H. Fu. Large-scale vehicle re-identification in urban surveillance videos. In *2016 IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE, 2016.

[88] X. Liu, W. Liu, T. Mei, and H. Ma. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *Proceedings of the European Conference on Computer Vision*, pages 869–884. Springer, 2016.

[89] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262–282, 2007.

[90] Z. Liu, D. Wang, and H. Lu. Stepwise metric promotion for unsupervised video person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[91] M. Long, H. Zhu, J. Wang, and M. I. Jordan. Deep transfer learning with joint adaptation networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2208–2217, 2017.

[92] C. C. Loy, C. Liu, and S. Gong. Person re-identification by manifold ranking. In *ICIP*, pages 3567–3571. IEEE, 2013.

[93] C. C. Loy, T. Xiang, and S. Gong. Multi-camera activity correlation analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1988–1995. IEEE, 2009.

[94] C. C. Loy, T. Xiang, and S. Gong. Time-delayed correlation analysis for multi-camera activity understanding. *International Journal of Computer Vision*, 90(1):106–129, 2010.

[95] W.-L. Lu, J.-A. Ting, J. J. Little, and K. P. Murphy. Learning to track and identify players from broadcast sports videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1704–1716, 2013.

[96] W. Luo, T.-K. Kim, B. Stenger, X. Zhao, and R. Cipolla. Bi-label propagation for generic multiple object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1290–1297, 2014.

[97] B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by fisher vectors for person re-identification. In *Proceedings of the European Conference on Computer Vision*, pages 413–422. Springer, 2012.

[98] B. Ma, Y. Su, and F. Jurie. Covariance descriptor based on bio-inspired features for person re-identification and face verification. *Image and Vision Computing*, 32(6-7):379–390, 2014.

[99] L. Ma, H. Liu, L. Hu, C. Wang, and Q. Sun. Orientation driven bag of appearances for person re-identification. *arXiv preprint arXiv:1605.02464*, 2016.

[100] L. Ma, X. Yang, and D. Tao. Person re-identification over camera networks using multi-task distance metric learning. *IEEE Transactions on Image Processing*, 23(8):3656–3670, 2014.

[101] X. Ma, X. Zhu, S. Gong, X. Xie, J. Hu, K.-M. Lam, and Y. Zhong. Person re-identification by unsupervised video matching. *Pattern Recognition*, 65:197–210, 2017.

[102] L. v. d. Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[103] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley. Least squares generative adversarial networks. In *Proceedings of IEEE International Conference on Computer Vision*, pages 2813–2821, 2017.

[104] N. Martinel, C. Micheloni, and G. L. Foresti. Kernelized saliency-based person re-identification through multiple metric learning. *IEEE Transactions on Image Processing*, 24(12):5645–5658, 2015.

[105] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato. Hierarchical gaussian descriptor for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1363–1372, 2016.

[106] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato. Hierarchical gaussian descriptors with application to person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[107] T. Matsukawa and E. Suzuki. Person re-identification using CNN features learned from combination of attributes. In *Proceedings of the International Conference on Pattern Recognition*, pages 2428–2433, 2016.

[108] J. Meng, S. Wu, and W. Zheng. Weakly supervised person re-identification. In *IEEE International Conference on Computer Vision and Pattern Recognition*. IEEE, 2019.

[109] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang. Pose-guided feature alignment for occluded person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 542–551, 2019.

[110] A. Mignon and F. Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[111] P. Morerio, J. Cavazza, and V. Murino. Minimal-entropy correlation alignment for unsupervised deep domain adaptation. *arXiv preprint arXiv:1711.10288*, 2017.

[112] B. Munjal, S. Amin, F. Tombari, and F. Galasso. Query-guided end-to-end person

search. *arXiv preprint arXiv:1905.01203*, 2019.

[113] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1):51–59, 1996.

[114] S. Paisitkriangkrai, C. Shen, and A. Van Den Hengel. Learning to rank in person re-identification with metric ensembles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1846–1855, 2015.

[115] L. Pang, Y. Wang, Y.-Z. Song, T. Huang, and Y. Tian. Cross-domain adversarial feature learning for sketch re-identification. In *Proceedings of the ACM Multimedia Conference on Multimedia Conference*, pages 609–617, 2018.

[116] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *Proceedings of the International Conference on Computer Vision*, pages 261–268, 2009.

[117] P. Peng, Y. Tian, T. Xiang, Y. Wang, M. Pontil, and T. Huang. Joint semantic and latent attribute modelling for cross-class transfer learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[118] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian. Unsupervised cross-dataset transfer learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[119] D. M. Powers. Evaluation: from precision, recall and f-measure to ROC, informedness, markedness and correlation. 2011.

[120] B. J. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary. Person re-identification by support vector ranking. In *Proceedings of the British Machine Vision Conference*, page 6, 2010.

[121] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 91–99, 2015.

[122] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Proceedings of the Workshop on Benchmarking Multi-Target Tracking*, 2016.

[123] P. M. Roth, M. Hirzer, M. Koestinger, C. Beleznai, and H. Bischof. Mahalanobis distance learning for person re-identification. In *Person Re-Identification*, pages 247–267. Springer, 2014.

[124] A. A. Salah, T. Gevers, N. Sebe, and A. Vinciarelli. Challenges of human behavior understanding. In *International Workshop on Human Behavior Understanding*, pages 1–12. Springer, 2010.

[125] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2013.

[126] W. J. Scheirer, L. P. Jain, and T. E. Boult. Probability models for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11):2317–2324, 2014.

[127] H. Schütze, C. D. Manning, and P. Raghavan. Introduction to information retrieval. In *Proceedings of the International Communication of Association for Computing Machinery Conference*, page 260, 2008.

[128] W. R. Schwartz and L. S. Davis. Learning discriminative appearance-based models using partial least squares. In *2009 XXII Brazilian Symposium on Computer Graphics and Image Processing*, 2009.

[129] G. Sharma and F. Jurie. Learning discriminative spatial representation for image classification. In *Proceedings of the British Machine Vision Conference*, pages 1–11. BMVA Press, 2011.

[130] Y. Shen, H. Li, T. Xiao, S. Yi, D. Chen, and X. Wang. Deep group-shuffling random walk for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[131] Y. Shen, H. Li, S. Yi, D. Chen, and X. Wang. Person re-identification with deep similarity-guided graph neural network. In *Proceedings of the European Conference on Computer Vision*, 2018.

[132] Z. Shi, T. M. Hospedales, and T. Xiang. Transferring a semantic representation for person re-identification and search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4184–4193. IEEE, 2015.

[133] C. Song, Y. Huang, W. Ouyang, and L. Wang. Mask-guided contrastive attention model for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[134] J. Song, Y. Yang, Y.-Z. Song, T. Xiang, and T. M. Hospedales. Generalizable person re-identification by domain-invariant mapping network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 719–728, 2019.

[135] C. Spampinato, Y.-H. Chen-Burger, G. Nadarajan, and R. B. Fisher. Detecting, tracking and counting fish in low quality unconstrained underwater videos. In *Proceedings of the Third International Conference on Computer Vision Theory and Applications - Volume 1: VISAPP*, pages 514–519, 2008.

[136] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao. Multi-task learning with low rank attribute embedding for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3739–3747, 2015.

[137] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian. Deep attributes driven multi-camera person re-identification. In *Proceedings of the European Conference on Computer Vision*, pages 475–491, 2016.

[138] A. Subramaniam, A. Nambiar, and A. Mittal. Co-segmentation inspired attention networks for video-based person re-identification. In *Proceedings of the IEEE Inter-*

national Conference on Computer Vision, pages 562–572, 2019.

[139] M. Sugiyama, S. Nakajima, H. Kashima, P. V. Buenau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Proceedings of the Advances in Neural Information Processing Systems*, 2008.

[140] Y. Suh, J. Wang, S. Tang, T. Mei, and K. Mu Lee. Part-aligned bilinear representations for person re-identification. In *Proceedings of the European Conference on Computer Vision*, pages 402–419, 2018.

[141] B. Sun and K. Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Proceedings of the European Conference on Computer Vision*, pages 443–450. Springer, 2016.

[142] Y. Sun, Q. Xu, Y. Li, C. Zhang, Y. Li, S. Wang, and J. Sun. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 393–402, 2019.

[143] Y. Sun, L. Zheng, W. Deng, and S. Wang. Svdnet for pedestrian retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3800–3808, 2017.

[144] 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 *Proceedings of the European Conference on Computer Vision*, 2018.

[145] S. Tang, M. Andriluka, B. Andres, and B. Schiele. Multiple people tracking by lifted multicut and person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3539–3548, 2017.

[146] Z. Tang, M. Naphade, S. Birchfield, J. Tremblay, W. Hodge, R. Kumar, S. Wang, and X. Yang. Pamtri: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 211–220, 2019.

[147] D. Tao, Y. Guo, M. Song, Y. Li, Z. Yu, and Y. Y. Tang. Person re-identification by dual-regularized kiss metric learning. *IEEE Transactions on Image Processing*, 25(6):2726–2738, 2016.

[148] G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13, 2007.

[149] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *Proceedings of the European Conference on Computer Vision*, pages 589–600. Springer, 2006.

[150] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4068–4076, 2015.

[151] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 4, 2017.

[152] R. Volpi, P. Morerio, S. Savarese, and V. Murino. Adversarial feature augmentation for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5495–5504, 2018.

[153] H. Wang, S. Gong, and T. Xiang. Unsupervised learning of generative topic saliency for person re-identification. In *Proceedings of the British Machine Vision Conference*, 2014.

[154] H. Wang, X. Zhu, T. Xiang, and S. Gong. Towards unsupervised open-set person re-identification. In *Proceedings of the IEEE International Conference on Image Processing*, pages 769–773, 2016.

[155] J. Wang, X. Zhu, S. Gong, and W. Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[156] P. Wang, B. Jiao, L. Yang, Y. Yang, S. Zhang, W. Wei, and Y. Zhang. Vehicle re-identification in aerial imagery: Dataset and approach. *arXiv preprint arXiv:1904.01400*, 2019.

[157] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu. Shape and appearance context modeling. In *Proceedings of IEEE International Conference on Computer Vision*, 2007.

[158] Y. Wang, L. Wang, Y. You, X. Zou, V. Chen, S. Li, G. Huang, B. Hariharan, and K. Q. Weinberger. Resource aware person re-identification across multiple resolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8042–8051, 2018.

[159] L. Wei, S. Zhang, W. Gao, and Q. Tian. Person transfer gan to bridge domain gap for person re-identification. *arXiv preprint arXiv:1711.08565*, 2017.

[160] L. Wei, S. Zhang, W. Gao, and Q. Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 79–88, 2018.

[161] Z. Wei-Shi, G. Shaogang, and X. Tao. Associating groups of people. In *Proceedings of the British Machine Vision Conference*, pages 23–1, 2009.

[162] K. Q. Weinberger and L. K. Saul. Fast solvers and efficient implementations for distance metric learning. In *Proceedings of The Tnternational Conference on Machine Learning*, 2008.

[163] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 2009.

[164] A. Wu, W.-S. Zheng, X. Guo, and J.-H. Lai. Distilled person re-identification: Towards a more scalable system. In *Proceedings of the IEEE Conference on Computer*

Vision and Pattern Recognition, pages 1187–1196, 2019.

[165] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1249–1258. IEEE, 2016.

[166] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3415–3424, 2017.

[167] X. Xin, J. Wang, R. Xie, S. Zhou, W. Huang, and N. Zheng. Semi-supervised person re-identification using multi-view clustering. *Pattern Recognition*, 88:285–297, 2019.

[168] C. Xu, S.-H. Hsieh, C. Xiong, and J. J. Corso. Can humans fly? action understanding with multiple classes of actors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2264–2273, 2015.

[169] Y. Xu, B. Ma, R. Huang, and L. Lin. Person search in a scene by jointly modeling people commonness and person uniqueness. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 937–940. ACM, 2014.

[170] M. Yamada and M. Sugiyama. Direct importance estimationiwith gaussian mixture models. *IEICE transactions on Information and Systems*, 2009.

[171] Y. Yan, Q. Zhang, B. Ni, W. Zhang, M. Xu, and X. Yang. Learning context graph for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2158–2167, 2019.

[172] B. Yang, C. Huang, and R. Nevatia. Learning affinities and dependencies for multi-target tracking using a crf model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1233–1240. IEEE, 2011.

[173] L. Yang and R. Jin. Distance metric learning: A comprehensive survey. *Michigan State University*, 2(2):4, 2006.

[174] W. Yang, H. Huang, Z. Zhang, X. Chen, K. Huang, and S. Zhang. Towards rich feature discovery with class activation maps augmentation for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1389–1398, 2019.

[175] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li. Salient color names for person re-identification. In *Proceedings of the European Conference on Computer Vision*, 2014.

[176] M. Ye, A. J. Ma, L. Zheng, J. Li, and P. C. Yuen. Dynamic label graph matching for unsupervised video re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[177] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Deep metric learning for person re-identification. In *Proceedings of the IEEE International Conference on Pattern Recognition*, pages 34–39. IEEE, 2014.

[178] H.-X. Yu, A. Wu, and W.-S. Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 994–1002, 2017.

[179] H.-X. Yu, A. Wu, and W.-S. Zheng. Unsupervised person re-identification by deep asymmetric metric embedding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[180] H.-X. Yu, W.-S. Zheng, A. Wu, X. Guo, S. Gong, and J.-H. Lai. Unsupervised person re-identification by soft multilabel learning. pages 2148–2157, 2019.

[181] H.-X. Yu, W.-S. Zheng, A. Wu, X. Guo, S. Gong, and J.-H. Lai. Unsupervised person re-identification by soft multilabel learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2148–2157, 2019.

[182] S. Zagoruyko and N. Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[183] D. Zapletal and A. Herout. Vehicle re-identification for automatic video traffic surveillance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 25–31, 2016.

[184] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1239–1248, 2016.

[185] Z. Zhang, C. Lan, W. Zeng, and Z. Chen. Densely semantically aligned person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 667–676, 2019.

[186] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

[187] L. Zhao, X. Li, Y. Zhuang, and J. Wang. Deeply-learned part-aligned representations for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3219–3228, 2017.

[188] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2528–2535, 2013.

[189] 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 *Proceedings of the European Conference on Computer Vision*, pages 868–884. Springer, 2016.

[190] L. Zheng, Y. Huang, H. Lu, and Y. Yang. Pose invariant embedding for deep person re-identification. *IEEE Transactions on Image Processing*, 2019.

[191] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference*

on Computer Vision, pages 1116–1124, 2015.

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

[193] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian. Person re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1367–1376, 2017.

[194] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong. Partial person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4678–4686, 2015.

[195] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz. Joint discriminative and generative learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2138–2147, 2019.

[196] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3754–3762, 2017.

[197] Z. Zheng, L. Zheng, and Y. Yang. A discriminatively learned CNN embedding for person re-identification. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 14(1):13, 2018.

[198] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3652–3661. IEEE, 2017.

[199] Z. Zhong, L. Zheng, S. Li, and Y. Yang. Generalizing a person retrieval model hetero- and homogeneously. In *Proceedings of the European Conference on Computer Vision*, pages 172–188, 2018.

[200] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 598–607, 2019.

[201] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang. Camera style adaptation for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5157–5166, 2018.

[202] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[203] X. Zhu, A. Bhuiyan, M. L. Mekhalfi, and V. Murino. Exploiting gaussian mixture importance for person re-identification. In *Proceedings of the IEEE International Conference on Advanced Video Signal-based Surveillance*, pages 1–6. IEEE, 2017.

[204] X. Zhu, P. Morerio, and V. Murino. Unsupervised domain adaptive person re-identification based on pedestrian attributes. In *Proceedings of the IEEE International Conference on Image Processing*, 2019.

[205] X. Zhu, X. Zhu, M. Li, V. Murino, and G. Shaogang. Intra-camera supervised person re-identification: A new benchmark. In *Proceedings of the IEEE International Conference on Computer Vision, Second Workshop and Challenge on Real-World Face and Object Recognition from Low-Quality Images and Videos*. IEEE, 2019.