2022

Learning Representations for Human Identification

Sinan Sabri
West Virginia University, sisabri@mix.wvu.edu

Learning Representations for Human Identification

Sinan Sabri

Dissertation submitted to the
Benjamin M. Statler College of Engineering and Mineral Resources
at West Virginia University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in
Computer Engineering

Gianfranco Doretto, Ph.D., Chair
Hanny Ammar, Ph.D.
Jeremy Dawson, Ph.D.
Jason Gross, Ph.D.
Katerina Goseva-Popstojanova, Ph.D.

Lane Department of Computer Science and Electrical Engineering

Morgantown, West Virginia
2022

Keywords: Human Identification, Person Re-Identification,
Face Recognition, Learning Representation, Deep Learning,
Low Rank Representation, Information Bottleneck

Abstract

Learning Representations for Human Identification

Sinan Sabri

Long-duration visual tracking of people requires the ability to link track snippets (a.k.a. tracklets) based on the identity of people. In lack of the availability of motion priors or hard biometrics (e.g., face, fingerprint, or iris), the common practice is to leverage soft biometrics for matching tracklets corresponding to the same person in different sightings. A common choice is to use the whole-body visual appearance of the person, as determined by the clothing, which is assumed to not change during tracking. The problem is challenging because distinct images of the same person may look very different, since no restrictions are imposed on the nuisance factors of variation, such as pose, illumination, viewpoint, background, and sensor noise, leading to very high intra-class variances, which make this human identification task still prone to high mismatch rates.

We introduce and study models for learning representations for human identification that aim at reducing the effects of nuisance factors. First, we introduce a modeling framework based on learning a low rank representation, which can be applied to face as well as whole-body images. The goal is to not only learn invariant representations for each identity, but also to promote a uniform inter-class separation to further reduce mismatch rates. Another advantage of the approach is a fast procedure for computing and comparing invariant representations for recognition and re-identification. Second, we introduce a learning framework for fusing representations of multiple biometrics for human identification. We focus on the face modality and clothing appearance and develop a representation fusion approach based on the Information Bottleneck method.

In the last part of the dissertation, we improve person re-identification by decreasing the effects of nuisance factors via multi-task learning. We design and combine improved versions of classification and distance metric losses. Classification losses improve their performance by imposing restrictions on the computation of their outputs. This makes their training harder. We mitigate this by investigating the combination of multiple tasks, such as attribute and metric learning, that might regularize the training while improving performance. Finally, we also include the explicit modeling of nuisance factors such as pose, to further improve the invariance of representations. For each model, we show the benefits of the proposed methods by characterizing their performance based on publicly available benchmarks, and by comparing them with the state of the art.

Dedication

This dissertation is dedicated to my parents for their unconditional love and support. I also dedicate this work to my beloved wife who has been a constant source of support and encouragement during the challenges of doctoral degree and life, and to my wonderful children, Mariam, Yazan, and Lena, whom brought to my life happiness and joy.

Acknowledgments

I want to especially thank my PhD advisor, Dr. Gianfranco Doretto, for his advice and support, as well as all my labmates at the Vision and Learning Group. I also acknowledge all my teachers during my student life for their encouragement and motivation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem Definition

There is a growing demand on enhancing public safety in recent years especially with the wide spread of surveillance cameras in many public places such as parks, shopping centers, stadiums, streets, etc. For instance, surveillance applications can be used to search, track and identify suspect individuals. These surveillance applications in one hand try to match the visual appearance of a person appears in one view with his/her visual appearance in other views over a period of time, these different views are non-overlapping. On the other hand, they try to distinguish the person of interest from other candidates that might have similar visual appearance to the person of interest. Relying on human to monitor and analyze surveillance videos is inefficient and time consuming. It is impossible for an operator to monitor several cameras that cover different views simultaneously with the purpose of find the person of interest. In addition to that, it is hard to maintain a consistent performance over time because human knowledge cannot be transferred from one operator to another. Thus, having an intelligent automotive video surveillance is an essential in order to detect, track and identify person of interest through different cameras. This has drawn attention to human identification task in computer vision [2].

Human identification [71, 53, 54, 56, 57] is the task of assigning the same identity to tightly cropped images of people, based solely on their whole body appearance information, with the assumption that clothing has not changed between sightings. Two ways have been used to tackle this problem, metric learning [71, 58, 53, 54, 56, 57, 59] and learning feature representation [60, 61, 62, 63, 64]. Metric learning uses training data to search for effective distance functions to compare people across different cameras. It aims to minimize the distance between same person images and maximize the distance between different person images. On the other hand, feature learning aims to learn a discriminative features representation for person images. The problem is challenging because distinct images of the same person may look very different, since no restrictions are imposed on the nuisance factors of variation, such as pose, illumination, viewpoint, background, and sensor noise, causing a high intra-class variance. Figure 1.1 shows the effects of these nuisance factors on person images and here are explanations of these factors:

- **Pose:** the whole body poses such as front, back or side, and with different moving actions like moving legs and arms in different direction has a great impact on person visual appearance. .
- **Illumination:** changes in intensity of the day light, illumination changes in color between indoor and outdoor environments.
- **Viewpoint:** the angle of view, different views of the same person lead to different details.
- **Background:** person images contain background regions due to irregular shape of persons in images. These background regions add noise to the learning model.

In order to address that challenge, the research landscape has evolved from developing feature-based models [60, 61] coupled with metric learning [65], to developing dedicated deep learning architectures [66] trained with classification and verification losses [67], to

developing specialized deep learning schemes [68, 69] aiming at extracting more robust feature embedding by leveraging powerful pretrained backbone architectures like ResNet-50 [70]. Among the more recent trends, there has been also the idea of learning feature embedding directly suitable for re-identification, by improving the ability to control how losses deal with intra-class and inter-class variances, while giving much less importance to the explicit modeling of the nuisance factors of variation. The underlying assumption is that recent powerful architectures should have the capacity to become invariant to nuisance factors, simply by teaching them how to minimize intra-class variation, and maximize inter-class separation.

1.2 Motivation and Challenges

In this dissertation, we are trying to improve the performance of a human identification model by learning representations of person images that minimize the effect of nuisance factors that have been explained in the previous section. Here we address the supervised learning where we train the proposed model on a training data that has person images with their corresponding labels. Then at the test time, we have the probe images from galley trying to match them with person images from the testing data. There is no overlapping between images of the same candidate in the probe and the gallery. In the next section we will discuss approaches used to address learning invariant embedding.

1.2.1 Learning Invariant Subspace Representations

Sparse coding has led to impressive performance even for image classification [5, 6]. However, sparse coding dictionary learning was shown to be sensitive to training samples corrupted by structural nuisance factors, such as occlusions, disguise, pose, lighting variations, and so on.

This has motivated the development of low-rank matrix decomposition approaches [4,

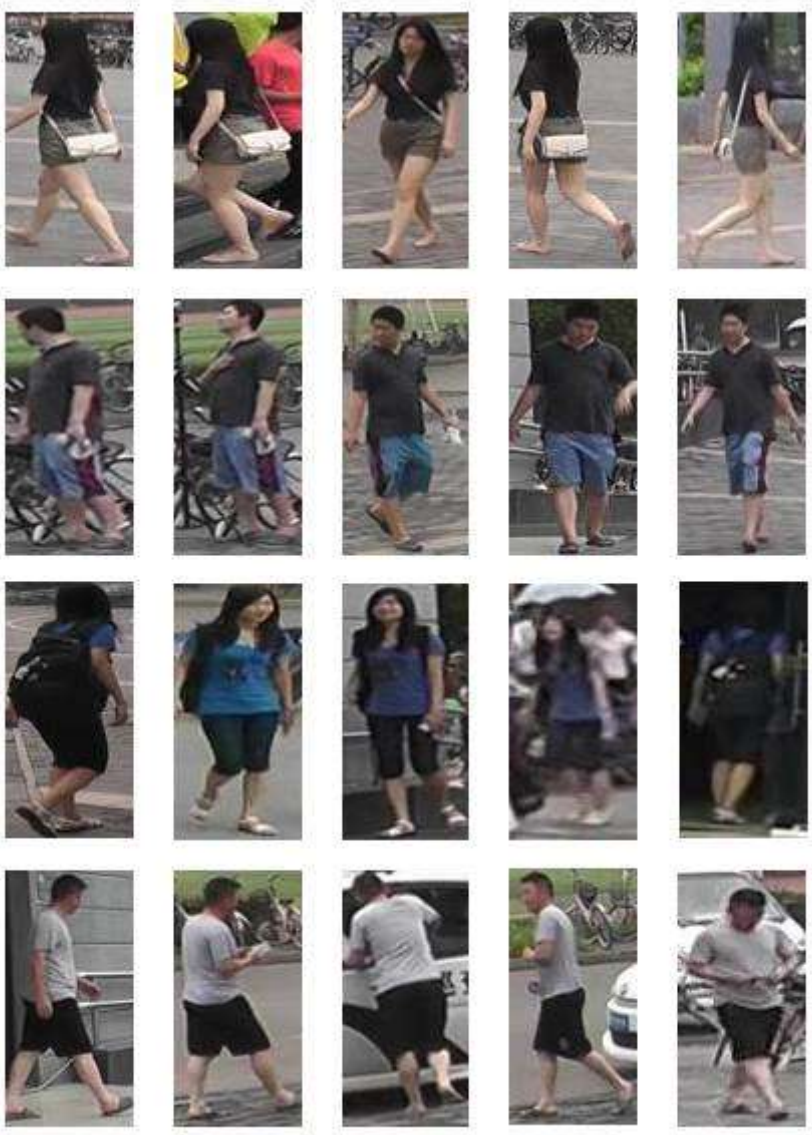


Figure 1.1: Example of nuisance factors on person images. Each row represents an identity.

11, 12, 13], which have the ability to learn a representational dictionary even in presence of corrupted data. Those approaches build a generative representation of the data that focusses on capturing all the information descriptive of an entity. This leads to complex training and testing for building robustness against, and filtering out unwanted data variations due to nuisance factors.

1.2.2 Non-Cooperative Human Identification

Searching for a person in a large video archive, possibly made of videos acquired by a network of surveillance cameras, is a fundamental task because it allows tracing down when and where a person was present in the scene. Large scale urban surveillance systems can tell a lot of required information about human such as when and where they enter a city, where they go, whether they are on the watch list and so on.

In such a scenario, quite often the human identification task can only rely on biometric traits acquired in unconstrained conditions from non-cooperative subjects. More specifically, this means that both gallery and probe images of someone's trait may be heavily corrupted by noise or other nuisance factors, such as the pose of an individual, or the illumination and occlusions. In such hostile conditions, human identification is typically attempted via face recognition (if enough image resolution is provided).

Although it has improved significantly even with corrupted probe and gallery images, there are still plenty of unfriendly scenarios where face recognition lack robustness. Hence, there is a critical need to reduce those cases in order to increase the effectiveness of human identification for searching unconstrained video archives.

1.2.3 Discriminative and Metric Embedding Learning

Recent deep learning approaches postulate that powerful architectures have the capacity to learn feature representations invariant to nuisance factors, by training them with losses that minimize intra-class variance and maximize inter-class separation, without modeling

nuisance factors explicitly. The dominant approaches use either a discriminative loss with margin, like the softmax loss with the additive angular margin, or a metric learning loss, like the triplet loss with batch hard mining of triplets.

We motivated by observing the limitations imposed by a margin based softmax loss onto the gradient flow that supervises the training of the embedding. We verified that adding a triplet loss as a regularizer serves as proxy for the missing gradient directions, and enables learning a better embedding. Moreover, the joint loss achieves its best performance when we do not require a margin in the softmax portion when it is used to expand the directions of the gradient flow.

1.3 Contributions and Dissertation Structure

In this dissertation, we propose approaches tackle problems explained in the previous section. Chapter two proposes a low-rank model to represent person images by their invariant components. Chapter three addresses the human identification problem when the probe and the gallery data acquired in unconstrained scenarios. Chapter four addresses the limitations imposed by a margin based softmax loss onto the gradient flow via proposing joint learning of softmax loss and triplet loss.

1.3.1 Learning Invariant Subspace Representations

In this chapter we present a low-rank modeling framework capable of capturing all the descriptive information of a person image (either a face image or a whole body image) that we denoted as the sufficient component, and emphases on learning a representation that is invariant to nuisance factors, denoted as the invariant component.

The main advantage of this approach is a fast procedure for computing and comparing invariant components for recognition and re-identification which can be achieved by a simple matrix multiplication. However, the main challenge of this approach is that different person

images may originate the same invariant component, thus preventing their discrimination. We will show that the proposed framework not only learns different invariant representations for different person images, but such representations promote a uniform inter-class separation. The approach couples simple geometry tools with advances in low-rank matrix recovery theory [4], and develops a supervised model for learning the proposed invariant representation, which spans an invariant subspace. Such subspace has to be orthogonal to the variation subspace, generated by data variation induced by nuisance factors on all the person images.

1.3.2 Non-Cooperative Human Identification

In this chapter we address the human identification problem with probe and gallery data acquired in unconstrained scenarios by developing a robust approach that fuses the face modality with the clothing appearance information. We extended our invariant component matching method in Chapter two for using kernel machines. The proposed method based on independent components and distance matching. Then, we developed a new approach for jointly exploiting face and whole body appearance for human identification. Our fusion approach based on Information Bottleneck (IB) [119] method. There is no such type of fusion has been studied.

We compare our proposed fusion approach with other standard fusion methods such as Sum Rule, Product Rule, Min Rule and Max Rule and we also compare our porposed approach on different state-of-the-art approaches. We demonstrated the performance of the joint face-clothing approach on our surveillance video archive. The proposed fusion approach improves the recognition rate over the-state-of-the-art approaches with standard fusion.

1.3.3 Discriminative and Metric Embedding Learning

The dominant deep learning approaches for human identification use either a discriminative loss with margin, like the softmax loss with the additive angular margin, or a metric learning loss, like the triplet loss with batch hard mining of triplets.

In this chapter, we improve person re-identification by combining the softmax and the triplet loss, to further improve the learning of a feature embedding. The intent is for the triplet loss to help the softmax further decrease intra-class variation, and increase inter-class distance by letting the triplet loss be the proxy for the gradient supervision that the embedding normalization has restricted, and we specify under what conditions this may happen. We also observe that the same strategy used to form the batch of triplets can be used in tandem with the softmax loss to prevent issues due to dataset imbalance, which are common in person re-identification. We perform an extensive evaluation of the proposed combination of losses on the latest person re-identification datasets. We found that not only this approach can deliver state-of-the-art performance, but the appropriate use of the combined loss does not require for the softmax component to use any margin, showing the importance of the contribution added by the triplet component.

In addition to that, we further improve person re-identification by integrating the attribute into our person re-identification model. Person attributes used in person re-identification and showed robustness against variation of viewpoint, illumination and pose. This is achieved via injecting M attribute classifiers set on part of the embedding. These attribute classifiers will help to extract attribute specific features that help improve the re-identification accuracy. Finally, we incorporate pose information into person re-identification model to overcome the problem of pose variation and to learn pose-invariant embedding.

1.4 Related Work

1.4.1 Invariant Subspace Representation

An entity (e.g., the vectorized version of the image pixels of a person) can be modeled by two additive components, sufficient component $s \in R^m$ (represents all the information necessary to recognize the entity) and variation component $v \in R^m$ (represents how the data point differs from the sufficient component by the effect of nuisance factors). Thus, data point can be modeled as

$$x \doteq s + v \tag{1.1}$$

Model (2.1) has been implicitly adopted by the most successful approaches to the face recognition problem. In particular, the SRC method [3] aims at "carefully" composing each class training data matrix in such a way that the selected samples are able to represent the salient components s_i's in the best possible way. The matching between a test point $x = s + v$, and a salient component s_i (i.e., the classification), is based on sparse coding and residual computation, and has demonstrated a remarkable robustness against the variation component v, leading to high recognition rates. The SRC approach has been further improved against potential corruptions of the test data point [41, 40]. For instance, [39] improves upon occlusions and computational cost, [42] robustifies the sparse coding problem by computing a sparsity-constrained maximum likelihood solution, [43] simultaneously handles the misalignment, pose and illumination invariance, and [44] addresses the problem of reducing the large amount of training data needed by SRC to be effective. To address the more general case where also the training data is highly affected by nuisance factors, and a "careful" composition of training data matrix is not possible, the SRC approach has been augmented in different ways. In [11] a low-rank matrix recovery [4] ap-

proach is designed for pre-processing the corrupted training data. After this step, the SRC method can be applied more effectively. Another approach, [34], proposes to apply sparse coding for modeling the sufficient component by learning a dictionary of prototypes, each of which, given by the average of the data in the class training data matrix , is meant to approximate s_i . In addition, sparse coding is also used for modeling the variation subspace. The concatenation of the prototype and the variation dictionaries form a new dictionary with which the SRC method can be applied more effectively.

1.4.2 Discriminative and Metric Embedding Learning

Human identification is a challenging task due to the nuisance factors of variation. Two ways have been used to tackle the human identification problem, discriminative metric learning [53, 73, 76, 63, 86, 88, 93] and discriminative feature representation [66, 58, 78, 92, 59]. Metric learning uses training data to search for effective distance functions to compare candidates across different cameras. [71] proposes one-shot metric learning by splitting the person re-identification metric into two components. Texture component trained on intensity images in order to make the learned embedding color invariant, while the color component is learned via patch based metric learning to address differences in camera color distributions. [58] proposes a pairwise metric learning trained on pairs of samples from different cameras in order to reduce the computational effort. [53] divides images in horizontal stripes and proposes a Local Maximal Occurrence (LOMO) descriptor to analyze the horizontal occurrence of local features and maximize the occurrence in order to make stable representation. The metric is learned on a discriminant low dimensional subspace using Cross-view Quadratic Discriminant Analysis (XQDA). [63] applies the Local Fisher Discriminant analysis for person Re-identification. [57] matches images in a discriminative null space of the training data to overcome the small sample size problem in person re-identification distance metric learning.

Feature representation is a key task in human identification. [60] uses HSV color his-

togram to construct effective feature representation. [61] investigates the role of different feature types given different appearance attributes using random forest. [64] uses Biologically Inspired Features (BIF) and covariance descriptors to improve the robustness to the illumination variation. [63] proposes a late fusion method at score level to identify feature effectiveness in a query-adaptive manner.

Recent works use deep learning to learn robust feature representations. [118] proposes to jointly learn features from multiple domains and then fine-tuning with domain guided drop out for the specific domain. [55] offers a deep convolutional architecture trained on pairs of images capable of learning features and similarity metric simultaneously. [114] combines the CRF model with DNN to learn more consistent multi-scale similarity metrics for person re-identification. [115] employs partition strategy on convolutional features. [78] learns embedding of the person image on a hypersphere manifold using spherical loss. [111] proposes a part aligned person representation by detecting body regions that are discriminative for person re-identification, while [75] learns full body and body parts features through a multi scale context aware network. Pose is used in [117, 80, 81, 82] to obtain person pose information. [116] addresses the limitation of CNNs in representing person images with large variations in body pose and scale by proposing a module to conclude the receptive fields according to the pose and scale of the input person image. [69] explores diverse discriminative visual cues without the assistance of pose estimation and human parsing by proposing Overlapped Activation Penalty (OAP) loss. [112] proposes a Fully Attention Block (FAB) plugged into a CNN to overcome the misalignment problem and to localize discriminative local features. Generative adversarial networks (GANs) [105] have been used in person re-identification. [109, 108, 68] aims to decompose pose information from image features via adversarial learning.

Person attributes used in person re-identification and showed robustness against variation of viewpoint, illumination and pose. [83] manually annotated person attributes for Market-1501 [84] dataset and DukeMTMC-ReID [85] dataset. It proposes attribute-person

recognition (APR) network capable of learning appearance and attributes features by combining identity classification task and attribute recognition task. [86] uses attribute classifier trained on a separate dataset to support the leaning of CNN features. [87] transforms attribute recognition from high level layer to mid-level layer. [88] jointly learn appearance and attribute representation via multi task learning. To learn discriminative person body parts, [89] utilizes person attribute information by integrating attribute features with identity and body part classification. [90] proposes a multi task network to learn identity part level representation and attribute global representation. [91] uses person attributes to detect attribute body parts in order to extract and refine local features capable of handling body part misalignment.

Chapter 2

Learning Invariant Subspace Representations

2.1 Introduction

Sparse representation and low-rank matrix decomposition approaches have been successfully applied to several computer vision problems. Approaches based on sparse representation [3] and low-rank matrix decomposition [4] have demonstrated great potential for addressing the problem of human identification, based on matching face images. Sparse coding has led to impressive performance even for image classification [5, 6]. Similarly, low-rank methods, after being applied to domains such as segmentation and grouping [7], tracking [8], and 3D visual recovery [9], are now also being used for classification [10]. For face recognition the sparse representation-based classification (SRC) method [3] has shown robustness with respect to a high degree of noise and occlusions in the test images. At the same time, sparse coding dictionary learning was shown to be sensitive to training samples corrupted by structural nuisance factors. This has motivated the development of low-rank matrix decomposition approaches [4, 11, 12, 13], which have the ability to learn a representational dictionary even in presence of corrupted data. Those approaches build a

generative representation of the data that focusses on capturing all the information descriptive of an entity. This leads to complex training and testing for building robustness against, and filtering out unwanted data variations due to nuisance factors.

This chapter presents a low-rank modeling framework capable of capturing all the descriptive information of a person image that we denoted as the sufficient component, and emphases on learning a representation that is invariant to nuisance factors, denoted as the invariant component. The main advantage of this approach is a fast procedure for computing and comparing invariant components for recognition and re-identification which can be achieved by a simple matrix multiplication. However, the main challenge of this approach is that different person images may originate the same invariant component, thus preventing their discrimination. We will show that the proposed framework not only learns different invariant representations for different person images, but such representations promote a uniform inter-class separation. The approach couples simple geometry tools with advances in low-rank matrix recovery theory [4], and develops a supervised model for learning the proposed invariant representation, which spans an invariant subspace. Such subspace has to be orthogonal to the variation subspace, generated by data variation induced by nuisance factors on all the person images.

While the framework is grounded on geometry, we will show how it relates to metric learning [24, 26, 21, 28], typically used for improving nearest neighbor (NN) classification based on the Euclidean distance. We will show that learning the invariant components is equivalent to learning the representatives of a set of entities (or classes), thus classification is based on identifying the nearest invariant component. Less intuitively, the same invariant components define a global metric, and also a local metric. This is important because local metric learning approaches [45, 46, 47, 31], improve upon global ones by taking into account the variability of the discriminative power of features across different neighborhoods. In particular, most of the approaches learn local metrics for different neighborhoods independently, and use regularization to avoid overfitting. Our framework learns the invariant

components, and therefore the local metrics jointly. In addition, their interpretation as a global metric is shown to promote uniform inter-class separation.

2.2 Invariant Subspace Learning

Let assume that a data point $x \in R^m$, representing an entity (e.g., the vectorized version of the image pixels oof a person, i.e. face or whole body), can be modeled by two additive components. The first one, $s \in R^m$, represents all the information necessary to recognize the entity (e.g., everything that describes the specific identity of the individual depicted by the person image). From a statistical point of view, we can imagine s to be the equivalent of a sufficient statistic for recognition, and we refer to it as the sufficient component. The second component, $v \in R^m$, represents how the data point differs from the sufficient component by the effect of nuisance factors, which are not descriptive of the entity. For instance, the image of a person might be modified by different lighting conditions, poses, occlusions, etc. We call variation subspace, $V \subset R^m$, the space where the variation component v is defined. It is assumed that v spans V as changes in nuisance factors affect a data point, which is modeled as

$$x \doteq s + v \tag{2.1}$$

If $P_v : R^m \rightarrow V$ is the projection operator mapping an m-dimensional vector onto V, x can be further decomposed as

$$x = (P_V s + v) + (s - P_V s) \tag{2.2}$$

In particular, the first component $a = PVs + v$, is defined in V, whereas the second component $b = s - P_V s$, is defined in the orthogonal complement of the variation space,

$V^{\perp}$.

The decomposition $x = a + b$ has the following property. Let us assume that x_1 and x_2 are two different points representing the same entity. According to (2.1), it must be that $x_1 = s + v_1$ and $x_2 = s + v_2$, because they have been affected by different nuisance factors. This means that $a_1 = P_V s + v_1$, and $a_2 = P_V s + v_2$; however, $b_1 = s - P_V s = b_2$, which highlights that the component b is invariant to the changes induced by the nuisance factors. We refer to the subspace where b is defined as the invariant subspace B, which will be a subspace of $V^{\perp}$.

Let assume that a set of n training data samples from N different entities, or object classes (e.g. images of faces or whole body appearances), are given, where each class i has n_i samples. Every sample x_j is modeled according to (1), and the data concatenated into a matrix $X = [X_1, X_2, ..., X_N] \in R^{m \times n}$, where $X_i \in R^{m \times n_i}$ is the training data matrix obtained by lining up the samples for class i.

Since every data point is modeled as $x_j = a_j + b_j$, the training data set X , can be decomposed by $X \doteq A + B$, where $A \in R^{m \times n}$ collects all the a_j's, and $B \in R^{m \times n}$ collects all the invariant components, b_j's. We assume that the variation subspace V has a finite dimension, which is lower than $\min\{m, n\}$. This is reasonable because it states that there are enough data for learning the variation subspace of interest, it allows avoiding over- fitting, and it makes the problem tractable. Therefore, attempting to recover A , which in turn allows recovering B , entails solving a low-rank matrix recovery problem.

In practice, the training data will also be affected by noise. We admit that a small percentage of the entries of X are corrupted by values not modeled by the variation and invariant components, which means that such noise should be sparse. This will account for data deviations unlikely to be captured by a finite dimensional linear subspace, such as those induced by image saturations, like image glare, or the presence of strong edges. Therefore, if $E \in R^{m \times n}$ is the matrix of sparse noise, the model for the training dataset is given by

$$X \doteq A + B + E \tag{2.3}$$

2.3 Low-rank Matrix Recovery

This section reviews the standard low-rank matrix recovery problem with sparse noise. Low-rank (LR) matrix recovery seeks to decompose a data matrix X into $A + E$, where A is a low-rank matrix and E is the associated sparse error. More precisely, given the input data matrix X , LR minimizes the rank of the matrix A while reducing $\| E \|_0$ to derive the low-rank approximation of X . Since the aforementioned optimization problem is NP-hard, [4] proposed to relax the original problem into the following tractable formulation

$$\min_{A,E} \|A\|_* + \alpha\|E\|_1 \qquad \text{s.t. } X = A + E \, . \tag{2.4}$$

In (2.4), the nuclear norm $\| A \|_*$ (i.e. the sum of the singular values) approximates the rank of A , and the ℓ_0-norm $\| E \|_0$ is replaced by the ℓ_1-norm $\| E \|_1$, which sums up the absolute values of the entries of E. It is shown in [4] that solving the relaxed version of the problem (2.3) is equivalent to solving the original low-rank matrix approximation problem, as long as the rank of A to be recovered is not too large and the number of errors in E is small (sparse). To solve the optimization problem (2.4) it is possible to apply the efficient method of augmented Lagrangian multipliers (ALM) [14]. In face recognition X represents the gallery of images of N subjects. By performing the low-rank matrix recovery (2.4), X gets decomposed into $A = [A_1, ..., A_N]$, and $E = [E_1, ..., E_N]$. The desired effect is for a subject i to produce a low-rank matrix A_i with columns that look very much alike and span a very narrow space around the sufficient component s_i [11]. The corresponding sparse matrix E_i is expected to pick up the variation components, caused by nuisance factors (e.g., occlusions, disguise, lighting variations, pose, etc.).

Unlike previous work, we do not learn a dictionary, and the columns of the low-rank matrix A are meant to span the variation space V , not the space of the sufficient components. Discriminability comes from learning the invariant components B , which leads to a very simple and efficient rule for classification, and can promote class separation with a supervised learning approach described in the following section.

2.4 Supervised Learning

To learn model (2.3), standard LR (2.4) is insufficient because we also need to learn the invariant components B . To do so, we need to take into account the geometric, and invariance constraints of (2.3).

2.4.1 Geometric Constraint

In particular, the invariant subspace should be included in the orthogonal complement of the variation subspace $V^{\perp}$. Therefore, A and B should satisfy the relationship

$$B^T A = 0 \tag{2.5}$$

2.4.2 Invariance Constraint

Given two data points $x_1 = a_1 + b_1 + e_1$ and $x_2 = a_2 + b_2 + e_2$, if they are representative of the same class i , the invariant components should be the same, i.e. $b_1 = b_2$. To express this in an algebraic form, b_1 and b_2 should be the solution to the linear system given by the equations $b_1 = \frac{1}{2}(b_1 + b_2)$, and $b_2 = \frac{1}{2}(b_1 + b_2)$. For n data points, where $B = [B_1, B_2, , B_N]$, the constraint on the invariant components would be $b_1 = b_2 = b_{n1}$, for $B_1, ...$, and $b_{n-n_N+1} = b_{n-n_N+2} = ... = b_n$, for B_N. This can still be expressed in an algebraic form, by generalizing the system of two linear equations to the following expression

$$B(I - Q) = 0, \tag{2.6}$$

where I is the identity matrix, and Q is a block-diagonal matrix, given by $Q \doteq \text{diag}\ (\frac{1}{n_1}1_{n_1}1_{n_1}^T, \frac{1}{n_2}1_{n_2}1_{n_2}^T, ..., \frac{1}{n_N}1_{n_N}1_{n_N}^T)$, and 1_{n_i} is a column vector with ones of length n_i.

In order to learn A and B, we propose to augment problem (2.4) with model (2.3), the geometric constraint (2.5), and the invariance constraint (2.6). In particular, to make the problem more tractable, the geometric and invariance constraints are relaxed to the penalty terms $\parallel B^T A \parallel_F^2$, and $\parallel B(IQ) \parallel_F^2$ in the following optimization problem

$$\begin{aligned} \min_{A,B,E} \|A\|_* + \alpha\|E\|_1 + \beta\|B(I-Q)\|_F^2 + \gamma\|B^\top A\|_F^2 \\ \text{s.t. } X = A + B + E\ , \end{aligned} \tag{2.7}$$

where $\parallel\ .\ \parallel_F$ indicates the Frobenius norm, and α ,β, and γ are penalty weights. Note that the addition of the invariance constraint (2.6) as a penalty, through Q injects the training dataset labeling information inside the learning problem, turning it into a supervised approach.

2.5 Optimization

In order to solve problem (2.7), we use the exact ALM method [14], and start by computing the augmented Lagrangian function $L(A, B, E, \lambda)$, given by:

$$
\begin{aligned}
L =& \|A\|_* + \alpha\|E\|_1 + \beta\|B(I-Q)\|_F^2 + \gamma\|B^\top A\|_F^2 \\
&+ \langle \lambda, X - A - B - E\rangle + \frac{\mu}{2}\|X - A - B - E\|_F^2 \\
=& \|A\|_* + \alpha\|E\|_1 + \beta\|B(I-Q)\|_F^2 + \gamma\|B^\top A\|_F^2 \\
&+ \frac{\mu}{2}\|X - A - B - E + \frac{\lambda}{\mu}\|_F^2 - \frac{1}{2\mu}\|\lambda\|_F^2 \\
=& \|A\|_* + \alpha\|E\|_1 + \beta\|B(I-Q)\|_F^2 + h(A, B, E, \lambda, \mu) \\
&- \frac{1}{2\mu}\|\lambda\|_F^2 \ ,
\end{aligned}
\tag{2.8}
$$

where $\langle X, Y\rangle \doteq trace(X^T Y)$, μ is a positive scalar, λ is a Lagrange multiplier matrix, and $h(A, B, E, , \mu) = \frac{\mu}{2} \parallel X - A - B - E + \frac{\lambda}{\mu} \parallel_F^2 + \gamma^T A \parallel_F^2$ is a quadratic convenience function. We optimize (2.8) with an alternating direction strategy, and at every outer iteration, A, B, and E are first iteratively updated until convergence; subsequently, γ and μ are updated.

Updating A_{k+1}: From the reduced augmented Lagrangian it is convenient to use the linearization technique of the LADMAP method [48], very effectively used also by other approaches [49, 50, 10], and replace the quadratic term h with its first order approximation, computed at iteration k, and add a proximal term giving the following update

$$
\begin{aligned}
A_{k+1} &= \arg\min_A \|A\|_* + < \nabla_A h(A_k, B_k, E_k, \lambda_k, \mu_k), \\
&\quad A - A_k > + \frac{\eta\mu_k}{2}\|A - A_k\|_F^2 \\
&= \arg\min_A \|A\|_* + \frac{\eta\mu_k}{2}\|A - (X - B_k - E_k + \frac{\lambda_k}{\mu_k} \\
&\quad - \gamma B_k B_k^\top A_k)\|_F^2 \ ,
\end{aligned}
\tag{2.9}
$$

where λ must be greater than $\parallel A \parallel_F^2$ [48]. The solution to (2.9) is obtained by applying the singular value thresholding algorithm [51], with the soft-thresholding shrinkage operator

$S_\epsilon(x)$, which is equal to: $x - \epsilon$ if $x > \epsilon$, $x + \epsilon$ if $x < -\epsilon$, and 0 elsewhere.

Updating E_{k+1}: From (2.8), the augmented Lagrangian reduces to

$$E_{k+1} = \arg\min_E \alpha\|E\|_1 + \frac{\mu_k}{2}\|E - (X - A_{k+1} - B_k + \frac{\lambda_k}{\mu_k})\|_F^2 \tag{2.10}$$

and the solution is still obtained with an instance of the singular value thresholding algorithm [51].

Updating B_{k+1}: This update is computed as

$$\begin{aligned} B_{k+1} = &\arg\min_B \frac{\mu_k}{2}\|X - A_{k+1} - E_{k+1} - B + \frac{\lambda_k}{\mu_k}\|_F^2 + \\ &\beta\|B(I - Q)\|_F^2 + \gamma\|B^\top A_{k+1}\|_F^2 \ . \end{aligned} \tag{2.11}$$

Note that the cost function in (2.11) is quadratic in B. Therefore, the update can be obtained by computing the partial derivative with respect to B of the cost function, and then setting it to zero. This leads to a Sylvester equation in B, given by

$$\begin{aligned} &\gamma A_{k+1}A_{k+1}^\top B + B\left((\beta + \frac{\mu_k}{2})I - 2\beta Q + \beta QQ\right) = \\ &\frac{\mu_k}{2}\left(X - A_{k+1} - E_{k+1} + \frac{\lambda_k}{\mu_k}\right) \ . \end{aligned} \tag{2.12}$$

Therefore, the update (2.11) can be computed with a standard Sylvester equation solver.

2.6 Classification

Given a test data point x, even if, strictly speaking, we are not in an instance-based learning setting, the obvious approach to perform classification is to compute a label y with a nearest-neighbor (NN) method, where $y = \arg\min_i \ d(x, B_i)$, and $d(.,.)$ is a suitable distance between x and the invariant matrix B_i , representing class i .

Following the strategy of recognition via Invariant Subspaces, from the invariant components B_i one can estimate $P_{B_i} : R^m \rightarrow B_i$, the operator that projects data points directly onto $B_i \subset B$, the invariant subspace for class i . Doing so has the advantage that the projection of x onto $V^\perp$ gives $b + P_{V^\perp}e$, whereas the projection of x onto B_i gives $b + P_{B_i}e$, and since $B_i \subset V^\perp$, it follows that $\parallel P_{B_i}e \parallel_F \leq \parallel P_{V^\perp}e \parallel_F$, which means a lower noise corruption. Therefore, we propose to use the following Frobenius norm $d_F(x, B_i) = n_i^{-\frac{1}{2}} \parallel B_i - P_{B_i}x1_{n_i}^T \parallel_F$. Note that if B_i can be approximated with $b_i 1_{n_i}^T$, as it normally should, then the distance computation is even faster, because given by

$$d_F(x, B_i) = \parallel b_i - P_{B_i}x \parallel_F \tag{2.13}$$

2.6.1 Local Metric Learning

Metric learning improves the performance of the NN classifier if used instead of the Euclidean metric. It has been applied effectively for classification [15], retrieval [16], person reidentification [17, 18], and widely for face verification [19, 20, 21, 22, 23]. Different aspects of metric learning have been investigated, like distance parameters selection, scalability, whether training data should be used in pairs [21], triplets [24] or quadruplets [25], or whether data undergoes a linear [26, 27], or nonlinear [28, 23, 29, 30] transformation.

The approach outlined before, which has been derived using geometry, is amenable to an interpretation from a metric learning perspective. Let us recall the definition of Mahalanobis distance between two points x_i and x_j , given by $d_M(x_i, x_j) = \sqrt{(x_i x_j)^T M (x_i x_j)}$,

where M is a symmetric positive semi-definite matrix. A *global* linear metric learning method learns a matrix M according to a specific criterion. Since the decomposition $M = L^T L$ is always possible, the Mahalanobis distance can be expressed also as $d_M(x_i, x_j) = \| L(x_i x_j) \|_F$.

Global metric learning methods learn the importance and correlation of different input features and take them into account for NN classification, regardless of the specific feature neighborhood where they are applied. Since discriminative power of input features might vary between different neighbors, learning a global metric may be suboptimal. This has motivated the development of local metric learning approaches [45, 46, 47, 31, 52], which increase the discriminative power of global Mahalanobis metric learning by learning a number of local metrics.

The proposed approach can be seen as a local metric learning approach, where for the neighborhood of each of the invariant components we learn a Mahalanobis metric. In particular, if $B_i = U_{B_i} S_{B_i} V_{B_i}^{\perp}$ is the singular value decomposition (SVD) of B_i , then the distance (2.13) can be rewritten as $d_F(x, B_i) = \| U_{B_i} U_{B_i}^{\perp} (x - b_i) \|_F$. This means that $d_F(x, B_i) = d_{M_i}(x, b_i)$, i.e., the Mahalanobis distance between x and b_i , with respect to $M_i = U_{B_i} U_{B_i}^{\perp}$. Therefore, learning a representation based on the invariant components B , is equivalent to learning a set of cluster centers b_i , and a set of Mahalanobis matrices M_i that act on the neighborhood of each center, and with which labels are assigned based on the NN rule $y = \arg\min_i d_{M_i}(x, b_i)$.

2.6.2 Class Separation

Most local approaches learn the metrics for each neighborhood independently [31] and require the addition of a form of regularization to avoid overfitting. In contrast, related to [32], our approach learns the metrics jointly, according to the constraints (2.5) and (2.6). While the first eliminates the effects of nuisance factors, the second ensures not only invariance, but also class separation. More specifically, since the invariance constraint (2.6)

can be rewritten as $Q = B^T(BB^T)^+B$, it is easy to realize that the Mahalanobis distance $d_M(b_i, b_j)$, with $M = (BB^T)^+/n$, between the invariant components b_i and b_j , for classes i and j, is such that

$$d_M(b_i, b_j) = \begin{cases} 0 & \text{if } i = j, \\ \sqrt{2N} & \text{otherwise} \end{cases} \tag{2.14}$$

where for simplicity it is assumed $n_i = n_j$. Without loss of generality, if we assume that the columns of B are zero mean, M is the inverse of the covariance of B (for a short discussion we do not address the rank deficiency of B, which leads to a reduced-rank metric, and to using the pseudoinverse $(BB^T)^+$). Therefore, (2.14) means that the invariant subspace B is such that two different sufficient components s_i and s_j originate two invariant components b_i and b_j that are different (i.e., $b_i = s_i P_{Bs_i} \neq s_j P_{Bs_j} = b_j$), and equidistant (i.e., $d_M(b_i, b_j) = \sqrt{2N} \ \ \forall i \neq j$), thus promoting a uniform class separation.

The observation above suggests also the use of a global Mahalanobis metric for NN classification, e.g., in the form of $d_M^2(x, B_i) = n_i^{-1} \sum_{b \subset B_i} d_M^2(x, b)$. However, it is more efficient to use the corresponding similarity measure $\kappa_M(b_i, b_j) = b_i^T(BB^T)^+ b_j$, which gives 0 if $i \neq j$, and 1 $\frac{1}{n_i}$ if $i = j$. Therefore, we propose the global Mahalanobis similarity measure defined as $\kappa_M(x, B_i) = 1_{n_i}^T B_i^T (BB^T)^+ x$, and the label assignment is done according to $y = \arg\max_i \ \kappa(x, B_i)$. If $B_i = b_i 1_{n_i}^T$, the similarity reduces to

$$\kappa_M(x, B_i) = n_i \ \ b_i^T (BB^T)^+ x \tag{2.15}$$

2.7 Experiments

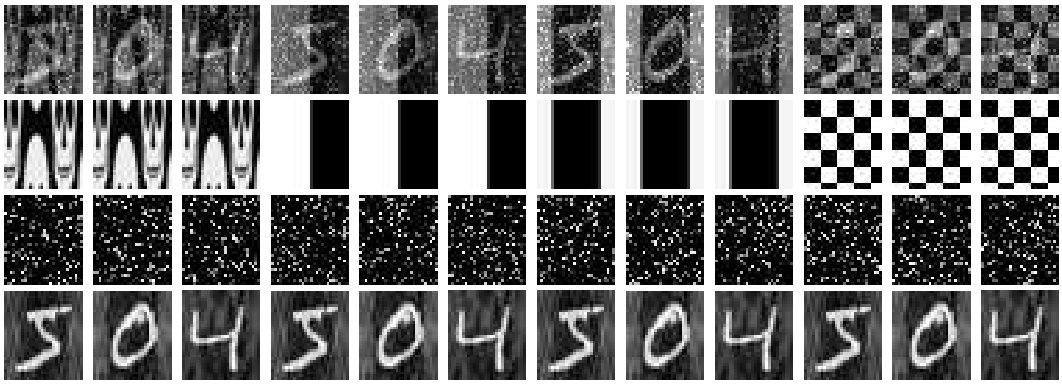
In order to validate the proposed method we have performed experiments on synthetic data, three face recognition datasets, and two person re-identification datasets. All the results were obtained with a grid search of the parameters α, β, and γ.

2.7.1 Synthetic Data

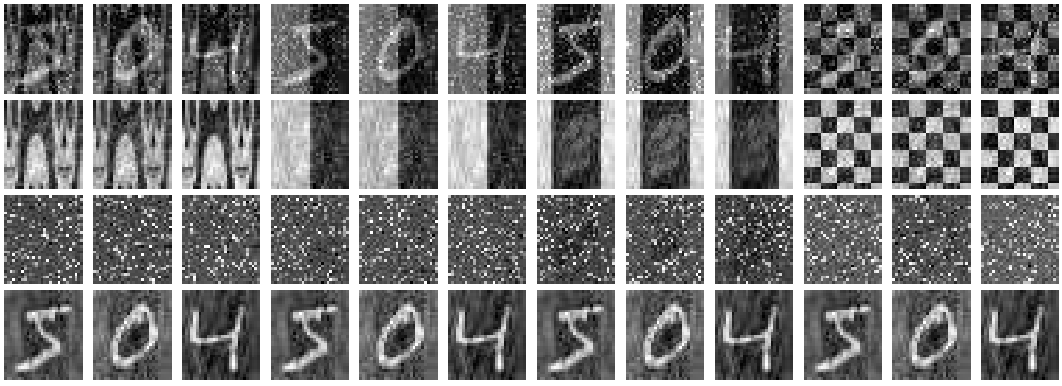
To empirically verify the convergence of our algorithm, we have created a synthetic dataset made of $n = 120$ images of 32×28 pixels, with $N = 10$ invariant components depicting digits, and with image patterns representing A. The synthetic A and B satisfy the constraints (2.5), and (2.6), and we have added sparse noise E, corrupting 20% of randomly selected pixels with values drawn from a uniform distribution between 0 and the largest possible pixel value in the image. Figure 1(a) shows the decomposition in A, E, and B of 12 synthetic data points, X (top row), and Figure 2.1 (b) shows the estimated decomposition of the same points. Visually, the recovered decomposition closely resembles the originals.

2.7.2 AR Dataset

For this face recognition dataset [33], we follow a protocol used also by other recent works [11, 10]. The dataset contains over 4,000 frontal images of 126 people's faces (70 men and 56 women), images are taken in two sessions and under different facial expressions, illumination conditions and occlusions. In each session 3 images are occluded by sunglasses, 3 by a scarf, and all are taken in different lighting conditions. The images have $165 \times 120 = 19,800$ pixels, and are converted into gray scale, and down-sampled by a factor of 4. As other authors did [11, 10, 34], we select a subset of 50 men and 50 women. Figure 2.2 illustrates 9 images taken from one subject in one session, along with the decomposition. The proposed algorithm effectively extracts the invariant component (bottom row), which is almost identical for every image, as expected. The second row from the top is a low-rank representation of the face images, and the second row from the bottom is sparse noise. Note how the low-rank representation, a, contains a significant amount of facial features. This is expected because it represents the additive contributions of the variation component, v , plus the projection, $P_V s$, of the sufficient component, s (which is



(a)



(b)

Figure 2.1: **Synthetic data.** (a) Decomposition of 12 synthetic data points. (b) Decomposition of the same 12 points with Algorithm 1. Top row: input points X. Second row: A components. Third row: Sparse errors E. Bottom row: Invariant components B. Columns with invariant components depicting the same digit belong to the same class. The digits appear "hazy" as a result of being orthogonal to the A components by construction.

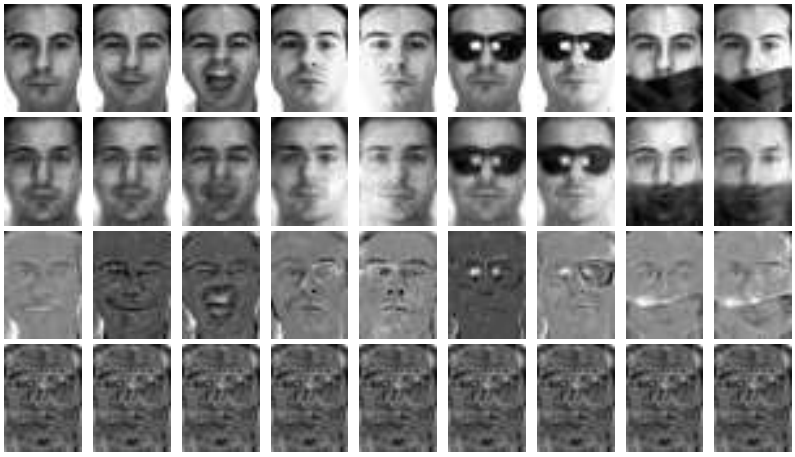


Figure 2.2: **AR dataset.** Decomposition results for the 9 images of one subject taken in one session. Row meanings are explained in Figure 2.1. Images are rescaled for better contrast and visualization.

essentially the face), onto the variation subspace, V , which is shared among all the classes.

Following [11, 10] we consider three scenarios, indicated as SUNGLASSES, SCARF and SUNGASSES+SCARF, where we do face recognition with highly corrupted training and testing data. For SUNGLASSES a subject in the training set is composed by p randomly selected face images occluded with sunglasses, and $8 - p$ neutral, all selected from session 1. The remaining $6 - p$ images occluded by sunglasses plus $6 + p$ neutral from both sessions, form 12 testing images per person. Note that face images with sunglasses are occluded about 20%. For the SCARF scenario, the data subdivision is identical only that we consider the face images occluded by a scarf, which produces occlusions of about 40%. For the SUNGLASSES+SCARF case, the difference is that for a given person, p images are occluded with sunglasses and p with the scarf, leaving 17 images for testing per person. Unlike previous work, that have shown results only for $p = 1$, here we also test the case for $p = 2$ and $p = 3$. The experiment has been repeated 5 times and the average recognition rates are plotted in Figure 2.3. The optimal penalty parameters were $\alpha = 1.5$, $\beta = 1000$, $\gamma = 0.9$. Unless otherwise specified, every result obtained in this section is with the distance (2.13), i.e., the local metric. Along with ours, we have also tested the

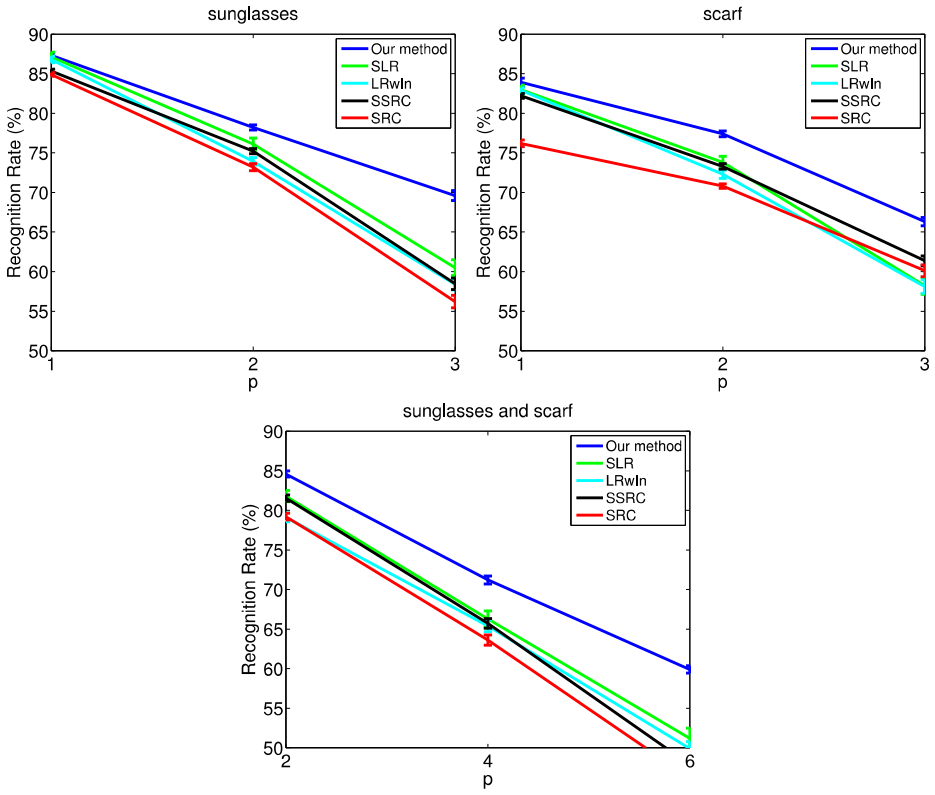
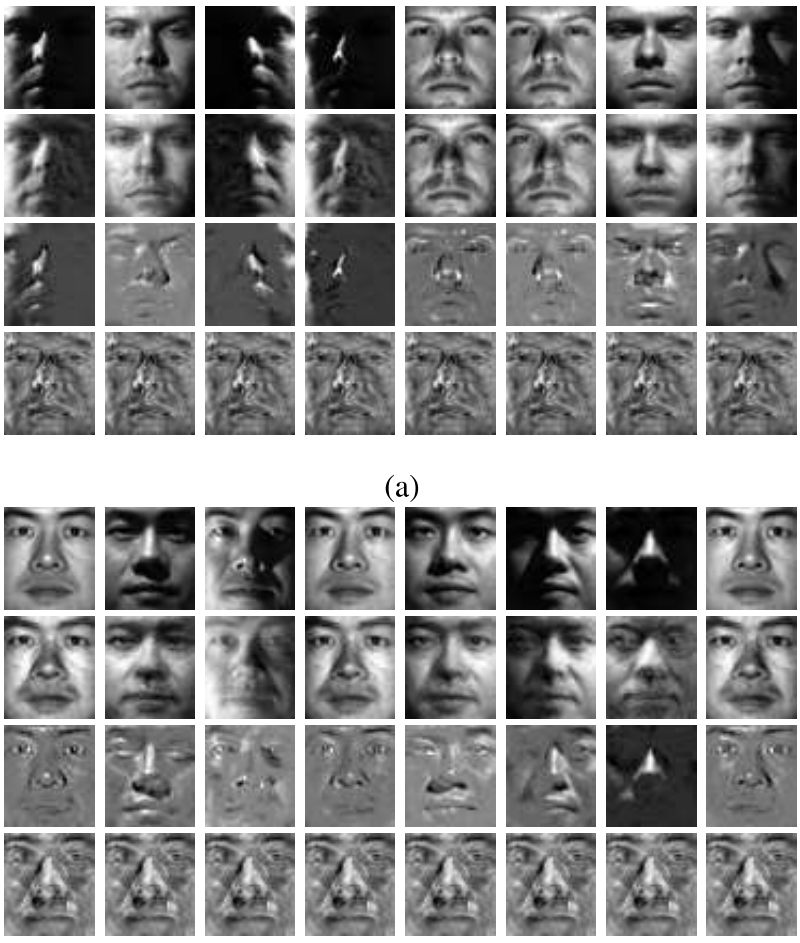


Figure 2.3: **AR dataset.** Recognition rates versus different numbers p, of corrupted training images per class for the three scenarios: sunglasses (top left), scarf (top right), sunglasses and scarf (bottom).

structured low-rank representation (SLR) approach [10], the low-rank with incoherence (LRwIn) approach [11], the superposed SRC (SSRC) approach [34], and the SRC [3]. We have reimplemented the SLR, the SSRC, and the LRwIn approaches. For the SRC we have used the code publicly available. Every approach was tested with input images with the same size, and with other parameters set at the peak of their performance. From Figure 2.3 it can be appreciated that the proposed approach demonstrates a superior robustness with respect to corruption in the training set as p increases. For instance, compared to the overall best competitor, which is SLR, for the SUNGLASSES+SCARF case, for $p = 1$ the improvement is 2.8%, for $p = 2$ is 4.9%, and for $p = 3$ is 8.7%.

2.7.3 Extende-Yale B Dataset

This face recognition dataset [35] contains tightly cropped face images of 38 subjects. Each of them has 59 to 64 images taken under varying lighting conditions, which in total add up to 2,414 images. The cropped images have $192 \times 168 = 32,256$ pixels. We randomly select 8, and in a subsequent experiment 32, training images for each person, and use the rest for testing in a recognition experiment. We repeat this 5 times and report the average



(a)

(b)

Figure 2.4: **Extended Yale B dataset.** Decomposition results on 2 subjects, and 8 training images per subject. First row is the original downsampled image (X), second row is the low rank matrix (A), third row is the sparse noise (E), and last row is the invariant subspace (B). Images are rescaled for better visualization.

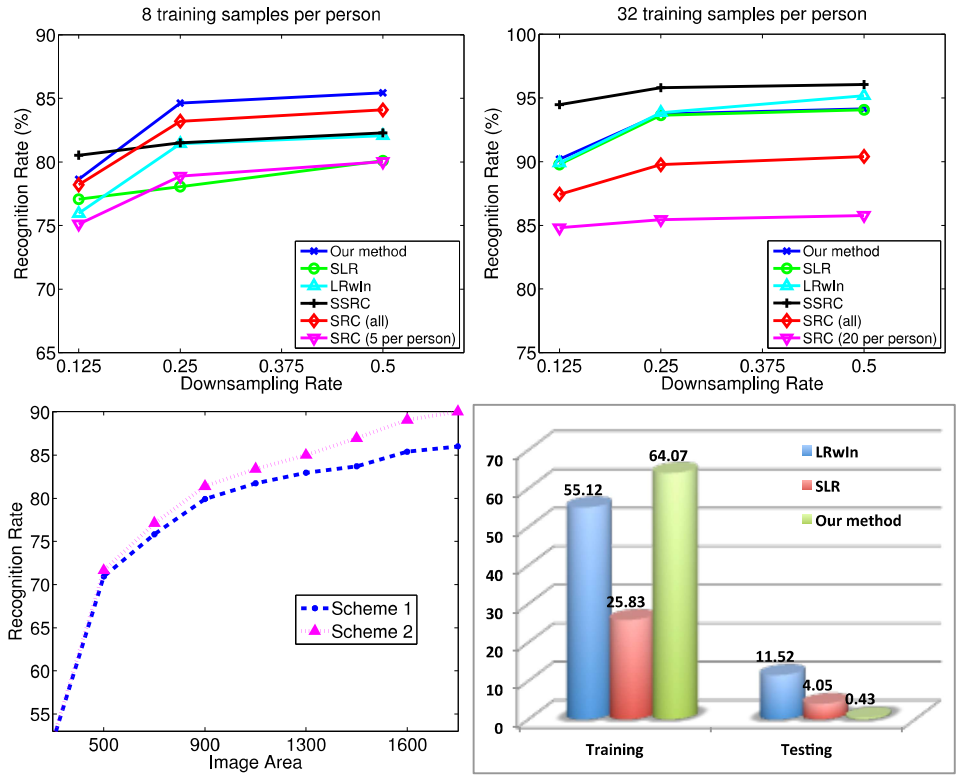


Figure 2.5: **Extended Yale B dataset.** First row: Recognition scores at different image downsampling rates for 8 and 32 training samples per subject. Second row: recognition rates obtained with the global metric (2.15) (Scheme 1) and the local metric (2.13) (Scheme 2) at various image resolutions and 32 training samples and running time in seconds of our Matlab implementations for training and testing.

recognition rate for the images down-sampled by a factor of 2, 4, and 8. For each of those conditions we also compare against the SLR [10], the LRwIn [11], the SSRC [34], and the SRC [3] approaches at the peak of their performance. For our approach the optimal penalty parameters were $\alpha = 0.9$, $\beta = 1000$, $\gamma = 0.01$. Figure 2.5 illustrates the comparison between the recognition rates. For the SRC, we also include what happens when the training set drops in size from 8 to 5, and from 32 to 20 training images. This experiment highlights that our approach compares favorably with the others especially when a smaller corrupted training dataset is available, and works on par with others (SLR and LRwIn) with lots of

Dataset	Euclidean	SRC	SSRC	LMNN	Ours
LFW	15.40 $\pm$0.50	36.91 $\pm$1.90	46.31 $\pm$2.43	46.90 $\pm$1.00	47.10 $\pm$1.50
AR	29.30 $\pm$0.50	63.60 $\pm$0.64	65.70 $\pm$0.61	62.80 $\pm$1.00	71.20 $\pm$0.52

Table 2.1: **Metric learning.** Comparison between metric learning and other methods on the LFW dataset [37], and on the AR dataset on the scenario SUNGLASSES+SCARF with $p = 2$.

training data. This is because our approach inherently attempts learning a global variation space, shared by all the training data. Even with fewer training images per person their aggregation allows learning the variation space better than in other approaches. SSRC, instead, is the best performer with lots of training data, since it can better learn the variation space for each individual.

Figure 2.5, right, also shows a comparison between the local metric approach (Scheme 2), based on (2.13), and the global metric approach (Scheme 1), based on (2.15), on a subset of the dataset with 32 training data points per person, against different image resolutions. As expected, the local metric learning approach, because it adapts to the invariant component where it operates on, is able to provide better performance. From a geometric perspective, the performance drop is justified by the fact that the global approach is not able to filter out as much noise as the local approach is capable of.

Figure 2.5, far right, shows a running time comparison between the Matlab implementations of ours, the SLR, and the LRwIn methods, running on a high-end PC. Our training procedure appears slightly more costly than the others, but, as anticipated, testing appears faster than SLR by a factor of 10, and faster than LRwIn by a factor of 25.

2.7.4 Metric Learning on LFW and AR Datasets

We have tested the large-margin nearest neighbor (LMNN) metric learning approach [36], SRC, SSRC, and ours on the Labeled Faces in the Wild (LFW) dataset [37]. Out of the

13,233 face images of 5749 unique individuals, we selected those with at least 10 images for a total of 143 people and 4174 face images, which were aligned using deep funneling [38], tightly cropped to include only face information, and resized to 106×96 pixels. For each subject, we randomly selected 7 images for training, and the rest were used for testing. The penalty parameters were $\alpha = 0.5$, $\beta = 1000$, $\gamma = 0.2$. The actual processing for both algorithms was repeated 10 times, and was done with the cropped images down-sampled by a factor of 4. In such a scenario with a highly non-linear variation space we obtained the results reported in Table 1, where we also provided results using the baseline Euclidean distance. We also run LMNN and our method on the AR dataset on the highly corrupted scenario given by SUNGLASSES+SCARF with $p = 2$. Table 2-2 reports the results, which shows that our method performs better especially when robustness against corrupted samples in the gallery is needed.

2.7.5 i-LIDS MCTS Dataset

This dataset contains 476 whole body person images of 119 people captured by multiple non overlapping surveillance cameras. There are 4 images on average per person. We excluded subjects who had only 1 or 2 images. All the images are normalized to 128 × 64 pixels. This dataset is used for person re-identification across camera views. Unlike faces that can be aligned, and have approximately a 9D linear illumination variability subspace, people images from unconstrained environments are highly misaligned, and this dataset pushes the proposed approach beyond limits. It has been added to show to what extent performance can degrade. Figure 2.6 shows the decomposition of 4 training images of two people. Figure 2.7, instead, shows the cumulative matching curves (CMC) with 30 and with 80 people in the training set. In a CMC curve, a rank r matching rate indicates the percentage of test (or probe) images with correct matches found in the top r ranks against the people in the training (or gallery) set. The penalty parameters were α = 1, β = 100, γ =

Figure 2.6: **i-LIDS MCTS dataset.** Decomposition results for 2 subjects under 4 different viewpoints. First row shows the original image data (X), second row is the low rank matrix (A), third row is the sparse noise (E), and last row is the identity subspace (B). Images are rescaled for better contrast and visualization

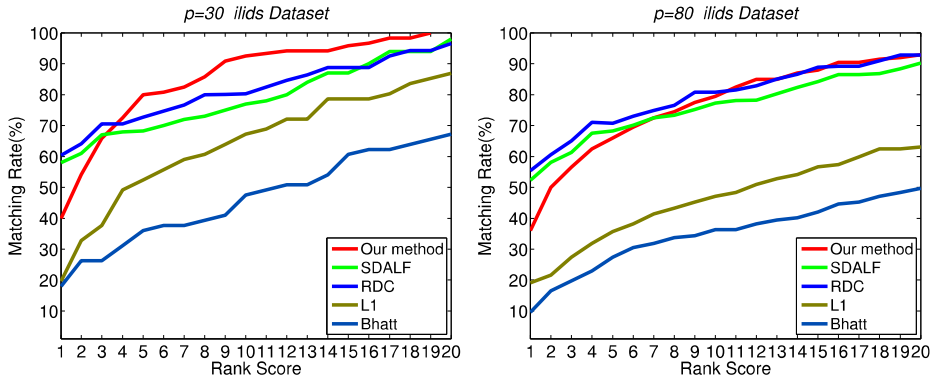


Figure 2.7: **i-LIDS MCTS dataset.** Cumulative matching curves obtained with 30 and 80 individuals. The plots refer to our method, the SDALF [95], the RDC [96], and to the matching done, as in [96], with a color histogram and the L1 norm (L1), and the Bhattacharyya distance (Bhatt).
1 for 30 subjects, and $\alpha = 2$, $\beta = 100$, $\gamma = 0.1$ for 80 subjects.

To compare our results with two state-of-the-art approaches, namely the relative distance comparison (RDC) [96] (a metric learning method), and SDALF [95] (a descriptor matching method), we have used 3 images in the training set, and 1 image in the testing set per person. We run the experiment four times to make sure an image is on average part of the probe set once. For SDALF we learned the signature from 3 images in the training set for fair comparison. In such extreme conditions, the CMC curves show that the performance of the approach has degraded more gracefully than expected, and with this limited dataset it is capable of keeping up with state-of-the- art approaches at higher matching ranks. However, performance is expected to degrade further, as the variation space becomes more nonlinear.

2.7.6 CAVIAR4REID Dataset

CAVIAR4REID dataset is a person re-identification dataset extracted from CAVIAR which is a person tracking and detection dataset. It contains multiple images for each person extracted from a real scenario (shopping centre) where re-identification is necessary due to the

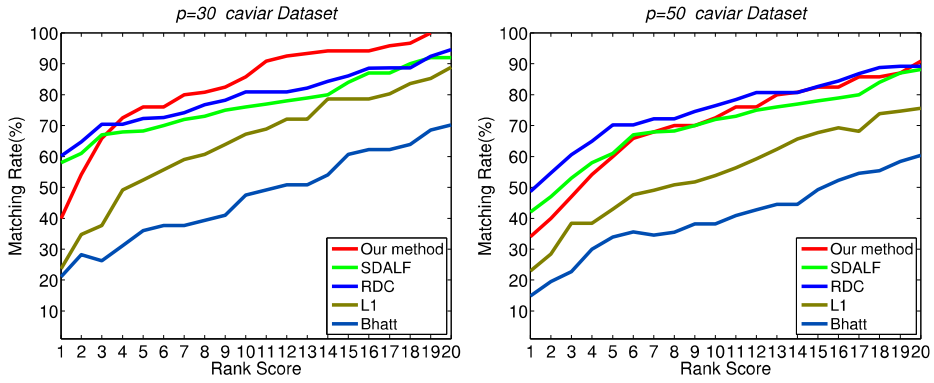


Figure 2.8: **CAVIAR4REID dataset.** Performance comparison using CMC curves on the CAVIAR4REID dataset. This is a new dataset for evaluating person re-identification algorithms. For this dataset we followed the same protocol used for the i-LIDS dataset. The left plot has been obtained with 30 people in the training set (gallery), and the right plot with 50 people.

presence of multiple cameras. it has severe pose variations and broad changes in resolution where the minimum and maximum size of images are 17 x 39 and 72 x 144 respectively. The same experiment performed on the i-LIDS MCTS dataset has been repeated on the CAVIAR4REID dataset. Figure 2.8 reports results similar to those in Figure 2.7.

2.8 Conclusion

we present a low-rank modeling framework which can be applied to face as well as whole-body images. This framework can capture all the descriptive information of a human image, and emphases on learning a representation that is invariant to nuisance factors. The framework is not only learned different invariant representations for different identities, but such representations promote a uniform inter-class separation. Another advantage of this framework is a fast procedure for computing and comparing invariant components for recognition and re-identification.

Chapter 3

Non-Cooperative Human Identification

3.1 Introduction

Searching for a person in a huge video library, may be made up of recordings captured by a network of surveillance cameras, is an important task because it allows investigators to determine when and where an individual was present at a scene. For example, it can save hours of human video inspection time while looking for and tracing criminals, such as those involved in the Boston Marathon bombings. In such a situation, the human identification task must frequently rely on biometric traits such as face and gait that obtained from non-cooperative subjects under unconstrained conditions. More specifically, this means that noise or other nuisance factors, such as an individual's pose or illumination and occlusions, can substantially corrupt both gallery and probe images of an individual's trait. If enough image resolution is provided, human identification is usually performed in such a hostile conditions using face recognition.

Even that face recognition has improved tremendously in the presence of corrupted gallery and probe images, there are still a number of unfavorable conditions in which face recognition fails. As a result, reducing those cases is crucial in order to improve the efficacy of human identification for searching unconstrained video archives.

In this chapter we propose to improve human identification by fusing the face modality with a pseudo-modality, such as an individual's clothing appearance. Using various biometric modalities to improve identification robustness is an effective methodology. In addition to the face, other feasible possibilities for the considered scenario include gait and clothing appearance. It is not possible to extract and characterize human gait unless certain parameters are met, which are usually excessively restrictive. Clothing appearance is not considered as a biometric trait but its usefulness in matching identities of human that have not changed their clothes between different sightings has been demonstrated [124, 125], which is the reason why it was chosen.

We extended our invariant component matching method in chapter two for using kernel machines. The proposed method based on independent components and distance matching. Then, we developed a new approach for jointly exploiting face and whole body appearance fusion for human identification based on Information Bottleneck (IB) [119] method that extracts the latent information from the data. There are two main goals for the information bottleneck (IB) method, the first one is to compress as possible as the data (e.g., the embedding of a human image). The second one is to preserve all the necessary information that is relevant to the task (e.g., predicting identities of human images).

3.2 The information bottleneck method

Here we summarize the information bottleck method [119] that was extended to the multivariate case in [122]. We are given a set of random variables $\mathbf{X} = X_1, ..., X_n$, distributed according to a known $p(\mathbf{X})$, a set of latent variables $\mathbf{T} = T_1, ..., T_k$, and a Bayesian network with graph G_{in} over $\mathbf{X} \cup \mathbf{T}$, defining which subset of $\mathbf{X}$ is compressed by which subset of $\mathbf{T}$. Another Bayesian network, G_{out}, also defined over $\mathbf{X} \cup \mathbf{T}$, is given and represents which conditional dependencies and independencies we desire $\mathbf{T}$ to be able to generate. The joint distribution is known and given by

$$q(\mathbf{X}, \mathbf{T}) \;\doteq\; q(\mathbf{T}|\mathbf{X})\;\; p(\mathbf{X}) \tag{3.1}$$

The compression requirements defined by G_{in}, and the desired independencies defined by G_{out}, are incompatible in general. Therefore, the multivariate IB method computes the optimal **T** by searching for the distribution $q(T|X)$, where **T** compresses **X** as much as possible, while the distance from $q(X, T)$ to the closest distribution among those consistent with the structure of G_{out} is minimal. This idea is implemented with the *multi-information* of **X**, which is the information shared by $X_1, ..., X_n$, i.e.,

$$I(\mathbf{X}) = D_{KL}[\; p(\mathbf{X}) \parallel p(X_1) \; ... \; p(X_n) \;], \tag{3.2}$$

where D_{KL} indicates the Kullback-Leibler divergence [123]. Therefore, the multivariate IB method looks for $q(\mathbf{T}|\mathbf{X})$ that minimizes the functional

$$L\;[\; q(\mathbf{T}|\mathbf{X}) \;] \;=\; I^{G_{in}}(\mathbf{X}, \mathbf{T}) + \gamma \;(\; I^{G_{in}}(\mathbf{X}, \mathbf{T}) - I^{G_{out}}(\mathbf{X}, \mathbf{T}) \;) \tag{3.3}$$

where γ strikes a balance between compression and the ability to satisfy the independency requirements of G_{out}. The multi-information I^G with respect to a Bayesian network G defined over $\mathbf{X} \sim \mathrm{p}(\mathbf{X})$ is computed as in [122], i.e.,

$$I^G(\mathbf{X}) = \sum_i I(X_i; \mathbf{Pa}^G_{X_i}), \tag{3.4}$$

where $I(X_i; \mathbf{Pa}^G_{X_i})$ is the mutual information between X_i and $\mathbf{Pa}^G_{X_i}$, the set of variables that are parents of X_i in G.

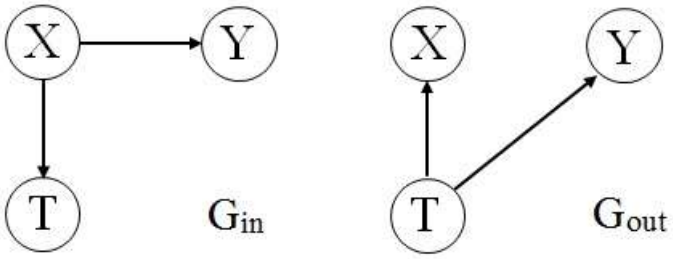


Figure 3.1: **Information Bottleneck.** Structural representation of G_{in} and G_{out} used by the original two-variable information bottleneck method [119].

Let us refer to Figure 3.1 for an example, where $\mathbf{X} = X, Y$, and $\mathbf{T} = T$. We interpret X as the main data we want to compress, and from which we would like to predict the relevant information Y. This is achieved by first compressing X into T, and then predicting Y from T. In G_{in} the edge $X \rightarrow Y$ indicates the relation defined by $p(X, Y)$. Moreover, since T will compress X, this is indicated by the edge $X \rightarrow T$, establishing that T is completely determined given the variable it compresses. The graph G_{out} instead, reflects the idea that we would like T to capture from X all the necessary information to perform the best possible prediction of Y. This means that knowing T makes X and Y independent, or equivalently that $I(X; Y|T) = 0$. To evaluate (3.3), instead, we obtain

$$I^{G_{in}} = I(T; X) + I(Y; X) \tag{3.5}$$

$$I^{G_{out}} = I(X; T) + I(Y; T) \tag{3.6}$$

By plugging (3.5) and (3.6) into (3.3), since $I(Y; X)$ is constant, the functional for learning the optimal representation for T is given by

$$L\ [\ q(T|X)\ ]\ =\ I(T; X) - \gamma\ I(X; T) - \gamma\ I(Y; T) \tag{3.7}$$

where γ strikes a balance between compressing X and imposing the independency requirements.

3.3 IB fusion for human identification

We want to leverage the IB method (3.7) because it provides a sound principle, grounded on information theory, for extracting information T from the invariant component of the main view X that is not only the most relevant for predicting Y (representing identity labels), but also minimizes $I(X;Y|T)$, which means that knowing T leaves with X minimal information about Y. This suggests that T is the representation of choice for predicting Y. However, while IB explicitly defines the compression map, T, by searching for $q(T|X)$, the computation of $q(Y|T)$ is much harder in general. For this reason, we introduce a modified IB method that is designed to fuse face and clothing appearance for the purpose of human identification.

We observe that by interpreting γ as a Lagrange multiplier, the last term in (3.7) corresponds to the constraint $I(Y;T) \geq constant$, enforcing T of carrying at least a certain amount of information about Y . Ultimately, such information should be used for classification purposes, by predicting Y through a function $\tilde{f} : T \rightarrow Y$. Therefore, we replace the constraint on $I(Y;T)$ with the risk associated to $\tilde{f}(T)$ according to a loss function ℓ. Thus (3.7) is modified into

$$L\ [\ q(T|X)\ \tilde{f}]\ =\ I(T;X) - \gamma\ I(X;T) + \beta\ E[\ell(\tilde{f}(T)), Y)] \tag{3.8}$$

where $E[.]$ denotes statistical expectation, and β balances the risk versus the compression requirements. Note that the modified IB criterion (3.8) is general, and could be used with any classifier.

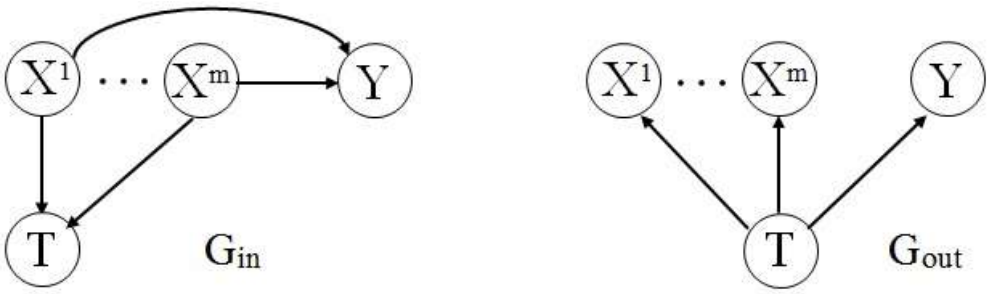


Figure 3.2: Information Bottleneck Extension to Fusion.

We are interested in designing a framework for fusing face modality with clothing appearance pseudo-modality based on information bottleneck method. Therefore, we propose to divide human whole body images into face and five horizontal stripes in addition to the whole body image. Thus, the main view of the i-th identity will be expressed by $X_i = \{X_i^B, X_i^F, X_i^{S1}, X_i^{S2}, X_i^{S3}, X_i^{S4}, X_i^{S5}\}$ (see Figure 3.2) and the T_i that compresses X_i will be

$$T_i = \phi(X_i; A) = \sum_{j=1}^{m} c^j A^j X_i^j \tag{3.9}$$

where c^j is a constant, A^j is the transition matrix on X^j that composed of the conditional probabilities between X^j and T. The decision function, $\tilde{f}$, is given by

$$Y = \textit{sign}\left(\langle w, T\rangle + b\right) \tag{3.10}$$

where $\langle ., . \rangle$ is a dot product, w defines the margin and b is an offset. Therefore, we drive the following classifier learning formulation using the hinge loss function

$$\min_{A,w,b,\xi_i} \sum_{j=1}^{m} I(X^j, T) + \frac{\beta}{2}||w||^2 + \frac{C}{N}\sum_{i=1}^{N}\xi_i$$

$$\text{s.t. } y_i\left(\langle w, T\rangle + b\right) \geq 1 - \xi_i,$$

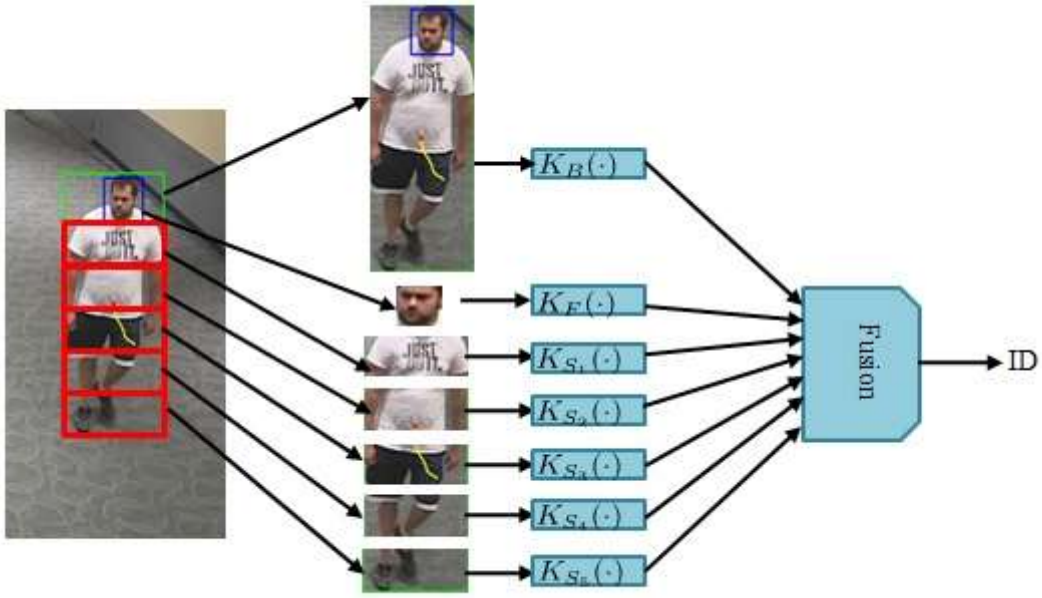


Figure 3.3: Information Bottleneck Fusion Approach.

$$\xi_i \geq 0, \quad \forall_i \in \{1, ..., N\}.$$

(3.11)

where C is the usual parameter to control the slackness. Figure 3.3 shows our IB fusion architecture where $K(x) = P_{B_i}x$, developed in Chapter 2, computes the invariant component for the whole body image, the face and the five horizontal stripes.

3.4 Optimization

When A is known, (3.11) is a soft-margin SVM problem. Instead, when the SVM parameters are known, (3.11) becomes

$$\min_{A} \sum_{j=1}^{m} I(X^j, T) + \frac{C}{N} \sum_{i=1}^{N} \xi_i$$

$$\text{s.t. } \xi_i = max\ \{0, 1 - y_i(\langle w, \phi(X_i; A)\rangle\ +\ b)\}. \tag{3.12}$$

Since the soft-margin problem is convex, if also (3.11) is convex, then an alternating direction method is guaranteed to converge. In general, the mutual information in (3.11) are convex functions of $q(T|X)$ [123]. The last term is also convex, however, the constraints define a non-convex set due to the discontinuity of the hinge loss function. Smoothing the hinge loss turns (3.11) into a convex problem, and allows to use an alternating direction method with variable splitting combined with the augmented Lagrangian method. Thus by splitting A^j into two variables U and V, we can set $f(U) = I(X^j, T)$, $g(V) = \frac{C}{N}\sum_{i=1}^{N}\xi_i$ and then solving

$$\min_{A^j} \{f(U) - g(V) : U - V = 0\} \tag{3.13}$$

For smoothing the hinge loss we use the Nesterov smoothing technique [127], used also in [128], which requires choosing a proximal function, and then computing the smoothed slack variables in this way $\xi_{i,\sigma} = max_{0\leq u_i\geq 1} u_i(1 - y_i w^T\ Ax_i) - \frac{\sigma}{2}||wx_i^T||_\infty u_i^2$, which gives

$$\xi_{i,\sigma} = \begin{cases} 0 & y_i w^T Ax_i > 1, \\ (1 - y_i w^T Ax_i) - \frac{\sigma}{2}||wx_i^T||_\infty & y_i w^T Ax_i < 1, \\ \frac{(1 - y_i w^T Ax_i)}{2\sigma||wx_i^T||_\infty} & \text{otherwise.} \end{cases} \cdot \tag{3.14}$$

where σ is a smoothing parameter. In this way, the minimization can be carried out with the Fast Alternating Linearization Method (FALM) [129]. This allows simpler computations, and has performance guarantees when $\bigtriangledown_f$ and $\bigtriangledown_g$ are Lipschitz continuous, which is the case, given the smoothing technique that we used.

FALM splits the minimization of the augmented Lagrangian function into two simpler functions to be minimized alternatively, which are given by

$$Q_g(U, V) = f(U) + g(V) + \langle \bigtriangledown_g(V), U - V \rangle + \frac{1}{\mu_g} D_{KL}(U||V) \tag{3.15}$$

$$Q_f(V, U) = f(U) + g(V) + \langle \bigtriangledown_f(U), V - U \rangle + \frac{1}{\mu_g} D_{KL}(V||U) \tag{3.16}$$

3.5 Experiments

In this section we decribe the results obtained by using our method for recognition on face and body, and how we improve it by using our proposed fusion approach.

Dataset: There is no publicly available datasets for face and whole body appearance in surveillance settings. Therefore, we have created our own video surveillance dataset (Figure 3.2) and then generated indivdual and face tracking annotations (Figure 3.3). Our dataset has the following properties:

- 133 individuals, each has from one to four outfits. There are 5 tracklets per identity.
- 10730 images. The average face resolution is 27 by 35 pixels, while the average whole body resolution is 88 by 226 pixels.

We applied [126] for frontal face detection. Faces are fine cropped and the size is normalized to 32 by 32 pixels. The whole body size is normalized to 50 by 128 pixels. We rejected identities with less than 10 frontal faces detected. Therefore, the total number of identities are 111, each is represented with more than 10 whole body and frontal faces images. The total number of images are 6811 whole bodies and 6811 frontal faces.

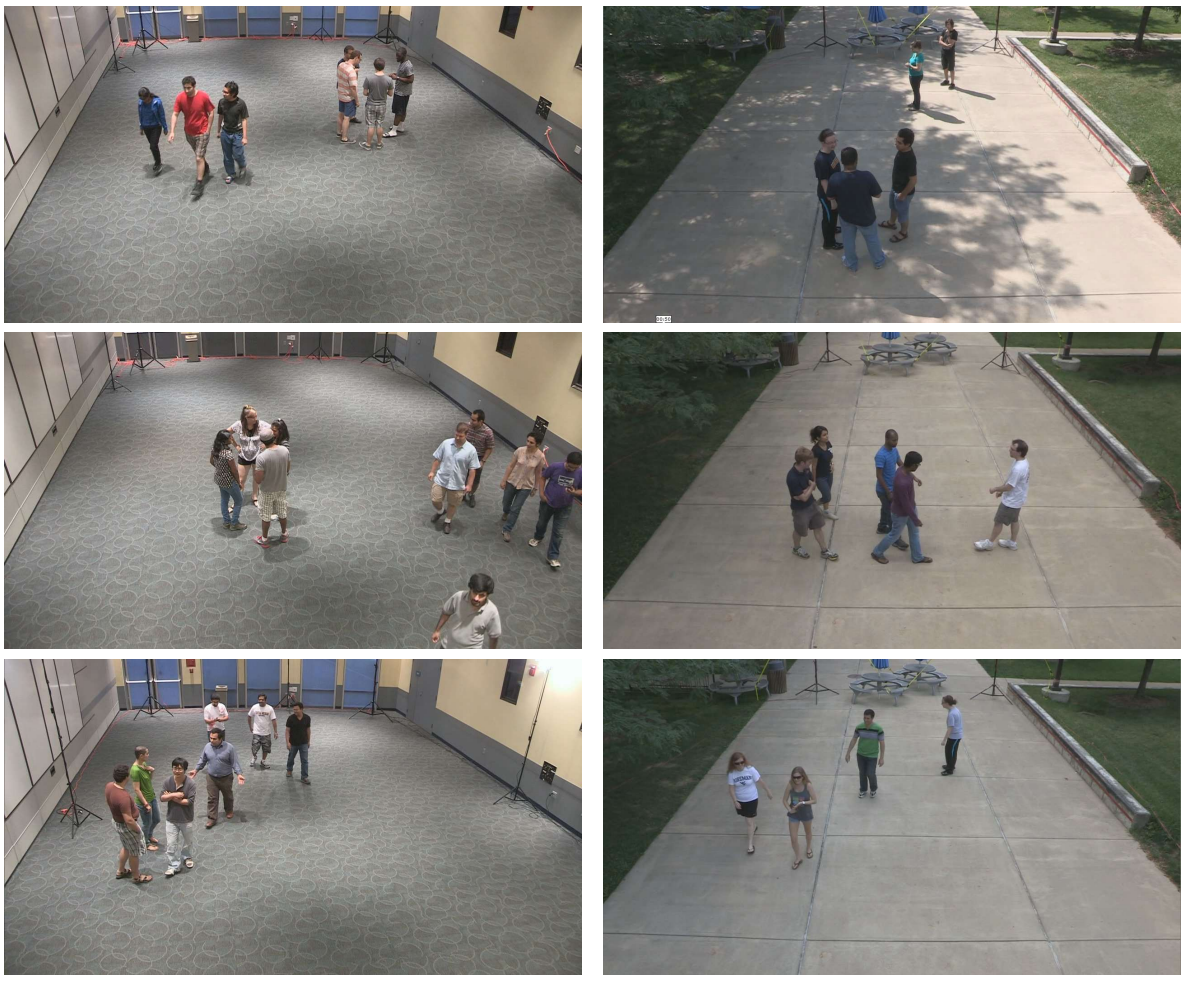


Figure 3.4: Example of our dataset. First column is indoor and second column is outdoor

Figure 3.5: Example of annotating our dataset. First column is indoor and second column is outdoor

Abbreviations: Below are the abbreviations of the proposed approach and the state-of-the-art approaches.

- **IC**: stands for Independent Component. Our approach based on independent component and distance matching as in Chapter two.
- **IC-IB Fusion**: Our fusion approach based on Information Bottleneck method that fuses the face and the body.
- **IC-IB Full Fusion**: Our full fusion approach based on Information Bottleneck method that fuses the face, the body and the stripes.
- **SRC**: Sparse Representation based Classification [3].
- **SSRC**: Superposed SRC [34].

Figures 3.6 and 3.7 show the CMC curve comparison between our method, IC, and SRC and SSRC on face only and body only respectively. However, tables 3.6 and 3.6 summarize the performance for rank-1, rank-10 and rank-20.Then, we applied the standard fusion rules, i.e. product, sum, min and max, on IC Face and IC Body as shown in figure 3.8 and table 3.6. It is clear that the best standard fusion rule is the product since it improve the performance by 3.60%, 6.31% and 3.60% on rank 1, 10 and 20 respectively. Figure 3.9 and table 3.6 show the CMC curve and the performance summary of Product rule fusion on IC Face and IC Body. We also applied the product rule fusion for the face and the body on SRC and SSRC, figure 3.10 and table 3.6. The performance on SRC and SSRC improved by 1.81% and 3.61% on rank-1, 9.91% and 4.5% on rank-10 through using the product rule fusion. In figure 3.11 and table 3.6, we applied our proposed fusion approach IC-IB on the face and the body only, the proposed fusion approach, IC-IB, able to improve the performance over IC-Product by 3.60%, 11.71% and 13.51% for rank 1, 10 and 20 respectively. In order to evaluate our proposed full fusion approach, i.e IC-IB Full Fusion, we first summarize the performance of IC on the face, body and stripes as shown in figure 3.12

and table 3.6. Then we compare them against IC-IB Full Fusion in figure 3.13. It is very clear that the proposed fusion approach (IC-IB Full Fusion) improved the performance. Figure 3.14 and table 3.6 compare IC-IB Full Fusion with IC-IB Fusion and product rule fusion on IC, SRC and SSRC. IC-IB Full Fusion able to improve the performance over SSRC-Product by 4.50%, 11.71% and 9.01% for rank 1, 10 and 20 respectively. Finally, we summarize the performance of IC Face, IC Body, the standard fusion rules and IC-IB Full Fusion in table 3.6. The IC-IB Full Fusion improve the performance over IC-Product by 11.71%, 14.72% and 10.36% for rank 1, 10 and 20 respectively.

3.6 Conclusion

we address the human identification problem with probe and gallery data acquired in unconstrained scenarios by developing a robust approach that fuses representations of multiple biometrics for human identification. We extended our invariant component matching method based on independent components and distance matching to develop a new approach for jointly exploiting face and whole-body appearance for human identification. Our fusion approach based on the Information Bottleneck method.

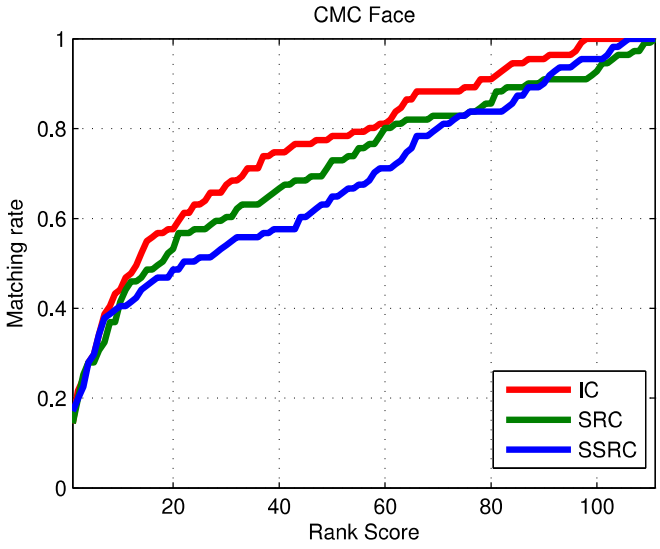


Figure 3.6: CMC comparison between our method, IC, and SRC and SSRC on face images only.

Summary of the Face Result				
Approach/Result	Rank 1	Rank 10	Rank 20	STD
IC	14.41	44.14	57.66	$\pm$2.20
SRC	14.41	41.14	53.15	$\pm$1.99
SSRC	17.12	40.54	48.65	$\pm$2.10

Table 3.1: The Performance by using 9 training samples and face features

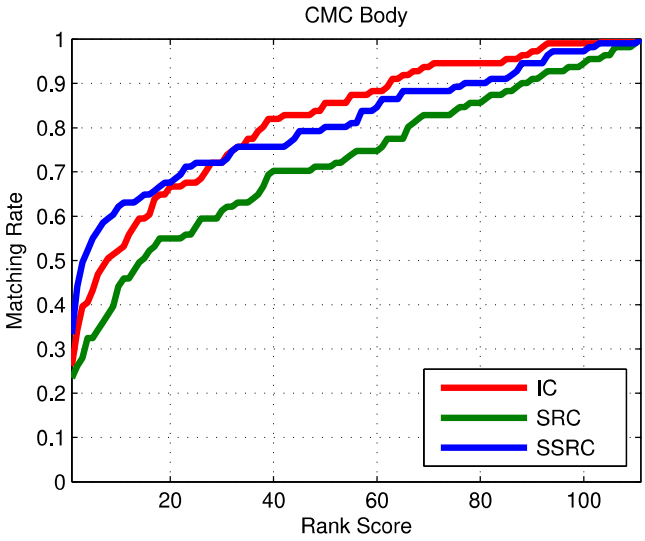


Figure 3.7: CMC comparison between our method, IC, and SRC and SSRC on body images only.

Summary of the Body Result				
Approach/Result	Rank 1	Rank 10	Rank 20	STD
IC	26.13	52.25	66.67	$\pm$1.74
SRC	23.42	44.14	54.95	$\pm$1.88
SSRC	33.33	62.16	67.57	$\pm$1.38

Table 3.2: The Performance by using 9 training samples and the body features

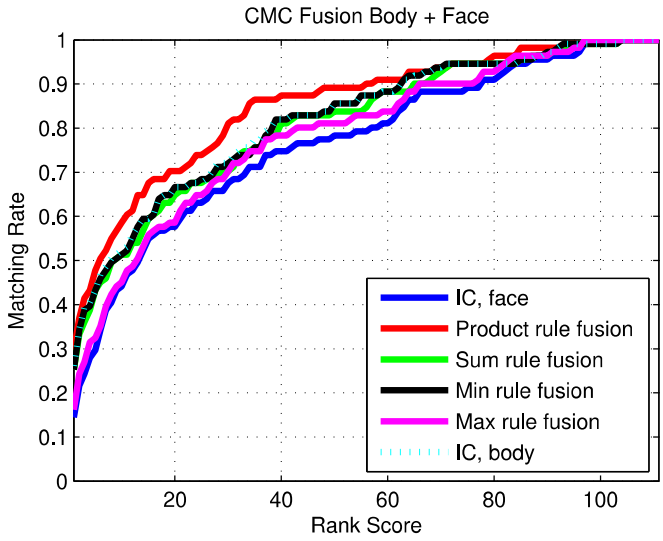


Figure 3.8: CMC comparison between our method, IC, on the face, the body and standards fusion rules.

Summary of the Fusion Result				
Approach/Result	Rank 1	Rank 10	Rank 20	STD
IC,Face	14.41	44.14	57.66	±2.02
IC,Body	26.13	52.25	66.67	±1.74
Product	29.73	58.56	70.27	±1.58
Sum	25.23	51.35	64.86	±1.78
Min	25.23	5135	6667	±1.76
Max	16.22	45.05	58.56	±1.99

Table 3.3: The Performance by using 9 training samples of the face and the body features

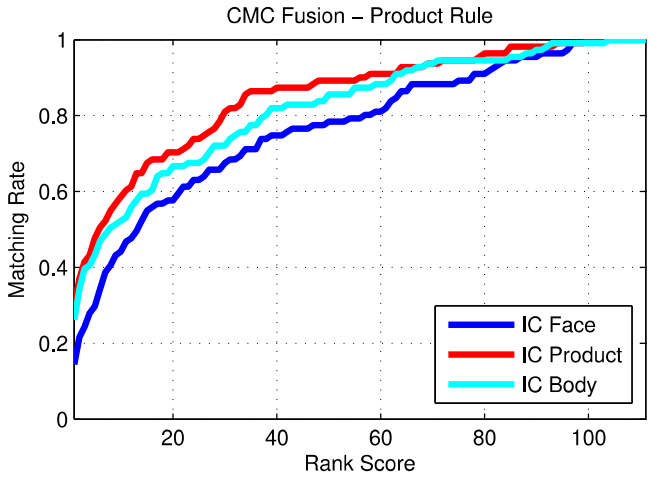


Figure 3.9: CMC comparison between our method, IC, on the face, the body and the fusion with product rule.

Summary of the Best Result by Product Fusion				
Approach/Result	Rank 1	Rank 10	Rank 20	STD
IC, Face	14.41	44.14	57.66	±2.02
IC, Body	26.13	52.25	66.67	±1.78
Ic, Product	29.73	58.56	70.27	±1.58

Table 3.4: The Performance of product rule fusion with face and body features

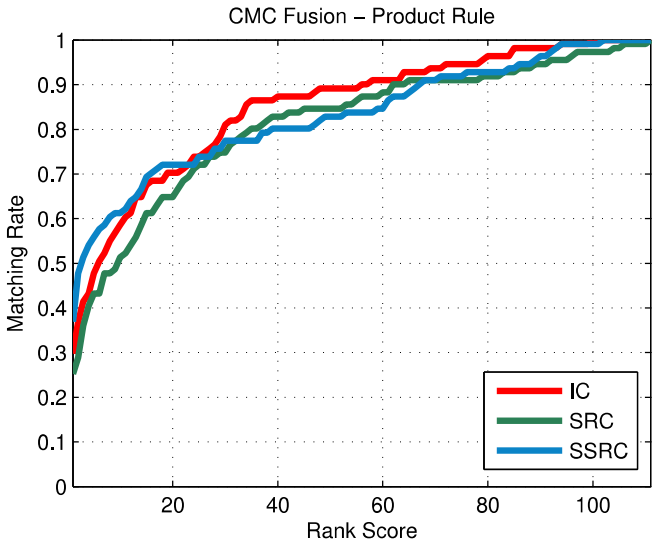


Figure 3.10: CMC comparison between our method, IC, SRC and SSRC using product rule fusion.

Summary of the Product Fusion				
Approach/Result	Rank 1	Rank 10	Rank 20	STD
IC	29.73	58.56	70.27	± 1.58
SRC	25.23	51.35	64.86	± 1.69
SSRC	36.94	61.26	72.07	± 1.35

Table 3.5: The Performance of product rule fusion on different approaches

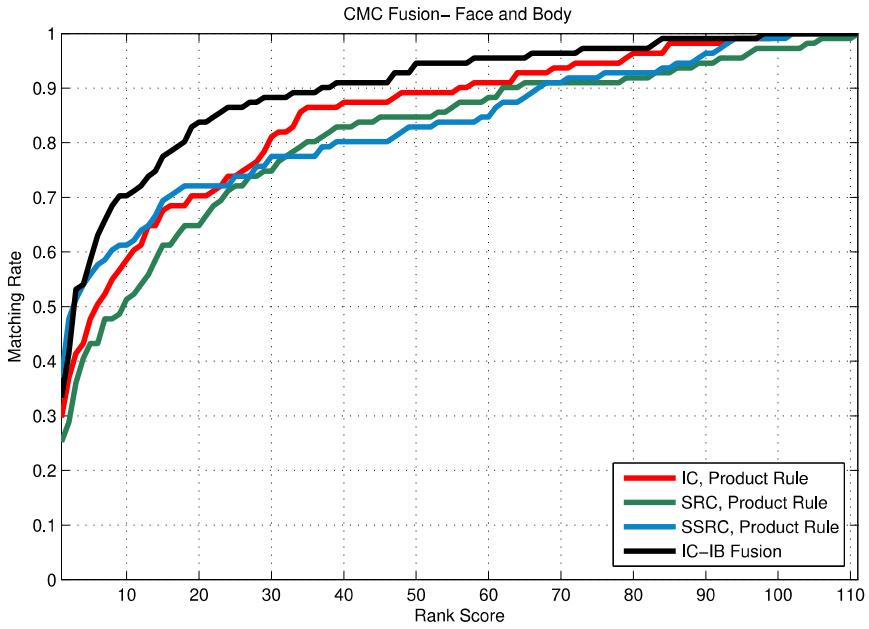


Figure 3.11: CMC comparison between our method, IC, SRC and SSRC using product rule fusion and IC-IB fusion.

Face and Body Fusion				
Approach/Result	Rank 1	Rank 10	Rank 20	STD
IC, Product rule	29.73	58.56	70.27	± 1.58
SRC, Product rule	25.23	51.35	64.86	± 1.69
SSRC, Product rule	36.94	61.26	72.07	± 1.35
IC-IB Fusion	33.33	70.27	83.78	± 1.26

Table 3.6: Face and Body Fusion by Different Approaches

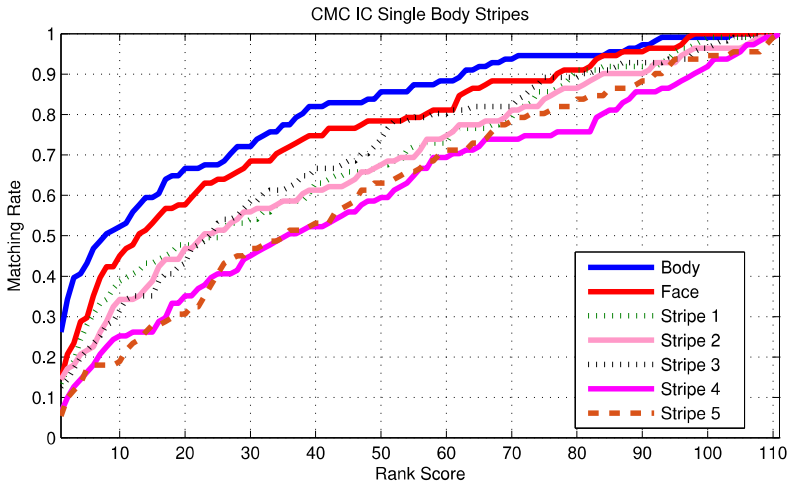


Figure 3.12: CMC comparison between our method, IC, on face, body and stripes.

Summary of the Result for each Stripe				
Feature/Result	Rank 1	Rank 10	Rank 20	STD
Body	26.13	52.25	66.67	±17.39
Face	14.41	45.05	57.66	±1.34
S1	13.51	38.74	47.75	±1.05
S2	14.41	34.23	46.85	±1.02
S3	12.61	31.53	43.24	±0.94
S4	6.30	25.23	35.14	±0.78
S5	5.41	18.92	30.63	±0.73

Table 3.7: The Performance of IC on All Strips

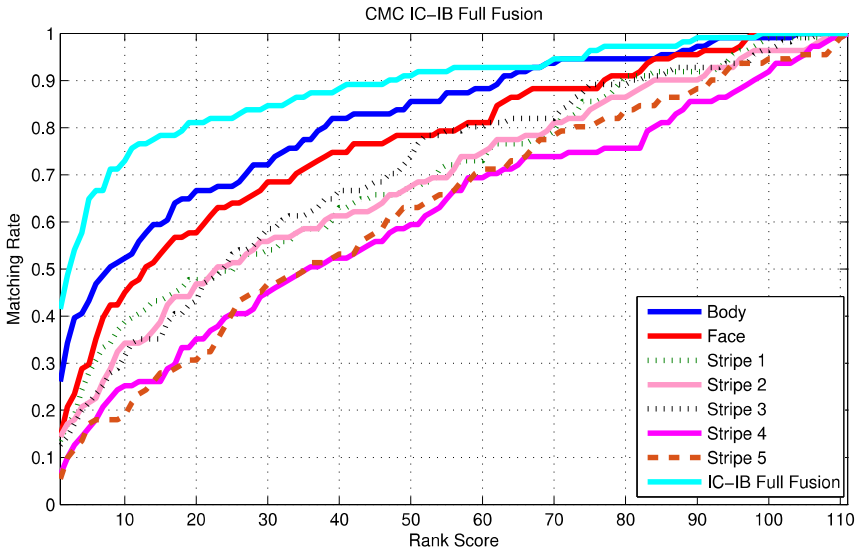


Figure 3.13: CMC comparison between our method, IC, on face, body, stripes and IC-IB full fusion.

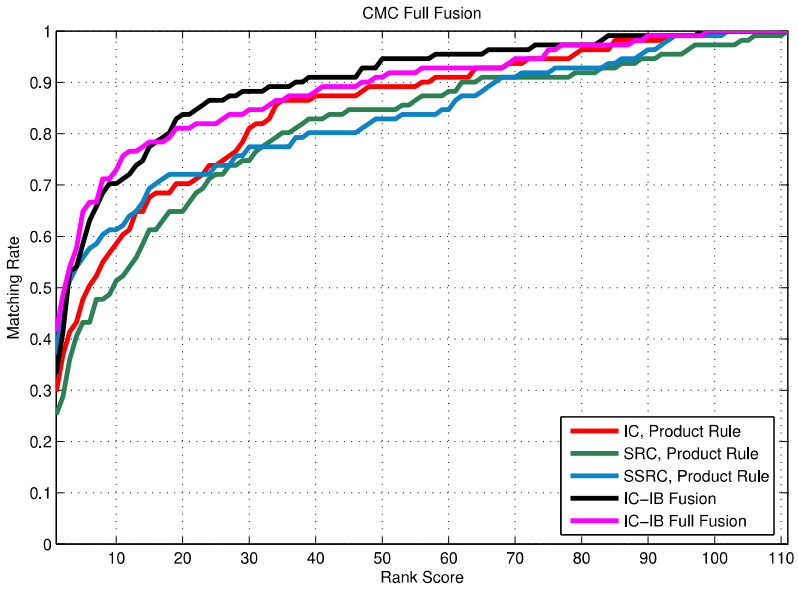


Figure 3.14: CMC comparison between IC-IB fusion, IC-IB full fusion and product rule fusion on IC, SRC and SSRC.

Full Fusion				
Approach/Result	Rank 1	Rank 10	Rank 20	STD
IC, Product rule	29.73	58.56	70.27	±0.1.58
SRC, Product rule	25.23	51.35	64.86	±1.70
SSRC, Product rule	36.94	61.26	72.07	±1.35
IC-IB Fusion	33.33	70.27	83.78	±1.27
IC-IB Full Fusion	41.44	72.97	81.08	±1.13

Table 3.8: Full Fusion

Summary of All Fusion Result				
Approach/Result	Rank 1	Rank 10	Rank 20	STD
IC,Face	14.41	44.14	57.66	± 2.02
IC,Body	26.13	52.25	66.67	± 1.74
Product	29.73	58.56	70.27	± 1.58
Sum	25.23	51.35	64.86	± 1.78
Min	25.23	51.35	66.67	± 1.76
Max	16.22	45.05	58.56	± 1.99
IC-IB Full Fusion	41.44	72.97	81.08	± 1.34

Table 3.9: Summary of All Fusion Result

Chapter 4

Discriminative and Metric Embedding Learning

4.1 Introduction

Person re-identification is a challenging task because of the high intra-class variance induced by the unrestricted nuisance factors of variations such as pose, illumination, viewpoint, background and sensor noise. Recent approaches postulate that powerful architectures have the capacity to learn feature representations invariant to nuisance factors, by training them with losses that minimize intra-class variance and maximize inter-class separation, without modeling nuisance factors explicitly.

There are two main lines of work in this area. The first one has focused on improving the triplet loss derived from metric learning [98] While introduced within the context of face recognition [72], the triplet loss has been leveraged for person re-identification [73, 74]. It directly attempts to pull together feature embedding of the same identity, while pushing apart those of different ones. Moreover, its computational complexity has improved dramatically, by efficiently mining large amounts of meaningful triplets [75]. The second line of work has focused on improving the softmax loss used for classification. By interpret-

ing the normalized weights of a softmax layer as the centroids representative of a given class or identity, within the context of face recognition, several works have shown how it is possible to control intra-class and inter-class distances by normalizing feature embedding, using cosine similarity, and adding margins to the loss [97, 76, 77]. However, we note that restricting the embedding to live on a hypersphere limits the gradient flow supervising the embedding under training. Thus, potentially generating a performance gap.

In this chapter, we improve person re-identification by combining the softmax and the triplet loss, to further improve the learning of a feature embedding. The intent is for the triplet loss to help the softmax further decrease intra-class variation, and increase inter-class distance by letting the triplet loss be the proxy for the gradient supervision that the embedding normalization has restricted, and we specify under what conditions this may happen. We also observe that the same strategy used to form the batch of triplets can be used in tandem with the softmax loss to prevent issues due to dataset imbalance, which are common in person re-identification. We perform an extensive evaluation of the proposed combination of losses on the latest person re-identification datasets. We found that not only this approach has competitive performance with state-of-the-art, but the appropriate use of the combined loss does not require for the softmax component to use any margin.

In addition to that, we further improve person re-identification by integrating the attribute into our proposed person re-identification. Person attributes used in person re-identification and showed robustness against variation of viewpoint, illumination and pose. This is achieved via injecting M attribute classifiers set on part of the embedding. These attribute classifiers will help to extract attribute specific features that help further improve the re-identification performance. Finally, we also include the explicit modeling of nuisance factors such as pose, to further improve the invariance of representations

4.2 Proposed Approach

For the person re-identification task, given a tightly cropped image sample of a person, I, we are seeking to learn a feature embedding $f_\theta(I)$, defined by the set of parameters θ, which is as invariant as possible to the nuisance factors of variations. Those include pose, illumination, viewpoint, as well as background and sensor noise. Rather than attempting to model nuisance factors, current deep neural network architectures have shown the promise to cope with their effects, by shifting the focus on designing clever training practices, as well as loss functions.

4.2.1 Classification Loss

A successful strategy for learning the embedding f_θ is through the use of a classification loss such as the categorical cross-entropy, which entails adding a softmax layer after the embedding. This leads to the loss

$$\mathcal{L}_S(\theta, W, b) = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{W_{y_i}^\top x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^\top x_i + b_j}} , \tag{4.1}$$

where $x_i = f_\theta(I_i) \in R^d$ is the embedding of I_i, which has identity y_i. Moreover, $W = [W_1, \cdots, W_c] \in R^{d \times c}$ and $b = [b_1, \cdots, b_c]$ are the weights and biases of the softmax layer, while N is the batch size.

Given two images I_i and I_j of the same identity, i.e., $y_i = y_j$, the softmax loss (4.1) will strive to make the target logit y_i be the highest for both images. While this should encourage $f_\theta(I_i)$ and $f_\theta(I_j)$ to be close, in general, there is not an explicit effort to impose $f_\theta(I_i) = f_\theta(I_j)$. This leads to a performance gap, given the large intra-class variability of the person re-identification task due to nuisance factors, which easily cause identity miss-classifications.

Within the context of face recognition, the issue above has been mitigated by taking

several steps. First, every logit is produced by comparing the input against ℓ^2-normalized weights [97, 76, 100], i.e. $\|W_j\| = 1$. This reduces by one the degrees of freedom by which two different logits could become equal, when activated by images I_i and I_j respectively, each of which depicting the same identity, i.e., $y_i = y_j$. Second, the input of every logit is also ℓ^2-normalized [100, 92, 93, 76], i.e. $\|x_i\| = 1$, and rescaled to a temperature value s. This further reduces the degrees of freedom by which different logits could become equal, by imposing the embeddings to be defined on the hypersphere of radius s. Also, this naturally suggests using cosine similarity as the metric for comparison between inputs and weights.

While input and weight normalizations positively contribute towards reducing intra-class variability, it is possible to further reduce the spread of the embeddings of the samples with same identity. This is done by introducing a margin in the cosine similarity, $\cos(\alpha)$, of the target logit. Doing so would further pull the embeddings closer to make up for the loss of similarity induced by the margin. In the literature there are at least three basic ways to add a cosine similarity margin, either by acting multiplicatively [97] or addictively [77] on the angle α, or addictively [76] on the similarity. By combining all three methods, the cosine similarity becomes $(\cos(m_1\alpha + m_2) - m_3)$, where m_1, m_2, m_3 are the three types of margin, respectively. In the ArcFace work [77] it is shown that for face recognition the *additive angular* margin [77] is the most effective, which reduces (4.1) to

$$\mathcal{L}_{AM}(\theta, W) = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{s\cos(\alpha_{y_i}+m)}}{e^{s\cos(\alpha_{y_i}+m)} + \sum_{j\neq y_i} e^{s\cos\alpha_j}} \ , \tag{4.2}$$

where we have set $b = 0$ for simplicity, as in [97]. In the experiments we explore the effectiveness of (4.2) for person re-identification. Its action should be to minimize the intra-class variation, while the denominator attempts to maximize the inter-class discrepancy by distancing the weights on the unit hypersphere.

4.2.2 Metric Learning Loss

By learning a metric embedding we directly train a function f_θ that maps images of the same identity as close as possible, effectively minimizing the intra-class variability of the embeddings, while images of different identities are mapped far away, creating a large inter-class discrepancy. In [98] they developed a margin based approach for k-nearest neighbor classification, which has then inspired the *triplet loss* formulation of FaceNet [72] as follows

$$\mathcal{L}_T(\theta) = \sum_{\substack{a,p,n \\ y_a = y_b \neq y_n}} [m + D(f_\theta(I_a), f_\theta(I_p)) - D(f_\theta(I_a), f_\theta(I_n))]_+ \tag{4.3}$$

where $D(\cdot,\cdot)$ denotes a suitable distance, and $[\cdot]_+$ is the hinge function, but other surrogates could be used, such as the softplus function $\ln(1+\exp(\cdot))$. The triplet loss (4.3) operates by ensuring that the distance between a positive sample I_p and an anchor I_a, which have same identities, is smaller than the distance between the ancor and a negative sample I_n, which has a different identity, at least by a margin m. When the loss is optimized over a large combination of triplets (I_a, I_p, I_n), it pulls embeddings of the same identity, while pushing apart those with different identities.

The challenges in using the triplet loss are related to the cubic growth in the number of triplets as the dataset size grows, and in forming meaningful triplets. It turns out that the embedding can quickly learn how to correctly map easy triplets, e.g., where the positive and anchor samples have a similar pose, and the negative sample is a person dressed up in completely different colors. Conversely, focussing on selecting very hard triplets may not be very useful too, because we would teach the embedding how to map outlier cases, while overlooking how to handle well "average" cases. This is why it is important to efficiently mine moderate positives and negatives [72, 94].

As described in [75], it turns out that there is an effective way to address both of the issues above. Triplets can be formed out of selecting P identities, and K samples per identity, with a total of PK samples in a batch. Because of this, we are not restricted to form only $PK/3$ triplets. Instead, every sample can be used as anchor, and the associated positives and negatives are picked to be the hardest. Since we are operating only within a batch, these hard selections will not be outliers, but mostly non-trivial moderate cases. In addition, this approach avoids the overhead induced by mining moderate cases from the full dataset processed by the latest update of f_θ. This procedure, named *batch hard* [75], changes (4.3) into

$$\mathcal{L}_{BH}(\theta) = \sum_{i=1}^{P}\sum_{a=1}^{K} [m + \max_{p} D(f_\theta(I_a^i), f_\theta(I_p^i)) - \min_{j,n,j\neq i} D(f_\theta(I_a^i), f_\theta(I_n^j))]_+ \ . \tag{4.4}$$

4.2.3 Joint Classification and Metric Loss

Besides evaluating the additive angular magin softmax loss (4.2) and the batch hard triplet loss (4.4) on the most recent re-identification datasets, we plan to study their contribution into a joint loss

$$\mathcal{L}_{AMBH}(\theta, W) = \mathcal{L}_{AM}(\theta, W) + \gamma\mathcal{L}_{BH}(\theta) \ . \tag{4.5}$$

where γ is a hyperparameter balancing the relative strengths of the losses.

There are a couple reasons that motivate the exploration of the join loss (4.5). The first one comes from observing that a major drawback of the loss (4.2) is that the gradient $\nabla_\theta\mathcal{L}_{AM}$ is proportional to the gradient $\nabla_\theta\tilde{f}_\theta$, where $\tilde{f}_\theta \doteq f_\theta/\|f_\theta\|$ because of the ℓ^2 normalization of the softmax inputs. Since $\tilde{f}_\theta$ lives on the unit hypersphere, the gradient $\nabla_\theta\tilde{f}_\theta$ will always be tangent to it. Therefore, no gradients perpendicular to the hypersphere

will be back-propagated to supervise f_θ for reducing the intra-class variability of the embedding, while maximizing the inter-class discrepancy. This issue suggests that adding a regularizing term to the loss (4.2), which allows orthogonal gradients to flow back could increase the hypothesis space of the embedding f_θ, and better become invariant to nuisance factors of variation.

By adding (4.4) to (4.2) as in (4.5), we are addressing the issue highlighted above. Indeed, the intent of (4.4) and (4.2) is the same, but in (4.4) we do not have the requirement for f_θ to be ℓ^2 normalized. Hence, by picking a distance $D(\cdot, \cdot)$ that does not normalize the embedding, (4.5) enables the gradient to flow in all directions. In [77] they attempted merging (4.2) with (4.3) without success, but there they used a distance with a normalized embedding, which we advocate not to use in this case. In our experiments, we picked $D(\cdot, \cdot)$ to be the Euclidean distance.

The second reason for using (4.5) comes from the composition of a batch, which has certain requirements because of the batch hard mining. We note that the same batch made of $N = PK$ samples can actually be used for the loss (4.2). More importantly, this approach may prevent issues related to imbalanced data. Since datasets for person re-identification may have identities with a lot more samples than others, sampling a constant number of identities, from which we sample a constant number of images, imposes the embedding to be trained uniformly across the identities, rather than being under/over trained on some of them. In all of our experiments we sample the batches as it is done for the batch hard mining, regardless of the loss that we use.

Finally, we note that since the loss (4.4) exercises a set of push-pull forces, it might be that when used as in (4.5), the effect of the additive angular margin in (4.2) could become less relevant. Indeed, this is one of our conclusions.

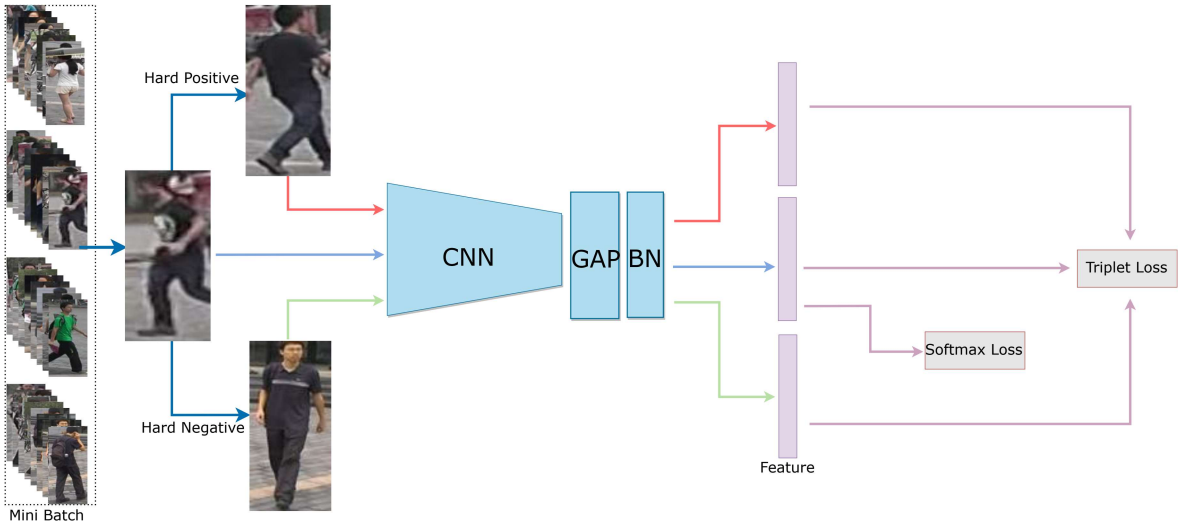


Figure 4.1: **Architecture.** Simple graphical description of the joint optimization of a softmax and a triplet loss. We use (4.2) as softmax, and (4.4) as triplet. The former normalizes the features and the layer weights, and the latter does not.

4.2.4 Network Architecture

As in most of the recent literature on person re-identification, we use a pretrained ResNet-50 [70] as backbone network. We simply discard the fully connected layer, and we change the stride of the last convolutional stage from 2 to 1. We then add a global average pooling (GAP) layer and a batch normalization layer (BN). At this point weight normalization and ℓ^2 feature normalization is applied before entering the additive angular margin softmax loss (4.2), while no normalization is needed for the batch hard triplet loss component (4.4). Figure 4.1 is a simple exemplification of the architecture.

During testing, unless otherwise specified, we perform all the experiments with the ℓ^2 normalized embedding $f_\theta / \| f_\theta \|$, and re-identification is done via cosine similarity.

4.3 Experiments

4.3.1 Datasets

we evaluate our proposed model on three Person Re-ID datasets:

Market-1501: [84] contains 32668 images of 1501 identities captured by six cameras, fives are high resolution cameras and one is low resolution camera. Pedestrians are detected by a deformable detector model (DPM) [101]. There are 751 identities with 12936 images for training and another 750 identifies for testing where gallery set have 19732 images and query have 3368 images.

DukeMTMC-ReID: is a subset of the DukeMTMC dataset [85] for image-based re-identification. It has 36441 images of 1812 identities captured by eight high resolution cameras. There are 16522 images of 702 identities in the training set, 2228 images of the other 702 identifies in the query set and 17661 images of 1110 identities (702 identities are the same as in query set and the others are distractors) in the test set.

MSMT17: [102] is the most recent and challenging person Re-ID dataset. It contains 126441 images of 4101 identities captured by 15 cameras, 12 outdoor and 3 indoor cameras. Faster RCNN [103] is used as a pedestrian bounding box detector. The training and testing ration is 1:3. The training set contains 32624 images of 1041 identities. The testing set contains 93820 images of 3060 identities. From the 93820 images in testing set, 11659 images are used as query images and the other 82161 are used as gallery images.

4.3.2 Implementation Details

We implemented our approach with PyTorch [104], and for the backbone network we use ResNet-50 [70] that is pre-trained on ImageNet [106]. The fully connected layer in ResNet-50 is discarded and the stride of the last stage of ResNet-50 backbone is changed from 2 to 1. We add a global average pooling and batch normalization after the last convolutional layer of ResNet-50. The dimensionality of the embedding features is 2048.

Each mini-batch consists of $P \cdot K$ images where P represents the number of different identities and K represents number of images per identity. Batch size is set to 32 where P is set to 4 and K is set to 8. γ is 0.43, 0.5 and 0.4 for Market-1501, DukeMTMC-reID and MSMT17 respectively.

For Market-1501 datasets, input image size is 256×128 while it is 288×144 for DukeMTMC-reID and MSMT17 datasets. For data augmentation, training images are randomly flipped and erased. The model trained using Adam optimizer with default hyper parameters for 150 epochs. The learning rate is linearly increased from 10^{-5} to 10^{-3} for first 20 epochs to help the network to bootstrap. Then learning rate sets to 10^{-3} after epoch 20 and it decreases to 10^{-4} and 10^{-5} after epochs 90, 130 respectively.

4.3.3 Comparison with State-of-the-Art Methods

The performance is evaluated by CMC (Cumulative Matching Characteristic) and mAP (Mean Average Precision) after computing the matching score between the probe image and gallery images. We discard the score if the probe image and galley image are from the same view. Post-processing technique, re-ranking [107] and multiquery fusion [84], are not used to present the results.

To show the performance of our proposed approach, we compare it with the state-of-the-art methods on three person re-identification dataset. Table 4.1 shows the results on Market-1501, our approach outperforms most of the state-of-the-art and the performance is close to the best i.e. DG-Net [68] in terms of Rank-1 and mAP. Table 4.2 shows the results on DukeMTMC-reID, our approach outperforms the state-of-the-art by achieving 89.2% rank-1 and 76.7% mAP. Table 4.3 shows results on MSMT17, the proposed approach outperforms DG-Net [68] by a gap of 0.9%, 0.9%, 0.7% and 1.1% for rank-1, rank-5, rank-10 and mAP respectively.

Method	Backbone	Rank-1	mAP
PN-GAN [108]	ResNet	89.4	72.6
FD-GAN [109]	ResNet	90.5	77.7
Part-aligned [110]	GoogleNet	91.7	79.6
Mancs [112]	ResNet	93.1	82.3
SGGNN [113]	ResNet	92.3	82.8
DeepCRF [114]	ResNet	93.5	81.6
PCB [115]	ResNet	93.8	81.6
AA-Net [12]	ResNet	93.9	83.4
IA-Net [116]	ResNet	94.4	83.1
SphereReID [78]	ResNet	94.4	83.6
CAMA [69]	ResNet	94.7	84.5
DG-Net [68]	ResNet	94.8	86.0
AM0BH (Our)	ResNet	$94.6_{\pm 0.21}$	$85.9_{\pm 0.28}$
AM0BHsp (Our)	ResNet	$94.4_{\pm 0.15}$	$85.8_{\pm 0.19}$

Table 4.1: Comparison with the-sate-of-the-art methods on the Market-1501 datasets.

Method	Backbone	Rank-1	mAP
PN-GAN [108]	ResNet	73.6	53.2
FD-GAN [109]	ResNet	80.0	65.4
SGGNN [113]	ResNet	81.1	68.2
PCB [115]	ResNet	83.3	69.2
SphereReID [78]	ResNet	83.9	68.5
Part-aligned [110]	GoogleNet	84.4	69.3
DeepCRF [114]	ResNet	84.9	69.5
Mancs [112]	ResNet	84.9	71.8
CAMA [69]	ResNet	85.8	72.9
DG-Net [68]	ResNet	86.6	74.8
IA-Net [116]	ResNet	87.1	73.4
AA-Net [89]	ResNet	87.7	74.3
AM0BH (Our)	ResNet	$89.2_{\pm 0.40}$	$76.7_{\pm 0.26}$
AM0BHsp (Our)	ResNet	$88.8_{\pm 0.31}$	$77.4_{\pm 0.3}$

Table 4.2: Comparison with the-sate-of-the-art methods on the DukeMTMC-ReID datasets.

Method	Backbone	Rank-1	Rank-5	Rank-10	mAP
PCB [115]	ResNet	68.2	81.2	85.5	40.4
IA-Net [116]	ResNet	75.5	85.5	88.7	46.8
DG-Net [68]	ResNet	77.2	87.4	90.5	52.3
AM0BH (Our)	ResNet	$78.1_{\pm 0.40}$	$88.3_{\pm 0.13}$	$91.2_{\pm 0.19}$	$53.4_{\pm 0.32}$
AM0BHsp (Our)	ResNet	$78.3_{\pm 0.27}$	$88.4_{\pm 0.11}$	$91.2_{\pm 0.14}$	$53.4_{\pm 0.40}$

Table 4.3: Comparison with the-sate-of-the-art methods on the MSMT17 datasets.

4.3.4 Ablation Study

In the ablation study we compare different loss combinations on all three datasets used in 4.3.3: **Market-1501:**, **DukeMTMC-ReID:**, and **MSMT17:**. Results of the ablation study are presented in the Tables 4.4, 4.5 and 4.6.

First, we examining additive angular margin softmax loss [77] and batch hard triplet loss

Additive angular margin softmax loss. We start with examining softmax loss (4.2) when margin is set to 0 (row AM0 in the Tables 4.4, 4.5 and 4.6), this case is equivalent to the loss described in [78]. Then, we use additive angular margin softmax loss when margin is set to 0.5, as in [77] (row AM in the Tables 4.4, 4.5 and 4.6). We observe significant improvement of all metrics, which proves that additive angular margin have a positive effect on the performance when softmax loss is used alone.

Batch hard triplet loss. We continue by examining batch hard triplet loss (row BH in the Tables 4.4, 4.5 and 4.6), which shows significantly worse performance compared to AM or AM0.

Additive angular margin softmax loss and Batch hard triplet loss. We analyze four cases: combination of AM0 and BH, combination of AM and BH, combination of AM0 and BH with feature normalization for BH, combination of AM0 and BH with softplus instead of hinge loss.

Combination of AM0 and BH. Combination of softmax loss (4.2) with margin set to 0 and batch hard triplet loss is presented in the row AM0BH of the TTables 4.4, 4.5 and 4.6. We observe improvement on almost all metrics compared to AM0 or BH individually, which signifies that they complement each other. To verify the reasons behind such complementing described in 4.2.3 we examine other variations of their combinations as follows.

Combination of AM and BH. Combination of additive angular margin softmax loss (4.2) with margin set to 0.5 and batch hard triplet loss is presented in the row AMBH of the Tables 4.4, 4.5 and 4.6. We observe that it does not provide improvement compared to

AM0BH, which signifies that angular margin became less relevant since BH exercises a set of push-pull forces and achieves the same effect as angular margin.

Combination of AM0 and BH with normalized features. Combination of softmax loss (4.2) with margin set to 0 and batch hard triplet loss with normalized features is presented in the row AM0BH1 of the Tables 4.4, 4.5 and 4.6. Because of the feature normalization prior to BH, no gradients perpendicular to the hypersphere will be back-propagated to supervise f_θ. Significantly decreased performance suggests, as described in 4.2.3, that orthogonal gradients flow could increase the hypothesis space of the embedding f_θ.

Combination of AM0 and BH with softplus. Combination of softmax loss (4.2) with margin set to 0 and batch hard triplet loss with softplus function instead of hinge loss is presented in the row AM0BHsp of the Tables 4.4, 4.5 and 4.6. It shows overall similar performance to AM0BH, but slightly higher mAP, while slightly lower rank1-10 metrics. We make a speculative hypothesis that hinge loss concentrates only on the triplets within the margin, ignoring the tail of the distribution, which is beneficial for rank-1 - rank-10 metrics, while with softplus the whole distribution of triples is accounted in the loss, which is beneficial for the mAP metric.

Therefore, here we presented ablation study that supports motivation behind picked loss as described in 4.2.3.

Loss	Rank1	Rank5	Rank-10	mAP
AM0	$92.86_{\pm 0.34}$	$97.57_{\pm 0.12}$	$98.47_{\pm 0.04}$	$83.67_{\pm 0.17}$
AM	$94.16_{\pm 0.23}$	$98.04_{\pm 0.21}$	$98.92_{\pm 0.08}$	$84.54_{\pm 0.22}$
BH	$84.74_{\pm 0.23}$	$94.64_{\pm 0.36}$	$96.74_{\pm 0.13}$	$67.40_{\pm 0.46}$
AM0BH1	$94.28_{\pm 0.13}$	$97.90_{\pm 0.20}$	$98.80_{\pm 0.07}$	$84.52_{\pm 0.16}$
AMBH	$93.29_{\pm 0.40}$	$97.80_{\pm 0.11}$	$98.70_{\pm 0.14}$	$84.00_{\pm 0.09}$
AM0BH(Our)	$94.64_{\pm 0.21}$	$98.22_{\pm 0.16}$	$99.02_{\pm 0.11}$	$85.90_{\pm 0.28}$
AM0BHsp(Our)	$94.42_{\pm 0.15}$	$98.22_{\pm 0.19}$	$99.02_{\pm 0.13}$	$85.76_{\pm 0.19}$

Table 4.4: **Ablation study**. Shows the effect of different loss combinations on Market-1501.

Loss	Rank1	Rank5	Rank-10	mAP
AM0	$87.74_{\pm 0.29}$	$94.06_{\pm 0.17}$	$95.62_{\pm 0.35}$	$74.60_{\pm 0.30}$
AM	$88.31_{\pm 0.18}$	$94.34_{\pm 0.48}$	$95.78_{\pm 0.24}$	$75.51_{\pm 0.16}$
BH	$81.60_{\pm 0.49}$	$91.02_{\pm 0.38}$	$93.46_{\pm 0.18}$	$65.16_{\pm 0.31}$
AM0BH1	$88.02_{\pm 0.18}$	$93.96_{\pm 0.27}$	$95.48_{\pm 0.26}$	$74.56_{\pm 0.32}$
AMBH	$88.18_{\pm 0.46}$	$94.62_{\pm 0.17}$	$96.19_{\pm 0.18}$	$76.71_{\pm 0.21}$
AM0BH(Our)	$89.20_{\pm 0.40}$	$94.72_{\pm 0.33}$	$96.26_{\pm 0.17}$	$76.68_{\pm 0.26}$
AM0BHsp(Our)	$88.79_{\pm 0.31}$	$94.85_{\pm 0.22}$	$96.32_{\pm 0.19}$	$77.42_{\pm 0.3}$

Table 4.5: **Ablation study**. Shows the effect of different loss combinations on DukeMTMC-reID.

Loss	Rank1	Rank5	Rank-10	mAP
AM0	$77.50_{\pm 0.25}$	$87.80_{\pm 0.30}$	$90.78_{\pm 0.18}$	$51.82_{\pm 0.19}$
AM	$78.20_{\pm 0.28}$	$88.15_{\pm 0.42}$	$91.15_{\pm 0.35}$	$53.00_{\pm 0.49}$
BH	$56.34_{\pm 0.92}$	$73.26_{\pm 0.99}$	$79.30_{\pm 0.70}$	$30.86_{\pm 0.67}$
AM0BH1	$77.46_{\pm 0.13}$	$87.64_{\pm 0.15}$	$90.60_{\pm 0.19}$	$51.86_{\pm 0.22}$
AMBH	$77.93_{\pm 0.47}$	$88.07_{\pm 0.47}$	$91.07_{\pm 0.38}$	$52.70_{\pm 0.53}$
AM0BH(Our)	$78.14_{\pm 0.40}$	$88.34_{\pm 0.13}$	$91.24_{\pm 0.19}$	$53.44_{\pm 0.32}$
AM0BHsp(Our)	$78.26_{\pm 0.27}$	$88.38_{\pm 0.11}$	$91.20_{\pm 0.14}$	$53.36_{\pm 0.40}$

Table 4.6: **Ablation study**. Shows the effect of different loss combinations on MSMT17.

P	Rank-1	Rank-5	Rank-10	mAP
4	94.64 $\pm$0.21	98.22 $\pm$0.16	99.02 $\pm$0.11	85.90 $\pm$0.28
8	94.22 $\pm$0.26	98.12 $\pm$0.26	98.94 $\pm$0.15	84.88 $\pm$0.26

Table 4.7: Performance on Market-1501 using different values of P. K is fixed to 8.

P	Rank-1	Rank-5	Rank-10	mAP
4	89.20 $\pm$ 0.40	94.72 $\pm$ 0.33	96.26 $\pm$ 0.17	76.68 $\pm$ 0.26
8	87.98 $\pm$ 0.37	94.64 $\pm$ 0.18	96.18 $\pm$ 0.18	75.38 $\pm$ 0.25

Table 4.8: Performance on DukeMTMC-reID using different values of P. K is fixed to 8.

P	Rank-1	Rank-5	Rank-10	mAP
4	78.14 $\pm$ 0.40	88.34 $\pm$ 0.13	91.24 $\pm$ 0.19	53.44 $\pm$ 0.32
8	77.92 $\pm$ 0.37	88.16 $\pm$ 0.32	91.04 $\pm$ 0.17	53.22 $\pm$ 0.23

Table 4.9: Performance on MSMT17 using different values of P. K is fixed to 8.

K	Rank-1	Rank-5	Rank-10	mAP
4	93.78 ±0.22	97.88 ±0.08	98.76 ±0.05	83.86 ±0.32
8	94.64 ±0.21	98.22 ±0.16	99.02 ±0.11	85.90 ±0.28
16	94.72 ±0.19	98.34 ±0.09	98.98 ±0.04	86.14 ±0.19

Table 4.10: Performance on Market-1501 using different values of K. P is fixed to 4.

K	Rank-1	Rank-5	Rank-10	mAP
4	87.44 ± 0.27	94.14 ± 0.29	95.96 ± 0.21	74.82 ± 0.18
8	89.20 ± 0.40	94.72 ± 0.33	96.26 ± 0.17	76.68 ± 0.26
16	88.88 ± 0.13	94.58 ± 0.37	95.82 ± 0.22	76.12 ± 0.22

Table 4.11: Performance on DukeMTMC-reID using different values of K. P is fixed to 4.

K	Rank-1	Rank-5	Rank-10	mAP
4	74.94 ± 0.18	86.02 ± 0.18	89.38 ± 0.08	49.82 ± 0.19
8	78.14 ± 0.40	88.34 ± 0.13	91.24 ± 0.19	53.44 ± 0.32
16	78.16 ± 0.25	88.30 ± 0.19	91.08 ± 0.24	52.74 ± 0.41

Table 4.12: Performance on MSMT17 using different values of K. P is fixed to 4.

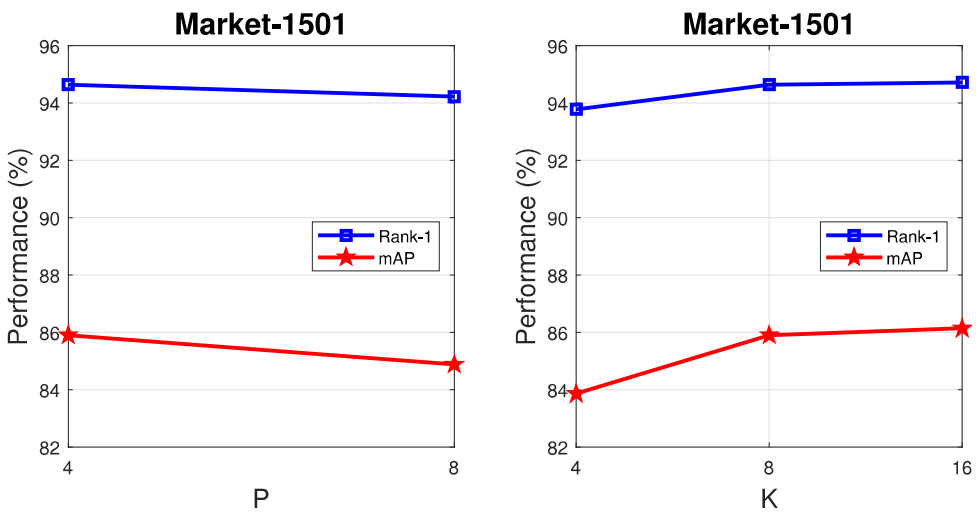


Figure 4.2: **Ablation study**. Shows the effect of P and K on the performance of AM0BH on Market-1501. **Left:** K is fixed to 8. **Right:** P is fixed to 4.

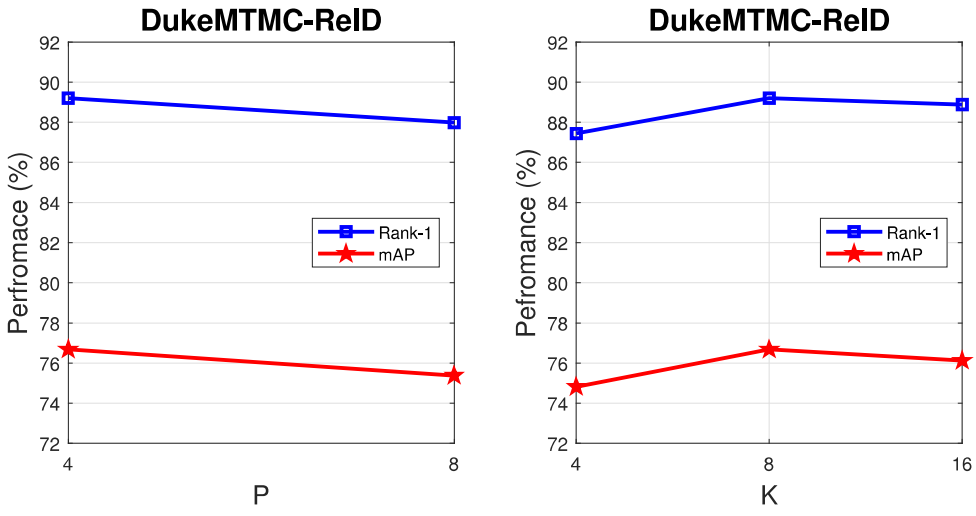


Figure 4.3: **Ablation study**. Shows the effect of P and K on the performance of AM0BH on DukeMTMC-reID. **Left:** K is fixed to 8. **Right:** P is fixed to 4.

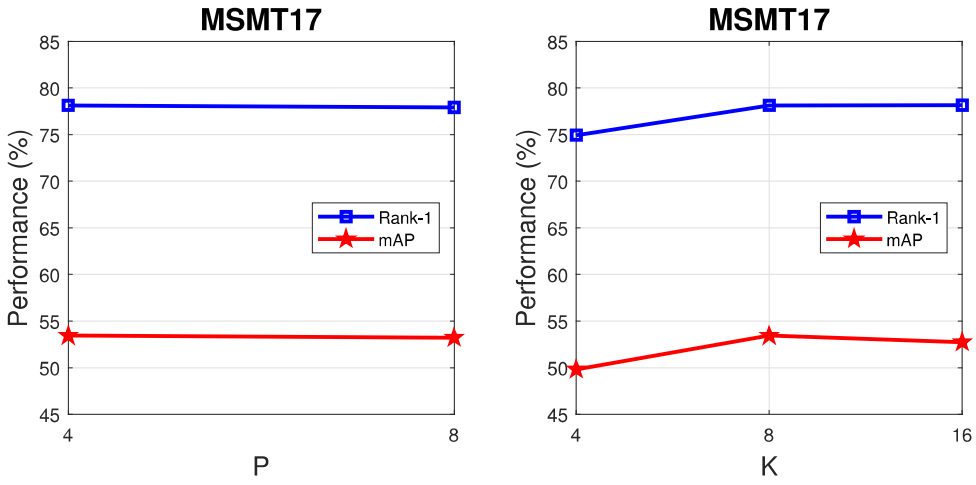


Figure 4.4: **Ablation study**. Shows the effect of P and K on the performance of AM0BH on MSMT17. **Left:** K is fixed to 8. **Right:** P is fixed to 4.

4.4 Person Attributes to Improve Person Re-Identification

We propose using pedestrian attributes to enable the person re-identification model to learn more discriminative embedding. Person attributes used in person re-identification and showed robustness against variation of viewpoint, illumination and pose. This can be achieved via integrating the attribute into our person re-identification model that proposed in this chapter via injecting M attribute classifiers set on part of the embedding. These attribute classifiers will help to extract attribute specific features that helps improving the re-identification accuracy.

4.4.1 Proposed Approach

We further push the training of the embedding to become invariant to the nuisance factors by leveraging the attribute labels. Assuming that a person image I is described by M binary attributes, we train the embedding $f_\theta(I)$ to be predictive of those attributes. Also in this case we follow the strategy of normalizing the input and weights to reduce intra-class variability and maximize the correct prediction of the attributes. Note that for every at-

tribute the problem is binary, meaning that each attribute is either present or not. However, in order to implement the normalization strategy, and leverage the additive angular margin loss, we cannot treat it like a multi-label problem where we use the binary cross-entropy loss for every attribute. Instead, we use a categorical cross-entropy loss for every attribute where the number of categories is 2.

To integrate the attributes into person re-identification, M binary attribute classifiers are trained on the attribute labels, based on part of the embedding in order to extract attribute specific features that help also the re-identification task. Thus, the embedding f_θ is divided into two parts $f_{\theta_{id}}$ and f_{θ_a} , where the first one is reserved for the identity classifier, while the other one is for attribute classifiers. Therefore (4.2) become:

$$\mathcal{L}_{AM_{id}}(\theta_{id}, W_{id}) = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cos(\alpha_{id_{y_i}}+m)}}{e^{s\cos(\alpha_{id_{y_i}}+m)}+\sum_{j\neq y_i}e^{s\cos\alpha_{id_j}}}\ , \tag{4.6}$$

On the other end, f_{θ_a} is the input of the M binary attribute classifiers based on the additive angular margin loss [77]. Therefore the corresponding loss for this pool of classifiers become:

$$\mathcal{L}_{AM_{attr}}(\theta_a, W_a) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{M}\log\frac{e^{s\cos(\alpha_{a_{a_k}}+m)}}{e^{s\cos(\alpha_{a_{a_k}}+m)}+e^{s\cos\tilde{\alpha}_{a_k}}}\ , \tag{4.7}$$

where $\tilde{\alpha}_{a_k}$ indicates the input to the cosine corresponding to the option where the attribute is not present. The final multi-task classification loss imposed on the identities and attributes is the sum of (4.6) and (4.7). Also, note that $f_{\theta_{id}}$ and f_{θ_a} actually share a significant amount of weights, since they only differ for the weights in the last layer, leading up to the two embedding components θ_{id} and θ_a . They are indicated in that way to limit the introduction of more notation.

The final loss is set by combining multi-task loss that includes the identity and the attribute components (4.6) and (4.7), respectively with the metric loss (4.4). This leads to

the full re-identification training model, given by

$$\mathcal{L}_{AMBH_{Attr}}(\theta, W) = \mathcal{L}_{AM}(\theta_{id}, W_{id}) + \lambda \mathcal{L}_{AM_{Attr}}(\theta_a, W_a) + \gamma \mathcal{L}_{BH}(\theta_{id}) \ . \tag{4.8}$$

4.4.2 Network Architecture

We use a pretrained ResNet-50 [70] as backbone network. We simply discard the fully connected layer, and we change the stride of the last convolutional stage from 2 to 1. We then add a global average pooling (GAP) layer and a batch normalization layer (BN). At this point weight normalization and ℓ^2 feature normalization is applied before entering the additive angular margin softmax losses (4.6) and (4.7), while no normalization is needed for the batch hard triplet loss component (4.4). Figure 4.5 shows a graphical description of the architecture for integrating the attribute in person re-identification.

During testing, unless otherwise specified, we perform all the experiments with the ℓ^2 normalized embedding $f_\theta / \| f_\theta \|$, and re-identification is done via cosine similarity for integrating the attribute in person re-identification.

4.4.3 Experiments

We evaluate our proposed model on two Person Re-ID datasets, Market-1501 [84] and DukeMTMC-ReID [85]. We use the data provided in [83] as attribute labels for Market-1501 and DukeMTMC. These attributes are manually annotated at the identity level. There are 27 attributes for Market-1501 and 23 attributes for DukeMTMC. Some examples of attributes include: gender, hair length, carrying backpack, carrying handbag, wearing hat, different upper body and lower body clothing colors, length and type of lower body cloth-

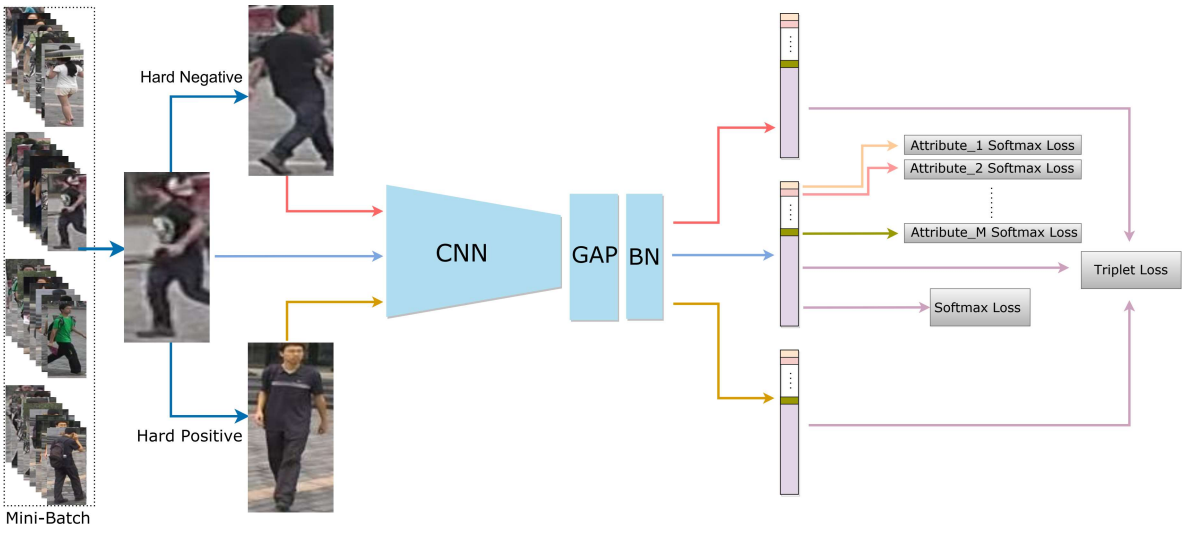


Figure 4.5: **Architecture.** Simple graphical description of the joint optimization of the multi-task re-identification loss, based on a multi-class classification (4.6) (for the identities), a multi-label classification (4.7) (for attributes), and a metric learning loss (4.4). The former (4.6) and (4.7) normalize the features and the layer weights, and the latter (4.4), does not.

ing, shoe type and shoes color.

We implemented our approach with PyTorch [104], and for the backbone network we use ResNet-50 [70] that is pre-trained on ImageNet [106]. The fully connected layer in ResNet-50 is discarded and the stride of the last stage of ResNet-50 backbone is changed from 2 to 1. We add a global average pooling and batch normalization after the last convolutional layer of ResNet-50. The dimensionality of the embedding features is 2048.

Each mini-batch consists of $P \cdot K$ images where P represents the number of different identities and K represents number of images per identity. Batch size is set to 32 where P is set to 4 and K is set to 8. For Market-1501 datasets, input image size is 256×128 while it is 288×144 for DukeMTMC-reID datasets.

For data augmentation, training images are randomly flipped and erased.The model trained using Adam optimizer with default hyper parameters for 150 epochs. The learn-

ing rate is linearly increased from 10^{-5} to 10^{-3} for first 20 epochs to help the network to bootstrap. Then learning rate sets to 10^{-3} after epoch 20 and it decreases to 10^{-4} and 10^{-5} after epochs 90, 130 respectively.

The 2048 embedding features are divided into f_{θ_a} and $f_{\theta_{id}}$. f_{θ_a} has size $M \cdot Q$, where M is the total number of attributes and Q is the size of the input features to each of the attribute classifiers. M is 27 for Market-1501 and 23 for DukeMTMC-reID. Q is set to 16. The rest of the embedding features, $f_{\theta_{id}}$ = 2048 - $M \cdot Q$, is the input to the identity classifier and triplet loss. In (4.8) we set λ to 0.25 and 0.2 for Market-1501 and DukeMTMC-reID, respectively, while γ is set to 0.54 and 0.33 for Market-1501 and DukeMTMC-reID, respectively.

We compare our proposed approach with the state-of-the-art methods on Market-1501 datasets and DukeMTMC-ReID datasets. Table 4.13 shows the results on Market-1501, our approach outperforms the state-of-the-art and $AM0BH_{Attr}$ further improves the performance over $AM0BH$ by 0.3% and 0.4% for rank-1 and mAP respectively. Table 4.14 shows the results on DukeMTMC-reID, our approach outperforms the state-of-the-art and $AM0BH_{Attr}$ able to improve the performance over $AM0BH$ by 0.1% and 0.7% for rank-1 and mAP respectively.

Also we present the ablation study that supports the addition of the attribute classification task to further improve the robustness against nuisance factors. We study the influence of Q and batch size on the performance of $AM0BH_{Attr}$. Tables 4.15 and 4.16 show the effect of the size, Q, of the embedding features fed to each attribute classifier. The same information is plotted in Figure 4.6. The best performance is achieved when Q is 16. Tables 4.17 and 4.18 and Tables 4.19 and 4.20 instead show the effect of the batch size by examining different values of P and K respectively on the performance. The batch size is PK, where P is number of different persons and K is number of different samples per person in each batch. Figure 4.7, and Figure 4.8 show the same information of Tables 4.17 and 4.18 and Tables 4.19 and 4.20, respectively.

Method	Backbone	Rank-1	mAP
PN-GAN [108]	ResNet	89.4	72.6
FD-GAN [109]	ResNet	90.5	77.7
Part-aligned [110]	GoogleNet	91.7	79.6
Mancs [112]	ResNet	93.1	82.3
SGGNN [113]	ResNet	92.3	82.8
DeepCRF [114]	ResNet	93.5	81.6
PCB [115]	ResNet	93.8	81.6
AA-Net [12]	ResNet	93.9	83.4
IA-Net [116]	ResNet	94.4	83.1
SphereReID [78]	ResNet	94.4	83.6
CAMA [69]	ResNet	94.7	84.5
DG-Net [68]	ResNet	94.8	86.0
AM0BH (Our)	ResNet	94.6$_{\pm 0.21}$	85.9 $_{\pm 0.28}$
AM0BH$_{\text{Attr}}$ (Our)	ResNet	94.9 $_{\pm 0.13}$	86.3 $_{\pm 0.10}$

Table 4.13: Comparison with the-sate-of-the-art methods on the Market-1501 datasets.

Method	Backbone	Rank-1	mAP
PN-GAN [108]	ResNet	73.6	53.2
FD-GAN [109]	ResNet	80.0	65.4
SGGNN [113]	ResNet	81.1	68.2
PCB [115]	ResNet	83.3	69.2
SphereReID [78]	ResNet	83.9	68.5
Part-aligned [110]	GoogleNet	84.4	69.3
DeepCRF [114]	ResNet	84.9	69.5
Mancs [112]	ResNet	84.9	71.8
CAMA [69]	ResNet	85.8	72.9
DG-Net [68]	ResNet	86.6	74.8
IA-Net [116]	ResNet	87.1	73.4
AA-Net [89]	ResNet	87.7	74.3
AM0BH (Our)	ResNet	89.2 $\pm$0.40	76.7$\pm$0.26
AM0BH$_{\text{Attr}}$ (Our)	ResNet	89.3 $\pm$0.19	77.4 $\pm$0.14

Table 4.14: Comparison with the-sate-of-the-art methods on the DukeMTMC-reID datasets.

Q	Rank-1	Rank-5	Rank-10	mAP
4	94.70 $\pm$ 0.27	98.32 $\pm$ 0.17	99.02 $\pm$ 0.12	86.00 $\pm$ 0.16
8	94.60 $\pm$ 0.20	98.32 $\pm$ 0.10	99.02 $\pm$ 0.05	85.92 $\pm$ 0.05
16	94.89 $\pm$ 0.13	98.26 $\pm$ 0.15	98.98 $\pm$ 0.12	86.26 $\pm$ 0.10
24	94.40 $\pm$ 0.28	98.35 $\pm$ 0.07	99.00 $\pm$ 0.00	85.80 $\pm$ 0.14
32	94.47 $\pm$ 0.21	98.20 $\pm$ 0.08	98.95 $\pm$ 0.06	85.80 $\pm$ 0.18

Table 4.15: Shows the effect of Q on the performance of $AM0BH_{Attr}$ on Market-1501. Q is the size of embedding features fed to each attribute classifier.

Q	Rank-1	Rank-5	Rank-10	mAP
4	88.55 ± 0.39	94.52 ± 0.22	96.10 ± 0.22	76.95 ± 0.13
8	88.82 ± 0.44	94.72± 0.24	96.14 ± 0.14	76.98 ± 0.15
16	89.29 ± 0.19	94.72 ± 0.24	96.17 ± 0.16	77.35 ± 0.14
24	88.85 ± 0.49	94.70 ± 0.14	96.25 ± 0.35	77.15 ± 0.21
32	88.97 ± 0.05	94.57 ± 0.15	96.18 ± 0.15	77.28 ± 0.17

Table 4.16: Shows the effect of Q on the performance of $AM0BH_{Attr}$ on DukeMTMC-reID. Q is the size of embedding features fed to each attribute classifier.

P	Rank-1	Rank-5	Rank-10	mAP
4	94.89 ±0.13	98.26 ±0.15	98.98 ±0.12	86.26 ±0.12
6	94.50 ±0.19	98.24 ±0.19	99.04 ±0.05	86.20 ±0.12
8	94.14 ±0.29	98.10 ±0.14	98.92 ±0.11	85.54 ±0.17
10	93.94 ±0.27	98.00 ±0.07	98.92 ±0.04	84.78 ±0.16

Table 4.17: Performance of AM0BH$_{\textbf{Attr}}$ on Market-1501 using different values of P. K is fixed to 8.

P	Rank-1	Rank-5	Rank-10	mAP
4	89.29 ± 0.19	94.72 ± 0.24	96.17 ± 0.16	77.35 ± 0.14
6	88.42 ± 0.54	94.60 ± 0.14	96.04 ± 0.13	76.66 ± 0.17
8	87.98 ± 0.16	94.50 ± 0.16	96.14 ± 0.17	76.14 ± 0.21
10	87.74 ± 0.62	94.50 ± 0.19	96.18 ± 0.23	75.68 ± 0.22

Table 4.18: Performance of AM0BH$_{\textbf{Attr}}$ on DukeMTMC-reID using different values of P. K is fixed to 8.

K	Rank-1	Rank-5	Rank-10	mAP
4	92.98 ±0.12	97.68 ±0.22	98.65 ±0.10	82.68 ±0.15
6	94.22 ±0.28	98.02 ±0.04	98.94 ±0.09	84.80 ±0.26
8	94.89 ±0.13	98.26 ±0.15	98.98 ±0.12	86.26 ±0.10
10	94.76 ±0.26	98.02 ±0.05	98.94 ±0.09	84.80 ±0.26
12	94.84 ±0.13	98.24 ±0.17	99.00 ±0.19	86.42 ±0.24
14	94.58 ±0.16	98.34 ±0.09	98.98 ±0.08	86.46 ±0.22

Table 4.19: Performance of AM0BH$_{\textbf{Attr}}$ on Market-1501 using different values of K. P is fixed to 4.

K	Rank-1	Rank-5	Rank-10	mAP
4	86.90 ± 0.22	93.90 ± 0.35	95.78 ± 0.17	75.02 ± 0.30
6	88.66 ± 0.26	94.74 ± 0.24	96.18 ± 0.19	76.80 ± 0.14
8	89.29 ± 0.19	94.72 ± 0.24	96.17 ± 0.16	77.35 ± 0.14
10	88.66 ± 0.26	94.74 ± 0.24	96.18 ± 0.19	76.80 ± 0.14
12	89.02 ± 0.37	94.58 ± 0.08	96.30 ± 0.14	77.18 ± 0.28
14	88.76 ± 0.49	94.72 ± 0.35	96.14 ± 0.23	77.10 ± 0.12

Table 4.20: Performance of AM0BH$_{\textbf{Attr}}$ on DukeMTMC-reID using different values of K. P is fixed to 4.

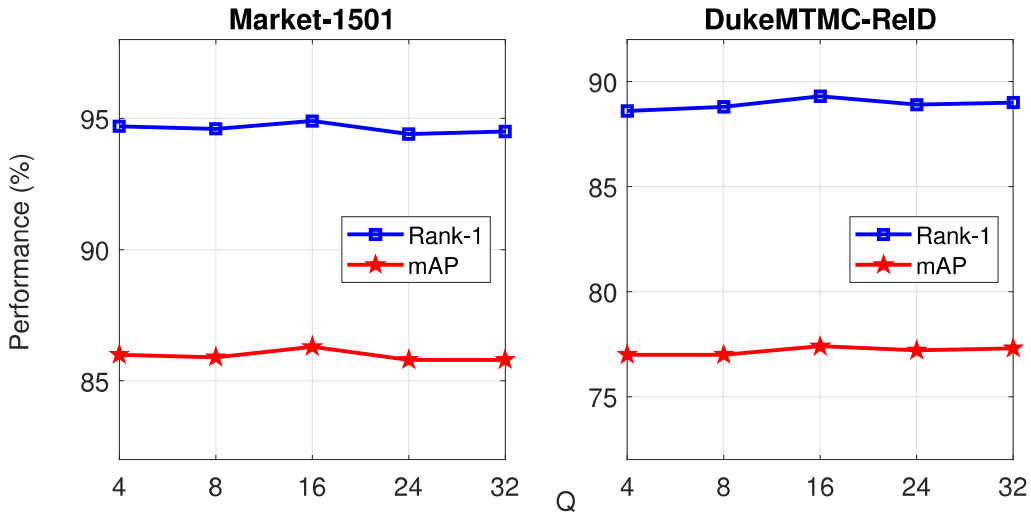


Figure 4.6: **Ablation study**. Shows the effect of Q on the performance of AM0BH$_{\textbf{Attr}}$. Q is the size of embedding features fed to each attribute classifier.

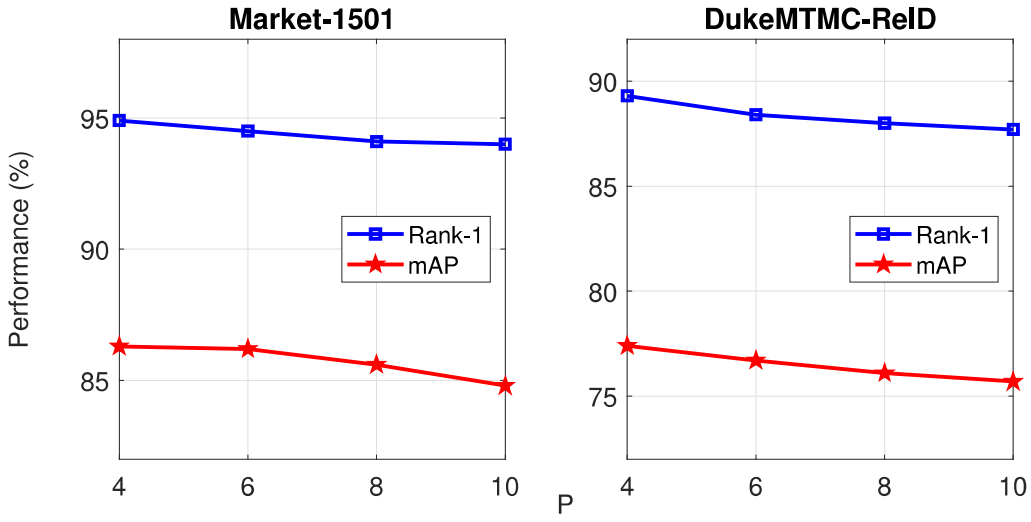


Figure 4.7: **Ablation study**. Shows the effect of P on the performance of AM0BH$_{\textbf{Attr}}$. K is fixed to 8.

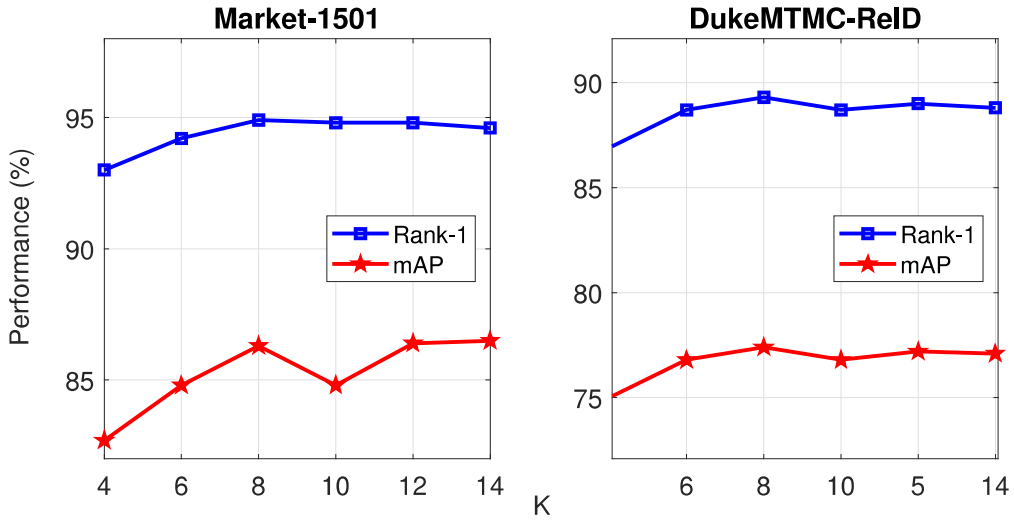


Figure 4.8: **Ablation study**. Shows the effect of K on the performance of AM0BH$_{\textbf{Attr}}$. P is fixed to 4.

4.5 Learning Pose-Invariant Embedding

Large variations introduced by human pose is a serious problem for any robust person re-identification model. Pose variations lead to various person appearance depending the on view point of the camera which make person body parts position non predictable in the bounding box. Having pose landmarks, some person re-identification methods tackle this problem either by generating person images or aligning body parts. We will incorporate those pose landmarks via deep learning model in order to overcome the problem of pose variation and learn pose-invariant embedding. Our proposed model capable of disentangling the embedding features into identification related features and pose unrelated features.

4.5.1 Proposed Approach

We further push the training of the embedding to become more invariant to the pose variations by incorporating the pose landmarks. Human pose estimation models aim to localize

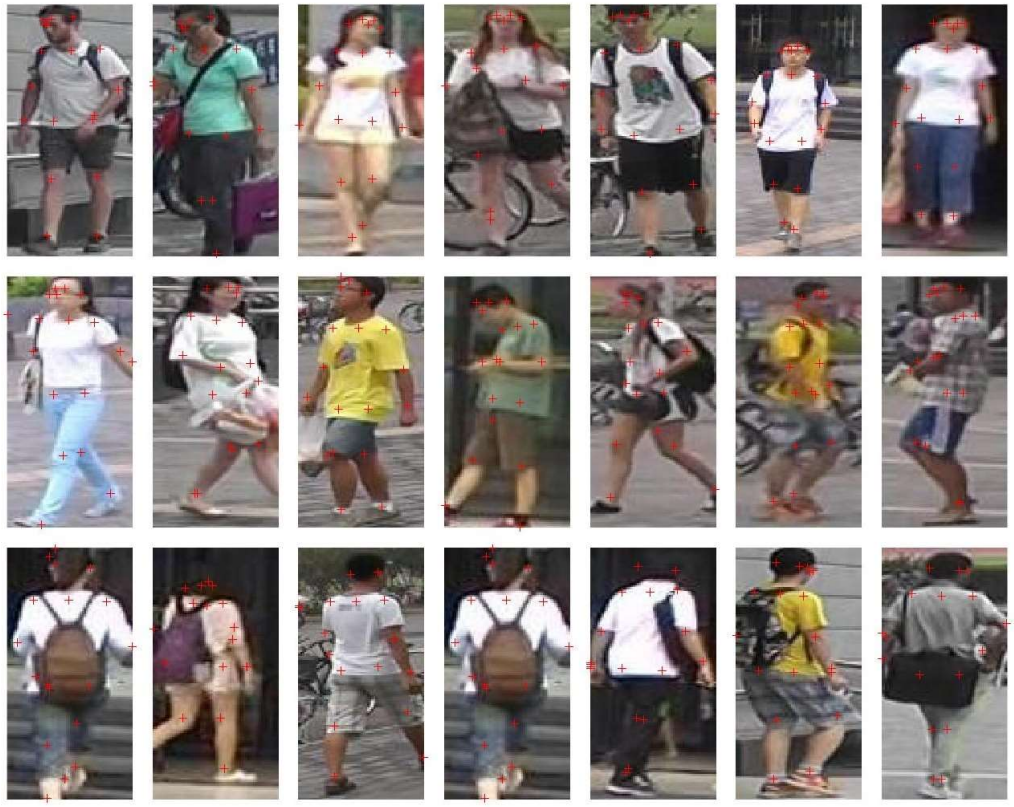


Figure 4.9: Example of human body parts estimations.

human body parts such as eyes, nose, ears, shoulders, elbows, wrists, hips, knees and ankles. Figure 4.9 shows an example of human body parts estimations. Using these pose landmarks, we train our proposed model to disentangle the embedding $f_\theta(I)$ into identification related features $f_{\theta_{id}}(I)$ and pose unrelated features $f_{\theta_p}(I)$. Also in this case we follow the strategy of normalizing the input and weights to reduce intra-class variability and maximize the correct prediction of identities.

To integrate the pose into person re-identification, we generate negative images have the same pose with training images. First. we generate 18 landmarks for each person image using a pose estimation model [141]. Then, we use an image generator model [142] to generate these negative samples. Therefore, triplets will become I_a, I_p and I_{Gn} where I_a is anchor sample, I_p is the hard positive sample but with different pose, while I_{Gn} is the hard

negative generated sample but share the same pose with the anchor sample. The intent is to train the model not just to classify identities but also to be invariant to pose variations. Thus, the embedding is divided into two parts $f_{\theta_{id}}$ and f_{θ_p} , where the first one is reserved for the identity classifier, while the other one is for pose. Therefore the classification loss become:

$$\mathcal{L}_{AM_{id}}(\theta_{id}, W_{id}) = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cos(\alpha_{id_{y_i}}+m)}}{e^{s\cos(\alpha_{id_{y_i}}+m)}+\sum_{j\neq y_i}e^{s\cos\alpha_{id_j}}} \tag{4.9}$$

Two triplet losses are used. The first triplet loss is to train the function $f_{\theta_{id}}$ that maps images of the same identity (I_a and I_p) as close as possible, effectively minimizing the intra-class variability of the embeddings, while images of different identities (I_a and I_{Gn}) are mapped far away, creating a large inter-class discrepancy. While the second triplet loss is to train functions f_{θ_p} that maps images of the same pose (I_a and I_{Gn}) as close as possible, effectively minimizing the intra-class variability of the embeddings, while images of different poses (I_a and I_p) are mapped far away, creating a large inter-class discrepancy. Thus the embedding f_θ will be separated into identification related features and pose unrelated features. The two triplet losses are:

$$\begin{aligned}\mathcal{L}_{BH_{id}}(\theta_{id}) = \sum_{i=1}^{P}\sum_{a=1}^{K} [m &+ \max_{p} D(f_{\theta_{id}}(I_a^i), f_{\theta_{id}}(I_p^i)) \\ &- \min_{j,n,j\neq i} D(f_{\theta_{id}}(I_a^i), f_{\theta_{id}}(I_{Gn}^j))]_+\end{aligned} \tag{4.10}$$

$$\begin{aligned}\mathcal{L}_{BH_{p}}(\theta_{p}) = \sum_{i=1}^{P}\sum_{a=1}^{K} [m &+ \max_{p} D(f_{\theta_{p}}(I_a^i), f_{\theta_{p}}(I_{Gn}^i)) \\ &- \min_{j,n,j\neq i} D(f_{\theta_{p}}(I_a^i), f_{\theta_{p}}(I_n^j))]_+\end{aligned} \tag{4.11}$$

The final loss is set by combining the identity classification loss (4.9) with the metric losses (4.10) and (4.11). This leads to the full re-identification training model, given by:

$$\mathcal{L}_{AMBH_{P.I.}}(\theta, W) = \mathcal{L}_{AM}(\theta_{id}, W_{id}) + \lambda \mathcal{L}_{BH_{id}}(\theta_{id}) + \gamma \mathcal{L}_{BH_p}(\theta_p) \tag{4.12}$$

4.5.2 Network Architecture

We use a pretrained ResNet-50 [70] as backbone network. We simply discard the fully connected layer, and we change the stride of the last convolutional stage from 2 to 1. We then add a global average pooling (GAP) layer and a batch normalization layer (BN). At this point weight normalization and ℓ^2 feature normalization is applied before entering the identity classification loss (4.9), while no normalization is needed for the batch hard triplet losses components (4.10) and (4.11). Figure 4.10 shows a graphical description of the architecture for integrating pose in person re-identification.

During testing, unless otherwise specified, we perform all the experiments with the ℓ^2 normalized embedding $f_{\theta_{id}}/\|f_{\theta_{id}}\|$, and re-identification is done via cosine similarity.

4.5.3 Experiments

We evaluate our proposed model on Market-1501 [84] datasets. We use OpenPose [141] to extract 18 pose landmarks and then use PoseStylizer [142] to generate hard negative images. We implemented our approach with PyTorch [104], and for the backbone network we use ResNet-50 [70] that is pre-trained on ImageNet [106]. The fully connected layer in ResNet-50 is discarded and the stride of the last stage of ResNet-50 backbone is changed from 2 to 1. We add a global average pooling and batch normalization after the last convolutional layer of ResNet-50. The dimensionality of the embedding features is 2048.

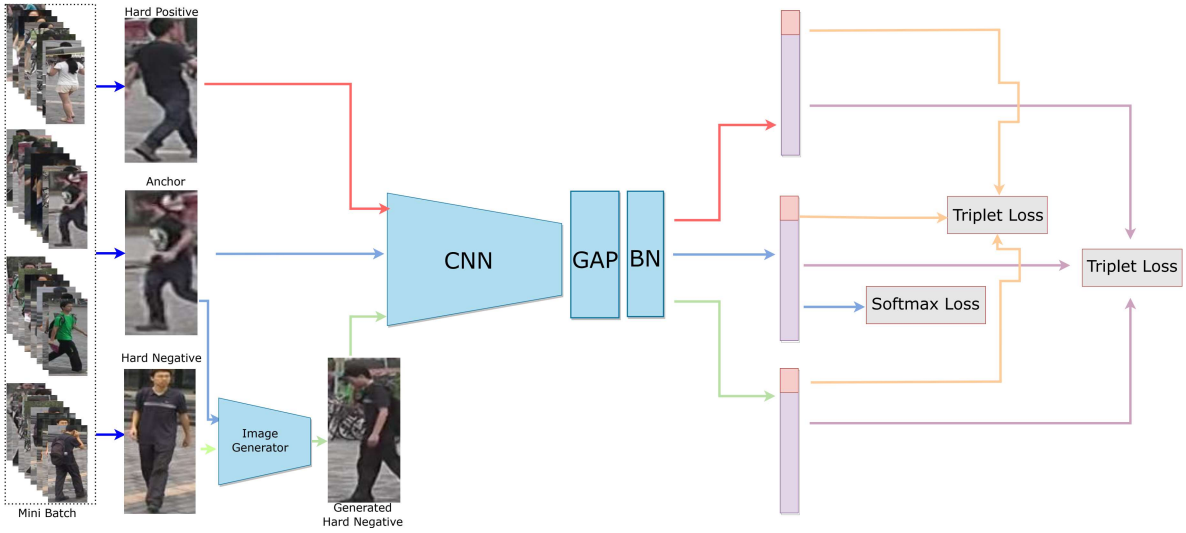


Figure 4.10: **Architecture.** Simple graphical description of the joint optimization of the pose invariant re-identification loss, based on a multi-class classification (4.9) (for the identities) and a metric learning losses (4.10) and (4.11). The former (4.9) normalize the features and the layer weights, and the latters (4.10) and (4.11) do not.

Each mini-batch consists of $P \cdot K$ images where P represents the number of different identities and K represents number of images per identity. Batch size is set to 32 where P is set to 4 and K is set to 8. For Market-1501 datasets, input image size is 256×128.

For data augmentation, training images are randomly flipped and erased. The model trained using Adam optimizer with default hyper parameters for 165 epochs. The learning rate is linearly increased from 10^{-5} to 10^{-3} for first 20 epochs to help the network to bootstrap. Then learning rate sets to 10^{-3} after epoch 20 and it decreases to 10^{-4}, 10^{-5}, 10^{-6} and 10^{-7} after epochs 90, 130, 150 and 160 respectively. For better features initialization, the model trained on (4.5) for the first 60 epochs.

The 2048 embedding features are divided into $f_{\theta_{id}}$ and f_{θ_p} . f_{θ_p} has size of 128 which is the input to the pose triplet loss. The rest of the embedding features, $f_{\theta_{id}}$, is the input to the identity classifier and the identity triplet loss. In (4.12) we set λ to 0.21 and γ to 0.21.

Method	Backbone	Rank-1	mAP
PN-GAN [108]	ResNet	89.4	72.6
FD-GAN [109]	ResNet	90.5	77.7
Part-aligned [110]	GoogleNet	91.7	79.6
Mancs [112]	ResNet	93.1	82.3
SGGNN [113]	ResNet	92.3	82.8
DeepCRF [114]	ResNet	93.5	81.6
PCB [115]	ResNet	93.8	81.6
AA-Net [12]	ResNet	93.9	83.4
IA-Net [116]	ResNet	94.4	83.1
SphereReID [78]	ResNet	94.4	83.6
CAMA [69]	ResNet	94.7	84.5
DG-Net [68]	ResNet	94.8	86.0
AM0BH (Our)	ResNet	94.6$_{\pm 0.21}$	85.9 $_{\pm 0.28}$
AM0BH$_{\text{P.I.}}$ (Our)	ResNet	94.9 $_{\pm 0.04}$	86.4 $_{\pm 0.24}$

Table 4.21: Comparison with the-sate-of-the-art methods on the Market-1501 datasets.

We compare our proposed approach with the state-of-the-art methods on Market-1501 datasets. Table 4.21 shows the results on Market-1501, our approach outperforms the state-of-the-art and $AM0BH_{P.I.}$ further improves the performance over $AM0BH$ by **0.3%** and **0.4%** for rank-1 and mAP respectively.

Also we study the influence of f_{θ_p} size on the performance of $AM0BH_{P.I.}$. Table 4.22 shows the effect of the f_{θ_p} size, the embedding features fed to the pose triplet loss. The same information is plotted in Figure 4.11.

f_{θ_p} size	Rank-1	Rank-5	Rank-10	mAP
64	94.60 ± 0.14	98.25 ± 0.07	99.00 ± 0.14	86.00 ± 0.14
96	94.45 ± 0.22	98.03 ± 0.15	98.83 ± 0.05	85.75 ± 0.24
128	94.87 ± 0.04	98.30 ± 0.08	99.03 ± 0.08	86.36 ± 0.24
160	94.53 ± 0.22	98.03 ± 0.15	98.03 ± 0.05	85.75 ± 0.24
192	94.65 ± 0.07	98.10 ± 0.14	98.90 ± 0.14	86.05 ± 0.21
224	94.35 ± 0.07	98.20 ± 0.00	98.80 ± 0.00	86.05 ± 0.21
256	94.75 ± 0.07	98.35 ± 0.07	99.1 ± 0.00	86.25 ± 0.07

Table 4.22: Shows the effect of f_{θ_p} size on the performance of AM0BH$_{\mathbf{P.I.}}$. f_{θ_p} is the size of embedding features fed to the pose triplet loss.

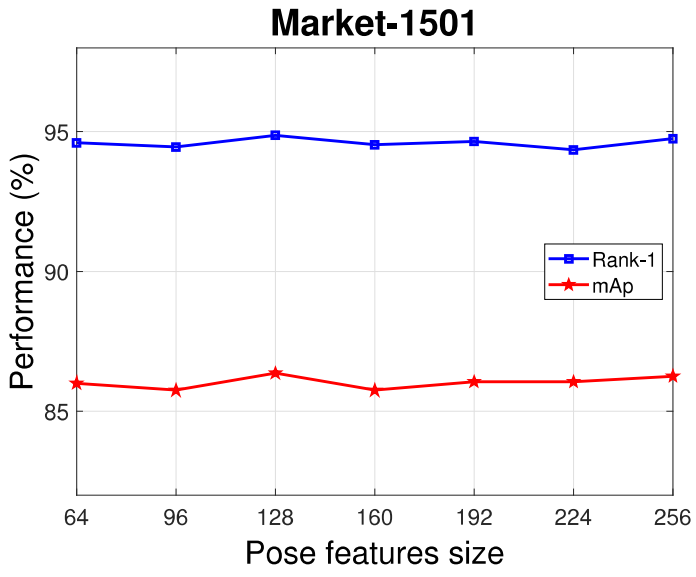


Figure 4.11: **Ablation study**. Shows the effect of f_{θ_p} size on the performance of AM0BH$_{\mathbf{P.I.}}$. f_{θ_p} is the size of embedding features fed to the pose triplet loss.

4.6 Conclusion

we improve person re-identification by combining the softmax with the additive angular margin and the triplet loss, to further improve the learning of a feature embedding. The intent is for the triplet loss to help the softmax further decrease intra-class variation and increase inter-class distance by letting the triplet loss be the proxy for the gradient supervision that the embedding normalization has restricted, and we specify under what conditions this may happen. We also observe that the same strategy used to form the batch of triplets can be used in tandem with the softmax loss to prevent issues due to dataset imbalance, which are common in person re-identification. We further extend this part to improve the performance of person re-identification model via two approaches. In the first approach, we improve the learning of a holistic representation in the form of a feature embedding for person re-identification, by combining the softmax and the triplet loss, and by learning multiple discriminative tasks, including not only the identity classification but also the attribute prediction. Adding the discriminative task of predicting attributes improves invariance to nuisance factors. While in the second approach, we incorporate pose information into person re-identification model to overcome the problem of pose variation and to learn pose-invariant embedding.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this dissertation, we explored models for learning representations for human identification that aim at reducing the effects of nuisance factors of variation, such as pose, illumination, viewpoint, background, and sensor noise. The problem is challenging because distinct images of the same person may look very different leading to a high intra-class variance which make human identification task still prone to high mismatch rates.

In chapter two, we present a low-rank modeling framework which can be applied to face as well as whole-body images. This framework can capture all the descriptive information of a human image, and emphases on learning a representation that is invariant to nuisance factors. The framework is not only learned different invariant representations for different identities, but such representations promote a uniform inter-class separation. Another advantage of this framework is a fast procedure for computing and comparing invariant components for recognition and re-identification.

In chapter three, we address the human identification problem with probe and gallery data acquired in unconstrained scenarios by developing a robust approach that fuses representations of multiple biometrics for human identification. We extended our invariant

component matching method based on independent components and distance matching to develop a new approach for jointly exploiting face and whole-body appearance for human identification. Our fusion approach based on the Information Bottleneck method.

In chapter four, we improve person re-identification by combining the softmax with the additive angular margin and the triplet loss, to further improve the learning of a feature embedding. The intent is for the triplet loss to help the softmax further decrease intra-class variation and increase inter-class distance by letting the triplet loss be the proxy for the gradient supervision that the embedding normalization has restricted, and we specify under what conditions this may happen. We also observe that the same strategy used to form the batch of triplets can be used in tandem with the softmax loss to prevent issues due to dataset imbalance, which are common in person re-identification. We further extend this part to improve the performance of person re-identification model via two approaches. In the first approach, we improve the learning of a holistic representation in the form of a feature embedding for person re-identification, by combining the softmax and the triplet loss, and by learning multiple discriminative tasks, including not only the identity classification but also the attribute prediction. Adding the discriminative task of predicting attributes improves invariance to nuisance factors. While in the second approach, we incorporate pose information into person re-identification model to overcome the problem of pose variation and to learn pose-invariant embedding.

5.2 Future work

In Chapter 3, we introduced a learning framework for fusing representations of multiple biometrics for human identification. We focus on the face modality and clothing appearance, and develop a representation fusion approach based on the Information Bottleneck method. For future work, we would like to investigate deep learning for fusing these multiple biometrics for human identification on large-scale datasets.

In Chapter 4, we improved person re-identification by decreasing the effects of nuisance factors via deep learning. We design and combine improved versions of classification and distance metric losses. Then, we further extended the proposed framework via two approaches. First, we added the task of prediction attributes. Second, we incorporated pose information into person re-identification model. For future work, we would like to investigate combining these two approaches, i.e. attribute prediction and pose information, to learn more robust representation.

The majority of person re-identification models use supervised learning on small-scale labeled datasets. Since the target domain might be very different from the small-scale training dataset, deploying these trained models to a real environment will reduce the model performance. In addition to that, labeling large-scale surveillance videos in order to facilitate supervised learning is unfeasible. Therefore, optimizing an adapted person re-identification model for real environment scenario is crucial for improving person re-identification approaches. It is challenging because it is easy to acquire large-scale unlabeled target dataset but it is hard train a deep learning model without annotated labels. Having labelled source data and unlabeled target data, the unsupervised domain adaptation (UDA) for person re-identification tries to learn a discriminative representation for the unlabeled target data.

Some feature representation learning approaches [131, 132] have been developed to learn discriminative features by exploring invariant cross-view person information in order to acquire domain-invariant features across different views. [133] Proposed an unsupervised camera-aware domain adaptation approach to decrease the discrepancy among different domains. [134] presented a patch-based unsupervised learning framework. To learn a generalizable person Re-ID model, [135] proposed a domain-invariant mapping network. [136] proposed an instance-guided context rendering to enable supervised learning in target domain by transferring source person identities into target contexts. [137] recommend learning domain-invariant representation through pose-guided image translation.

For future work. we interested to propose a deep learning unsupervised model for person

re-identification capable to handle cross domain variation between labeled train domain and unlabeled target domain.

References

[1] Farzad Siyahjani, Ranya Almohsen, Sinan Sabri, and Gianfranco Doretto. A supervised lowrank method for learning invariant subspace. In ICCV, 2015.

[2] Zheng, Liang, Yi Yang, and Alexander G. Hauptmann. Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984, 2016.

[3] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. IEEE TPAMI, 31(2):210–227, 2009.

[4] E. Cand´es, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? Journal ACM, 58(3), 2011.

[5] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Discriminative learned dictionaries for local image analysis. In IEEE CVPR, pages 1–8, 2008.

[6] E. Elhamifar and R. Vidal. Robust classification using structured sparse representation. In IEEE CVPR, pages 1873–1879, 2011.

[7] B. Cheng, G. Liu, J. Wang, Z. Huang, and S. Yan. Multi-task lowrank affinity pursuit for image segmentation. In IEEE ICCV, pages 2439–2446, 2011.

[8] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja. Low-rank sparse learning for robust visual tracking. In IEEE ECCV, pages 470–484, 2012.

[9] J. Lee, B. Shi, Y. Matsushita, I. Kweon, and K. Ikeuchi. Radiometric calibration by transform invariant low-rank structure. In IEEE CVPR, pages 2337–2344, 2011.

[10] Y. Zhang, Z. Jiang, and L. S. Davis. Learning structured low-rank representations for image classification. In IEEE CVPR, pages 676– 683, 2013.

[11] C.-F. Chen, C.-P. Wei, and Y.-C. Wang. Low-rank matrix recovery with structural incoherence for robust face recognition. In IEEE CVPR, pages 2618–2625, 2012.

[12] Q. Zhang and B. Li. Mining discriminative components with lowrank and sparsity constraints for face recognition. In KDD, pages 1469–1477, 2012.

[13] R. Vidal and P. Favaro. Low rank subspace clustering (LRSC). Pattern Recognition Letters, 43:47–61, 2014.

[14] Z. Lin, M. Chen, and Y. Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. arXiv preprint arXiv:1009.5055, 2010.

[15] T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka. Distance-based image classification: Generalizing to new classes at near-zero cost. In IEEE TPAMI, 35(11):2624–2637, 2013.

[16] S. Hoi, W. Liu, M. Lyu, and W.-Y. Ma. Learning distance metrics with contextual constraints for image retrieval. In IEEE CVPR, pages 2072–2078, 2006.

[17] W.-S. Zheng, S. Gong, and T. Xiang. Reidentification by relative distance comparison. In IEEE TPAMI, 35(3):653–668, 2013.

[18] M. Hirzer, P. M. Roth, M. Kostinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In IEEE ECCV, pages 780–793, 2012.

[19] Z. Cui, W. Li, D. Xu, S. Shan, and X. Chen. Fusing robust face region descriptors via multiple metric learning for face recognition in the wild. In IEEE CVPR, pages 3554–3561, 2013.

[20] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In IEEE ICCV, pages 498–505, 2009.

[21] A. Mignon and F. Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In IEEE CVPR, pages 2666–2672, 2012.

[22] H. V. Nguyen and L. Bai. Cosine similarity metric learning for face verification. In ACCV, pages 709–720. Springe, 2011.

[23] J. Hu, J. Lu, and Y.-P. Tan. Discriminative deep metric learning for face verification in the wild. In IEEE CVPR, pages 1875–1882, 2014.

[24] M. Schultz and T. Joachims. Learning a distance metric from relative comparisons. In NIPS, 2004.

[25] M. Law, N. Thome, and M. Cord. Quadruplet-wise image similarity learning. In IEEE ICCV, pages 249–256, 2013.

[26] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information theoretic metric learning. In ICML, pages 209–216, 2007.

[27] M. Kostinger, M. Hirzer, P. Wohlhart, P. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In IEEE CVPR, pages 2288–2295, 2012.

[28] X. Cai, C. Wang, B. Xiao, X. Chen, and J. Zhou. Deep nonlinear metric learning with independent subspace analysis for face verification. In ACM Multimedia, pages 749–752, 2012.

[29] I. W. Tsang, J. T. Kwok, C. Bay, and H. Kong. Distance metric learning with kernels. In ICANN, pages 126–129. Citeseer, 2003.

[30] D.-Y. Yeung and H. Chang. A kernel approach for semi supervised metric learning. In IEEE Trans. Neural Networks, 18(1):141–149, 2007.

[31] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In JMLR, 10:207–244, 2009.

[32]] J. Wang, A. Woznica, and A. Kalousis. Parametric local metric learning for nearest neighbor classification. In NIPS, pages 1610–1618, 2012.

[33] A. M. Martinez and R. Benavente. The AR face database. Technical Report 24, CVC, 1998.

[34] W. Deng, J. Hu, and J. Guo. In defense of sparsity based face recognition. In IEEE CVPR, pages 399–406, 2013.

[35] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: illumination cone models for face recognition under variable lighting and pose. In IEEE TPAMI, 23(6):643–660, 2001.

[36] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In JMLR, 10:207–244, 2009.

[37] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In IEEE ECCV, 2008.

[38] G. Huang, M. Mattar, H. Lee, and E. G. Learned-Miller. Learning to align from scratch. In NIPS, pages 764-772, 2012.

[39] M. Yang and L. Zhang. Gabor feature based sparse representation for face recognition with gabor occlusion dictionary. In IEEE ECCV, pages 448–461, 2010.

[40] M. Yang, L. Zhang, J. Yang, and D. Zhang. Robust sparse coding for face recognition. In IEEE CVPR, pages 625–632, 2011.

[41] R. He, W. S. Zheng, and B. G. Hu. Maximum correntropy criterion for robust face recognition. In IEEE TPAMI, 33(8):1561–1576, 2011.

[42] Y. Meng, D. Zhang, Y. Jian, and D. Zhang. Robust sparse coding for face recognition. In IEEE CVPR, pages 625–632, 2011.

[43] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma. Toward a practical face recognition system: Robust alignment and illumination by sparse representation. In IEEE TPAMI, 34(2):372–386, 2012.

[44] W. Deng, J. Hu, and J. Guo. Extended SRC: Under sampled face recognition via intraclass variant dictionary. In IEEE TPAMI,34(9):1864–1870, 2012.

[45] T. Hastie and R. Tibshirani. Discriminant adaptive nearest neighbor classification. In IEEE TPAMI, 18(6):607–616, 1996.

[46] M. Bilenko, S. Basu, and R. J. Mooney. Integrating constraints and metric learning in semi-supervised clustering. In ICML, pages 81–88, 2004.

[47] A. Frome, Y. Singer, and J. Malik. Image retrieval and classification using local distance functions. In NIPS, volume 19, pages 417–424, 2007.

[48] Z. Lin, R. Liu, and Z. Su. Linearized alternating direction method with adaptive penalty for low-rank representation. arXiv, 2011.

[49] L. Ma, C. Wang, B. Xiao, and W. Zhou. Sparse representation for face recognition based on discriminative low-rank dictionary learning. In IEEE CVPR, pages 2586–2593, 2012.

[50] L. Zhuang, H. Gao, Z. Lin, Y. Ma, X. Zhang, and N. Yu. Non- negative low rank and sparse graph for semi-supervised learning. In IEEE CVPR, pages 2328–2335, 2012.

[51] J. Cai, E. Candés, and Z. Shen. A singular value thresholding algorithm for matrix completion. In SIAM Journal on Optimization, 20(4):1956–1982, 2010.

[52] J. Bohne, Y. Ying, S. Gentric, and M. Pontil. Large margin local metric learning. In IEEE ECCV, pages 679–694, 2014.

[53] S. Liao, Y. Hu, X. Zhu, S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In IEEE CVPR, pages 2197–2206, 2015.

[54] S. Pedagadi, J. Orwell, S. Velastin, B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In IEEE CVPR, pages 3318–3325, 2013.

[55] E. Ahmed, M. Jones, T. K. Marks. An improved deep learning architecture for person re-identification. In IEEE CVPR, pages 3908–3916, 2015.

[56] F. Xiong, M. Gou, O. Camps, M. Sznaier. Person re-identification using kernel-based metric learning methods. In IEEE ECCV, Springer, pages 1–16, 2014.

[57] L. Zhang, T. Xiang, S. Gong. Learning a discriminative null space for person re-identification. In IEEE CVPR, pages 1239–1248, 2016.

[58] M. Hirzer, P. M. Roth, M. Köstinger, H. Bischof. Relaxed pairwise learned metric for person re-identification. In IEEE ECCV, Springer, pages 780–793, 2012.

[59] W.-S. Zheng, S. Gong, T. Xiang Reidentification by relative distance comparison. In IEEE transactions on pattern analysis and machine intelligence, pages 653–668, 2012.

[60] M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In IEEE CVPR, pages 2360–2367, 2010.

[61] C. Liu, S. Gong, C. C. Loy, X. Lin. Person re-identification: What features are important?. In IEEE ECCV, Springer, pages 391–401, 2012.

[62] Z. Shi, T. M. Hospedales, T. Xiang. Transferring a semantic representation for person re-identification and search. In IEEE CVPR, pages 4184–4193, 2015.

[63] L. Zheng, S. Wang, L. Tian, F. He, Z. Liu, Q. Tian. Query-adaptive late fusion for image search and person re-identification. In IEEE CVPR, pages 1741–1750, 2015.

[64] B. Ma, Y. Su, F. Jurie. Bicov: a novel image representation for person re-identification and face verification. In British Machine Vision Conference, 2012.

[65] Wei-Shi Zheng, Shaogang Gong, Tao Xiang. Person re-identification by probabilistic relative distance comparison. In IEEE CVPR, Vol. 0, pages 649–656, 2011.

[66] Z. Zheng, L. Zheng, Y. Yang. A discriminatively learned CNN embedding for person re-identification. arXiv:1611.05666, 2016.

[67] W. Li, X. Zhu, S. Gong. Person re-identification by deep joint learning of multi-loss classification. In Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI'17, AAAI Press, pages 2194–2200, 2017.

[68] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, J. Kautz. Joint discriminative and generative learning for person re-dentification. In IEEE CVPR, pages 2138–2147, 2019.

[69] W. Yang, H. Huang, Z. Zhang, X. Chen, K. Huang, S. Zhang. Towards rich feature discovery with class activation maps augmentation for person re-identification. In IEEE CVPR, pages 1389–1398, 2019.

[70] K. He, X. Zhang, S. Ren, J. Sun. Deep residual learning for image recognition. In IEEE CVPR, pages 770–778, 2016.

[71] S. Bak, P. Carr. One-shot metric learning for person re-identification. In IEEE CVPR, pages 2990–2999, 2017.

[72] F. Schroff, D. Kalenichenko, J. Philbin. Facenet: A unified embedding for face recognition and clustering. In IEEE CVPR, pages 815–823, 2015.

[73] S. Ding, L. Lin, G. Wang, H. Chao. Deep feature learning with relative distance comparison for person re-identification. Pattern Recognition. pages 2993–3003, 2015.

[74] D. Cheng, Y. Gong, S. Zhou, J. Wang, N. Zheng. Person re-identification by Multi-Channel Parts-Based CNN with improved triplet loss function. In IEEE CVPR, pages 1335–1344, 2016.

[75] A. Hermans, L. Beyer, B. Leibe. In defense of the triplet loss for person Re-Identification. arXiv:1703.07737, 2017.

[76] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, W. Liu. CosFace: Large margin cosine loss for deep face recognition. In IEEE CVPR, 2018.

[77] J. Deng, J. Guo, N. Xue, S. Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In IEEE CVPR, 2019.

[78] X. Fan, W. Jiang, H. Luo, M. Fei. Spherereid: Deep hypersphere manifold embedding for person re-identification. In Journal of Visual Communication and Image Representation pages 51–58, 2019.

[79] D. Li, X. Chen, Z. Zhang, K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In IEEE CVPR, pages 384–393, 2017.

[80] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, Q. Tian. Pose-driven deep convolutional model for person re-identification. In IEEE ICCV, pages 3960–3969, 2017.

[81] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In IEEE CVPR, pages 1077–1085, 2017.

[82] L. Zheng, Y. Huang, H. Lu, Y. Yang. Pose-invariant embedding for deep person re-identification. In IEEE Transactions on Image Processing, pages 4500–4509, 2019.

[83] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, Y. Yang. Improving person re-identification by attribute and identity learning. In Pattern Recognition, pages 151–161, 2019.

[84] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian. Scalable person re-identification: A benchmark. In IEEE ICCV, pages 1116–1124, 2015.

[85] Z. Zheng, L. Zheng, Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In IEEE ICCV, pages 3754–3762, 2017.

[86] A. Schumann, R. Stiefelhagen. Person re-identification by deep learning attribute-complementary information. In IEEE CVPR workshops, pages 20–28, 2017.

[87] G. Zhang, J. Xu. Person re-identification by mid-level attribute and part-based identity learning. In Asian Conference on Machine Learning, In PMLR, pages 220–231, 2018.

[88] J. Liu, Z.-J. Zha, H. Xie, Z. Xiong, Y. Zhang. Ca3net: Contextual-attentional attribute-appearance network for person re-identification. In Proceedings of the ACM international conference on Multimedia, pages 737–745, 2018.

[89] C.-P. Tay, S. Roy, K.-H. Yap. Aanet: Attribute attention network for person re-identifications. In IEEE CVPR, pages 7134–7143, 2019.

[90] P. Chikontwe, H. J. Lee. Deep multi-task network for learning person identity and attributes. In IEEE Access, pages 60801–60811, 2018.

[91] S. Li, H. Yu, R. Hu. Attributes-aided part detection and refinement for person re-identification. Pattern Recognition, 2020.

[92] R. Ranjan, C. D. Castillo, R. Chellappa. L2-constrained softmax loss for discriminative face verification. arXiv:1703.09507, 2017.

[93] F. Wang, W. Liu, H. Liu, J. Cheng. Additive margin softmax for face verification. arXiv:1801.05599, 2018.

[94] H. O. Song, Y. Xiang, S. Jegelka, S. Savarese. Deep metric learning via lifted structured feature embedding. In IEEE CVPR, pages 4004–4012, 2016.

[95] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In IEEE CVPR, pages 2360–2367, 2010.

[96] W.-S. Zheng, S. Gong, and T. Xiang. Reidentification by relative distance comparison. In IEEE Pattern Analysis and Machine Intelligence, pages 653–668, 2013.

[97] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, L. Song. SphereFace: Deep hypersphere embedding for face recognition. CVPR, 2017.

[98] K. Q. Weinberger, L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In Journal of machine learning research , 2009.

[99] F. Schroff, D. Kalenichenko, J. Philbin. Facenet: A unified embedding for face recognition and clustering. In IEEE CVPR, pages 815–823, 2015.

[100] F. Wang, X. Xiang, J. Cheng, A. L. Yuille. NormFace: L2 hypersphere embedding for face verification, arXiv:1704.06369, 2017.

[101] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, D. Ramanan. Object detection with discriminatively trained part-based models. In IEEE transactions on pattern analysis and machine intelligence, pages 1627–1645, 2009.

[102] L. Wei, S. Zhang, W. Gao, Q. Tian. Person transfer gan to bridge domain gap for person re-identification. In IEEE CVPR, pages 79–88, 2018.

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

[104] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer. Automatic differentiation in pytorch. 2017.

[105] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio. Generative adversarial nets. In Advances in neural information processing systems, pages 2672–2680, 2014.

[106] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In IEEE CVPR, pages 248–255, 2009.

[107] Z. Zhong, L. Zheng, D. Cao, S. Li. Re-ranking person re-identification with k-reciprocal encoding. In IEEE CVPR, pages 1318–1327, 2017.

[108] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, X. Xue. Pose-normalized image generation for person re-identification. In IEEE ECCV, pages 650–667, 2018.

[109] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang, et al. FD-GAN: Pose-guided feature distilling gan for robust person re-identification. In Advances in neural information processing systems, pages 1222–1233, 2018.

[110] Y. Suh, J. Wang, S. Tang, T. Mei, K. Mu Lee, Part-aligned bilinear representations for person re-identification, In IEEE ECCV, pages 402–419, 2018.

[111] L. Zhao, X. Li, Y. Zhuang, J. Wang, Deeply-learned part-aligned representations for person re-identification, In IEEE ICCV, pages 3219–3228, 2017.

[112] C. Wang, Q. Zhang, C. Huang, W. Liu, X. Wang, Mancs: A multi-task attentional network with curriculum sampling for person re-identification, In IEEE ECCV, pages 365–381, 2018.

[113] Y. Shen, H. Li, S. Yi, D. Chen, X. Wang, Person re-identification with deep similarity-guided graph neural network, In IEEE ECCV, pages 486–504, 2018.

[114] D. Chen, D. Xu, H. Li, N. Sebe, X. Wang, Group consistent similarity learning via deep crf for person re-identification, In IEEE CVPR, pages 8649–8658, 2018.

[115] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond part models: Person retrieval with refined part pooling and a strong convolutional baseline, In IEEE ECCV, pages 480–496, 2018.

[116] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, X. Chen, Interaction-and-aggregation network for person re-identification, In IEEE CVPR, pages 9317–9326, 2019.

[117] J. Miao, Y. Wu, P. Liu, Y. Ding, Y. Yang, Pose-guided feature alignment for occluded person re-identification, In IEEE ICCV, pages 542–551, 2019.

[118] T. Xiao, H. Li, W. Ouyang, X. Wang, Learning deep feature representations with domain guided dropout for person re-identification, In IEEE CVPR, pages 1249–1258, 2016.

[119] N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. arXiv, 2000.

[120] S. Motiian and G. Doretto. Information bottleneck domain adaptation with privileged information for visual recognition In IEEE ECCV, Springer, pages 630–647, 2016.

[121] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto. Information bottleneck learning using privileged information for visual recognition. In IEEE CVPR, pages 1496–1505, 2016.

[122] N. Slonim, N. Friedman, and N. Tishby. Multivariate information bottleneck. In Neural Computation, pages 1739–1789, 2006.

[123] T. M. Cover and J. A. Thomas. Elements of Information Theory. Wiley and Sons, Inc., 1991.

[124] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. H. Tu. Shape and appearance context modeling. In IEEE ICCV, pages 1–8, 2007.

[125] G. Doretto, T. Sebastian, P. Tu, and J. Rittscher. Appearance-based person reidentification in camera networks: problem overview and current approaches. In Journal of Ambient Intelligence and Humanized Computing, 2011.

[126] X. Zhu and D. Ramanan. Face detection, pose estimation,and landmark localization in the wild. In IEEE CVPR, 2012.

[127] Y. Nesterov. Smooth minimization of non-smooth functions. In Mathematical Programming, pages 127–152, 2005.

[128] T. Zhou, D. Tao, and X. Wu. NESVM: A fast gradient method for support vector machines. In ICDM, pages 679–688, 2010.

[129] D. Goldfarb, S. Ma, and K. Scheinberg. Fast alternating linearization methods for minimizing the sum of two convex functions. In Mathematical Programming, pages 349–382, 2013.

[130] C. J. Lin. Projected gradient methods for nonnegative matrix factorization. In Neural Computation, pages 2756–2779, 2007.

[131] H.-X. Yu, A. Wu, and W.-S. Zheng, Unsupervised person reidentification by deep asymmetric metric embedding. In IEEE transactions on pattern analysis and machine intelligence, vol. 42, no. 4, pages 956–973, 2020.

[132] Z. Liu, D. Wang, and H. Lu. Stepwise metric promotion for unsupervised video person re-identification. In IEEE ICCV, pages 2429–2438, 2017.

[133] L. Qi, L. Wang, J. Huo, L. Zhou, Y. Shi, and Y. Gao. A novel unsupervised camera-aware domain adaptation framework for person reidentification. In IEEE ICCV, pages 8080–8089, 2019.

[134] Q. Yang, H.-X. Yu, A. Wu, and W.-S. Zheng. Patch-based discriminative feature learning for unsupervised person re-identification. In IEEE CVPR, pages 3633–3642, 2019.

[135] J. Song, Y. Yang, Y.-Z. Song, T. Xiang, and T. M. Hospedales. Generalizable person re-identification by domain-invariant mapping network. In IEEE CVPR, pages 719–728, 2019.

[136] Chen, Y., Zhu, X., Gong, S.: Instance-guided context rendering for cross-domain person re-identification. In IEEE ICCV, 2019.

[137] Li, Y.J., Lin, C.S., Lin, Y.B., Wang, Y.C., Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation. In IEEE ICCV, 2019.

[138] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool. Pose guided person image generation. In NeurIPS, 2017.

[139] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz. Disentangled person image generation. In IEEE CVPR, 2018.

[140] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In IEEE CVPR, 2018.

[141] Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In IEEE CVPR, 2017

[142] S. Huang, H. Xiong, Z.-Q. Cheng, Q. Wang, X. Zhou, B. Wen, J. Huan, and D. Dou. Generating person images with appearance-aware pose stylizer. In IJCAI, 2020,