NORTHWESTERN UNIVERSITY

Learning Visual Matching From Small-Size Samples

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Electrical Engineering and Computer Science

By

Jiahuan Zhou

EVANSTON, ILLINOIS

December 2018

ABSTRACT

Learning Visual Matching From Small-Size Samples

Jiahuan Zhou

Visual matching is an important and fruitful research topic in computer vision area. Starting from the early face recognition, super-resolution, object tracking to the most recent person re-identification, cross-model retrieval, visual matching plays an important role as the core component in these tasks. The quality of visual matching directly and largely influence of the ultimate performance of these tasks.

This dissertation concentrates on developing effective and efficient visual matching learning algorithms to facilitate the critical small-size sample challenges in visual matching, that only very few labeled positive, even only one sample is available for a particular instance. A specific human-centric visual matching task, image-based person re-identification, is adopted to evaluate our proposed works. The goal of person re-identification generally refers to evaluating the similarity of a probe image from an unknown identity against a set of gallery images with known identities. The gallery images may be obtained from different cameras at a different time. Person re-identification still remains a critical yet very challenging task in video surveillance due to the general difficulties of the large and complex variations in the visual appearances

of a person under various views, poses, illumination and occlusion conditions. Besides the aforementioned difficulties, another critical issue that the very few labeled positive samples of one identity and severely imbalanced negative samples significantly constricts the quality of learning visual matching in person re-identification.

This dissertation presents various effective and efficient techniques to address the critical small-size sample challenges in visual matching across images: a global metric learning algorithm based on a novel proposed similarity constraint, termed reference constraint, only needs few-shot positive samples for learning without any requirement of negative samples; an online local metric adaptation algorithm which is adoptable to any feature descriptors and any global metrics by using only one positive and extra unlabeled negative samples for metric learning; an extended online joint multi-metric learning method to learn multiple sharing-based joint Mahalanobis metrics for the given unlabeled data, no supervision label is requirement; and a two-stage hierarchical local metric adaptation algorithm to joint enhance the local discriminant of both unlabeled and labeled data. All the aforementioned methods aim to solve the severe small-size sample problem by relaxing the requirement of a large number of labeled positives for learning. Extensive experiments under different task setting on different datasets have validated the effectiveness and efficiency of the proposed approaches in the domain of image-based person re-identification.

Acknowledgements

I would like to express my sincere gratitude and thanks to my advisor Prof. Ying Wu for the continuous support of my Ph.D. study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research. I would like to thank him for encouraging my research and for allowing me to grow as a research scientist. His advice on both research as well as on my career have been priceless.

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Aggelos K. Katsaggelos and Prof. Thrasyvoulos N. Pappas, for their encouragement, insightful comments, and challenging questions.

My sincere thanks also go to Dr. Gang Hua for offering me the summer internship opportunities in their groups and leading me working on diverse exciting projects.

I thank my fellow labmates in Northwestern Vision Group: Jiang Wang, Zhuoyuan Chen, Pei Yu, Yin Xia, Chen Jiang, Wei Tang and Xiangyun Zhao for the stimulating discussions, for the sleepless nights we were working together before deadlines, and for all the fun we have had in the last four years. I would also like to thank my friends in Northwestern, Shengxin Zha, Jing Wang, Jue Lin, Yanran Wang, and friends in China, Jianhui Wang, Bing Su, Weijun Zhang and Xu Zou for the discussion during my whole Ph.D. period.

Last but not least, I would like to thank my parents, Yong Zhou and Xiangli Liu, for giving birth to me at the first place and supporting me spiritually throughout my life.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

1.1. Background

1.1.1. Visual Matching Across Images

Visual matching is an important and fruitful research topic in computer vision area. Starting from the early face recognition, super-resolution, object tracking to the most recent person re-identification, cross-model retrieval, visual matching plays an important role as the core component in these tasks. The quality of visual matching directly and largely influence of the ultimate performance of these tasks.

The definition of visual matching is aiming to evaluate how similar a given image x and a target image y is. In other words, visual matching across images needs to measure whether the given images x and y are matched based on some matching metric $\mathbf{M}$. From this definition we can see, there are two main important factors in visual matching: the feature representation for the images and the measurement metric for matching. As for the feature representations, for different visual matching tasks and different data, different feature descriptors can be used, either handcrafted features or the learned features from training samples. Once the feature representation is determined, another key-point in visual matching is the matching metric. Generally, the Euclidean distance is directly used for visual matching due to its simplicity and flexibility. In order to improve the visual matching performance, a discriminative metric, e.g., Mahalanobis distance, is learned from training samples and used for visual matching.

This dissertation concentrates on developing effective and efficient visual matching learning algorithms for a robust matching performance. A specific human-centric visual matching task, image-based person re-identification, is evaluated by our proposed works. The term "re-identification" is firstly defined by Alvin Plantinga in 1961 as: "To re-identify a particular, then, is to identify it as (numerically) the same particular as one encountered on a previous occasion". For person re-identification, the goal of it generally refers to evaluating the similarity of a probe image from an unknown identity against a set of gallery images with known identities. The gallery images may be obtained from different cameras at a different time. Person re-identification still remains a critical yet very challenging task in video surveillance due to the general difficulties of the large and complex variations in the visual appearances of a person under various views, poses, illumination and occlusion conditions. Besides the aforementioned difficulties, another critical issue is known as the small-size sample challenge, that only very few labeled positive, even only one sample is available for a particular instance, significantly constricts the quality of learning visual matching in person re-identification. Such small-size positive samples and large-scale negative samples for one identity caused a severely imbalanced data distribution, the learning of visual matching is dominated by the large-scale negatives which results in poor learning quality.

1.2. Human Centric Visual Matching: Person Re-Identification (P-RID)

In this dissertation, we focus on a human-centric visual matching task, called Person re-identification (P-RID), which is a critical yet very challenging task in video surveillance [**90**]. It generally refers to evaluate the similarity between a probe image of an unknown person against a set of gallery candidates with known identities. The gallery images are usually taken from

different camera-views at different times. Based on different input probes, the P-RID can be mainly categorized into three main branches. If the input probe is represented by a single image, we call it single-shot person re-identification (SsP-RID), in where the visual matching is performed as an image-to-image matching. If the given probe is a set of multiple discrete images for the same identity, or a set of continuous image frames from a video sequence, both of them are called multi-shot person re-identification (MsP-RID). Therefore the visual matching is a set-to-set matching or sequence-to-sequence matching respectively for them. This dissertation will cover the research of all three scenarios in person re-identification.

Research efforts have been devoted to single-shot person re-identification [**50, 52, 51, 112, 119, 103**] in recent years. However, besides viewpoint changes, the quality of the only given probe image can be severely degraded by various unpredictable conditions such as illumination changes, partial occlusion, low-resolution, etc. Thus SsP-RID still remains a very challenging problem. In fact, practical scenarios in video surveillance can provide continuous video or multiple images for the same person, which has motivated the research of multi-shot person re-identification [**29, 79, 95, 109, 66**] that utilizes multiple images for the same person from the same camera-view, expecting to improve the performance. The performance of visual matching in person re-identification is mainly influenced by two key factors, feature representation of images and matching measurement metric. In the past years, there are abundant researches focusing on these two directions. Robust hand-crafted feature descriptors [**51, 104**] are designed to minimize the within-identity variation of visual appearances. Due to the limitation of human designation, more and more attention has been paid to learning discriminative and robust feature embeddings to facilitate visual matching. On the other hand, another branch of works aims to learn discriminative matching metrics instead of using the traditional Euclidean

distance metric [**52, 50, 112, 103, 4**]. Such methods generally propose to pull the same-identity samples closer as well as push different-identity samples far away so that various constraints, e.g., contrastive constraint, triplet constraint, structure constraint, etc, are proposed to minimize the within-class scatters and maximize the between-class scatters. Therefore promising performance has been achieved by metric learning-based approaches. In recent years, due to the success of deep learning in various computer vision tasks, deep neural network-based approaches [**85, 40, 96, 92**] dominate the person re-identification area due to the surprising performance. Besides, the metric learning idea is incorporated into the deep neural network to learn a deep metric from large-scale training data, which is known as deep metric learning [**108**].

In our research, we mainly focus on how to learn robust and discriminative matching metrics to facilitate visual matching in person re-identification. However, besides the aforementioned general challenges of appearance variation, a critical challenge in learning visual matching is the small-size sample issue, that for one specific instance, there are very few positives available for learning but much more negatives provided. Such issue is even more severe in instance-level visual matching tasks, like single-shot person re-identification, there is usually one single image provided for each identity, but hundreds of thousands of images from the other identities are given. Such extremely imbalanced data distribution will largely influence the learning quality since the learning is dominated by the large-scale negative constraints and the power of small-size positive constraints is suppressed.

1.3. Organization

Various novel methods are proposed to address the critical visual matching problem, focusing on different human-centric tasks such as single-shot person re-identification, multi-shot person re-identification, etc. Therefore this dissertation is organized as follows:

- Chapter 2 demonstrates a novel global metric learning approach to tackle to multi-shot person re-identification problem by utilizing only the given few positive samples. Although the given positive and negative samples are extremely imbalanced, our method tackle such issue by using a novel proposed reference constraint to facilitate metric learning. In Sec. 2.2, we briefly review the existing metric learning-based works to solving multi-shot person re-identification problem. Sec. 2.3 demonstrates the details of our proposed novel learning constraint, reference constraint, and three optimal transport learning-based schemes are designed to automatically generate different reference constraints. A ridge-regression based metric learning method with closed-form solution is proposed to utilize the generated reference constraint for metric learning. A theoretical analysis of the generalization ability of our proposed reference-based metric learning method is presented in Sec. 2.4. Extensive experiments are shown in Sec. 2.5 and discussions are given in Sec. 2.6.
- Chapter 3 presents a novel online local metric adaptation methods by using only one positive for learning to tackle the special challenge on the online testing stage of person re-identification, that only one positive sample is given for visual matching. Some existing solution for online learning of person re-identification and related researches are introduced in Sec. 3.2. Then Sec. 3.3 presents the proposed novel online metric adaptation algorithm by using only one positive and unlabeled negative samples. Three

theoretical sound justifications of the proposed method are demonstrated in Sec. 3.4, including both the asymptotic and practical scenarios. Sec. 3.5 presents various experiments to verify the effectiveness and efficiency of the proposed method. The chapter concludes in Sec. 3.6.

- Chapter 4 describes a novel unsupervised joint multi-metric learning methods from unlabeled samples for online local adaptation. Instead of performing individual learning for each sample as in Chapter 3, the visual similarity relationship of unlabeled samples are utilized for joint learning. Therefore, a novel sharing-based multi-metric learning algorithm is introduced in Sec. 4.3. Some justifications of the proposed method and comparisons with the related works are presented in Sec. 4.3.2. Sec. 4.5 shows extensive experimental results to support the proposed method. Finally, conclusions and discussions are made in Sec. 4.6.
- Chapter 5 presents a novel two-stage hierarchical local metric learning method for bi-directional local discriminant enhancement of the given samples consisting of both labeled and unlabeled samples. In Sec. 5.2, some related works about online re-ranking of visual matching and the state-of-the-art methods are introduced. Our proposed method is presented in Sec. 5.3. Extensive experimental results of our proposed method are shown in Sec. 5.4 and conclusions are made in Sec. 5.5.
- Chapter 6 summarizes the dissertation by four typical small-size sample learning problems, i.e., global metric learning from few positives, instance-level local metric adaptation from one-shot positive, group-level local metric learning from unlabeled samples and a hierarchical local metric learning scheme from both labeled and unlabeled samples.

1.4. Contribution

Inspired by the insufficient existing solutions and inherent challenges to the different person re-identification problems in visual matching, this dissertation proposes various metric learning methods based on small-size samples, including few-shot positives, only one positive, and unlabeled samples, etc. As illustrated, all these small-size sample settings are the critical issues that must be addressed to facilitate the learning of visual matching as well as make the applications of visual matching in the intelligent video surveillance better. To summarize, the following contributions have been made in the dissertation:

- A novel learning constraint called reference constraint is proposed to facilitate the poor and difficult metric learning solution caused by the large-scale and imbalanced constraints used before. Our proposed reference constraint aims to associate the samples from the same class to one or multiple By utilizing our proposed reference constraint, a ridge-regression based global metric learning from few positives and no negatives is proposed to learn a discriminative metric. A closed-form solution can be obtained for our metric learning objective so that the learning is not only effective and also efficient compared with the related metric learning approaches [**52, 51, 112**].
- In order to address the two critical issues of the proposed global metric learning work, including the failure on one-shot positive scenario which is an extremely challenging small-size sample setting and the failure on handling the hard negative distractors, a novel online instance-specific local metric learning is proposed by using only one positive but a large number of negatives for learning. Our proposed online local metric adaptation algorithm can be applied to any offline learned baselines on any features, and an efficient optimization solution is proposed to our method which requires very

trivial online learning cost. Three theoretical sound justifications guarantee the improvement of our method under both the asymptotic scenario and practical learning scenario.

- A novel online joint multi-metric learning algorithm is designed via learning from unlabeled samples. By considering the given matching samples as an unlabeled batch-shot query set, the intrinsic visual similarity sharing relationships among the samples can be utilized by mining different sharing-subsets. For each sharing-subset, the samples share the same visual similarity relationship are grouped from which a joint Mahalanobis metric is learned to jointly adapt the local distribution of all the subset samples. Compared with our instance-specific local metric adaptation work, our joint multi-metric learning algorithm is not only more effective for matching performance improvement but also has lower online learning cost.
- A novel bidirectional local discriminant enhancement from the combination of few-shot labeled and unlabeled samples is proposed to perform a two-stage hierarchical local metric adaptation for both the probe and gallery samples in visual matching. Unlike the previous method which only focus on the local discriminant enhancement of the given matching probes, the local discriminant of the gallery samples is also enhanced by our method so that the "hard" gallery distractors which are indistinguishable will be well tackled by our method.

An interesting note is that for our four proposed works, although they rely on different approaches to learn a discriminative matching metric to facilitate the severe small-size sample problem in visual matching, actually they have strong research connection to each other. The Chapter 3, Chapter 4 and Chapter 5 can be readily implemented on Chapter 2 by using it as

the baseline model. Chapter 4 aims to address the intrinsic issue in Chapter 3 by learning from sharing, and Chapter 5 focuses on how to improve the performance of Chapter 3 and Chapter 4 by a bi-directional local discriminant enhancement strategy.

CHAPTER 2

Learning From Few-Shot Positives: Global Metric Learning From Reference Constraint

2.1. Introduction

In this chapter, we propose a novel algorithm to solve the multi-shot person re-identification (MsP-Rid) [**95, 29, 109, 79**] problem in computer vision area. One common solution to MsP-Rid is to treat the multiple images as a sequence of consecutive frames which prefers to utilize the temporal information or motion to extract more sophisticated features for identification. In practice, the motion information may not be discriminative nor reliable enough for MsP-Rid. Firstly, the dynamic background and temporal misalignment of the image sequences impede the reliable motion pattern estimation [**57**] (Fig. 2.1(a)). Secondly, motion patterns may not be discriminative enough for identification since different persons may walk in the same walking pattern [**95**] (Fig. 2.1(d) and (e)). Because MsP-Rid is a non-contextual long-term identification problem, the same person may exhibit different walking behaviors at different times. As shown in Fig. 2.1(b) and (c), a person is walking with luggage captured by one camera. At a different time, the same person is viewed by another camera but without the luggage. Such large intra-class variation in motion and dynamics across different camera-views along a long time duration is very difficult to handle. As a result, the performance of this approach is still far from satisfactory even additional motion/dynamics features are utilized.

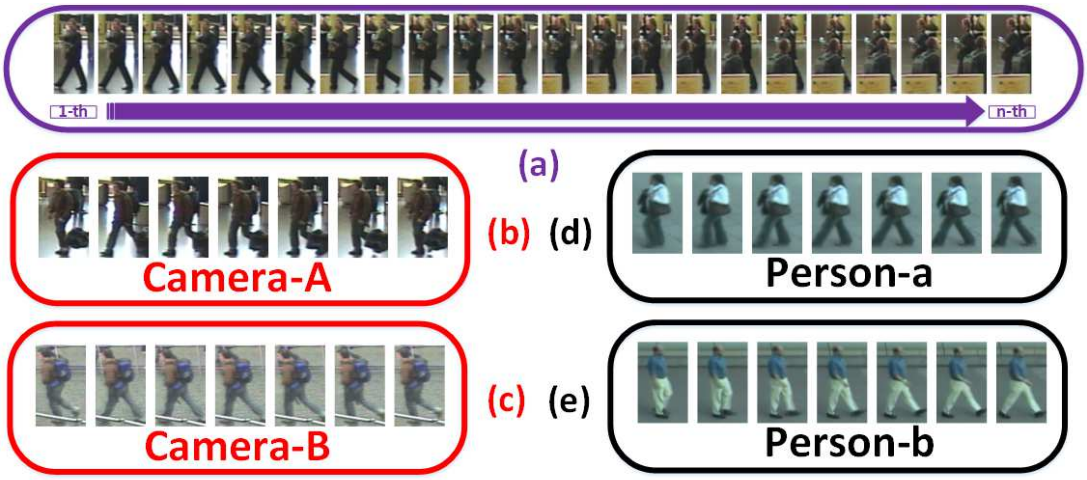


Figure 2.1. **(a)** The background occlusion completely conceals the motion information on the legs; **(b)** & **(c)** Even for the same person, the walking behavior can be very different; **(d)** & **(e)** For different persons, they may share very similar walking patterns.

Another approach [**41, 49**] treats the multiple images as separate samples, paying more attention to the variations in their visual appearances. Efforts have been made to design specific appearance features [**51, 104, 57**], but there is still room for performance improvement. Recent methods have been focused on learning discriminative visual metrics to facilitate identification. Many such methods [**51, 122, 95, 52**] learn a global Mahalanobis-like distance metric that reduces the intra-class variation and enlarges the inter-class variation. In practice, there are several difficulties to be overcome. Firstly, these methods use pair-wise [**52**] or triple-wise [**122**] data similarity and dissimilarity constraints. The scale order of such constraints is quadratic $O(n^2)$ or cubic $O(n^3)$ to the number of data points n. As a result, these constraints can be enormous, and it is computationally demanding to obtain optimal solutions that satisfy all these constraints. When adopting computationally-feasible but sub-optimal solutions, their performances suffer significantly. In addition, although having more samples sounds appealing, not all of them are

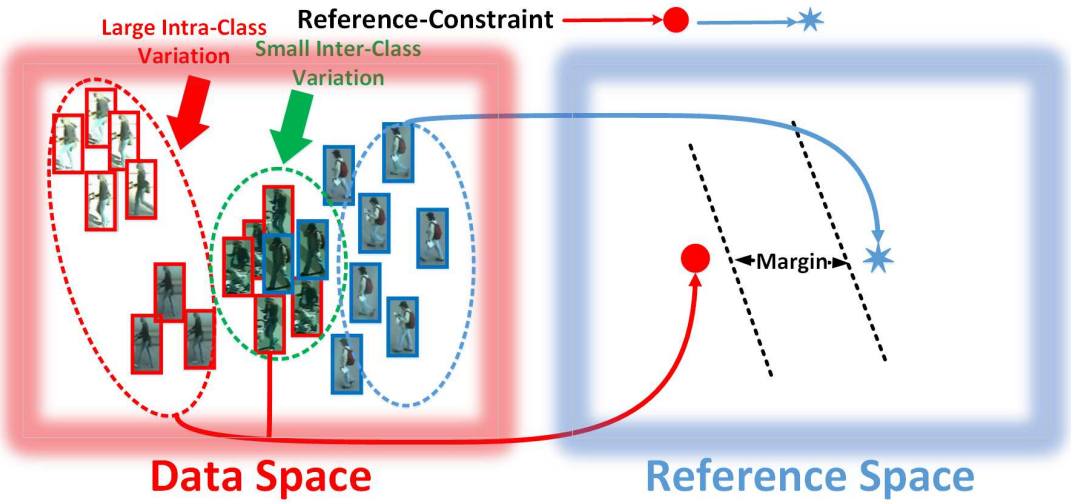


Figure 2.2. The proposed **reference constraint** correlates the original indiscriminative same class data to the common discriminative reference points (note: there can be multiple reference points to handle the multiple-mode distribution of same class data).

actually necessary or helpful for learning. The computational complexity induced by the redundant samples will largely slow down the optimization process in learning, and a small portion of "adverse" inputs will significantly jeopardize the learning quality [**76**]. Moreover, in practice, the positive and negative samples are significantly imbalanced. As the learning can be largely dominated by the negative pairs [**52**], it leads to unstable and non-discriminative learning results.

To overcome these difficulties, in this chapter, we propose a novel type of similarity constraints which assigns given sample points to a set of pre-determined points with explicit meanings, as shown in Fig. 2.2. We call the pre-determined points *references*, and the constraints between the original samples and the references *reference constraints*. Such reference points are automatically generated based on different criteria. Several optimal transport-based schemes for determining the reference points and assignments are proposed and studied. The proposed

reference constraints can be readily used for a regressive metric learning model [**18, 71**] to learn a discriminative metric with a closed-form solution.

Our contributions are three-fold. (1) In contrast to the existing methods that use a $O(n^2)$ or $O(n^3)$ number of constraints, our method only uses a linear $O(n)$ number of reference constraints, which is much easier to deal with. (2) The proposed reference constraints can be readily used for a general regression-based string-to-string mapping framework [**18**] for metric learning, the closed-form solution and its general non-linear version can be easily obtained. (3) Compared with the state-of-the-art MsP-Rid methods based on appearance features, our method significantly outperforms them by a large margin in terms of both identification accuracy and running speed. Besides, even no temporal information is used, our model still achieves comparable even better performance against the ones using both appearance and temporal features. Extensive experiments have demonstrated the superiority of our method on several multi-image benchmarks including the CAVIAR [**15**], the PRID 2011 [**35**], the iLIDS-VID [**95**] and the Market-1501 [**119**] datasets.

The rest of this chapter is organized as follows: Sec. 2.2 briefly reviews the existing metric learning-based works to solving multi-shot person re-identification problem. In Sec. 2.3, we demonstrates our proposed novel learning constraint, reference constraint, and three optimal transport learning-based schemes are designed to automatically generate different reference constraints. A ridge-regression based metric learning method with closed-form solution is proposed to utilize the generated reference constraint for metric learning. A theoretical analysis of the generalization ability of our proposed reference-based metric learning method is presented in Sec. 2.4. Extensive experiments are shown in Sec. 2.5 and discussions are given in Sec. 2.6.

2.2. Previous Work

In this section, we would like to give a brief overview of P-Rid problem, especially for the multiple-image based ones. For a though survey, feel free to check [**7, 90**].

Person re-identification problem is closely related to various research topics such as tracking, identification, etc. Firstly, re-identification approaches could be categorized into three groups based on a classical categorization system in [**79**]: short-term, contextual long-term and non-contextual long-term re-identification. The so-called short-term re-identification is known as the famous tracking problem that is to associate the target frame by frame based on appearance features. Contextual long-term re-identification aims to differentiate the target from the other extractors by learning online models based on the context of a single static camera in the scene. Since the identification is only restricted in the same camera, the contextual information is available to give aid to identify the same target. The last problem, non-contextual long-term re-identification, is exactly the recent person re-identification problem that the identification is applied across arbitrary cameras and viewpoints. The camera topology is unknown and the photo capturing time can vary along a pretty long-term time span. On the other hand, based on the data type to deal with, existing person re-identification works can be categorized into four different scenarios: single image, multiple images without motion information, multiple temporally-aligned video fragments and the whole unaligned video. For each type of algorithm, either specific appearance/motion feature is designed or learned for view-point or illuminant invariant representation, or learning robust metrics or sub-spaces for matching across-cameras. In the following overview, we are only interested in the last three scenarios.

A novel multi-task maximally collapsing metric learning (MtMCML) model was proposed by Ma *et al.* [**62**] for multi-image re-identification in camera networks. Simonnet *et al.* [**79**] utilized the widely-used Dynamic Time Warping (DTW) algorithm in action recognition to solve the video-based person re-identification problem. Wang *et al.* [**95**] proposed a walking cycle extraction method to further divide the video into multiple aligned walking sequences. Then a sequence fragment selection and ranking framework are adopted for person matching. Liu *et al.* [**57**] built a new spatio-temporal appearance representation for video-based person re-identification in order to handle the temporal alignment problem. Karanam *et al.* [**41**] aimed to learn a dictionary that is capable of discriminatively and sparsely encoding features representing different people. You *et al.* [**109**] introduced a top-push distance learning model (TDL) to select the most discriminative video feature for robust matching. Li *et al.* [**49**] presented a subspace learning algorithm maximizing the Fisher criterion for discriminative feature extraction. Recently, a large-scale video benchmark dataset for person re-identification, named Motion Analysis and Re-identification Set (MARS), is proposed by Zheng *et al.* [**118**] which makes deep neural network methods available to solve this problem. McLaughlin *et al.* [**66**] proposed a novel recurrent neural network (RNN) to tackle video-based person re-identification problem. Both color and optical flow map are used as input for network training to provide both the appearance and motion information.

2.3. Our Solution: Learning from Reference Constraints

2.3.1. Problem Setup

In this work, we aim to learn a discriminative positive semi-definite (PSD) Mahalanobis metric $\mathbf{M} = \mathbf{L}\mathbf{L}^T$ by utilizing the proposed reference constraints. Given a labeled dataset $\mathcal{S} =$

$\{(x_i, l_i)\}_{i=1}^n$, we construct a new learning set $\mathcal{S}_r = \{(x_i, r_i)\}_{i=1}^n$, where x_i is the data point, $l_i \in \mathcal{L} = \{1, 2, 3, ..., c\}$ is its label and r_i is the associated reference point to x_i determined by its label l_i (details see Sec. 2.3.2). For the sake of convenience, let's denote $\mathbf{X} = (x_1, x_2, ..., x_n)^T$ and $\mathbf{R} = (r_1, r_2, ..., r_n)^T$. It's worth mentioning that the reference point set $\mathbf{R}$ can be drawn from another distribution $\mathcal{D}'$ so that $\mathbf{R} \subseteq \mathbb{R}^{d'}$. If $d' \ll d$, the learned Mahalanobis metric $\mathbf{M}$ automatically perform the dimension reduction on the original samples.

2.3.2. Automatic Reference Constraint Generation

In this section, we will show how to automatically generate the reference constraints under a general optimal transport (OT) framework [**91**]. The motivation of regressing the original given data $\mathbf{X}$ to a reference set $\mathbf{R}$ is the poor discriminative power of $\mathbf{X}$ can be enhanced by the "good quality" reference set $\mathbf{R}$, then the coupling between $\mathbf{X}$ and $\mathbf{R}$ can be modeled as an optimal transport procedure [**19, 70, 21**]:

$$\arg\min_{\mathcal{T}} \quad \langle \mathcal{T}, \mathbf{C} \rangle_{\mathcal{F}} + \mathcal{G}(\mathcal{T}) \tag{2.1}$$

where $\mathcal{T}$ is the optimal transformation, $\mathbf{C}$ is the cost matrix between $\mathbf{X}$ and $\mathbf{R}$. The first transport cost term is the Frobenius dot product between $\mathcal{T}$ and $\mathbf{C}$, and $\mathcal{G}(\mathcal{T})$ is a regularization term to constrain $\mathcal{T}$. In the following, three different schemes are proposed to automatically determine $\mathbf{R}$ and find optimal $\mathcal{T}$ based on different $\mathbf{C}$ and $\mathcal{G}(\mathcal{T})$.

2.3.2.1. R from Camera Viewpoint Alignment. The major challenge for P-Rid is rendered by the large appearance variation due to the camera viewpoint changes. Identifying the same

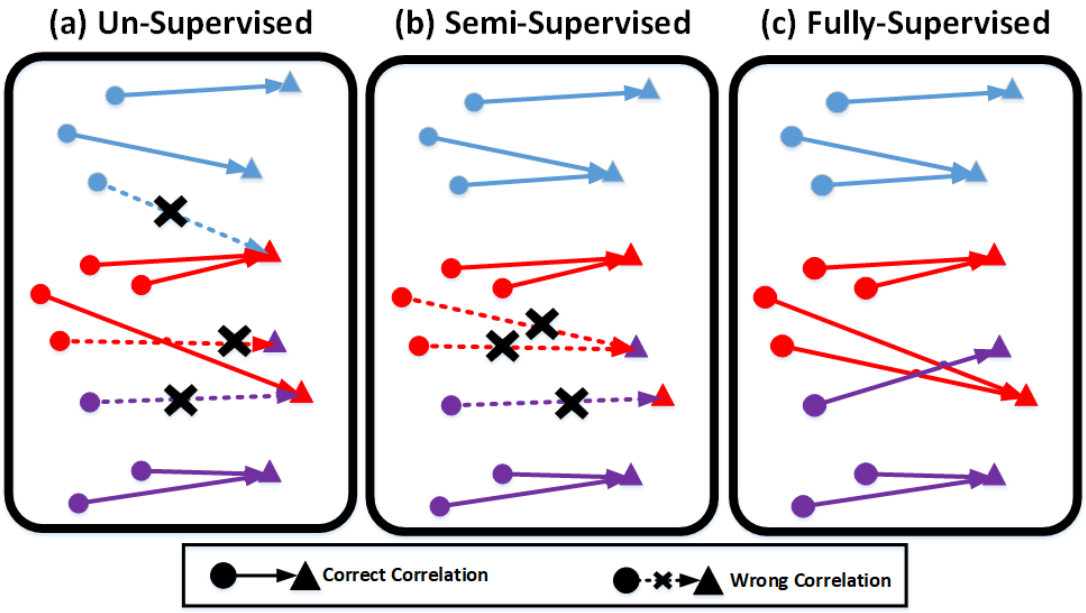


Figure 2.3. **(a)** is the result of an unsupervised OT method [**21**]; **(b)** is a semi-supervised OT method [**19**]; **(c)** is our proposed supervised OT method with cross-bin cost function Eqn. 2.2. Different colors (Red, Blue, Purple) represent different classes, and different shapes mean different distributions.

person across a significant viewpoint change is difficult because of the visually spatial misalignment [**76**]. An intuitive idea to generate **R** is to directly re-align the data from different camera viewpoints.

The alignment can be achieved via a supervised optimal transport learning. Traditionally, OT methods are un-supervised since no class label information is used. Hence the correlations between two distributions are completely unconstrained (Fig. 2.3(a)) which will be problematic in the P-Rid problem, where the identity label is given for each sample. In [**19**], a novel semi-supervised OT method is proposed to utilize the label of source data while the labeling for target distribution is unknown. Under this condition, although one target sample is not assigned to the source samples from different classes, the mis-matching between different classes still exists

(Fig. 2.3(b)). In contrast to these methods, we propose a novel cross-bin cost function $\mathbf{C}_A$ to fulfill the fully supervised learning requirement of P-Rid:

$$\mathbf{C}_A(i,j) = \|x_i^{\mathcal{A}} - x_j^{\mathcal{B}}\|_2 I(l_i^{\mathcal{A}} = l_j^{\mathcal{B}}) + \infty \cdot I(l_i^{\mathcal{A}} \neq l_j^{\mathcal{B}}) \tag{2.2}$$

where $I(\cdot)$ is a binary indicator, $x_i^{\mathcal{A}}$ is the i^{th} sample with class label $l_i^{\mathcal{A}}$ from camera space $\mathcal{A}$, so as the $x_j^{\mathcal{B}}$. Therefore we formulate the alignment between two camera viewpoint spaces via an optimal transport $\mathcal{T}_A$ as Eqn. 2.3:

$$\begin{aligned} \arg\min_{\mathcal{T}_A} \quad & \langle \mathcal{T}_A, \mathbf{C}_A \rangle_{\mathcal{F}} + \frac{1}{\lambda} \sum_{i,j} \mathcal{T}_A(i,j) \log \mathcal{T}_A(i,j) \\ & + \eta \sum_j \sum_c \|\mathcal{T}_A(l_i^{\mathcal{A}} = c, j)\|_q^p \end{aligned} \tag{2.3}$$

where the λ and η are the regularization parameters. The second regularization aims to compute the entropy of the transport $\mathcal{T}_A$. The third sparsity regularization is to group the samples from the same class together that $\mathcal{T}_A(l_i^{\mathcal{A}} = c, j)$ corresponds to the j^{th} column of $\mathcal{T}_A$ where the label is c. The desired optimal transport $\mathcal{T}_A$ is a matrix with the same size as $\mathbf{C}_A \in \mathbb{R}^{|\mathcal{A}| \times |\mathcal{B}|}$.

By utilizing the proposed $\mathbf{C}_A$, the transport cost term will be optimal only if the transports are restricted within the same class samples. The mis-matching occurred in the existing methods ([**19, 70, 21**]) as illustrated in Fig. 2.3(a) and (b) can be avoided, and thus a clean transport flow can be achieved (Fig. 2.3(c)). The objective Eqn. 2.3 can be efficiently solved via the alternation between the Sinkhorn-Knopp algorithm[**21**] and the Majoration-Minimization strategy[**19**]. The parameters of l_q-norm in the third term are $p = \frac{1}{2}$ and $q = 1$. Once the optimal transport $\mathcal{T}_A$ is learned, the corresponded reference constraint set is $\mathbf{R} = \mathbf{X}\mathcal{T}_A$

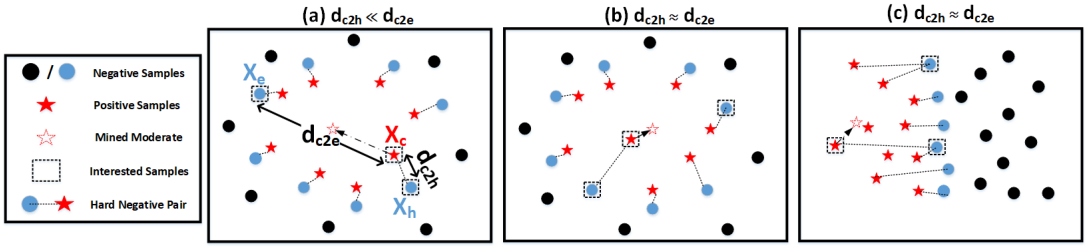


Figure 2.4. Moderate positive mining for a local unimodal data distribution.

2.3.2.2. R from Class-based Discriminative Space. An efficient and straightforward idea is to explicitly determine the **R** in a class-based discriminative space (CDS). Let $u_i \in \mathbb{R}^{|\mathcal{L}|}$ be a unit vector ($1 \leq i \leq |\mathcal{L}|$) in a $|\mathcal{L}|$-dimensional feature space, $\mathbf{R} = \{u_i\}_{i=1}^{|\mathcal{L}|}$ contains all such u_i. The optimal transport from **X** to **R** can be modeled as optimizing:

$$\arg\min_{\mathcal{T}_C} \quad \langle \mathcal{T}_C, \mathbf{C}_C \rangle_{\mathcal{F}} \tag{2.4}$$

with $\mathbf{C}_C(x_i, u_j) = 0 \cdot I(\#l_i = j) + \infty \cdot I(\#l_i \neq j)$ that $\#l_i$ is the label index. Obviously, a naive optimal solution to $\mathcal{T}_C$ is

$$\mathcal{T}_C\left(x_i, l_i\right) = u_{\#l_i} \tag{2.5}$$

that all the samples in **X** from the same class $\#l_i$ will be transported into one single point $u_{\#l_i}$ in **R** to guarantee a zero within-class distance, and large distances between the collapsed points can be explicitly guaranteed to avoid mixing classes after transformation. If the class number $|\mathcal{L}|$ is much smaller than the dimensionality d of **X**, $\mathcal{T}_C$ is equivalent to learn a lower-dimensional embedding where the samples drawn from different classes become much more discriminative.

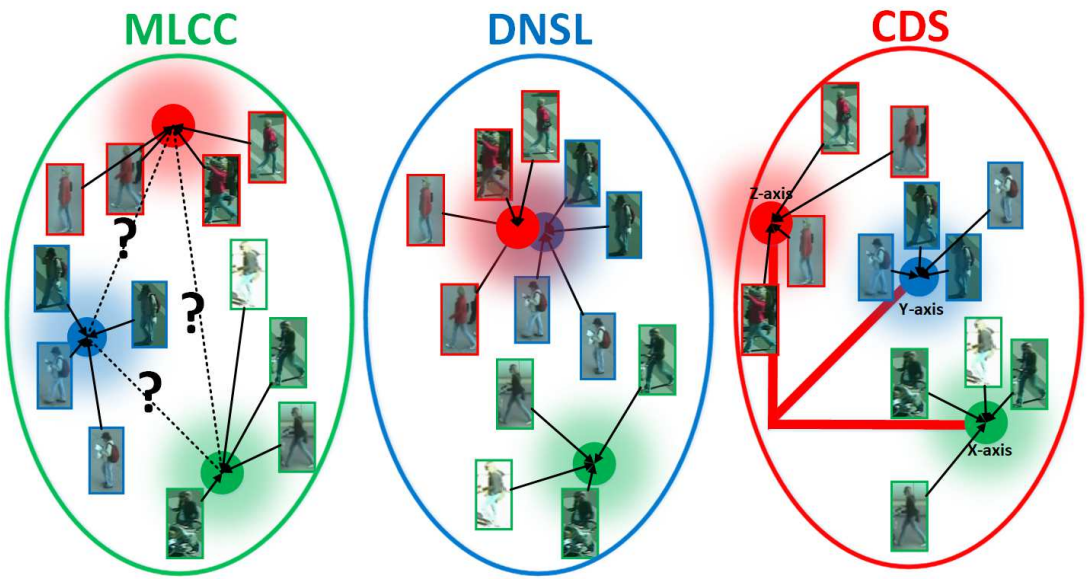


Figure 2.5. The comparison of three related algorithms: MLCC [**30**], DNSL [**112**] and our CDS method.

Optimality of R from CDS: The similar idea of our CDS is shared by many existing works like the well-known metric learning algorithm MLCC [**30**] and a recently state-of-the-art P-Rid algorithm DNSL [**112**]. As illustrated by Fig. 2.5, all the three approaches will collapse the same class samples into one single point in the projected space, so as to enforce the within-class distance to be zero. However, three methods have completely different strategies to handle the between-class distance. Let's take the Fisher discriminant criterion $\mathcal{J}(\mathbf{L}) = \dfrac{\mathbf{L}^T\mathbf{S}_b\mathbf{L}}{\mathbf{L}^T\mathbf{S}_w\mathbf{L}}$ into consideration. The larger the $\mathcal{J}(\mathbf{L})$ is, the more discriminative the learned projection $\mathbf{L}$ is. All of MLCC, DNSL and CDS will give us zero within-class scatter $\mathbf{L}^T\mathbf{S}_w\mathbf{L} = 0$, but MLCC simply omits the between-class scatter part, DNSL only requires $\mathbf{L}^T\mathbf{S}_b\mathbf{L} > 0$. Our CDS will strictly require $\mathbf{L}^T\mathbf{S}_b\mathbf{L} = c$ to a constant margin.

2.3.2.3. R from Local Moderate Positive Mining. Another approach to obtain good quality $\mathbf{R}$ is from the intrinsic distribution of $\mathbf{X}$ directly which is inspired by the SMOTE algorithm

for imbalanced learning [**12**]. We propose to mine a set of "moderate" representations from **X** which are conceptually not too close to the hard negatives around the classification boundary, but also convey enough discriminative information.

A moderate positive mining (MPM) algorithm is proposed to mine the references **R** in a local manner. Denote by $\mathcal{X}_c = \{x_i^c\}$ for a subset containing all the samples from class c, and by $\mathcal{X}_{\bar{c}} = \{x_i^{\bar{c}}\}$ for a subset including different class samples. For each x_i^c in $\mathcal{X}_c$, its corresponded "hardest" negatives $\{x_{i,h}^{\bar{c}}\}_{i=1}^{|\mathcal{X}_c|}$ are obtained from $\mathcal{X}_{\bar{c}}$. The pair $(x^c, x_h^{\bar{c}}) = \max_i d(x_i^c, x_{i,h}^{\bar{c}})$ with the largest distance to its "hardest" negative is retrieved. Then another sample $x_e^{\bar{c}}$ that is farthest away from x^c is retrieved from the obtained hardest negative set $\{x_{i,h}^{\bar{c}}\}_{i=1}^{|\mathcal{X}_c|}$ which is the "easiest-hardest" negative for x^c. Finally, the reference points for all $\mathcal{X}_c$ is the synthetic point:

$$r^c = \frac{1}{2}(1 + \frac{d_{c2h}}{d_{c2e}})x^c + \frac{1}{2}(1 - \frac{d_{c2h}}{d_{c2e}})x_e^{\bar{c}} \tag{2.6}$$

where the weighting parameter $d_{c2e} = d(x^c, x_e^{\bar{c}})$ and $d_{c2h} = d(x^c, x_h^{\bar{c}})$ [1]. Various conditions of d_{c2h} and d_{c2e} are shown in Fig. 2.4 which indicates our MPM algorithm can always mine the moderate representations no matter how the local data distribution is. Finally, by solving a similar Eqn. 2.4 with $\mathbf{C}_M(x_i, r^j) = \|x_i - r^j\|_2^2 \cdot I(\#l_i = j) + \infty \cdot I(\#l_i \neq j)$, the optimal transport to associate **X** to **R** is:

$$\mathcal{T}_M(x_i, l_i) = r^{\#l_i} \tag{2.7}$$

Since real-world data generally exhibit multiple-mode distribution due to various complicated conditions, in order to eliminate the influence of the high-density modes, firstly we adopt Mean-shift clustering [**17**] to $\mathcal{X}_c$ to divide $\mathcal{X}_c$ into several sub-class clusters, thus each cluster

[1] It is obvious that $d_{c2h} \leq d_{c2e}$ is alway true.

bears a unimodal distribution. Then the proposed MPM algorithm is further performed to these unimodal clusters. Therefore even for the same class data $\mathcal{X}_c$, they may be assigned different moderate points as references.

2.3.3. Metric Learning from R via Regression

Once the reference set **R** is determined, we aim to learn a positive semi-definite (PSD) Mahalanobis metric $\mathbf{M} = \mathbf{L}\mathbf{L}^T$ by solving the following regularized regression problem [**18, 71**]:

$$\mathbf{L}^* = \min_{\mathbf{L}} \frac{1}{n}\|\mathbf{X}\mathbf{L} - \mathbf{R}\|_{\mathcal{F}}^2 + \lambda\|\mathbf{L}\|_{\mathcal{F}}^2 \tag{2.8}$$

where the λ is a weighting parameter to balance the two terms. The closed-form solution to objective Eqn. 2.8 can be derived.

Theorem 1. *The optimal solution of objective Eqn. 2.8 has a closed form, as shown in the following two equivalent solutions:*

$$\boldsymbol{L} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda n\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{R} \tag{2.9}$$

$$\boldsymbol{L} = \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T + \lambda n\boldsymbol{I})^{-1}\boldsymbol{R} \tag{2.10}$$

PROOF. Compute the derivative of Eqn. 2.8:

$$\frac{\partial f(\mathbf{L}, \mathbf{X}, \mathbf{R})}{\partial \mathbf{L}} = 2\left(\frac{1}{n}\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)\mathbf{L} - \frac{2}{n}\mathbf{X}^T\mathbf{R} \tag{2.11}$$

By setting this derivative to zero we can obtain:

$$\mathbf{L} = (\mathbf{X}^T\mathbf{X} + \lambda n\mathbf{I})^{-1}\mathbf{X}^T\mathbf{R} \tag{2.12}$$

□

Theorem 2. *The optimal solutions Eqn. 2.9 and Eqn. 2.10 of objective Eqn. 2.8 are exactly equivalent.*

PROOF. For Eqn. 2.9, we perform Taylor expansion to the $(\mathbf{X}^T\mathbf{X} + \lambda n\mathbf{I})^{-1}\mathbf{X}^T$ part:

$$\begin{aligned}
(\mathbf{X}^T\mathbf{X} + \lambda n\mathbf{I})^{-1}\mathbf{X}^T &= \frac{1}{\lambda n}(\mathbf{I} + \frac{1}{\lambda n}\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\
&= \frac{1}{\lambda n}\sum_{1}^{\infty}(-1)^n(\frac{1}{\lambda n})^n(\mathbf{X}^T\mathbf{X})^n\mathbf{X}^T \\
&= \frac{\mathbf{X}^T}{\lambda n}\sum_{1}^{\infty}(-1)^n(\frac{1}{\lambda n})^n(\mathbf{X}\mathbf{X}^T)^{n-1}\mathbf{X}\mathbf{X}^T \\
&= \frac{\mathbf{X}^T}{\lambda n}\sum_{1}^{\infty}(-1)^n(\frac{1}{\lambda n})^n(\mathbf{X}\mathbf{X}^T)^n \\
&= \frac{\mathbf{X}^T}{\lambda n}(\mathbf{I} + \frac{1}{\lambda n}\mathbf{X}\mathbf{X}^T)^{-1} \\
&= \mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \lambda n\mathbf{I})^{-1}
\end{aligned} \tag{2.13}$$

Therefore Eqn. 2.13 proves that Eqn. 2.9 and Eqn. 2.10 are exactly the same solution for the proposed objective Eqn. 2.8. □

From Eqn. 2.9, we obtain the Mahalanobis metric $\mathbf{M}$:

$$\mathbf{M} = \mathbf{L}\mathbf{L}^T = (\mathbf{X}^T\mathbf{X} + \lambda n\mathbf{I})^{-1}\mathbf{X}^T\mathbf{R}\mathbf{R}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda n\mathbf{I})^{-1} \tag{2.14}$$

As we can see from Eqn. 2.14, the bottleneck to compute the metric kernel $\mathbf{M}$ is the inversion of a $d \times d$ matrix, where d is the data dimension. In the case of a large d, appropriate dimension reduction techniques are needed before learning.

2.3.4. Non-Linear Extension by Kernelization

The linear model in Sec. 2.3.3 may not be powerful enough to handle complicated metrics, but we can extend it to a nonlinear form via kernelization.

Assume a kernel function is $K(x, x') = \phi(x)^T \phi(x')$ where the $\phi(x)$ is a nonlinear projection function. For the learning set $\mathbf{X}$, we are able to compute the kernel distance matrix $K_{\mathbf{X}} \in \mathbb{R}^{n \times n}$, where the element k_{ij} is equal to $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$. Rewrite $K_{\mathbf{X}} = \phi(\mathbf{X})^T \phi(\mathbf{X})$, where $\phi(\mathbf{X}) = (\phi(x_1), \phi(x_2), ..., \phi(x_n))^T$. So the kernelized version of $\mathbf{L}$ is defined as $\mathbf{L}_K = \phi(\mathbf{X})^T (K_{\mathbf{X}} + \lambda n \mathbf{I})^{-1} \mathbf{R}$, which can be easily obtained by kernelizing Eqn. 2.10. Therefore the kernelized Mahalanobis metric $\mathbf{M}_K$ is written as:

$$\mathbf{M}_K = \phi(\mathbf{X})^T (K_{\mathbf{X}} + \lambda n \mathbf{I})^{-1} \mathbf{R}\mathbf{R}^T (K_{\mathbf{X}} + \lambda n \mathbf{I})^{-1} \phi(\mathbf{X}) \tag{2.15}$$

The squared Mahalanobis distance between x and x' can be easily computed by:

$$d^2_{\mathbf{M}_K}(x, x') =$$

$$\phi(x)^T \mathbf{M}_K \phi(x) + \phi(x')^T \mathbf{M}_K \phi(x') - 2\phi(x)^T \mathbf{M}_K \phi(x')$$

that each term can be written as:

$$\phi(x)^T \mathbf{M}_K \phi(x) =$$

$$K_{\mathbf{X}}(x)^T (K_{\mathbf{X}} + \lambda n \mathbf{I})^{-1} \mathbf{R}\mathbf{R}^T (K_{\mathbf{X}} + \lambda n \mathbf{I})^{-1} K_{\mathbf{X}}(x)$$

where $K_{\mathbf{X}}(x) = (K(x, x_1), K(x, x_2), ..., K(x, x_n))^T$. For the kernelized version $\mathbf{M}_K$, we need to compute the inversion of a $n \times n$ matrix where n is the number of samples.

2.4. Generalization Ability Analysis

For our objective Eqn. 2.8, the empirical error risk is $\mathcal{E}(\mathbf{L}, \mathcal{S}_r) = \frac{1}{n}\|\mathbf{XL} - \mathbf{R}\|_{\mathcal{F}}^2$, which is to measure how close the projected samples $\mathbf{XL}$ are to the reference points $\mathbf{R}$ after learning. We still care about how large the true error risk $\mathcal{E}(\mathbf{L}, \mathcal{D}_{\mathbf{R}}) = \mathbb{E}_{(x_i,r_i)\sim\mathcal{D}_{\mathbf{R}}}\|x^T\mathbf{L} - r^T\|_2^2$ is for the whole data distribution $\mathcal{D}_{\mathbf{R}}$. Here, we prove that once a low empirical error $\mathcal{E}(\mathbf{L}, \mathcal{S}_r)$ can be obtained, with a very high probability, a low true error $\mathcal{E}(\mathbf{L}, \mathcal{D}_{\mathbf{R}})$ is bounded [**8**].

Theorem 3. *Assume $\|r\|_2 \leq B_r$ for any $r \in \mathcal{R}$, and $\|x\|_2 \leq B_x$ for any $x \in \mathcal{X}$. With probability $1 - \delta$, for any matrix $\boldsymbol{L}$ which is the optimal solution of Eqn.2.8 with stability $\beta = \frac{8B_x^2B_r^2}{\lambda n}\left(1 + \frac{B_x}{\sqrt{\lambda}}\right)^2$, we have:*

$$\|\mathcal{E}(\boldsymbol{L}, \mathcal{D}_{\boldsymbol{R}}) - \mathcal{E}(\boldsymbol{L}, \mathcal{S}_r)\| \leq \left(1 + \left(2n + \frac{\lambda n}{8B_x^2}\right)\sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right)\beta \tag{2.16}$$

As shown by Theorem. 3, with a convergence rate $O(1/\sqrt{n})$, the difference between empirical error risk and true error risk converges to zero. The proof of Theorem. 3 can be found in our supplementary materials. More specifically, if a zero-empirical error can be obtained during training $\mathcal{E}(\mathbf{L}, \mathcal{S}_r) \approx 0$, the true error risk over the whole unknown distribution will approach to 0 with convergence rate $O(1/\sqrt{n})$.

2.5. Experiment Results

2.5.1. Experimental Setup

Dataset. To evaluate our proposed method, we conduct thorough experiments on four widely-used multi-shot benchmarks: the CAVIAR [**15**], the PRID 2011 [**35**], the iLIDS-VID [**95**] and the Market-1501 [**119**] datasets. The CAVIAR dataset contains 1220 images of 72 individuals from two non-overlapped cameras in a shopping mall. For the 72 individuals, 50 of them appear in both camera views and the remaining 22 persons only appear in one camera view. Each identity has 10 to 20 images and the resolutions vary from 17×39 to 72×144. The PRID 2011 dataset consists of video pairs recorded from two static surveillance cameras. There are 385 persons recorded in camera view A, as well as 749 persons in camera view B. Among all the persons, 200 persons are recorded in both camera views. The videos in PRID 2011 have 5 to 675 image frames, with an average of 100 for each. The iLIDS-VID dataset is generated from images captured in a busy airport arrival hall so the videos suffer severe occlusions caused by people and luggages. 600 videos of 300 randomly sampled people are recorded so that each person has one pair of videos from two different non-overlapped camera views. The video in iLIDS-VID is comprised of 23 to 192 image frames, with an average of 73 for each. The Market-1501 [**119**] is the latest and biggest benchmark dataset to date which contains 32668 bboxes of 1501 identities. Each person is recorded by six cameras at most, and two at least.

Feature. In all the experiments, only the image-level appearance feature descriptor is utilized. The high-dimensional feature LOMO [**51**] is adopted as the visual feature representation. Since it is not practical to directly use such a high dimensional feature in metric learning, we

employ principal component analysis (PCA) to reduce the feature dimension to a reasonable scale, 2000 dimensions.

Setting. To conduct fair comparisons, we follow the same experimental protocols as in [**103, 13, 95, 109**]. For the 50 persons who are captured by both cameras in CAVIAR, we randomly select 14 of them for training [2] and the remaining 36 persons are used for testing. As for the PRID 2011, we only utilize the 200 persons who appear in both cameras. For iLIDS-VID, the 300 persons are randomly divided into 150 for training and the other 150 for testing, so that there are $p = 36$, $p = 100$ and $p = 150$ individuals in the test sets of CAVIAR, PRID 2011 and iLIDS-VID respectively. As for the Market-1501 dataset, the pre-determined 12936 images from 750 identities are used for training, and the other 19732 images from disjointed identity set are for testing. In order to get statistically reliable results, 10 times random-splitting procedures are repeated to report the average performance. The **multi-shot** evaluation is adopted to report the Cumulated Matching Characteristic (CMC) results. The weighting parameter λ in Eqn. 2.8 is chosen as $\lambda = 0.01$ for all the experiments, which empirically produces both small training errors and stable solutions.

State-of-the-art. For the comparison experiments, we select three state-of-the-art metric learners: MLAPG [**52**], XQDA [**51**], DNSL [**112**] whose code is publicly available and the feature descriptor can be replaced. We compare our method with the above approaches under the completely same experimental setting and using the same LOMO feature. In addition, the results reported in the most recent papers are also presented for a thorough comparison.

[2]Training set of CAVIAR also includes the other 22 single-camera-view persons, so totally 36 persons are used for training)

Method	**Ave Time**	Method	**Ave Time**
$\mathcal{T}_C$-L	**0.03**	$\mathcal{T}_C$-K	**0.17**
$\mathcal{T}_M$-L	**0.37**	$\mathcal{T}_M$-K	**0.54**
$\mathcal{T}_A$-L	**4.86**	$\mathcal{T}_A$-K	**4.14**

Table 2.1. Comparison of training time (seconds) on CAVIAR. **-L** meas linear model, and **-K** means kernelized model.

2.5.2. The Learning Efficiency Analysis

In order to validate the learning efficiency of the used reference-driven regression scheme, a running cost experiment is firstly conducted on a small-size dataset, CAVIAR. Different reference generation schemes are tested for both linear and kernelized learning scenarios. Table. 2.1 shows the average training time of 10 random trials on CAVIAR. All the experiments are conducted on the same desktop PC with an Intel i7-2600 @3.40GHz CPU and 8G memory.

As we analyzed in Sec. 2.3.3, the computational complexity of learning the metric $\mathbf{M}$ is quadratic to the training sample number n or data dimension d. Table. 2.2 shows the comparison results of training time with other state-of-the-art learners on the large-size benchmark, Market-1501. All the experiments are conducted on a remote server with an Intel i7-5930K @3.50GHz CPU and 32G memory. [3] Compared with the other metric learners, our models are the most efficient except the $\mathcal{T}_A$-based ones which are a little slower than the kLFDA. This is because the optimization procedure requires computing the cost matrix $\mathbf{C}$ which is pretty time-consuming for a large number of data. And it is worth mentioning that the DNSL [**112**] also has a closed-form solution, but it requires many times of SVD operation for the kernelized data matrix, which is indeed time-consuming.

[3]The overall training time of our method includes the reference constraint generation, data kernelization and metric learning steps.

Method	XQDA	MLAPG	kLFDA	DNSL	$\mathcal{T}_C$-L
Training Time	3233.8	2732.8	995.2	3149.7	1.32
Method	$\mathcal{T}_M$-L	$\mathcal{T}_A$-L	$\mathcal{T}_C$-K	$\mathcal{T}_M$-K	$\mathcal{T}_A$-K
Training Time	290.08	1194.2	166.29	446.78	1319.2

Table 2.2. Comparison of training time (seconds) on Market-1501.

Method	$\frac{1}{n}\|\mathbf{XL}-\mathbf{R}\|_{\mathcal{F}}^2$	Method	$\frac{1}{n}\|\mathbf{XL}-\mathbf{R}\|_{\mathcal{F}}^2$
$\mathcal{T}_C$-L	**0.189**	$\mathcal{T}_C$-K	**1.1e-04**
$\mathcal{T}_M$-L	**0.261**	$\mathcal{T}_M$-K	**1.4e-04**
$\mathcal{T}_A$-L	**0.256**	$\mathcal{T}_A$-K	**1.3e-04**

Table 2.3. The average empirical training error on CAVIAR.

2.5.3. Empirical Training Error Verification

Theorem. 3 proves that with a sufficient number of samples, a low empirical error $\mathcal{E}(\mathbf{L}, \mathcal{S}_r)$ guarantees a low true risk $\mathcal{E}(\mathbf{L}, \mathcal{D}_r)$ with high probability. In the experiments, we study how large the empirical training error $\frac{1}{n}\|\mathbf{XL}-\mathbf{R}\|_{\mathcal{F}}^2$ actually is after learning. Taking the CAVIAR dataset as an example, we quantitatively verify that a low empirical training error can be obtained by our proposed algorithm. For a fair comparison, the training data are firstly normalized by $\{\hat{x_i} = x_i/\|x_i\|^2\}_{i=1}^n$ to get a constant-1 l_2-norm. The average training error of 10 random trials on the CAVIAR dataset under different algorithm settings is shown in Table. 2.3. The non-linear model has a much smaller training error than the linear ones since the non-linearity introduced by kernelization is able to better fit the high-dimensional feature space. The visualization result of affinity matrix refinement is shown in our supplementary material.

Input	Method	PRID 2011				iLIDS-VID				From
		R=1	R=5	R=10	R=20	R=1	R=5	R=10	R=20	
Image	TDL	30.20	59.10	74.00	88.40	9.81	27.52	46.10	62.19	[**109**]
	MLAPG(lomo)	45.60	58.20	63.80	69.80	30.54	45.58	53.02	60.78	[**52**]
	XQDA(lomo)	47.50	60.20	66.20	72.00	30.66	44.48	51.84	59.53	[**51**]
	DNSL(lomo)	51.00	63.40	68.60	74.10	24.44	34.11	39.68	46.85	[**112**]
	DVDL	40.60	69.70	77.80	85.60	25.90	48.20	57.30	68.90	[**41**]
	Salience	25.80	43.60	52.60	62.00	10.20	24.80	35.50	52.90	[**116**]
	KISSME	28.54	59.78	72.13	83.26	10.67	28.33	39.80	57.00	[**42**]
	LFDA	26.40	56.07	69.89	81.12	7.80	23.93	36.47	50.80	[**69**]
	LADF	8.20	20.45	29.89	42.25	4.33	14.00	21.20	32.13	[**50**]
	LDA	27.64	58.09	69.66	82.47	10.27	27.40	39.80	55.27	[**26**]
Video	SMP	80.90	95.60	98.80	99.40	41.70	66.30	74.10	80.70	[**59**]
	DGM+IDE	56.40	81.30	88.00	96.40	36.20	62.80	73.60	82.70	[**107**]
	CNN+KISS	69.90	90.60	-	98.20	48.80	75.60	-	92.60	[**118**]
	TDL	56.74	80.00	87.64	93.59	56.33	87.60	95.60	98.27	[**109**]
	Co&LBP+DVR	37.60	63.90	75.30	88.30	34.50	56.70	67.50	77.50	[**95**]
	KISSME	34.38	61.68	72.13	81.01	36.53	67.80	78.80	87.07	[**42**]
	LFDA	43.70	72.80	81.69	90.89	32.93	68.47	82.20	92.60	[**69**]
	LADF	47.30	75.50	82.69	91.12	39.00	76.80	89.00	96.80	[**50**]
	LDA	15.84	41.46	55.51	70.67	42.06	79.13	89.40	94.47	[**26**]
Linear	$\mathcal{T}_C$**-L**	**70.10**	**79.10**	**83.30**	**87.10**	**44.67**	**57.33**	**63.33**	**68.67**	Ours
	$\mathcal{T}_M$**-L**	**64.80**	**77.00**	**80.20**	**84.30**	**38.67**	**56.67**	**61.67**	**70.67**	Ours
	$\mathcal{T}_A$**-L**	**70.40**	**80.90**	**85.60**	**88.40**	**42.67**	**58.67**	**63.33**	**72.07**	Ours
Kernel	$\mathcal{T}_C$**-K**	**66.90**	**77.10**	**80.80**	**84.60**	**37.33**	**47.73**	**54.53**	**60.67**	Ours
	$\mathcal{T}_M$**-K**	**65.10**	**77.30**	**78.70**	**85.30**	**39.33**	**56.00**	**59.33**	**65.74**	Ours
	$\mathcal{T}_A$**-K**	**70.90**	**78.70**	**82.70**	**87.30**	**42.00**	**52.67**	**60.03**	**66.67**	Ours

Table 2.4. Comparison results on PRID 2011 and iLIDS-VID under the **multi-shot** and **video-based** matching settings.

2.5.4. Extensive Comparisons on Benchmarks

Due to the page limitation, the full CMC curves of comparison results are shown in the supplementary material.

Experiments on CAVIAR: Although the CAVIAR is a multi-shot dataset, most existing methods use it under the single-shot setting [**13, 58, 103**]. Due to the success of SsP-Rid on

Method	R@1	R@5	R@10	R@20
MLAPG(lomo)[**52**]	50.00	71.85	84.25	93.11
XQDA(lomo)[**51**]	51.18	75.59	90.33	96.86
DNSL(lomo)[**112**]	53.54	77.17	86.61	94.69
SSCDL-S[**58**]	49.10	80.20	93.50	97.90
MLAPG(lomo)-S[**52**]	40.60	71.70	83.30	95.70
XQDA(lomo)-S[**51**]	42.20	69.90	82.50	95.50
DNSL(lomo)-S[**112**]	47.60	75.66	87.37	96.20
MFA-χ^2-S[**103**]	40.20	70.20	83.90	95.10
EPKFM-S[**13**]	40.10	65.60	78.00	90.50
PCCA-χ^2_{RBF}-S[**103**]	33.20	65.90	81.90	95.20
LFDA-S[**69**]	32.00	56.30	70.70	87.40
LADF-S[**50**]	30.30	62.80	78.00	92.60
$\mathcal{T}_C$**-L**	**65.25**	**86.49**	**91.89**	**96.33**
$\mathcal{T}_M$**-L**	**70.90**	**88.73**	**93.24**	**98.36**
$\mathcal{T}_A$**-L**	**68.73**	**87.84**	**94.21**	**97.88**
$\mathcal{T}_C$**-K**	**66.80**	**88.61**	**94.02**	**97.30**
$\mathcal{T}_M$**-K**	**73.36**	**88.32**	**93.03**	**97.95**
$\mathcal{T}_A$**-K**	**61.02**	**84.36**	**92.47**	**96.72**

Table 2.5. Comparison results on CAVIAR under the **multi-shot** matching setting. **'-S'** means the single-shot result.

CAVIAR, we would like to also report the state-of-the-art single-shot results, including SSCDL [**58**], MFA-χ^2 [**103**], EPKFM [**13**], PCCA-χ^2_{RBF} [**103**], LADF [**50**] and LFDA [**69**]. It can be observed from Table. 2.5 that the proposed method outperforms the existing state-of-the-art algorithms with a significant improvement in both multi-shot and single-shot settings. For our models, the kernelized cases are slightly better than the linear cases except for the $\mathcal{T}_A$. The $\mathcal{T}_M$-K model performs the best, with a 37% relative improvement compared to the best player DNSL on Rank-1 accuracy. This is because the complex multi-modal data distribution of CAVIAR can be well captured by the $\mathcal{T}_M$ reference constraints.

Experiments on PRID 2011: The recent state-of-the-art results on PRID 2011 are shown in Table. 2.4. As we can see, all of our proposed reference-based methods consistently outperform

Method	Sing-Q	Multi-Q	From
	R@1	R@1	
Baseline	35.84	44.36	[**119**]
MLAPG(lomo)	38.80	61.33	[**52**]
XQDA(lomo)	44.80	55.82	[**51**]
DNSL(lomo)	51.73	57.70	[**112**]
KISSME(lomo)	40.50	N/A	[**119**]
MFA-χ^2(lomo)	45.67	N/A	[**103**]
kLFDA(lomo)	51.37	52.67	[**103**]
Hist-Loss	59.47	N/A	[**89**]
$\mathcal{T}_C$**-L**	**57.73**	**68.27**	Ours
$\mathcal{T}_M$**-L**	**54.67**	**64.53**	Ours
$\mathcal{T}_A$**-L**	**51.07**	**72.40**	Ours
$\mathcal{T}_C$**-K**	**63.20**	**73.87**	Ours
$\mathcal{T}_M$**-K**	**60.93**	**70.40**	Ours
$\mathcal{T}_A$**-K**	**56.03**	**68.93**	Ours

Table 2.6. Comparison results on Market-1501.

the state-of-the-art multi-shot based methods with a large margin. For the most important Rank-1 evaluation, the proposed $\mathcal{T}_A$-K model improves the performance with an impressive relative 39.0% improvement against the best player, DNSL. Although no temporal feature is used in our models, we are still able to achieve comparable, even better performance against the state-of-the-art video-based approaches which use both the temporal and appearance features together for learning.

Experiments on iLIDS-VID: For the iLIDS-VID dataset, the methods tested on the PRID 2011 benchmark are also compared here. As shown in Table. 2.4, our models achieve a significant improvement on Rank-1 evaluation against the other multi-shot based approaches, whose best Rank-1 performance is only 30.66%. Even compared to the video-based methods, our models still achieve comparable performances on Rank-1 accuracy. For the Rank-20 accuracy rate, the multi-shot based methods, including ours, can not compete against the video-based

methods. Because a lot of images in the iLIDS-VID dataset suffer severe occlusion from the background, which significantly deteriorates the appearance features, and thus degrades the identification rate. Under video-based setting, such bad influence might have been alleviated by considering the whole sequence as one probe/gallery.

Experiments on Market-1501: The comparison results on the Market-1501 benchmark are presented in Table. 2.6. The baseline [**119**] uses the BoW-based features and l_2-Norm distance. Besides, the state-of-the-art results based on the same LOMO feature are also included here for comparison (their detailed experimental settings might be slightly different). A recently proposed deep embedding-based method, Hist-Loss [**89**] is also compared. As can be seen, no matter under the single-shot or multi-shot scenarios, our methods outperform the others with a large margin improvement. On the Rank-1 evaluation, the proposed $\mathcal{T}_C$-K model improves the state-of-the-art from 59.47% to 63.20%.

2.6. Discussion

In this chapter, we propose a novel solution to the important yet challenging MsP-RID problem. In contrast to the existing metric learning-based MsP-RID methods which rely on the data similarity/dissimilarity constraints produced by both positive and negative samples, a novel linear-scaled constraint, called *reference constraint*, is proposed which assigns the given samples to the pre-determined reference points. Three different optimal transport-based schemes are proposed and studied to automatically generate the discriminative reference constraints. A regression-based metric learning model with a closed-form solution can be adopted to learn

a discriminative distance metric from the proposed reference constraints efficiently and effectively. Extensive experiments on the widely-used multi-shot benchmarks have clearly shown that our proposed approach is superior to the state-of-the-art algorithms.

However, there still are two main issues remaining in learning a global metric from our proposed reference constraints. The first one is that our proposed global learning method can not handle the extremely challenging small-size sample setting, that only one positive is available for each identity. Since the proposed reference constraint can not be constructed by using only one positive. Although some other constraints, like contrastive constraint, can work under such one-shot positive situation, only one positive constraint can be generated but hundreds of thousands negative constraints can be obtained, thus the learning of visual matching is totally dominated by the imbalanced constraints. Another critical issue is learning a single global metric can not handle the hard negative distractors in visual matching which are around the classification boundary. Since the learned single metric is not able to capture all the different instance-specific characteristics for different samples. Therefore, the above two issues motivate our work in Chapter 3.

CHAPTER 3

Learning From One-Shot Positive: Efficient Online Local Metric Adaptation From Negative Samples

3.1. Introduction

Recent attempts, including our work in Chapter 2, were based on learning a visual metric to better capture the visual similarities [**52, 51, 62, 122, 123**], and reported encouraging results. These methods typically attempt to train a faithful global metric offline, hoping to cover the enormous visual appearance variations so as to directly use it online for all test probes. The training data for such metric learning are generally sample pairs: a *positive* pair refers to two images of the same identity, and a *negative* pair otherwise. These methods usually demand a huge set of positive/negative training pairs to facilitate the learning. In practice, although it is relatively easy to collect negative pairs, it is in general difficult to obtain many positive pairs for a specific person. Therefore, the metrics learned from insufficient positive training data are likely to be biased. In addition, most methods [**52, 51, 123**] aim to learn a positive semi-definite (PSD) Mahalanobis metric, but it is computationally intensive to learn such a strictly PSD metric, while ignoring the PSD constraint leads to unstable and noisy metrics [**52**].

In contrast to these methods, this chapter advocates a different paradigm: **shifting part of the metric learning to the online local metric adaptation**. Specifically, for each online probe at the testing time, our new approach learns a dedicated local metric with a nominal computational cost. Combining a global metric with local metric adaptation achieves an adaptive

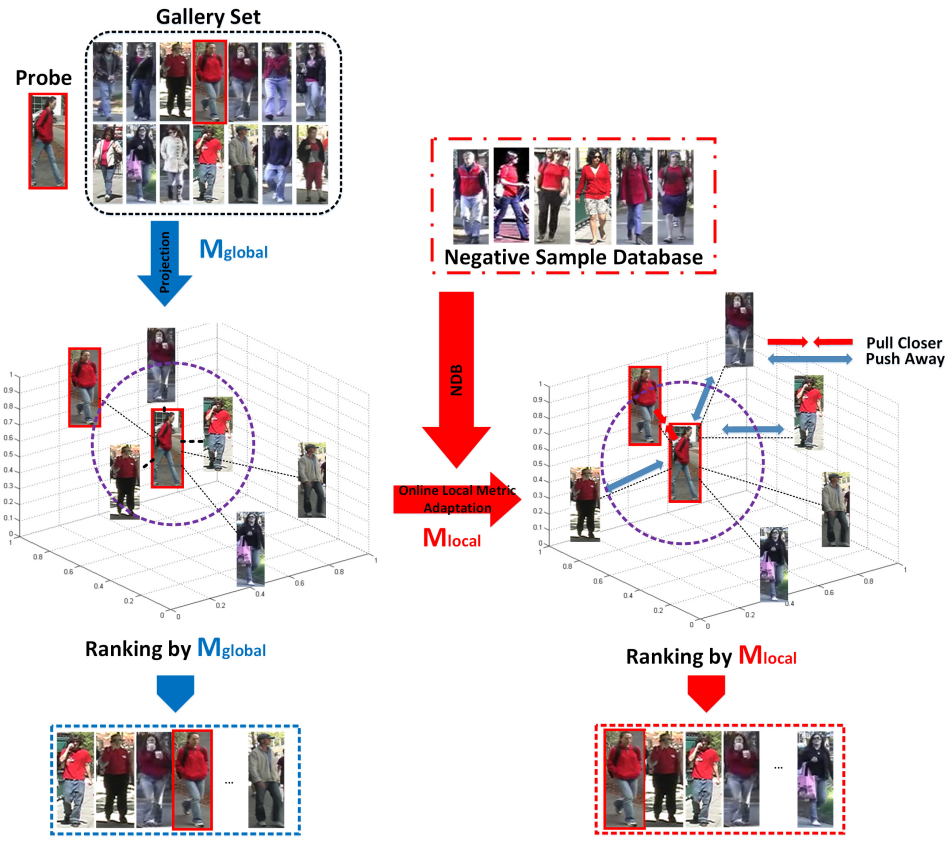


Figure 3.1. The overall idea of our proposed online local metric adaptation algorithm. Unlike existing methods that learn a single global metric for all probes, we exploit negative samples to learn a dedicated local metric for each online probe.

nonlinear metric. In our approach, its online learning is special, because there are no positive training pairs available at all for the testing probe, as its identity is unknown.

An attractive property of our proposed method is that it only uses negative data from a negative sample database (NDB) for adaptation learning. We call it **OL-MANS** for short of **O**nline **L**ocal **M**etric **A**daptation via **N**egative **S**amples. For a given testing probe, a subset of samples from NDB are selected to form informative negative pairs with this testing probe. These utilized samples from NDB are visually similar to the probe, but are guaranteed to have different

identities from the probe (at least with a very large probability). These negative samples provide effective local discrimination for further constraining the local metric tuning, by pushing away local false positives (shown in Fig. 3.1). For each testing probe, our method learns a strictly PSD local metric efficiently, via solving a kernel SVM problem. Comparing to offline global learning, the computational cost of the proposed online adaptation is negligible. Moreover, our method is generally applicable to be used on top of any global metric.

Another significant property of our proposed OL-MANS is that it is justified and backed up with a theoretical guarantee to improve the performance of the underlying global metric. This chapter gives in-depth theoretical analysis to well justify the proposed method. We first prove that this new method guarantees the reduction of classification error asymptotically when there are an infinite number of training data. Then we pursue the best approximation of the asymptotic case by using a finite number of training data, since we can prove that the learning objective of the proposed local metric adaptation is equivalent to the optimal approximation of the asymptotic case. In addition, we also provide consistency and sample complexity analysis to guarantee the generalization ability of our proposed OL-MANS method. These theoretical analyses indicate that the learned local metric is bound to improve the P-Rid performance. These properties have been confirmed to be very effective and practical by our extensive experiments and comparative studies on almost all the important P-Rid benchmarks (VIPeR, QMUL GRID, CAVIAR, iLIDS, P-Rid 450S, CUHK Campus, CUHK03 and Market-1501).

The rest of this chapter is organized as follows: Section. 3.2 summarizes the previous works on P-Rid and metric learning. We describe our proposed algorithm in Section. 3.3, and illustrate its performance on many benchmark datasets in Section. 3.5. In Section. 3.4, we theoretically analyze some important properties of our proposed algorithm.

3.2. Previous Work

3.2.1. Person Re-identification

Global Learning in P-Rid: Existing metric learning-based P-Rid methods either learn a single global metric or a local discrimination to facilitate identification. Zheng *et al.* [**122**] proposed a relative distance comparison method (PRDC) to maximize the probability of a positive pair to have a smaller distance than a negative pair. Hirzer *et al.* [**36**] relaxed the PSD constraint to simplify the computation. Liao *et al.* [**51**] learned a discriminant subspace and a global distance metric simultaneously for dimension reduction and optimal dimensionality. A logistic metric learning called MLAPG was proposed by Liao *et al.* [**52**] for a global PSD metric via an asymmetric sample weighting strategy.

Local Learning in P-Rid: Many methods are based on local learning strategies. Zhang *et al.* [**111**] formulated the P-Rid problem as a local distance comparison problem to handle the multi-modal distributions of the visual appearances. Li *et al.* [**50**] proposed the Locally-Adaptive Decision Functions (LADF) which integrates a traditional distance metric with a local decision rule. Pedagadi *et al.* [**69**] employed the Local Fisher Discriminant Analysis (LFDA) which combines the fisher discriminant analysis (FDA) and Local Preserving Projections (LPP) to exploit the local geometrical information of samples. Liong *et al.* [**53**] developed a regularized local metric learning (RLML) method to combine global and local metrics, so as to utilize the local data distribution to alleviate over-fitting. Zhang *et al.* [**114**] proposed LSSCDL to learn a specific SVM classifier for each training sample, then the weight parameters of a new sample can be inferred. A novel multi-task maximally collapsing metric learning (MtMCML) model was proposed by Ma *et al.* [**62**].

Figure 3.2. The improvement of ranking result by our OL-MANS on VIPeR [**31**]. BLUE boxes: input probes, RED: gallery targets. For each case, the top row is the result from the baseline [**52**], and the bottom row is our result. (Best view in color and enlarged)

Deep Learning in P-Rid: Recently, deep learning has shown excellent performance for P-Rid. The convolutional neural network-based P-Rid approaches aim to integrate the feature extraction and metric learning into one end-to-end framework, in which a neural network is built to extract from each pedestrian image a feature that satisfies certain ranking criterion. Li *et al.* [**47**] firstly utilized deep learning method to extract more effective and discriminative features to facilitate P-Rid. Ding *et al.* [**23**] proposed a scalable deep feature learning model for P-Rid via relative distance comparison based on triplet loss. Shi *et al.* [**76**] proposed a novel moderate positive mining method to embed robust deep metric for P-Rid. Ustinova *et al.* [**89**] suggested a new loss for learning deep embeddings and demonstrate competitive results of the new loss on a number of P-Rid datasets.

In contrast to the methods learning a global metric, our proposed method is mainly focused on learning local metrics specifically adaptive to individual testing probes. Different from RLML that requires clustering in advance to obtain the local data distributions, our new

approach does not need clustering but is rather instance-based learning, and thus avoiding the risk of inaccurate clustering results. Also note that MtMCML learning still follows the global manner although it learns different metrics for different cameras. In contrast to LADF that needs a large number of positive sample pairs to drive the local decision function learning, our new approach only uses negative sample pairs which are much easier to obtain. LSSCDL also requires a lot of positive training pairs for offline learning, but ours performs online learning per probe without the requirement of positive pairs. By considering the deep learning-based method as a global mapping learning framework, our proposed OL-MANS method can be readily applied on top of it to further boost the performance.

3.2.2. Metric Learning

Metric learning is an active research area and Mahalanobis distance learning becomes more and more important since [**102**]. Among these approaches, the Large Margin Nearest Neighbor Learning (LMNN) [**99**] has reported outstanding performances. It learns a Mahalanobis metric to improve the k-Nearest neighbor classifier, by pulling together the data from the same class, while pushing away data from different classes by a large margin. Besides LMNN, Information Theoretic Metric Learning (ITML) [**22**] and Logistic Discriminant Metric Learning (LDML) [**32**] are other effective methods in Mahalanobis distance learning.

There have been recent advances in learning local metric. Discriminant Adaptive Nearest Neighbor classification (DANN) [**33**] attempts to learn local metrics by shrinking neighborhoods in directions orthogonal to the local decision boundaries and enlarging the neighborhoods parallel to the boundaries. Generative Local Metric Learning (GLML) [**67**] learns local metrics by minimizing the expected classification error of nearest neighbor classifier to alleviate

overfitting. Unlike these methods that learn a number of local unrelated metrics, Parametric Local Metric Learning (PMLM) [**94**] exploits an instance-based learning strategy, aiming to find an adaptive local Mahalanobis metric for each data point. It shares a similar idea as in the Exemplar-SVMs [**63**], which is a conceptually simple but surprisingly powerful method to learn a discriminative object classifier. Exemplar-SVMs is defined over a single positive instance and millions of negatives, and thus the learned classifier is specific to its positive exemplar.

Our proposed approach attempts to achieve online adaptive local metric tuning on top of a global metric. For this approach, the choices for the global metric are flexible, and its novelty lies in the proposed local metric tuning method that only exploits negative data. It is related to but sufficiently different from Large Margin Nearest Neighbor Learning (LMNN) [**99**], Generative Local Metric Learning (GLML) [**67**] and Exemplar-SVMs [**63**]. Compared to LMNN, the new method does not demand positive samples and it is much more weakly supervised. Our method is completely different from PMLM, as PMLM uses global constraints and cannot work with weakly supervised data as in the proposed method. In addition, our proposed method is different from Exemplar-SVMs because we learn local Mahalanobis metrics that are positive semi-definite.

3.2.3. Online Re-ranking

Re-ranking technique [**106, 105, 55, 44, 28, 2, 16, 127**] has been initially studied in the instance retrieval tasks, recently the P-Rid community has paid tremendous attention to how to boost re-identification accuracy via re-ranking. Some works [**93**] requires human inter-action to derive re-ranking. Li *et al.* [**45**] tried to re-rank the initial result by analyzing both the relative and direct information of near neighbor of the images. Garcia *et al.* [**28**] proposed an unsupervised

re-ranking model by taking advantage of the content and context information in the ranking list. Leng *et al.* [**44**] aimed to re-evaluate the initial ranking list by performing a novel bidirectional ranking algorithm with a fusion similarity of both content and contextual similarity. Ye *et al.* [**105**] revised the rank list by considering both the nearest neighbors of the global and local features. And in [**106**], in order to compute both the similarity and dissimilarity, a k-nearest neighbor set is adopted to different methods. Zhong *et al.* [**127**] proposed a k-reciprocal encoding method to re-rank the initial re-ID rank list based on a hypothesis that gallery image is similar to the probe in the k-reciprocal nearest neighbors, it is more likely to be a true match.

Both our proposed method and re-ranking share the same appealing online manner, but our algorithm outperforms the re-ranking by several unique merits which will be detailedly discussed in Sec. 3.3.5.

3.3. Our Solution: Online Local Metric Adaptation From Negative Samples

3.3.1. Problem Setup

A single-shot P-Rid dataset consists of n pairs of identity images $\{(x_i^p, x_i^g)\}_{i=1}^n$ collected from two different disjoint cameras: x_i^p is from the probe camera and x_i^g is from the gallery camera. The index $i = \{1, 2, ..., n\}$ represents the identity label of n different persons. For training and testing in P-Rid, all identity pairs can be divided into two disjoint subsets $\{u_1, u_2, ..., u_{m'}\}$ and $\{v_1, v_2, ..., v_m\}$ where $n = m + m'$ and

$$\begin{aligned} \mathbf{X}_{train} &= \mathbf{X}_{train}^p \cup \mathbf{X}_{train}^g = \left\{x_{u_i}^p\right\}_{i=1}^{m'} \cup \left\{x_{u_i}^g\right\}_{i=1}^{m'} \\ \mathbf{X}_{test} &= \mathbf{X}_{test}^p \cup \mathbf{X}_{test}^g = \left\{x_{v_i}^p\right\}_{i=1}^{m} \cup \left\{x_{v_i}^g\right\}_{i=1}^{m} \end{aligned} \tag{3.1}$$

So that $\mathbf{X}_{train}$ is used as the training set and $\mathbf{X}_{test}$ is the test set. In our algorithm, an additional negative sample database, denoted by $\mathbf{Y}^{neg} = \{y_i\}_{i=1}^{k}$, is needed, and will be discussed shortly in Sec. 3.3.3.

3.3.2. Conventional Global Metric Learning

Conventional learning-based P-Rid methods [**51, 122, 95, 52**] aim to learn a single global Mahalanobis distance metric $\mathbf{M}_G$ by using the training set $\mathbf{X}_{train}$. The learned metric $\mathbf{M}_G$ projects the original samples into another feature space, and the matching between one probe x_i^p and one gallery image x_j^g at test stage is measured by:

$$d_{\mathbf{M}_G}\left(x_i^p, x_j^g\right) = \left\|x_i^p - x_j^g\right\|_{\mathbf{M}_G}^2 = \left(x_i^p - x_j^g\right)^T \mathbf{M}_G \left(x_i^p - x_j^g\right) \tag{3.2}$$

where $\mathbf{M}_G = \mathbf{W}^T\mathbf{W} \succeq 0$ needs to be positive semi-definite, as $\mathbf{W}$ is the learned projection. Different methods adopt different loss functions to learn $\mathbf{M}_G$, and a good solution to $\mathbf{M}_G$ should align the similarity structure in the projected feature space, so as to pull the samples from the same identity group closer and to make different identities more discriminative. Due to the fact that the global metric does not aim to fit the local distributions for all the samples specifically, it may lead to large biases and distortions in some places in the feature space. As illustrated in Fig. 3.1, our new approach puts an instance-based online local metric adaptation on top of the global metric.

3.3.3. Instance-Specific OL-MANS

In this section, we propose an online local metric adaptation algorithm called *OL-MANS* to adaptively adjust the metric dedicated to specific test probes with minimum online training by utilizing only negative training samples.

Specifically, for a probe image $x^p_{v_i}$ in the probe set $\mathbf{X}^p_{test}$, we aim to learn a local Mahalanobis distance $\mathbf{M}^i_L$ only using the samples in a negative sample database $\mathbf{Y}^{neg}$ as training data. This negative sample database provides rather faithful negative samples to the tests with a large probability. There are many ways to collect $\mathbf{Y}^{neg}$, e.g., data from a different benchmark can be used, or false positive matches from images that do not contain humans. The insight here is that all such negative samples are "hard negatives" for the probes. In this research, we have investigated how $\mathbf{Y}^{neg}$ influences the performance in the experiment section.

As the global projection $\mathbf{W}$ learned by the global metric learning maps $\mathbf{X}^p_{test}$ to a low dimensional subspace $\hat{\mathbf{X}}^p_{test} = \mathbf{W}\mathbf{X}^p_{test} = \left\{\hat{x}^p_{v_i}\right\}_{i=1}^m$, we propose to further adjust the local similarity for each specific $\hat{x}^p_{v_i}$ by an online learned local metric $\mathbf{M}^i_L$ which is solely learned from $\mathbf{Y}^{neg}$. In other word, for one probe $\hat{x}^p_{v_i}$, only the negative samples and the probe are used to learn its local adaptation $\mathbf{M}^i_L$.

We propose to pursue an optimal PSD Mahalanobis metric $\mathbf{M}^i_L$ for the local adaptation, by maximizing the distance to the closest (or "hardest" conceptually) negative sample of $\hat{x}^p_{v_i}$, as shown in Fig. 3.3:

$$\mathbf{M}^i_L = \arg\max_{\mathbf{M}^i_L \succeq 0} \left(\min_{1 \le j \le k} \left(\hat{x}^p_{v_i} - \hat{y}_j\right)^T \mathbf{M}^i_L \left(\hat{x}^p_{v_i} - \hat{y}_j\right) \right) \tag{3.3}$$

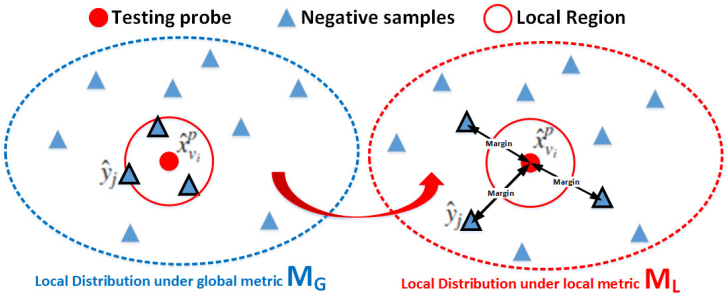


Figure 3.3. The local metric $\mathbf{M}_L^i$ for $\hat{x}_{v_i}^p$ can push the closest negative sample $\hat{y}_j$ of $\hat{x}_{v_i}^p$ away from the local region $\Omega(\hat{x}_{v_i}^p)$

where $\hat{y}_j = \mathbf{W}y_j$ is the projected negative sample based on the global metric. We regularize $\mathbf{M}_L^i$ for a stable solution. This can be done via minimizing the norm under a fixed margin constraint, instead of maximizing the margin under a fixed norm constraint [**25**], so the alternative objective is:

$$\begin{aligned} &\mathbf{M}_L^i = \arg\min_{\mathbf{M}_L^i} \frac{1}{2}\left\|\mathbf{M}_L^i\right\|^2 \\ &sub\ to:\ \left(\hat{x}_{v_i}^p - \hat{y}_j\right)^T \mathbf{M}_L^i \left(\hat{x}_{v_i}^p - \hat{y}_j\right) \geq 2,\ \forall 1 \leq j \leq k \\ &\mathbf{M}_L^i \succeq 0 \end{aligned} \tag{3.4}$$

where the constant 2 is arbitrary only for manipulation convenience. While this is a convex semi-definite programming problem, it can be very slow for high dimensional data, even for the state-of-the-art PSD solvers.

In the proposed OL-MANS approach, we relax the PSD constraint requiring $\mathbf{M}_L^i \succeq 0$, but we prove below that the relaxed objective is equivalent to a kernel SVM problem with a quadratic kernel. And thus the solution is still a PSD metric. In addition, it can be readily solved with off-the-shelf SVM solvers such as LIBSVM [**10**]. More importantly, we also prove that

this learning objective is equivalent to the best approximation to the asymptotic classification error, which is proved to be lower than the global metric (details see Sec. 3.4).

Theorem 4. *The solution to Eqn. 3.4 is equivalent to a kernel SVM with* $k(x,y) = \langle x,y\rangle^2$ *on* $\{\tilde{y}_0, \tilde{y}_1, \tilde{y}_2, ..., \tilde{y}_k\}$ *where* $\tilde{y}_j = \hat{x}^p_{v_i} - \hat{y}_j$ *(for* $j \geq 1$*), and* $\tilde{y}_0 = \hat{x}^p_{v_i} - \hat{x}^p_{v_i} = 0$.

PROOF. Define auxiliary labels by:

$$\zeta_j = \begin{cases} -1, \ j = 0 \\ 1, \ j \neq 0 \end{cases} \tag{3.5}$$

so the objective Eqn. 3.4 can be rewritten as:

$$\begin{aligned} &\mathbf{M}^i_L = \arg\min_{\mathbf{M}^i_L} \frac{1}{2}\left\|\mathbf{M}^i_L\right\|^2 \\ &sub\ to: \ \zeta_j\left(\tilde{y}_j^T \mathbf{M}^i_L \tilde{y}_j - 1\right) \geq 1, \forall\ 0 \leq j \leq k \end{aligned} \tag{3.6}$$

Eqn. 3.6 is exactly an SVM problem with quadratic kernel and with bias fixed to one. Next we prove the solution to objective Eqn. 3.6 is exactly the same as that to the original objective Eqn. 3.4. Consider the dual of the SVM, the optimal solution $\mathbf{M}^i_L$ has the form:

$$\mathbf{M}^i_L = \sum_{j=0}^{k} \alpha_j \zeta_j \tilde{y}_j \tilde{y}_j^T, \ \alpha_j \geq 0 \tag{3.7}$$

Since $\tilde{y}_j \tilde{y}_j^T$ is PSD for $j \geq 1$ ($\tilde{y}_0 \tilde{y}_0^T = 0$) and $\zeta_j = 1$ for $j \geq 1$, so we have:

$$\mathbf{M}^i_L = \sum_{j=0}^{k} \alpha_j \zeta_j \tilde{y}_j \tilde{y}_j^T = \sum_{j=1}^{k} \alpha_j \tilde{y}_j \tilde{y}_j^T \succeq 0 \tag{3.8}$$

□

It is obvious that the positive semi-definiteness of $\mathbf{M}_L^i$ is guaranteed even if no PSD constraint is explicitly imposed in our learning objective Eqn. 3.6.

3.3.4. Person Re-identification via OL-MANS

At the online test stage, for a probe $x_{v_i}^p$ from $\mathcal{X}_{test}^p$ and one gallery image $x_{v_j}^g$ from $\mathcal{X}_{test}^g$, our method combines a global metric $\mathbf{M}_G$ (with flexible choices) with our local metric adaptation $\mathbf{M}_L^i$ to achieve an adaptive nonlinear metric:

$$\begin{aligned} d(x_{v_i}^p, x_{v_j}^g) &= d_{\mathbf{M}_G}(x_{v_i}^p, x_{v_j}^g) + \lambda d_{\mathbf{M}_L}(x_{v_i}^p, x_{v_j}^g) \\ &= (x_{v_i}^p - x_{v_j}^g)^T \mathbf{W}^T (\mathbf{I} + \lambda \mathbf{M}_L^i) \mathbf{W} (x_{v_i}^p - x_{v_j}^g) \end{aligned} \tag{3.9}$$

where $\mathbf{M}_G = \mathbf{W}^T\mathbf{W}$ is an learned global metric and $\mathbf{M}_L^i$ is the local metric adaptation specific for $x_{v_i}^p$. λ is the weighting parameter which can be decided by cross-validation. In our work, we set λ by Eqn. 3.10 in all the experiments which can be explained in Sec. 3.5.

$$\lambda = \max_{1 \le j \le m'} \left(d_{\mathbf{M}_G}(x_{v_i}^p, y_{v_j}^g) \right) / \max_{1 \le j \le m'} \left(d_{\mathbf{M}_L}(x_{v_i}^p, y_{v_j}^g) \right) \tag{3.10}$$

At first, we find that even simply using only the learned local metric for re-identification, the results are still much better than using the original global metric. Further, when combining the global and local metrics, we are able to obtain much better and more stable performances. The reason behind it can be explained by the idea of boosting. Either the global metric or the local metric can be considered as a "weak" classifier for P-Rid, and their combination forms a "stronger" classifier. As proved by the boosting theory, this combination is able to improve the classification error.

3.3.5. OL-MANS vs Re-ranking

Both our proposed OL-MANS algorithm and the re-ranking technique can be readily combined with the other P-Rid methods in the online phase without modifying the original P-Rid framework. But our OL-MANS owns more unique merits than re-ranking in both the efficiency and effectiveness facets. More comparison experiment results will be presented in Sec. 3.5.

Data: Most re-ranking methods require no additional training samples, but utilize the given testing probe and gallery sets to help refine the ranking. In contrast, our OL-MANS takes advantage of a set of easily-available negative samples, based on which it finds online adaptation for the optimal local metric.

Effectiveness: The effectiveness of re-ranking depends heavily on the quality of the initial ranking list (if the true match is not in the top-k ranks). It may hurt the initial rank result, because the true match may have a lower rank after re-ranking if the false matches are included in the top-k list. Thus re-ranking may degrade the performance. The performance of our OL-MANS model relies on the quality of the set of negative data, as illustrated by Theorem. 5, even if the quality of the given NDB is pretty bad (no hard negatives are provided), OL-MANS still won't degrade the original performance. Comparing to re-ranking, our OL-MANS has a unique and plausible advantage: it does not degrade the performance of the original methods (the original global metric) in theory. As indicated in the objective Eqn. 3.4, when the negative samples are not good (i.e., they are already far away from the positive point under the original global metric), the learned local metric $\mathbf{M}_L$ will be the same as the original global metric $\mathbf{M}_G$, since the constraints in Eqn.4 have already been fulfilled by $\mathbf{M}_G$. So OL-MANS won't give a worse performance than the original method. As described in Sec. 3.4, our theoretical analysis has

shown that asymptotically our negative-augmented approach always improves the identification performance, and can be very close to the Bayesian error.

Efficiency: Another merit of our OL-MANS compared with re-ranking is its high efficiency. OL=MANS is very efficient even if there are a lot of negative samples available for local adaptation. Because the learned local metric $\mathbf{M}_L$ is only related to a handful set of hard negatives, not all the negatives. In contrast, other methods, such as re-ranking (depend on data number and nearest neighbor number k), transfer learning, domain adaptation techniques, are usually time-consuming.

3.4. Theoretical Analysis and Justification

We first prove that the asymptotic error by using the proposed OL-MANS is bound to be lower than that without. When the negative samples are truly hard negative ones, the asymptotic error by using OL-MANS can be very close to the Bayesian error (Sec. 3.4.1). Besides this theoretically meaningful result, we prove that this strong asymptotic error can actually best approximated by using finite data, which is practically also meaningful. More importantly, we prove that this approximation is actually achieved by OL-MANS (Sec. 3.4.2). We also present its consistency and sample complexity analysis in Sec. 3.4.3.

3.4.1. Asymptotic Error is Reduced

The core of P-Rid is indeed a two-class (ω_+ and ω_-) 1-Nearest neighbor (NN) classification problem by using the gallery set $\mathcal{D}$. If there is infinite number of data, it is well-known that its asymptotic error $\mathcal{P}(e|x)$ is bounded by 2 times the Bayesian error [**20**]:

$$\mathcal{P}^* \leq \mathcal{P}(e|x) = 2P(\omega_+|x)P(\omega_-|x) \leq 2\mathcal{P}^* \tag{3.11}$$

where $\mathcal{P}^*$ is the Bayesian error. In our work, we prove that by adding the hard negative samples x_a to $\mathcal{D}$ to form an augmented dataset $\mathcal{D}^a$, the asymptotic error $\mathcal{P}^a(e|x)$ by using $\mathcal{D}^a$ is always smaller than $\mathcal{P}(e|x)$:

$$\mathcal{P}^a(e|x) \leq \mathcal{P}(e|x) \tag{3.12}$$

Theorem 5. *For an input x, its NN is x' in $\mathcal{D}^a$. Define the probability that x' is an augmented data x_a, i.e., $x' \sim x_a$ as $P(x' \sim x_a) = q$; otherwise, x' is not an augmented data x_a, i.e., $x' \neg x_a$, $P(x' \neg x_a) = 1 - q$, where $0 \leq q \leq 1$. The asymptotic error $\mathcal{P}^a(e|x)$ by using $\mathcal{D}^a$ is:*

$$\mathcal{P}^a(e|x) = \frac{(2-q)\mathcal{P}(e|x)}{2 - 2q\mathcal{P}(e|x)} \leq \mathcal{P}(e|x) \tag{3.13}$$

PROOF. Denote by $\mathcal{D}$ the original data set (i.e., the gallery set), and by D^a the augmented data set by adding hard negatives. Except for the augmented hard negative data (denoted by x_a), the rest in $\mathcal{D}^a$ are the same as $\mathcal{D}$.

Let's consider the two-class 1-NN classification, without losing the generality. The asymptotic error for 2-class 1-NN using $\mathcal{D}$ is

$$\mathcal{P}(e|x) = 2P(\omega_+|x)P(\omega_-|x)$$

Let's consider the asymptotic error for 2-class 1-NN using $\mathcal{D}^a$. We denote it by $P^a(e|x)$. Our goal is to prove:

$$\mathcal{P}^a(e|x) \leq \mathcal{P}(e|x) \tag{3.14}$$

The prove is the following. For an input x, its nearest neighbor x' in $\mathcal{D}^a$ (denoted by $x \sim x'$) has two cases:

- case 1: the nearest neighbor of x' is an augmented data x_a, i.e., $x' \sim x_a$. Its probability $P(x' \sim x_a) = q$;
- case 2: the nearest neighbor of x' is not an augmented data x_a, i.e., $x' \neg x_a$. Its probability $P(x' \neg x_a) = 1 - q$.

We denote the nearest neighbor of x' in $\mathcal{D}$ by $x''(x')$. There are two cases. If x' is x_a, then its nearest neighbor in $\mathcal{D}$ is $x''(x_a)$ whose class label is ω_+ (because x_a are all hard negative samples). If x' is not x_a, then its nearest neighbor in $\mathcal{D}$ is $x''(x') = x'$.

Now we consider the asymptotic probability of assigning ω_+ to x'. In case 1, we need to guarantee both x and $x''(x')$ to be ω_+ (i.e., the hard negative data is actually useful). In case 2, we only need to guarantee x to be ω_+ (i.e., the hard negative data is not effective). So we have:

$$\phi(\omega_+|x') \propto P^2(\omega_+|x)q + P(\omega_+|x)(1-q)$$

Similarly, the asymptotic probability of assigning ω_- to x' is:

$$\phi(\omega_-|x') \propto P^2(\omega_-|x)q + P(\omega_-|x)(1-q)$$

Because $\phi(\omega_+|x') + \phi(\omega_-|x') = 1$, we have:

$$\begin{aligned} \phi(\omega_+|x') &= \frac{P^2(\omega_+|x)q + P(\omega_+|x)(1-q)}{(1-q) + q[1 - 2P(\omega_+|x)P(\omega_-|x)]} \\ \phi(\omega_-|x') &= \frac{P^2(\omega_-|x)q + P(\omega_-|x)(1-q)}{(1-q) + q[1 - 2P(\omega_+|x)P(\omega_-|x)]} \end{aligned} \tag{3.15}$$

Therefore, we can compute the 2-class 1-NN asymptotic error for x on $\mathcal{D}^a$:

$$\begin{aligned}\mathcal{P}^a(e|x) &= \phi(\omega_+|x')P(\omega_-|x) + \phi(\omega_-|x')P(\omega_+|x) \\ &= \frac{[2(1-q)+q]P(\omega_+|x)P(\omega_-|x)}{(1-q)+q[1-2P(\omega_+|x)P(\omega_-|x)]} \\ &= \frac{(2-q)\mathcal{P}(e|x)}{2(1-q)+2q(1-\mathcal{P}(e|x))} \\ &= \frac{(2-q)\mathcal{P}(e|x)}{2-2q\mathcal{P}(e|x)}\end{aligned} \tag{3.16}$$

Because $0 \leq q \leq 1$, it is easy to see:

$$\begin{aligned}if\ \mathcal{P}^a(e|x) \leq \mathcal{P}(e|x) &\Leftrightarrow \frac{(2-q)P(e|x)}{2-2qP(e|x)} \leq P(e|x) \\ &\Leftrightarrow (2-q) \leq 2-2qP(e|x) \\ &\Leftrightarrow 2qP(e|x) \leq q \\ &\Leftrightarrow P(e|x) \leq \frac{1}{2}\end{aligned} \tag{3.17}$$

Since the error rate $P(e|x) \leq \frac{1}{2}$ is always true, the proof in Eqn. 3.17 is true, and Theorem. 5 holds. □

Since q is the probability of $P(x' \sim x_a)$, $0 \leq q \leq 1$. If $q = 0$ which indicates that the augmented negative data are useless, then we have $P^a(e|x) = P(e|x)$. Another extreme is when $q = 1$ implying the negative data are abundant and effective to constrain the classification, then we have [1]

$$\mathcal{P}^a(e|x) = \frac{\mathcal{P}(e|x)}{2[1-\mathcal{P}(e|x)]} \leq \mathcal{P}(e|x) \tag{3.18}$$

[1] $\mathcal{P}(e|x) \leq \frac{1}{2}$ is always true.

In this case, when $\mathcal{P}(e|x)$ is very small, we have

$$\mathcal{P}^a(e|x) \simeq \frac{\mathcal{P}(e|x)}{2} \simeq \mathcal{P}^*(e) \tag{3.19}$$

The asymptotic error of our negative-augmented approach can be very close to the Bayesian error.

3.4.2. Finite Approximation to $\mathcal{P}^a(e|x)$

The asymptotic error $\mathcal{P}^a(e|x)$ in Eqn. 3.13 is only meaningful when the sample size is infinite, $n \to \infty$. However, in practice, only finite number of samples are available. To make it practically meaningful, we prove that it can be best approximated by the practical error rate $\mathcal{P}_n(e|x)$ (n is finite) by finding a local metric $\mathbf{M}_L$. And this local metric turns out to be the one for the proposed OL-MANS.

Still consider the 2-class 1-NN rule scenario (on the negative-augmented data $\mathcal{D}^a$). To make the notation less cluttered, here we use $\mathcal{P}(e|x)$ to indicate $\mathcal{P}^a(e|x)$ without confusion. Given a sample x and its nearest neighbor x' from the finite dataset containing n samples. The probability of error for x is:

$$\begin{aligned}\mathcal{P}_n(e|x) &= P(\omega_+|x)P(\omega_-|x') + P(\omega_-|x)P(\omega_+|x') \\ &= \mathcal{P}(e|x) + [P(\omega_+|x) - P(\omega_-|x)][P(\omega_+|x) - P(\omega_+|x')]\end{aligned}$$

Our goal is to find a best local metric $\mathbf{M}_x$ for x such that the conditional MSE:

$$\min_{\mathbf{M}_x} \mathbb{E}\{[\mathcal{P}_n(e|x) - \mathcal{P}(e|x)]^2|x\}$$

is minimized. Since $[P(\omega_+|x) - P(\omega_-|x)]$ is constant for a given x, so the minimization is equal to:

$$\min_{\mathbf{M}_x} \mathbb{E}\{[P(\omega_+|x) - P(\omega_+|x')]^2|x\} \tag{3.20}$$

Because $P(\omega_+|x') \simeq P(\omega_+|x) + \nabla P(\omega_+|x)^T(x' - x)$, Eqn. 3.20 is approximately equivalent to:

$$\min_{\mathbf{M}_x} \mathbb{E}\{\|\nabla P(\omega_+|x)^T(x' - x)\|^2|x\} \tag{3.21}$$

The core here is to compute the gradient of posterior $\nabla P(\omega_+|x)$. Recall our proposed OL-MANS approach, a local linear classifier $\mathbf{w}$ where $\mathbf{M}_x = \mathbf{w}\mathbf{w}^T$ is learned for sample x via a standard kernel SVM framework. So the posterior of x in a logistic sigmoid function form is:

$$P(\omega_+|x) = \frac{1}{1 + e^{\zeta_x(\mathbf{w}^T x + b) - \gamma}}, P(\omega_-|x) = 1 - P(\omega_+|x) \tag{3.22}$$

The gradient of $P(\omega_+|x)$ can be easily computed:

$$\nabla P(\omega_+|x) = \zeta_x P(\omega_+|x) P(\omega_-|x) \mathbf{w} \tag{3.23}$$

Substituting Eqn. 3.23 for $\nabla P(\omega_+|x)$ in Eqn. 3.21 gives us:

$$\begin{aligned} &\min_{\mathbf{M}_x} \mathbb{E}\{\|\zeta_x P(\omega_+|x) P(\omega_-|x) \mathbf{w}^T(x' - x)\|^2|x\} \\ &= \min_{\mathbf{M}_x} (x' - x)^T \mathbf{w}\mathbf{w}^T(x' - x) \end{aligned} \tag{3.24}$$

Recall our optimization objective Eqn. 3.6, for the positive samples, we have $1 - (x' - x)^T\mathbf{M}_x(x' - x) \geq 1$ which is equal to $(x' - x)^T\mathbf{M}_x(x' - x) \leq 0$. On the other hand, $(x -$

$x')^T\mathbf{M}_x(x - x') \geq 0$ is always true for a PSD $\mathbf{M}_x$, so $(x' - x)^T\mathbf{M}_x(x' - x) \equiv 0$ always holds. It is obvious Eqn. 3.24 is always optimized by adopting the local metric $\mathbf{M}_x$ learned by our algorithm Eqn. 3.6.

3.4.3. Consistency and Sample Complexity Analysis

A set of samples $\{x_0, x_1, ..., x_k\}$ is identically drawn from a D-dimensional space $\mathbb{D} \in \mathbb{R}^D$ where l_i is the label of x_i, then a paired sample set $S_k^{pair} = \{s_i\}_{i=1}^k = \{(x_0, x_i)\}_{i=1}^k$ of size k is formed. For our proposed objective Eqn. 3.6, the true risk over the whole distribution $\mathbb{D}$ and the empirical error based on S_k^{pair} are defined as:

$$Err^{\lambda}(\mathbf{M}_L, \mathbb{D}) = \mathbb{E}_{x_i, x_j \sim \mathbb{D}} \phi^{\lambda}(\mathbf{M}_L, (x_i, x_j))$$

$$Err^{\lambda}(\mathbf{M}_L, S_k^{pair}) = \frac{1}{k} \sum_{i=1}^{k} \phi^{\lambda}(\mathbf{M}_L, s_i)$$

where $\phi^{\lambda}(\mathbf{M}_L, s_i)$ is the hinge loss function:

$$\phi^{\lambda}(\mathbf{M}_L, s_i) = \lambda[\zeta_i \left((x_i - x_0)^T\mathbf{M}_L(x_i - x_0)\right) - \gamma_{\zeta_i}]_+$$

where $\zeta_i = -1$ if $l_i = l_0$ and 1 otherwise, $[A]_+ = \max(0, A)$ is the hinge loss and γ_{ζ_i} is the desired margin. The empirical risk minimizing metric based on S_k^{pair} can be readily defined as $\mathbf{M}_L^* = \arg\min_{\mathbf{M}_L} Err^{\lambda}(\mathbf{M}_L, S_k^{pair})$. Our goal is to compare the generalization performance of $\mathbf{M}_L^*$ over the unknown $\mathbb{D}$.

Theorem 6. *Let $\phi^{\lambda}(\boldsymbol{M}_L, s_i)$ be a distance-based loss function that is λ-Lipschitz in the first argument. Then with probability at least $1-\delta$ over $\{s_1, ..., s_k\}$ from an unknown B-bounded-support (each $(x, l) \sim \mathbb{D}, ||x|| \leq B$) distribution $\mathbb{D}$, we have:*

$$\begin{aligned} &\underset{\boldsymbol{M}_L \in \mathcal{M}}{Sup} \left[Err^{\lambda}(\boldsymbol{M}_L, \mathbb{D}) - Err^{\lambda}(\boldsymbol{M}_L, S_k^{pair}) \right] \\ &\leq O\left(\lambda B^2 \sqrt{D \ln(1/\delta)/k} \right) \end{aligned} \tag{3.25}$$

Theorem. 6 proves that to achieve an estimation error rate ϵ, $k = \Omega\left((\lambda B^2/\epsilon)^2 D \ln(1/\delta)\right)$ samples are sufficient. The brief proof is shown here. Let $\mathcal{P}$ be the probability measure induced by the random variable $(X; \mathcal{L})$, where $X := (x, x')$, $\mathcal{L} := 1[l = l']$. Define function class:$\mathcal{F} := \{X \mapsto \|x - x'\|_{\mathbf{M}_L}\}$ and consider our loss function $\phi^{\lambda}(\mathbf{M}_L, s_i) = \lambda[\zeta_i \left((x_i - x_0)^T \mathbf{M}_L (x_i - x_0)\right) - \gamma_{\zeta_i}]_+$ which is λ-Lipschitz in the first argument. Then, we are interested in bounding the quantity

$$\underset{(X;\mathcal{L}) \in \mathcal{P}}{Sup} \left[\phi^{\lambda}(f_{\mathbf{M}_L}(X), \mathcal{L}) - \frac{1}{k} \sum_{i=1}^{k} \phi^{\lambda}(f_{\mathbf{M}_L}(X_i), \mathcal{L}_i) \right]$$

Define $\hat{x}_i := x_0 - x_i$ for each pair s_i, then the Rademacher complexity[2] of our function class $\mathcal{F}$ (with respect to the distribution $\mathcal{P}$) is bounded, since (let $\sigma_1, \sigma_2, ..., \sigma_k$ denote independent uniform ± 1-valued random variables):

[2]See the formal definition of the Rademacher complexity in [**6**]

$$
\begin{aligned}
\mathcal{R}(\mathcal{F},\mathcal{P}) &:= \mathbb{E}_{X_i,\sigma_i}\left[\underset{f_{\mathbf{M}_L}\in\mathcal{F}}{Sup}\frac{1}{k}\sum_{i=1}^{k}\sigma_i f_{\mathbf{M}_L}(X_i)\right] \\
&= \frac{1}{k}\mathbb{E}_{X_i,\sigma_i}\underset{\mathbf{M}_L\in\mathcal{F}}{Sup}\left[\sum_{i=1}^{k}\sigma_i\hat{x}_i^T\mathbf{M}_L\hat{x}_i\right] \\
&= \frac{1}{k}\mathbb{E}_{X_i,\sigma_i}\underset{\mathbf{M}_L\in\mathcal{F},[a^{jk}]_{jk}=\mathbf{M}_L}{Sup}\left[\sum_{j,k}a^{jk}\sum_{i=1}^{k}\sigma_i\hat{x}_i^j\hat{x}_i^k\right]
\end{aligned} \tag{3.26}
$$

$$
\begin{aligned}
\mathcal{R}(\mathcal{F},\mathcal{P}) &\leq \frac{1}{k}\mathbb{E}_{X_i,\sigma_i}\underset{\mathbf{M}_L\in\mathcal{F}}{Sup}\left[\|\mathbf{M}_L\|_F\left(\sum_{j,k}\left(\sum_{i=1}^{k}\sigma_i\hat{x}_i^j\hat{x}_i^k\right)^2\right)^{1/2}\right] \\
&\leq \frac{\sqrt{D}}{k}\mathbb{E}_{X_i,i\in[k]}\left(\mathbb{E}_{\sigma_i,i\in[k]}\sum_{j,k}\left(\sum_{i=1}^{k}\sigma_i\hat{x}_i^j\hat{x}_i^k\right)^2\right)^{1/2} \\
&= \frac{\sqrt{D}}{k}\mathbb{E}_{X_i,i\in[k]}\left(\sum_{j,k}\sum_{i=1}^{k}(\hat{x}_i^j)^2(\hat{x}_i^k)^2\right)^{1/2} \\
&= \frac{\sqrt{D}}{k}\mathbb{E}_{X_i,i\in[k]}\left(\sum_{i=1}^{k}\|\hat{x}_i\|^4\right)^{1/2} \\
&= \frac{\sqrt{D}}{k}\mathbb{E}_{(x_0,x_i)\in(\mathbb{D}\times\mathbb{D}),i\in[k]}\left(\sum_{i=1}^{k}\|x_i-x_0\|^4\right)^{1/2} \\
&\leq \sqrt{\frac{D}{k}}\mathbb{E}_{(x_0,x_i)\in(\mathbb{D}\times\mathbb{D}),i\in[k]}\left(\|x_i-x_0\|^4\right)^{1/2} \leq 4B^2\sqrt{\frac{D}{k}}
\end{aligned} \tag{3.27}
$$

Recall that $\mathbb{D}$ has bounded support (with bound B). Thus, by noting that ϕ^λ is $8B^2$ bounded function that is λ-Lipschitz in the first argument, we can readily apply the Theorem.8 in [**6**] to obtain the desired uniform deviation bound. We can further generate Theorem. 6 to find a tighter bound.

Theorem 7. *Let $\boldsymbol{M}_L$ be any class of weighting metrics on the feature space $X = \mathbb{R}^D$, and define $d := Sup_{\boldsymbol{M}_L \in \mathcal{M}} \|\boldsymbol{M}_L\|_F^2$. Following the same parameter setting in Theorem. 6, we have:*

$$\begin{aligned} &\sup_{\boldsymbol{M}_L \in \mathcal{M}} \left[Err^{\lambda}(\boldsymbol{M}_L, \mathbb{D}) - Err^{\lambda}(\boldsymbol{M}_L, S_k^{pair}) \right] \\ &\leq O\left(\lambda B^2 \sqrt{d \ln(1/\delta)/k} \right) \end{aligned} \tag{3.28}$$

Let $\mathcal{P}$ be the probability measure induced by the random variable $(X; \mathcal{L})$, where $X := (x, x')$, $\mathcal{L} := 1[l = l']$. Define function class:

$$\mathcal{F} := \{X \mapsto \|x - x'\|_{\mathbf{M}_L}\}$$

Following the same steps in the proof of Theorem. 6, we can conclude that the Rademacher complexity of $\mathcal{F}$ is bounded. In particular,

$$\mathcal{R}_k(\mathcal{F}) \leq 4B^2 \sqrt{\frac{Sup_{\mathbf{M}_L \in \mathcal{M}} \|\mathbf{M}_L\|_F^2}{k}}$$

Finally, we note that ϕ^{λ} is λ-Lipschitz in the first argument, so that we can readily apply Theorem.8 in [**6**].

From Theorem. 7, we observe that if the learned metric $\mathbf{M}_L$ has a low metric learning complexity $d \ll D$, it can help sharpen the sample complexity result, yielding a dataset-dependent bound. Recall our objective Eqn. 3.6, $d := Sup_{\mathbf{M}_L \in \mathcal{M}} \|\mathbf{M}_L\|_F^2$ is already optimized via our proposed learning objective. Therefore, the bound is further tighter under the same number of samples.

3.5. Experiments

3.5.1. Experiment Settings

Data & Evaluation. We have performed thorough experiments and comparative studies to evaluate our method on eight most widely-used benchmark datasets: VIPeR [**31**], QMUL GRID [**60**], CAVIAR [**15**], iLIDS [**121**], PRID 450S [**74**], CUHK Campus [**46**], CUHK03 [**47**] and Market-1501 [**119**]. The last three large-scale datasets are pretty challenging due to the extremely complicated variance of person appearance and abundant distractors. For a fair comparison, the training data of each dataset are used as the negative training samples for itself $\mathbf{Y}^{neg} = \mathbf{X}_{train}$, so no more extra information is utilized in the experiment. For all the experiments, the single-shot evaluation setting (except for the CUHK Campus dataset where the multi-shot matching setting is applied) is adopted and all the average results of 10 random trials are shown in the form of Cumulated Matching Characteristic (CMC) curves.

Feature. The recently proposed high-dimensional feature LOMO [**51**] is adopted as the visual feature representation. Since it is not practical to directly use such a high dimensional feature (usually 26960-dim for the original LOMO feature) in metric learning, we employ principal component analysis (PCA) to reduce the feature dimension to a reasonable scale.

Baselines. For fair comparisons, several global metric learning approaches [**52, 51, 112**] whose code is available to access and the feature can be replaced are compared to our proposed method under the same experiment setting and using the same LOMO feature. Besides, the most recent state-of-the-art published results are also reported for a thorough comparison. For all the experiments, the global metric learner, MLAPG [**52**] is chosen as the underlying baseline so that our online local metric adaptation algorithm is applied on top of it.

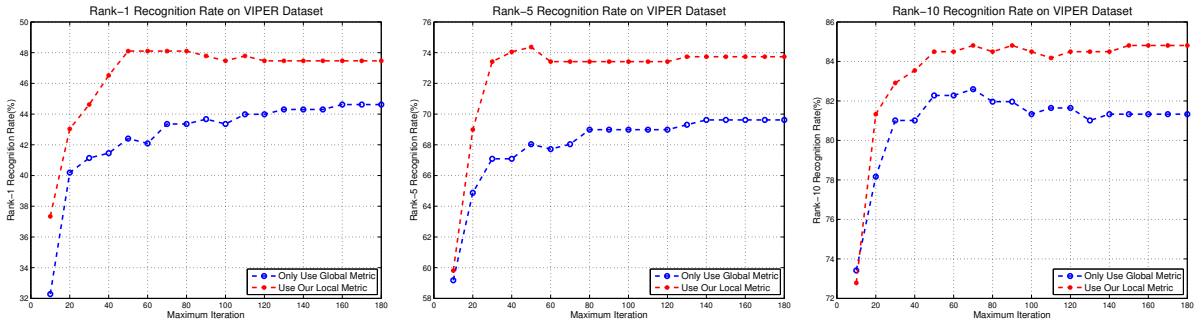


Figure 3.4. The influence of the quality of global metric. The x-axis means the maximum iteration time for global metric learning and the y-axis is the identification rate (Rank@1, Rank@5 and Rank@10 on VIPeR).

3.5.2. Influence of Global Metric Learning Quality

Our proposed OL-MANS algorithm is applied on top of a global metric $\mathbf{M}_G$, thus its overall performance may depend on the learning quality of adopted global metric learner. In order to verify whether our OL-MANS can always be helpful, global metrics obtained at various learning stages of a global metric learner [**52**] are tested, as in general the performance of a global metric learner improves with more training (e.g., more training iterations). As shown in Fig. 3.4, even the learned global metric does perform poorly (in its early training stages), our online local metric adaptation is able to consistently and significantly improve the performances by a large margin. This is because the local discriminative information introduced by hard negative samples is able to capture the specific crux of one identity which is quite helpful for identification.

3.5.3. Influence of Global Metric Learner Choice

An interesting question is whether our OL-MANS can always work for any global metric learners as promised. To verify it, we conduct the following experiment that different kinds of global

Methods	GRID		VIPeR	
	R@1	R@20	R@1	R@20
Euc	9.12	29.76	15.32	50.66
Euc+OL-MANS	20.88	45.12	21.99	56.11
XQDA[**51**]	12.96	43.52	38.99	91.94
XQDA+OL-MANS	29.20	50.96	43.54	92.15
MLAPG[**52**]	17.60	56.08	40.28	93.39
MLAPG+OL-MANS	30.16	59.36	44.97	93.64
DNSL[**112**]	15.12	53.12	40.19	93.54
DNSL+OL-MANS	28.96	56.96	43.67	93.61

Table 3.1. Comparison of identification rate with/without OL-MANS on VIPeR and GRID. All the experiments are under the same setting and use the same **LOMO** feature. **+OL-MANS** means implementing our OL-MANS on the original global metric learner. Red represents the better results.

metric learners, Euclidean distance, XQDA [**51**], MLAPG [**52**] and DNSL [**112**] are adopted as the underlying global metric that our OL-MANS algorithm will be readily applied on. For each learner, we compare the identification rates without and with our online local metric adaptation. The 10-run-average results on VIPeR and GRID datasets are reported in Table. 3.1, as well as the complete CMC curves in Fig. 3.5. We observe that for all the learners, our proposed online local metric adaptation algorithm is able to boost the identification performance with a significantly improvement, even double the identification accuracy (on GRID). Even for the most state-of-the-art global metric learner [**112**], applying our OL-MANS to it can still achieve a non-trivial improvement.

3.5.4. Influence of the Weighting Parameter λ

The parameter λ in Eqn. 3.9 is used to balance the underlying global metric and the learned local metric. Different λ will have different influences to the identification performance. We conducted an experiment on the VIPeR dataset to determine the value of λ, the results of which

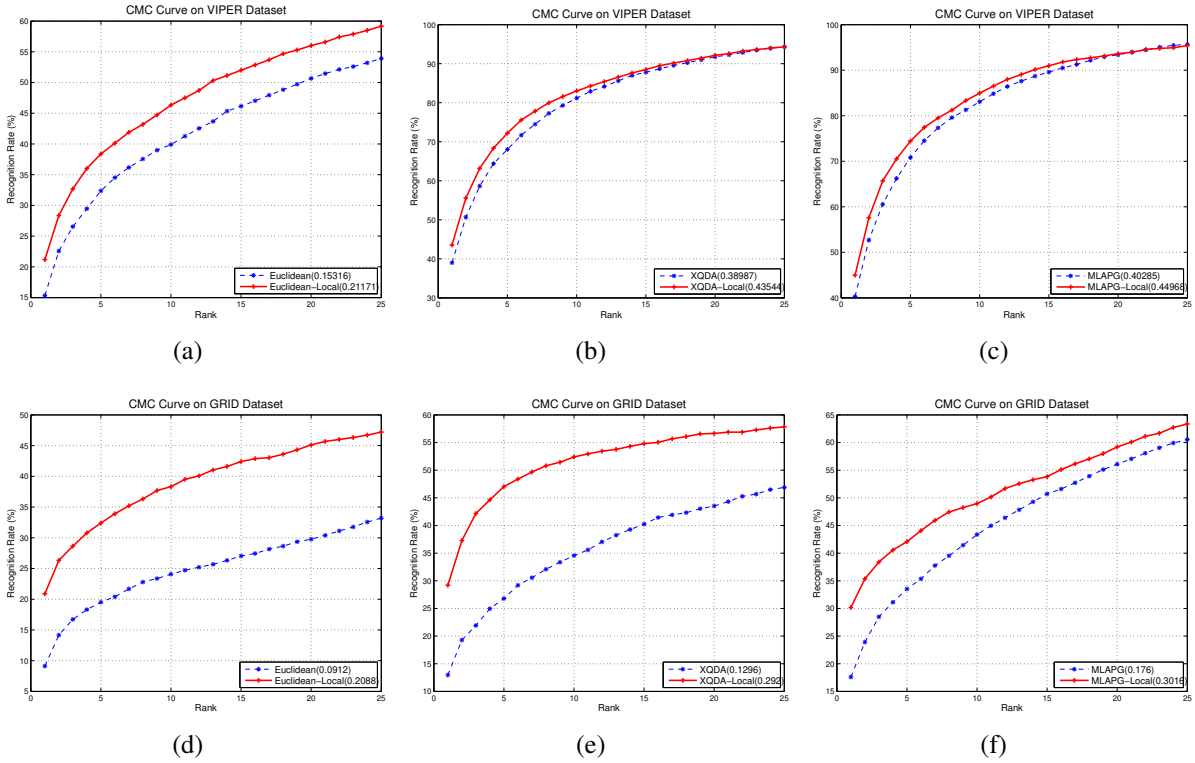


Figure 3.5. We conducted the experiment to figure out the influence of the choice of global metric learner. (a) and (d) are the results on VIPeR and GRID directly using the Euclidean distance; (b) and (e) are XQDA [**51**] results; (c) and (f) are MLAPG [**52**] results.

are shown in Fig.3.6. We need to point out some special λ values: The $\lambda = 0$ is the baseline result from [**52**] without our local metric tuning and $\lambda = max$ represents that λ is set as Eqn. 3.10. As we can see, setting $\lambda = \frac{\max\limits_{1\leq j\leq n-m}\left(D_G(x_{v_i}^p, x_{v_j}^g)\right)}{\max\limits_{1\leq j\leq n-m}\left(D_L(x_{v_i}^p, x_{v_j}^g)\right)}$ achieves the best result because it normalizes the norm scales of the global and local metric distances.

3.5.5. Influence of Negative Sample Database

For our OL-MANS, a negative sample database (NDB) is used to provide the negative training data. Because there are various strategies to collect NDB, we conduct the following experiments

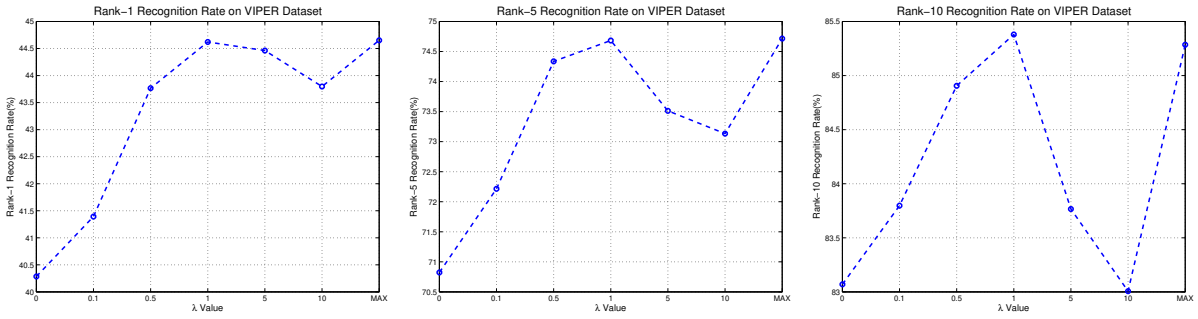


Figure 3.6. We conducted the experiment to determine the value of λ. The x-axis means the value of λ and the y-axis is the identification rate. Here are the identification results at Rank@1, Rank@5 and Rank@10 on VIPeR.

to investigate the influences of different NDB choices. The first two experiments are conducted on the VIPeR dataset [**31**] and the challenging CAVIAR [**15**] is used in the third experiment. Moreover, the global metric learning method proposed in [**52**] is adopted as the baseline method for the global metric learning $\mathbf{M}_G$.

Using the training data $\mathbf{X}_{train}$ from the same benchmark as negative sample: Here the training samples $\mathbf{X}_{train}$ in VIPeR which have different identities from P_i(the training data for global metric learning) are used as negative samples. It guarantees that the obtained NDB is clearly meaningful. The accuracy in P-Rid is given in Table.3.2 as **Our-SAME**.

Using different benchmark datasets as the NDB: In this experiment, we utilize other benchmarks as the NDB. The QMUL GRID [**60**] and CAVIAR [**15**] are combined into one dataset then used as the NDB in this experiment, so that we can guarantee that the identities of all the negative samples in the NDB are different from P_i. For each identity P_i, the k nearest negative samples are found in the NDB (under $\mathbf{M}_G$) and used for our OL-MANS. Different values of k (50, 100, 500) are chosen for further comparisons. The experiment results **Our-D-50/100/500** are shown in Table.3.2. Moreover, an additional experiment **Our-D-RAM** that

uses 50 random negative samples from the NDB for OL-MANS is compared. This experiment validates the insight of our method that the effective negative samples are those that are close to the probe in the feature space (e.g., strong false positives).

The NDB includes the false negative samples: We investigate how the "contamination" in the NDB impacts our proposed method. In this situation, some negative samples in the NDB are deliberately collected from the same identity to P_i. We call them *false negative samples*, in addition to the use of the probe image set of CAVIAR [**15**] as the rest of the NDB. Since there are multiple images of the same identity in CAVIAR, they can be considered as the false negative samples. The experiment results **Our-NoFN, Our-FN** are shown in Table. 3.2. **Our-NoFN** refers to a "clean" NDB with no false negative samples in it, and **Our-FN** refers to a "contaminated" NDB that includes false negative samples for all the probe images.

From Table. 3.2, it can be observed that **Our-SAME** performs the best because the negative data from the same benchmark dataset are most discriminative. Results on **Our-D-50/100/500** also largely outperform the baseline by consistent improvements. Moreover, the false negative sample may influence and degrade the performance of **Our-NoFN**, but not significantly. Nevertheless, a clean NDB with hard negatives is useful and effective.

3.5.6. Influence of Feature Descriptors

In the past yeas, deep features [**80, 80, 34**] have been widely used in many computer vision tasks no except of P-Rid. In this part, we compared various different feature descriptors for P-Rid problem to verify that the performance of our OL-MANS is independent of the choice of feature. Several hand-crafted features, LOMO [**51**] and deep features, CaffeNet [**43**], VGG-16 [**80**] and ResNet-50 [**34**] are examined. The above pre-trained CNN models from

Set	Method	R@1	R@5	R@10	R@20
VIPeR	Baseline [**52**]	40.73	69.94	82.34	92.37
	Our-D-RAM	39.87	70.51	82.28	91.77
	Our-SAME	44.97	74.43	84.97	93.64
	Our-D-050	42.63	73.63	84.81	93.54
	Our-D-100	43.04	73.86	84.30	93.42
	Our-D-500	42.53	73.89	84.15	93.35
CAVIAR	Baseline [**52**]	40.63	71.72	83.34	95.67
	Our-NoFN	51.68	76.36	86.38	96.55
	Our-FN	50.34	74.83	85.72	96.03

Table 3.2. Comparison of different NDBs on VIPeR (P=316) and CAVIAR (P=36).

Dataset	Method	Euclidean	MLAPG	XQDA	DNSL
VIPeR	LOMO	15.32/21.99	40.28/44.97	38.99/43.54	40.19/43.67
	CaffeNet	17.72/21.84	18.35/19.30	20.41/28.16	20.38/23.26
	VGG-16	20.25/26.27	20.25/23.73	23.45/29.02	23.86/26.52
	ResNet-50	22.78/27.22	23.42/26.58	31.93/40.47	33.70/38.01
GRID	LOMO	9.12/20.88	17.60/30.16	12.96/29.20	15.12/28.96
	CaffeNet	2.40/13.60	5.60/10.42	10.24/21.92	7.28/16.72
	VGG-16	6.40/18.44	7.20/16.84	12.72/21.52	10.24/17.36
	ResNet-50	12.84/23.22	12.40/19.12	21.44/34.96	17.36/29.44

Table 3.3. Comparison of different feature choices on VIPeR and GRID under different metrics (10-folds average Rank@1 performance is reported). For each result, the former one is the result **without** our OL-MANS, and the last one is our OL-MANS result.

which we have removed the final fully-connected (FC) layer are further fine-tuned by the large-scale Market-1501 datasets (**their R-1/mAP performances are: CaffeNet=44.31/0.24, VGG-16=63.93/0.425 and ResNet-50=77.22/0.561**), then they are used to extract the features for the other P-Rid datasets. As can be seen from Table. 3.3, the performance improvement by our OL-MANS method is independent from the used feature descriptors.

3.5.7. Local Metric Rank Analysis

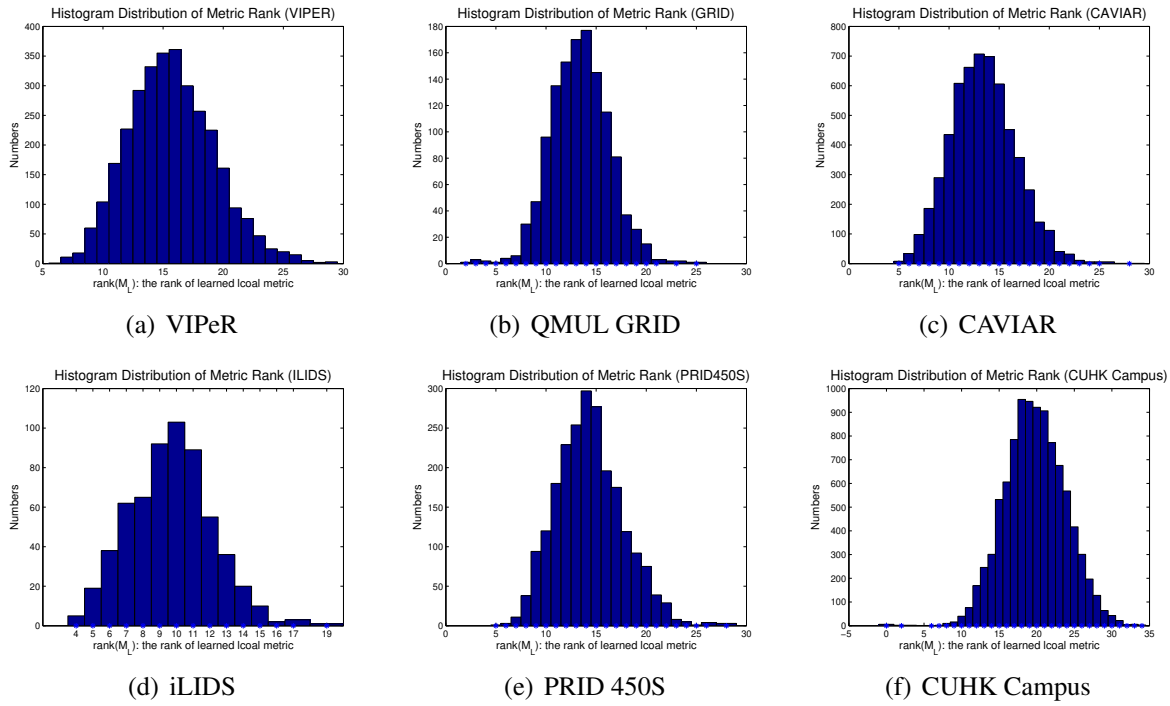
Feature descriptors used in our experiment are generally high dimensional in order to handle the complex appearance variations. In practice, most existing methods apply PCA blindly to reduce the feature dimension without clear justification and effectiveness. In contrast, our OL-MANS can be performed in the original high dimensional space while allowing the selection of a low rank local metric. The effectiveness of the low rank metric is also verified in [**52**] and [**39**].

For our proposed OL-MANS, we solve the original time-consuming positive semidefinite (PSD) problem by solving an efficient kernel SVM instead, as in Eqn. 3.6. The obtained local metric $\mathbf{M}_L^i$ is formed as Eqn. 3.29:

$$\mathbf{M}_L^i = \sum_{j=0}^{k} \alpha_j \zeta_j \tilde{y}_j \tilde{y}_j^T = \sum_{j=1}^{k} \alpha_j \tilde{y}_j \tilde{y}_j^T = \sum_{j=1}^{k'} \alpha_j^r (\tilde{y}_j^r)(\tilde{y}_j^r)^T \succeq 0 \tag{3.29}$$

where $\alpha_j^r \neq 0$. It is obvious that $\mathbf{M}_L^i$ is the linear combination of all the support vectors of $\{\tilde{y}_j^r\}_{j=1}^{k'}$. Therefore, the rank of $\mathbf{M}_L^i$ is bounded by the number of support vectors, k'. In practice, the local metric is constrained by the strong negative samples (the hard negatives). In other words, the coefficient vector $A = [\alpha_1, \alpha_2, ..., \alpha_k]$ should be sparse.

To validate this, we have conducted an experiment that we compute the ranks of all the learned local metrics for all the probes in different benchmarks (VIPeR, GRID, CAVIAR, iLIDS, P-Rid 450S and CUHK Campus). The result is presented in Fig. 3.7, where it is evident that almost all the learned local metrics are pretty low rank, even though the size of the negative database (NDB) is large. This negative database has over 10,000 negative samples, and more than 500 strong negative samples are generally selected for each datum to learn its local



(a) VIPeR (b) QMUL GRID (c) CAVIAR

(d) iLIDS (e) PRID 450S (f) CUHK Campus

Figure 3.7. Histogram distributions of metric rank for all the learned local metrics on different benchmark datasets.

metric. This clearly shows an advantage of our proposed method, as it allows us to work in a high dimensional space while most existing methods do not.

3.5.8. Training Cost Analysis and Comparison

Although every test probe needs to learn a local Mahalanobis metric at the test stage, solving a kernel SVM problem instead of solving the original PSD problem makes the learning efficient and largely reduces the training time. Table. 3.4 [3] provides a thorough comparison of average training time of various state-of-the-art metric learning-based methods on VIPeR dataset. Besides, Table. 3.5 shows the training time of different advanced global metric learners on a

[3]The total learning time of OL-MANS includes the local metric adaptation time and gallery ranking time for all probes.

Method	ITML [**22**]	MLAPG [**52**]	LADF [**50**]
Ave Time	20.5	25.8	31.7
Method	LMNN [**98**]	PRDC [**123**]	**OL-MANS**
Ave Time	152.9	394.6	4.8

Table 3.4. Average training time (seconds) on VIPeR.

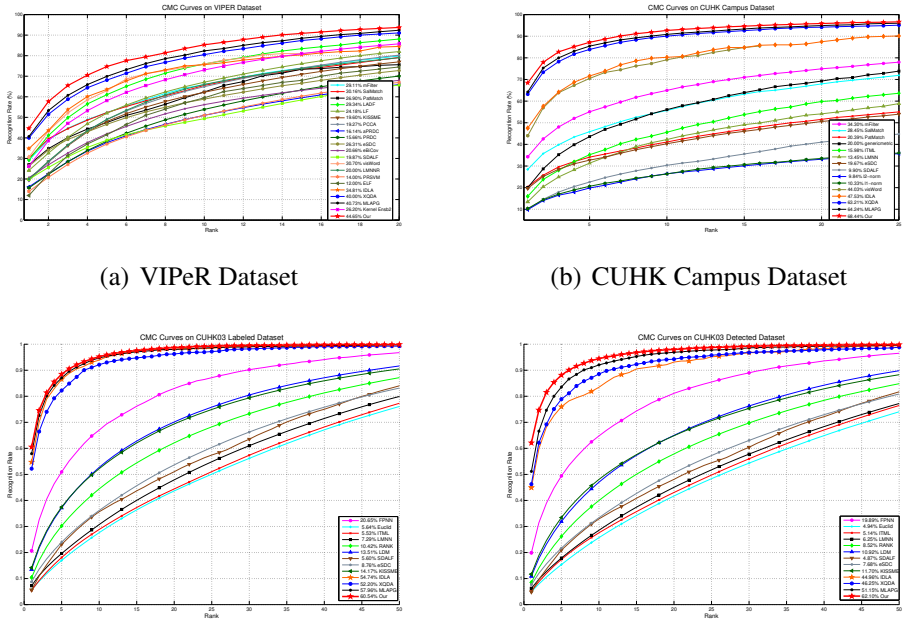
Method	XQDA [**51**]	MLAPG [**52**]	MFA [**103**]
Train Time	3233.8	2732.8	437.8
Method	kLFDA [**103**]	DNSL [**112**]	**OL-MANS**
Train Time	995.2	3149.7	19.60

Table 3.5. Training time (seconds) on Market-1501.

large-scale dataset, Market-1501. All the experiments are conducted on a remote server with an Intel i7-5930K @3.50GHz CPU and 32G memory. The total average training time of our method on VIPeR is only 4.81 seconds for the adaptation of all the 316 probes, much shorter than learning a single global metric in 25.82 seconds. For the large scale dataset Market-1501, the efficiency advantage of ours is much more pronounced. Our local metric adaptation time is $10 \sim 100$ times less than the other global metric learners. So the extra time spent in our local metric adaptation is indeed nominal compared with learning a global metric.

3.5.9. Comparison with Re-ranking Results

As we discussed in Sec. 3.3.5, both the re-ranking technique and our proposed OL-MANS can be applied to other P-Rid methods to further boost the identification accuracy. In this part, we will evaluate our proposed OL-MANS algorithm and a state-of-the-art re-ranking method [**127**] on the Market-1501 dataset. XQDA [**51**] and MLAPG [**51**] are selected as the baseline metric learners. The Rank@1 performance results are shown in Table. 3.6 which demonstrates that our OL-MANS our-performs the re-ranking method with a large margin. The re-ranking may



(a) VIPeR Dataset (b) CUHK Campus Dataset

(c) CUHK03 Labeled Dataset (d) CUHK03 Detected Dataset

Figure 3.8. Comparison of CMC curves and Rank@1 identification rates on benchmark **(a)** VIPeR, **(b)** CUHK Campus, **(c)** CUHK03-Labeled and **(d)** CUHK03-Detected datasets.

Methods	**Market-1501**	**CUHK03-Labeled**	**CUHK03-Detected**
XQDA[**51**]	45.87	49.7	44.6
XQDA[**51**]+Re-rank[**127**]	48.34	50.0	45.9
XQDA[**51**]+OL-MANS	51.87	56.41	52.34
MLAPG[**51**]	43.87	57.96	51.15
MLAPG[**51**]+Re-rank[**127**]	42.79	56.24	50.76
MLAPG[**51**]+OL-MANS	44.93	61.68	62.71

Table 3.6. Comparison between our proposed OL-MANS and the state-of-the-art re-ranking method under **single-shot** evaluation and **LOMO** feature. **Rank@1** result is reported. Red represents the best result.

not work even degrade the original performance while our OL-MANS can always boost the performance which has been theoretically guaranteed.

Method	R@1	R@5	R@10	R@20
Ours	44.97	74.43	84.97	93.64
LSSCDL[**114**]	42.66	-	84.27	91.93
DNSL[**112**]	42.28	71.46	82.94	92.06
MLAPG[**52**]	40.73	69.94	82.34	92.37
XQDA[**51**]	40.00	68.13	80.51	91.08
TMA[**64**]	39.88	-	81.33	91.46
KISSME[**42**]	34.81	60.44	77.22	86.71
ITML[**22**]	24.64	49.78	63.04	78.39
LMNN[**98**]	29.43	59.78	73.51	84.91
kCCA[**54**]	30.16	62.69	76.04	86.80
MFA[**103**]	38.67	69.18	80.47	89.02
kLFDA[**103**]	38.58	69.15	80.44	89.15

Table 3.7. Comparison results on VIPeR. **All the methods use the same LOMO feature**. RED is the best result and BLUE is the second best one.

3.5.10. Extensive Comparisons on Benchmarks

Experiments on VIPeR: The VIPeR dataset [**31**] is a widely used benchmark dataset for P-Rid. It contains 632 pedestrian image pairs taken from 2 different cameras in an outdoor environment. We follow the widely adopted experimental protocol on VIPeR: 632 pairs are randomly divided into half for training and the other half for testing. We conducted the comparison experiment under the same experiment setting and using the same LOMO feature and the results are reported in Table. 3.7. Our proposed algorithm achieves the best performances on all the ranks. For the important Rank@1 evaluation, our performance 44.97% outperforms the second best approach LSSCDL by 2.31%. This promising performance indicates that the proposed local metric adaptation method is consistently effective, several representative examples are shown in Fig. 3.2.

The Table. 3.8 which shows that our method is still the best one. One interesting observation is our performance at Rank@20 is a little bit lower than the latest TSRPR [**77**] method. This is

Method	R@1	R@5	R@10	R@20
Ours	44.97	74.43	84.97	93.64
SCNCD[**104**]	37.80	68.50	81.20	90.40
EPKFM[**13**]	36.80	70.40	83.70	91.70
K-Ensb2[**103**]	36.10	68.70	80.10	85.60
IDLA[**1**]	34.81	-	-	-
TSRPR[**77**]	31.10	68.60	82.80	94.90
kBiCov[**61**]	31.11	58.33	70.71	82.44
LADF[**50**]	30.22	64.70	78.92	90.44
SalMatch[**115**]	30.16	-	65.54	79.15
Mid-L-F[**117**]	29.11	-	65.95	79.87
MMCML[**62**]	28.83	59.34	75.82	88.51
eSDC[**116**]	26.74	50.70	62.37	76.36
SSCDL[**58**]	25.60	53.70	68.10	83.60
PRDC[**123**]	15.66	38.40	53.86	70.09

Table 3.8. More comparison results on VIPeR.

expected as our local metric becomes less effective when the true positive gallery image is far from the probe in the feature space. Nevertheless, our method still beats all the other approaches at Rank@20.

Experiments on QMUL GRID: The GRID dataset [**60**] contains 250 pedestrian image pairs taken from 8 disjoint camera views and 775 additional images that do not belong to the 250 persons. GRID is also a pretty tough dataset because of the large viewpoint variations and the low-resolution image quality. The experimental protocol for GRID is the same as [**52, 13, 62**]: we randomly divide the 250 identities into half for training and the other half for testing as well as the extra 775 images are used as distractors to enlarge the gallery set. The average performance of 10 random trials is provided in Table. 3.9. It can be clearly observed that the proposed algorithm outperforms all the existing algorithms at Rank@1 by a very significant 7.76% improvement on the identification rate. From the results we can see that the GRID

dataset is more challenging than the VIPeR dataset, but our proposed algorithm can still handle it well by adapting the local similarity structure of each probe.

Experiments on P-Rid 450S: The P-Rid 450S [**74**] is another recent dataset. It consists of 450 image pairs recorded from two different, static surveillance cameras. Since it is a more recent benchmark, few methods have been tested on it. We adopt the experimental protocol in [**77**]: all the images are normalized to 168×80 pixels and the persons are split to 225 for training and 225 for testing. The final results are presented in Table. 3.10, from which we can observe our proposed method achieves a pretty large improvement against the state-of-the-art approaches. The 22.41% increase of identification rate at Rank@1 verifies the strength of our adaptive online local metric tuning strategy.

Experiments on iLIDS: The iLIDS dataset [**121**] is generated from video images captured in a busy airport arrival hall so the images suffer from severe occlusions caused by people and luggage. With an average of 4 images for each person, it contains a total of 476 shots of 119 people captured by multiple non-overlapping cameras. We follow the experimental protocol in [**103**]: The persons are randomly split to 59 for training and 60 for testing. Under each partition, one image for each person in the testing set is randomly selected as the gallery image and the rest of the images are used as probes. Compared to the state-of-the-art in Table. 3.10, our proposed method is much better by 15.57% from the current best Rank@1 identification rate reported obtained from DFL-RDC [**23**].

Experiments on CAVIAR: The CAVIAR dataset [**15**] contains 1220 images of 72 individuals from 2 cameras in a shopping mall. For the 72 individuals, 50 of them appear in both camera views and the remaining 22 persons only appear in one camera view. Each identity has

Method	R@1	R@5	R@10	R@20
Ours	30.16	42.64	49.20	59.36
LSSCDL(LOMO)[**114**]	22.40	-	51.28	61.20
DNSL(LOMO)[**112**]	15.12	31.92	40.72	53.12
MLAPG(LOMO)[**52**]	17.60	33.52	43.36	56.08
XQDA(LOMO)[**51**]	12.96	26.80	34.56	43.52
EPKFM[**13**]	16.30	35.80	46.00	57.60
MtMCML[**62**]	14.08	34.64	45.84	59.84
PRDC[**123**]	9.68	22.00	32.96	44.32

Table 3.9. Comparison results on GRID.

10 to 20 images and the resolutions vary from 17×39 to 72×144. We use the same experimental protocol with [**103, 13, 50**] that the persons are randomly split to 36 for training and 36 for testing. It can be observed from Table. 3.10 that the proposed method outperforms all existing algorithms at Rank@1. We also have the second best performances at Rank@5, Rank@10 and Rank@20, except for the SSCDL [**58**] algorithm which slightly outperforms our method. The reasons might be due to the following two facts: [1] the SSCDL exploits the 22 persons appearing in only one camera as its specific unlabeled training data while our method does not employ this additional information; [2] the number of identities in this dataset is very small so the training data (only 36 identities) is not sufficient.

Experiments on CUHK Campus: The CUHK Campus dataset [**46**] consists of 971 persons captured from two camera views in a campus environment, two images per person in each camera view. We split the set to 485 for training and 486 for testing and multi-shot matching scenario is applied to CUHK Campus dataset for evaluation [**51, 1, 77, 103**]. We evaluate the performance by fusing scores of all the probe images of the same identity. As shown in Table. 3.11, the proposed method consistently outperforms other state-of-the-art methods in all Rank@1, Rank@5, Rank@10 and Rank@20 identification rates.

Benchmark	Method	R@1	R@5	R@10	R@20
PRID 450S	**Ours**	65.51	86.27	92.27	96.00
	XQDA [**51**]	61.42	-	90.84	95.33
	LSSCDL[**114**]	60.49	-	88.58	93.60
	TMA[**64**]	54.22	73.78	83.11	90.22
	TSRPR[**77**]	43.10	70.50	78.20	86.30
	SCNCD[**104**]	41.60	68.90	79.40	87.80
	KISSME[**42**]	33.00	-	71.00	79.00
iLIDS	**Ours**	67.67	86.17	90.17	95.00
	D-RDC[**23**]	52.10	68.20	78.00	88.80
	LTR[**68**]	50.34	-	-	-
	MLAPG[**52**]	49.83	67.50	88.17	93.33
	K-Ensb2[**103**]	40.30	66.70	78.10	89.60
	PRDC[**123**]	37.83	63.70	75.10	88.40
CAVIAR	**Ours**	51.78	76.57	86.82	96.62
	SSCDL[**58**]	49.10	80.20	93.50	97.90
	MLAPG[**52**]	40.63	71.72	83.34	95.67
	MFAχ^2[**103**]	40.20	70.20	83.90	95.10
	EPKFM[**13**]	40.10	65.60	78.00	90.50
	χ^2_{RBF}[**103**]	33.20	65.90	81.90	95.20
	LADF[**50**]	30.30	62.80	78.00	92.60
	LFDA[**69**]	32.00	56.30	70.70	87.40

Table 3.10. Comparison results on P-Rid 450S, iLIDS and CAVIAR.

Method	R@1	R@5	R@10	R@20
Ours	68.44	87.16	92.67	95.88
LSSCDL[**114**]	65.97	-	-	
DNSL(LOMO)[**112**]	64.98	84.96	89.92	94.36
MLAPG(LOMO)[**52**]	64.24	85.41	90.84	94.92
XQDA(LOMO)[**51**]	63.21	83.89	90.04	94.16
kFLDA(LOMO)[**103**]	54.63	80.45	86.87	92.02
MFA(LOMO)[**103**]	54.79	80.08	87.26	92.72
kCCA(LOMO)[**54**]	54.63	80.45	86.87	92.02
IDLA[**1**]	47.53	-	-	-
Mid-L-F[**117**]	34.30	-	64.96	74.94
TSRPR[**77**]	32.70	51.20	64.40	76.30
K-Ensb2[**103**]	24.00	38.90	46.70	55.40

Table 3.11. Comparison results on CUHK Campus.

Method	R@1	R@5	R@10	R@20
Ours	61.68	88.39	95.23	98.47
MLAPG(LOMO) [**52**]	57.96	87.09	94.74	98.00
XQDA(LOMO) [**51**]	52.20	82.23	92.14	96.25
DNSL(LOMO) [**112**]	58.90	85.60	92.45	96.30
DeepReID[**47**]	20.65	51.50	66.50	80.00
Im-Deep[**1**]	54.74	86.50	93.88	98.10

Table 3.12. Comparison results on CUHK03 **Labeled**.

Method	R@1	R@5	R@10	R@20
Ours	62.71	87.59	93.80	97.55
MLAPG(LOMO)[**52**]	51.15	83.55	92.05	96.90
XQDA(LOMO)[**51**]	46.25	78.90	88.55	94.25
DNSL(LOMO)[**112**]	53.70	83.05	93.00	94.80
DeepReID[**47**]	19.89	50.00	64.00	78.50
Im-Deep[**1**]	44.96	76.01	83.47	93.15

Table 3.13. Comparison results on CUHK03 **Detected**.

Experiments on CUHK03: The CUHK03 dataset [**47**] is a large-scale dataset which contains 13164 images of 1360 pedestrians. All the images are captured by six surveillance cameras over months. Each person is observed by two disjoint camera views with an average of 4.8 images in each view. Two kinds of data are provided: manually cropped pedestrian images and images detected with a state-of-the-art pedestrian detector. We follow the same experimental protocol [**47, 52, 51**]: splitting all the pedestrians into a training set of 1160 persons and a test set of 100 persons. The results in Table. 3.12 and Table. 3.13 show that for both two datasets, the proposed algorithm achieves the best performances at all ranks. It outperforms the second best approach by almost 10% in Rank@1 rate, even for the data under such a complicated practical situation, which is very significant.

Methods	Single-Q		Multi-Q	
	R@1	R@20	R@1	R@20
Baseline[**119**]	35.84	67.64	44.36	73.25
Kissme(LOMO)[**119**]	40.50	N/A	N/A	N/A
MFA-χ^2(LOMO)[**103**]	45.67	N/A	N/A	N/A
kLFDA(LOMO)[**103**]	51.37	N/A	52.67	N/A
Hist-Loss[**89**]	59.47	91.09	N/A	N/A
Euc(LOMO) [**119**]	32.93	63.87	40.33	69.40
Euc+OL-MANS	40.93	74.06	51.45	80.98
MLAPG(LOMO)[**52**]	43.87	88.40	61.33	96.40
MLAPG+OL-MANS	44.93	89.20	62.40	94.27
XQDA(LOMO)[**51**]	45.87	81.73	56.27	85.07
XQDA+OL-MANS	51.87	84.40	74.00	94.00
DNSL(LOMO)[**112**]	51.73	88.67	57.70	88.59
DNSL+OL-MANS	60.67	91.87	66.80	92.19

Table 3.14. Comparison results on Market-1501 under both the **single-shot** and **multiple-shot** evaluation settings. Red represents the better result.

Experiments on Market-1501: Market-1501 [**119**] is the largest image-based P-Rid benchmark dataset to date which contains 32668 bboxes of 1501 identities. Each person is recorded by six cameras at most, and two at least. For training and testing, the given fixed training and test set are utilized and both single-shot and multi-shot settings are used for evaluation. All the results are presented in Table. 3.14. We perform our online local metric adaptation algorithm to different global metric learners [**52, 51, 112**] based on the same LOMO features and experiment setting. As shown by the result, for all the global metric learners, a significant improvement on Rank@1 can be achieved by performing our OL-MANS algorithm to it, no matter under the single-shot or multi-shot evaluation setting.

3.6. Discussion

In this chapter, we proposed a novel online local metric adaptation algorithm to learn a dedicated Mahalanobis metric for each probe at the test stage. This new approach only uses negative samples for metric adaptation, which is practical in real situation. It largely reduces the demand for a large number of positive training data as in existing P-Rid methods, and it only incurs minimum computational costs to perform online training. In-depth theoretical analysis well justifies our algorithm and extensive experiments also demonstrate that our new approach consistently and significantly outperforms the state-of-the-art methods.

The main issue in our proposed OL-MANS method is that even multiple probe images for matching are given at once, no matter they are from the same identity or different identities, our OL-MANS will handle them individually, the relationships among the given images are simply ignored by OL-MANS. Such an individual-specific learning strategy is not the most efficient and effective way for online local metric adaptation. We expect the utilization of the intrinsic sharing information among samples could refine the OL-MANS to a better solution, that not only the performance can be further improved, but also the online metric learning burden can be largely reduced. So that we will present how we achieve this refined solution in Chapter 4.

CHAPTER 4

Learning From Unlabeled Samples: Joint Local Metric Adaptation From Sharing

4.1. Introduction

In the past years, most existing P-Rid researches focus on obtaining discriminative metrics and feature representations *offline* to better capture the variation of visual appearances [**52, 51, 62**]. However, limited and imbalanced labeled training samples and the distribution gap between training and testing data (the training and testing sets contain entirely different classes) severely constrain the performance. Therefore more attention has been paid to the post rank

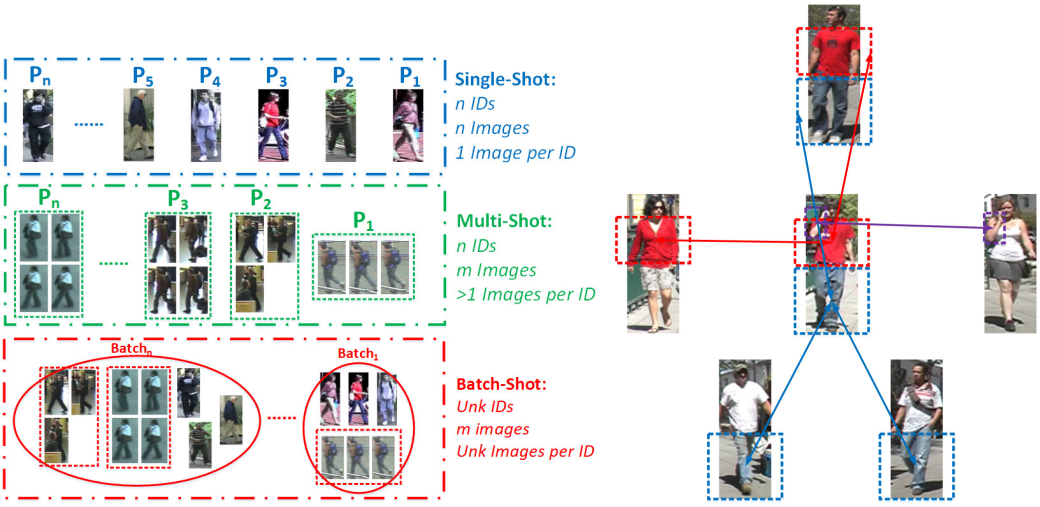


Figure 4.1. (a) The batch-shot setting is more practical during online testing phase. (b) Even no supervision information is available, the visual similarity sharing among queries is intrinsic.

refinement during *online* testing stage. Various re-ranking methods [**106, 28, 2, 127, 129, 5, 3**] have achieved promising performance for rank refinement of existing P-Rid baselines, which rely on either the initial ranking list or extra learning samples. However, no matter for the single-shot or multi-shot query (Fig. 4.1), the aforementioned online re-ranking methods always treat each query individually because of the lack of supervision information (identify label) of the given testing probes.

In practice, the all given probe data can be considered as a *batch* query set (Fig. 4.1(a)). These images in the batch usually exhibit different intrinsic sharing relationships in the visual similarity, but always be ignored and considered to be trivial by the previous re-ranking methods. As shown in Fig. 4.1(b), even no identity information is known, the red clothes, blue jeans, raised arms are actually shared visually among these images. Instead of treating each query separately, we aim to take advantage of such visual similarity sharing by considering the given query probes as a *batch* with the expectation that the identification performance should be better than dealing with them separately. The effectiveness of exploring sharing relationship among data has been verified by the Multi-Task Learning (MTL) [**9**] research. The success of MTL relies on discovering the latent sharing relationships among tasks, which cannot be found by learning each task independently. Learning from sharing is good at handling such condition that only a limited number of training data are available for each task by taking the sharing relationship as a kind of data augmentation. Such sharing strategy is particularly suitable for P-Rid in where each testing probe itself is the only positive sample available for learning.

In this chapter, we propose a novel **J**oint **M**ulti-**M**etric **L**earning algorithm for online re-ranking (Fig. 4.2). On the online testing stage, instead of re-ranking each query individually, we prefer to consider all the testing probes as a *batch* query set. By automatically mining different

sharing-subsets that each one contains a group of the visually similar queries, a joint local Mahalanobis metric is efficiently learned for all the probes in the subset. Therefore a series of *sharing-based* local Mahalanobis metrics for all the queries are learned to jointly adjust the local distributions. The contributions of this chapter are three-fold: (1) We propose a novel online joint multi-metric learning algorithm which extends the state-of-the-art re-ranking scheme to a more generalized and feasible model. A theoretical sound optimization approach is proposed for efficient optimization. (2) By considering the given query probes as a batch query set, the intrinsic visual sharing relationship is explored. Therefore the total number of the learned local metrics is largely reduced, as opposed to the linear growth $O(n)$ with independently learned local metrics. (3) A better re-ranking performance can be achieved via a multi-kernel late fusion scheme for the jointly learned metrics which has been verified by the extensive experiments on several P-Rid benchmarks.

This following sections in this chapter are: our proposed novel sharing-based multi-metric learning algorithm is introduced in Sec. 4.3. Some justifications of the proposed method and comparisons with the related works are presented in Sec. 4.3.2. Sec. 4.5 shows extensive experimental results to support the proposed method. Finally, conclusions and discussions are made in Sec. 4.6.

4.2. Related Work

4.2.1. Online Rank-Refinement For P-Rid

In recent years, online rank-refinement technique has attracted more attention in P-Rid community [**105, 28, 127, 5, 106, 3, 129**]. Garcia *et al.* [**28**] proposed an unsupervised re-ranking model by taking advantage of the content and context information in the ranking list. Ye *et al.* [**105**]

revised the ranking list by considering the nearest neighbors of both the global and local features. A ranking aggregation algorithm is proposed by Ye. *et al.* [**106**] to utilize both similarity and dissimilarity evidence from various baseline methods. Bai *et al.* [**3**] aimed to refine the ranking results by estimating the similarity between two samples in the context of other pairs of references under the proposed Supervised Smoothed Manifold (SSM). Zhong *et al.* [**127**] proposed a k-reciprocal encoding method for re-ranking by assuming that a gallery image is more likely to be a true match if it is similar to the probe in the k-reciprocal nearest neighbors. Zhou *et al.* [**129**] proposed a novel instance-specific local metric adaptation algorithm to learn different Mahalanobis metrics for different probes by using extra negative samples. Barman *et al.* [**5**] focused on how to make a consensus-based decision for retrieval by aggregating the ranking lists from multiple algorithms, only the matching scores are needed. Compared with the aforementioned online re-ranking approaches, the most important difference in our method is to fully utilize the similarity sharing information with a low learning burden by considering the given probes as a batch query, instead of treating them separately. More appealing merits of our method are discussed in Sec. 4.4 in detail.

4.2.2. Multi-Task Learning For P-Rid

Multi-task learning-based P-Rid methods [**27, 65, 83, 62**] are proposed to facilitate the small-size sample (SSS) issue in P-Rid [**112**] by learning from sharing. A multi-task maximally collapsing metric learning (Mt-MCML) model is proposed by Ma *et al.* [**62**] to jointly learn multiple Mahalanobis distance kernels for data from different distributions. Recently, McLaughlin *et al.* [**65**] made use of multi-task learning for deep network model design in order to prevent over-fitting to the small-size training data. Gao *et al.* [**27**] proposed a multi-task CNN combining

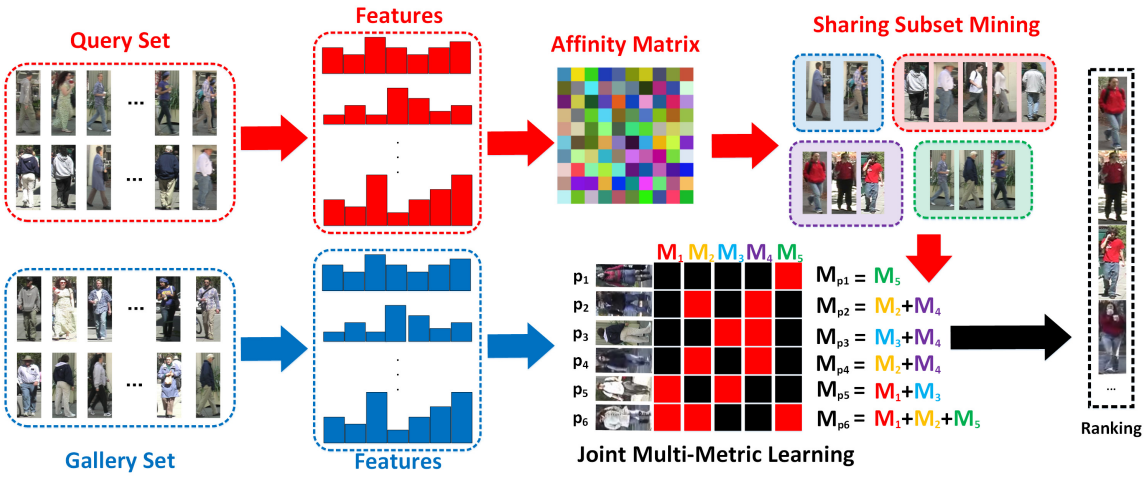


Figure 4.2. Unlike existing online re-ranking approaches, our proposed method aims to improve the ranking results by utilizing the sharing information among given queries to learn a set of dedicated local metrics for all the testing probes. The affinity matrix of given queries is computed based on their extracted features by a baseline method, then a series of sharing-subsets is automatically mined to cluster queries into different visually similar groups. Thus multiple sharing-subset specific metrics are jointly learned by our proposed algorithm which are further utilized by a multi-kernel late fusion module for re-ranking. **(Best view in color)**

the Softmax loss with Siameses loss for retrieval. A novel multi-task learning with low-rank attribute embedding (MTL-LORAE) framework is proposed by Su *et al.* [**83**] to address the multi-modal data distribution issue. However, these multi-task learning-based approaches only utilize the sharing relationships on the offline training stage, no instance-specific local adaptation is considered for the unseen testing samples. So the performance is indeed limited due to the shifted testing data distribution, that samples for testing are drawn from different classes. Our proposed method addresses this issue by jointly learning multiple local metrics for different testing queries via exploring the intrinsic sharing relationships among them.

4.3. Joint Multi-Metric Learning From Sharing

4.3.1. Definition of Problem

The given testing datasets contain: **a query set** $\mathcal{Q}$ and **a gallery set** $\mathcal{G}$

$$\begin{aligned} \mathcal{Q} &= \{(q_i, l_{q,i})\}_{i=1}^{n_q} \\ \mathcal{G} &= \{(g_i, l_{g,i})\}_{i=1}^{n_g} \end{aligned} \tag{4.1}$$

that q_i and g_i are feature representations by either hand-crafted [**104, 61, 51**] or offline learned (metric learning [**112, 52**], deep embedding [**43, 80, 87, 34**], etc). $l_{q,i}$ and $l_{g,i}$ are the corresponded class labels. All the samples in $\mathcal{Q}$ and $\mathcal{G}$ are drawn from c classes. Here we consider the *closed-set condition* that both the $\mathcal{Q}$ and $\mathcal{G}$ contain samples from all the c classes respectively.

P-Rid aims to obtain a ranking list of $\mathcal{G}$ for each query q_i based on the visual similarity between q_i and g_j:

$$d\left(q_i, g_j\right) = \|q_i - g_j\|^2 \tag{4.2}$$

Our goal is to refine the initial ranking results by boosting the rank of true-matches for q_i.

4.3.2. Unsupervised Sharing-Subset Mining

Although the supervision information (identity label) $\{l_{q,i}\}$ of $\mathcal{Q}$ is unknown on online stage, the intrinsic visual similarity among $\mathcal{Q}$ implies salient sharing relationships, regardless of the supervision information exists or not. Therefore an unsupervised sharing-subset (SSSet) mining algorithm (Alg.1) is proposed to automatically group $\mathcal{Q}$ into different sharing-subsets $\{\mathcal{R}_i\}_{i=1}^{n_r}$, in

where all the samples in $\mathcal{R}_i$ share strong visual similarity to each other. Typically such sharing-subset mining is a combinatorial problem suffering from exponential computation complexity $O(2^n)$, inspired by the boosting research [**88**], we solve the sharing-subset mining problem via an efficient heuristic greedy solution.

The affinity matrix $\mathbf{A} \in \mathbb{R}^{n_q \times n_q}$ of $\mathcal{Q}$ is defined as:

$$\mathbf{A}_{i,j} = \begin{cases} \exp\left(\frac{-d(q_i,q_j)}{2\sigma}\right), \ i \neq j \\ 0, \ i = j \end{cases} \tag{4.3}$$

where σ is the variance parameter of distance matrix from $\mathcal{Q}$ so that $\mathbf{A}_{i,j}$ represents the visual similarity between q_i and q_j. $\mathbf{A}$ is further normalized to a soft-max affinity distribution matrix $\mathbf{A}^s$:

$$\mathbf{A}^s_{i,j} = \frac{\mathbf{A}_{i,j}}{\sum_j \mathbf{A}_{i,j}} \tag{4.4}$$

that the i-th row of $\mathbf{A}^s$ represents the similarity distribution between q_i and the other samples in $\mathcal{Q}$. In order to mine the most reliable sharing relationships, a threshold Θ defined as the average affinity of the top-k nearest neighbors for all $\mathcal{Q}$ is used for outlier filtering:

$$\Theta = \frac{\sum_{i=1}^{n_q} \sum_{j=1}^{k} \mathbf{A}^s_{i,\mathcal{N}_i(j)}}{k \cdot n_p} \tag{4.5}$$

where $\mathcal{N}_i(j)$ is the index of j-th largest element in i-th row of $\mathbf{A}^s$. Therefore, a binary index map $\mathbf{B}$ is obtained by:

$$\mathbf{B}_{i,j} = \begin{cases} 1, \ \mathbf{A}^s_{i,j} \geq \Theta \\ 0, \ \mathbf{A}^s_{i,j} < \Theta \end{cases} \tag{4.6}$$

The non-zero $\mathbf{B}_{i,j}$ implies the strong similarity sharing relationship between q_i and q_j. Finally, each mined sharing-subset is one of the n_r non-zero rows of $\mathbf{B}$:

$$\mathcal{R}_i = \{q_j\},\ \forall \mathbf{B}_{i,j} = 1 \tag{4.7}$$

Algorithm 1 Unsupervised Sharing-Subset Mining

Require: The given query set $\mathcal{Q}$
Ensure: n_r sharing-subsets $\{\mathcal{R}_i\}_{i=1}^{n_r}$
1: Compute the normalized affinity matrix $\mathbf{A}^s$ of $\mathcal{Q}$ by Eqn. 4.3 and Eqn. 4.4;
2: Compute the threshold Θ by Eqn. 4.5;
3: Compute the binary index map $\mathbf{B}$ by Eqn. 4.6;
4: Mine each sharing-subset by Eqn. 4.7;
5: Return $\{\mathcal{R}_i\}_{i=1}^{n_r}$

4.3.3. Joint Multi-Metric Learning For $\mathcal{R}_i$

Once the SSSets $\{\mathcal{R}_i\}_{i=1}^{n_r}$ are obtained via Alg.1, our goal is to jointly learn n_r sharing-based Mahalanobis metrics from $\{\mathcal{R}_i\}_{i=1}^{n_r}$ in order to collapse the same-SSSet samples together meanwhile push the different-SSSet samples far away. So the designed objective is:

$$\begin{aligned}
&\mathbf{M}_r = \arg\min_{\mathbf{M}_r} \frac{1}{2}\|\mathbf{M}_r\|^2 \\
&w.r.t:\ \mathbf{M}_r \succeq 0 \\
&\left(q_i^{pos} - q_j^{neg}\right)^T \mathbf{M}_r \left(q_i^{pos} - q_j^{neg}\right) \geq 2,\ \forall q_i^{pos} \in \mathcal{R}_r, q_j^{neg} \in \bigcup_{k \neq r} \mathcal{R}_k \\
&\left(q_i^{pos} - q_j^{pos}\right)^T \mathbf{M}_r \left(q_i^{pos} - q_j^{pos}\right) = 0,\ \forall q_i^{pos}, q_j^{pos} \in \mathcal{R}_r
\end{aligned} \tag{4.8}$$

For the r-th SSSet $\mathcal{R}_r$, a specific Mahalanobis metric $\mathbf{M}_r$ is learned by Eqn. 4.8 which is shared by all the samples in $\mathcal{R}_r$. For simplicity, we reduce the size of inequality and equality constraints in Eqn. 4.8.

Theorem 8. *The objective Eqn. 4.8 has an exactly equivalent form by only keeping the constraints related to one anchor sample q^{pos} in $\mathcal{R}_r$, that q^{pos} can be any sample in $\mathcal{R}_r$. Therefore the equivalent form is Eqn. 4.9:*

$$
\begin{aligned}
&\boldsymbol{M}_r = \arg\min_{\boldsymbol{M}_r} \frac{1}{2}\|\boldsymbol{M}_r\|^2 \\
&w.r.t: \ \boldsymbol{M}_r \succeq 0 \\
&\left(q^{pos} - q_j^{neg}\right)^T \boldsymbol{M}_r \left(q^{pos} - q_j^{neg}\right) \geq 2, \ \forall q_j^{neg} \in \bigcup_{k \neq r} \mathcal{R}_k \\
&\left(q^{pos} - q_j^{pos}\right)^T \boldsymbol{M}_r \left(q^{pos} - q_j^{pos}\right) = 0, \ \forall q_j^{pos} \in \mathcal{R}_r
\end{aligned}
\tag{4.9}
$$

PROOF. Revisit Eqn. 4.8, the equality constraints in it propose to collapse all $q_i^{pos} \in \mathcal{R}_r$ together. Therefore keeping only the equality constraints related to the anchor sample q^{pos} achieves the same collapsing performance. So as to the inequality constraints in Eqn. 4.8. Finally we can reduce the constraint size by only considering q^{pos}. The re-written objective Eqn. 4.9 has only $O(n)$ linear-scale constraints, while the scale of constraints in the original objective Eqn. 4.8 is quadratic $O(n^2)$. □

The Eqn. 4.9 can be efficiently optimized by solving a much easier version [**25**]:

Theorem 9. *All the vectors $q^{pos} - q_j^{pos}$ can form a spanning space $\boldsymbol{S} = span(\sum_j \lambda_j (q^{pos} - q_j^{pos}))$. The Eqn. 4.9 is equivalent to replace $q^{pos} - q_j^{neg}$ by y_j, which is the projection of $q^{pos} - q_j^{neg}$ to $\boldsymbol{S}^{\perp}$, that $\boldsymbol{S}^{\perp}$ is the orthogonal space of $\boldsymbol{S}$.*

PROOF. Since $\mathbf{M}_r$ is positive semi-definite, the constraint $\left(q^{pos} - q_j^{pos}\right)^T \mathbf{M}_r \left(q^{pos} - q_j^{pos}\right) = 0$ is equivalent to $\mathbf{M}_r \left(q^{pos} - q_j^{pos}\right) = 0$ which means the $\mathbf{M}_r s = 0$ for all $s \in \mathbf{S}$. Project $q^{pos} - q_j^{neg}$ to $\mathbf{S}$ and $\mathbf{S}^{\perp}$ generates two orthogonal bases x_j and y_j respectively, so $q^{pos} - q_j^{neg} = x_j + y_j$. Replace the inequality constraints in Eqn. 4.9 by $x_j + y_j$:

$$\begin{aligned} & \left(q^{pos} - q_j^{neg}\right)^T \mathbf{M}_r \left(q^{pos} - q_j^{neg}\right) \\ & = \left(x_j + y_j\right)^T \mathbf{M}_r \left(x_j + y_j\right) \\ & = {y_j}^T \mathbf{M}_r y_j \end{aligned} \tag{4.10}$$

Now Eqn. 4.9 has an equivalent form as:

$$\begin{aligned} & \mathbf{M}_r = \arg\min_{\mathbf{M}_r} \frac{1}{2}\|\mathbf{M}_r\|^2 \\ & w.r.t: \ \mathbf{M}_r \succeq 0 \\ & {y_j}^T \mathbf{M}_r y_j \geq 2, \ \forall q_j^{neg} \in \bigcup_{k \neq r} \mathcal{R}_k \\ & \mathbf{M}_r s = 0, \ \forall s \in \mathbf{S} \end{aligned} \tag{4.11}$$

□

Finally, we prove that Eqn. 4.11 has the same solution to Eqn. 4.8 by solving a kernel SVM problem without its PSD constraint $\mathbf{M}_r \succeq 0$ and equality constraints, and the solution is still PSD.

Theorem 10. *The solution to Eqn. 4.8 is exactly the same as solving the Eqn. 4.11 by relaxing its equality and PSD constraints, since they are indeed off-the-shelf.*

PROOF. If we get rid of the equality and PSD constraints in Eqn. 4.11, the new form is:

$$\begin{aligned} &\mathbf{M}_r = \arg\min_{\mathbf{M}_r} \frac{1}{2}\|\mathbf{M}_r\|^2 \\ &w.r.t: \ {y_j}^T \mathbf{M}_r y_j \geq 2, \ \forall q_j^{neg} \in \bigcup_{k \neq r} \mathcal{R}_k \end{aligned} \tag{4.12}$$

Eqn. 4.12 is exactly the same form of the objective in [**129**]. As proved by the Theorem.1 in [**129**], the positive semi-definiteness of $\mathbf{M}_r$ is guaranteed even if no PSD constraint is explicitly imposed in Eqn. 4.12 since $\mathbf{M}_r = \sum \alpha_i \varphi(y_i) = \sum \alpha_i y_i \cdot y_i^T \succeq 0$. For the equality constraints in Eqn. 4.11, given a member s of $\mathbf{S}$, we have:

$$\mathbf{M}_r s = \left(\sum \alpha_i y_i \cdot y_i^T\right) s = \sum \alpha_i y_i \cdot (y_i^T s) = 0 \tag{4.13}$$

which proves that the solution to Eqn. 4.12 satisfies the equality constraints as well. □

4.3.4. Multi-Metric Late Fusion For P-Rid

Our proposed objective Eqn. 4.8 can be efficiently solved via Theorem. 10. Therefore a set of sharing-based local metrics $\{\mathbf{M}_r\}_{r=1}^{n_r}$ are readily obtained. For one query probe q_i, it may be contained by multiple sharing-subsets so that there will be multiple learned metrics $\mathbf{M}_r$ associated to q_i. The final metric $\mathbf{M}_{q_i}$ for q_i is a boosting-form multi-kernel late fusion [**82, 81**]:

$$\mathbf{M}_{q_i} = \sum_{r=1}^{n_r} \mathbf{B}_{r,i} \mathbf{M}_r \tag{4.14}$$

For a gallery image g_j, the distance between q_i and g_j is:

$$d_{\mathbf{M}_{q_i}}(q_i, g_j) = (q_i - g_j)^T \mathbf{M}_{q_i} (q_i - g_j) \tag{4.15}$$

4.4. Justification and Comparison

Compared with the closely related state-of-the-art re-ranking technique [**127**] and [**129**], our proposed algorithm owns more appealing merits which are verified by the comparison experiments in Sec. 4.5.6.

Ours vs [**129**]: As proved by Theorem. 10, [**129**] is just a special case of ours. Therefore our method outperforms [**129**] by **three** main advantages. (1) our method is more robust to overfitting than [**129**] since the sharing-subset is a kind of positive data augmentation. Compared with using the only one testing query as positive for learning, the usage of sharing-subset in our method will increase the diversity and reduce the variation of the given queries. (2) Our proposed method has better and more stable performance than [**129**], since each jointly learned metric $\mathbf{M}_r$ owns the exact same property as [**129**], so the error bound of $\mathbf{M}_r$ is reduced. The late fused metric kernel $\mathbf{M}_{q_i} = \sum_{r=1}^{n_r} \mathbf{B}_{r,i}\mathbf{M}_r$ for query q_i performs better than each single $\mathbf{M}_r$ in practical identification problems verified by previous research [**81, 82**]. (3) The last merit is the low computational burden of our method. [**129**] needs to learn n_q individual local metrics for all $\mathcal{Q}$ which is linearly ordered, while by utilizing visual similarity sharing, our method only needs to learn much fewer sharing-based joint metrics but achieves better performance (As shown in Sec. 4.5.5).

Ours vs [**127**]: Compared with re-ranking technique [**127**], our method takes advantage of the visual similarity sharing relationship among the unlabeled query set $\mathcal{Q}$ to find the optimal online local metric adaptation. (1) For **effectiveness** concern, [**127**] depends heavily on the quality of initial ranking list (if the true match is not in the top-k ranks). It may hurt the initial rank result, because the true match may have a lower rank after re-ranking if the false matches are included in the top-k list. Instead, the performance of our method relies on the quality of

Dataset	VIPeR	GRID	PRID450S	PRID2011	CUHK01	Mkt-1501	Duke
# identities	632	250	450	200	971	1501	1404
# cameras	2	8	2	2	2	6	8
# distractors	0	775	0	0	0	2793	408
# BBoxes	1264	500	900	>20000	1942	32668	36411

Table 4.1. The statistics of different P-Rid benchmark datasets.

the mining sharing-subsets [**129**]. For the worst case that the quality of sharing-subsets is pretty bad, no hard negatives are provided and the query set shares nothing within itself, our method still will not degrade the original performance but degenerate to the previous special case [**129**]. (2) Another merit of our method is its high **efficiency**. The optimization of our method is efficient even if there are a lot of query samples available. Because the learned sharing-based joint metric $\mathbf{M}_r$ is only related to a handful set of hard negatives in $\bigcup_{k \neq r} \mathcal{R}_k$.

4.5. Experiments

4.5.1. Experimental Settings

Datasets. Several P-Rid datasets are tested including the VIPeR [**31**], QMUL GRID [**60**], PRID 450S [**74**], PRID 2011 [**35**], CUHK 01 [**46**], Market-1501 [**119**] and DukeMTMC-reID [**124**]. The statistic details of the above datasets are summarized in Table. 4.1. For the VIPeR, GRID, PRID 450s and PRID 2011 datasets, we follow the widely adopted experimental protocol [**52, 13**] that all the persons in each dataset are randomly divided into half for training and the other half for testing, then 10-run-average performance is reported. For CUHK 01, which consists of 971 persons captured from two camera views with two images per person in each camera view, we split the persons into 485 for training and 486 for testing [**51, 103**]. For Market-1501 and DukeMTMC-reID, the given fixed training and testing sets are directly utilized.

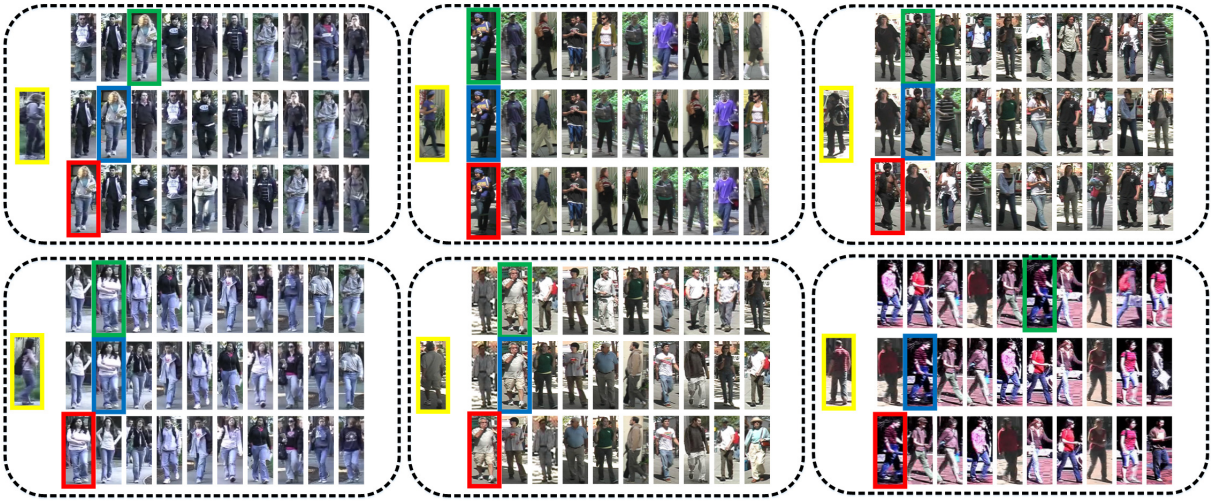


Figure 4.3. The comparison of re-ranking improvement on VIPeR. For each probe, its top-10 ranking results (from **left** to **right**) are retrieved by the baseline (1st row) [**52**], OL-MANS (2nd row) [**129**], and our method (3th row). **(Best view in color and enlarged)**

Evaluation. For a fair comparison, all the experiments are conducted in the same experimental setting. For evaluation, the single-shot evaluation setting is adopted [1] and all the results are reported in the form of Cumulated Matching Characteristic (CMC) at several selected ranks. For Market-1501 and DukeMTMC-reID, mean average precision (mAP) is also reported.

Features and Baselines. Both the hand-crafted and deep features are tested. The LOMO [**51**] feature is selected as the hand-crafted representation. Besides, several deep features, CaffeNet [**43**], VGG16 [**80**], GoogLeNet [**87**] and ResNet50 [**34**] are chosen as deep feature representatives. Several state-of-the-art global metric learning approaches, MLAPG [**52**], XQDA [**51**] and DNSL [**112**], are selected as the baselines. Finally, two recently proposed online re-ranking methods, [**129**] and [**127**] are compared with our algorithm.

[1] multi-shot evaluation is applied to CUHK 01

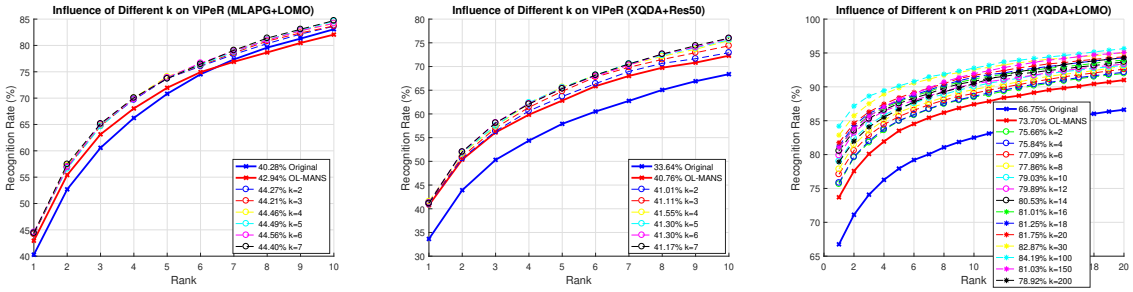


Figure 4.4. The influence of parameter k in Eqn. 4.5. The x-axis is the rank and the y-axis is the identification rate.

4.5.2. Influence of Parameter k in Eqn. 4.5

Recall Eqn. 4.5, the parameter k influences the quality of sharing-subset by influencing the threshold Θ, which will further influence the re-ranking performance of our method. If k is too small, the sharing information cannot be fully explored; while if k is too large, the sharing-subset will be polluted. Fig. 4.4 shows the influence of different choices of k on both the VIPeR and PRID 2011 datasets. If the given queries convey sufficient sharing information (PRID 2011), the larger the k is, the better the performance is. While for the single-shot VIPeR dataset, the testing probes are indeed visually discriminative to each other so small k will guarantee the purity of sharing information. In our experiments, we normally set $k = 6$.

4.5.3. Influence of Baseline Quality

Our method can be applied on top of any offline learned P-Rid baselines, thus its overall performance may depend on the learning quality of the baseline. In order to verify whether our proposed method can always be helpful, a global metric learner baseline MLAPG [**52**] is tested that different MLAPG results are obtained at different learning stages to serve as baselines. Fig. 4.5 shows even the learned baseline does perform poorly (in its early training stages), our

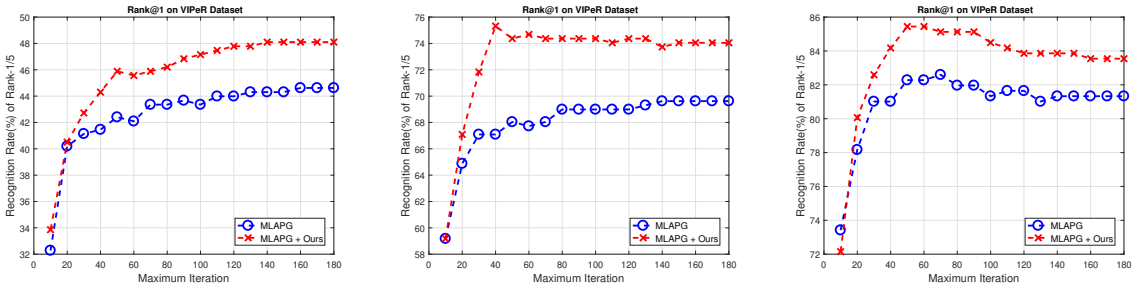


Figure 4.5. The influence of learning quality of $f(\cdot)$. The x-axis is the max-learning iteration and the y-axis is the identification rate.

algorithm is still able to consistently and significantly improve the performances at all ranks by a large margin.

4.5.4. Influence of Baseline Choice

To verify that our method should work for any offline learned baseline models and any kinds of features, we conduct the experiments that different combinations of baselines (Euclidean distance, XQDA, MLAPG and DNSL) and feature descriptors (illustrated in Sec. 4.5.1) are tested with(w/) and without(w/o) our proposed method. A thorough comparison result is shown in Table. 4.2, no matter for the hand-crafted feature (LOMO) or the learned deep features, no matter on the small-size GRID dataset or the larger-scale Market-1501 benchmark, our method can always achieve non-trivial improvement on Rank@1 performance, even double the accuracy on GRID.

4.5.5. Influence of Computational Cost

We conduct experiments to verify that the total number of jointly learned metrics by our method is sub-linear. Fig. 4.6 shows the number of learned local metrics and how many probes are

Dataset	**Rank@1**	Euclidean	MLAPG[**52**]	XQDA[**51**]	DNSL[**112**]
VIPeR	LOMO[**51**]	15.3→22.3	40.3→44.6	39.0→44.3	40.2→43.6
	CaffeNet[**43**]	18.4→22.4	19.0→19.5	23.1→27.9	23.1→24.5
	VGG16[**80**]	17.4→20.1	20.0→20.2	23.5→29.3	24.9→27.1
	GoogLeNet[**87**]	22.3→25.8	24.8→26.6	21.7→29.8	25.9→29.4
	ResNet50[**34**]	22.0→25.7	23.7→25.8	33.6→41.6	34.1→39.5
Dataset	**Rank@1**	Euclidean	MLAPG[**52**]	XQDA[**51**]	DNSL[**112**]
GRID	LOMO[**51**]	08.2 →21.5	16.5 →27.9	16.5→33.0	14.6→31.0
	CaffeNet[**43**]	07.4 →13.5	08.1 →11.4	12.2→23.6	11.6→19.1
	VGG16[**80**]	06.3 →12.2	07.5 →10.2	09.9→20.7	09.8→16.6
	GoogLeNet[**87**]	15.3 →21.1	14.2 →17.9	15.4→29.4	15.4→22.6
	ResNet50[**34**]	10.3 →18.6	10.3 →16.2	23.1→39.9	18.6→31.9
Dataset	**Rank@1(mAP)**	Euclidean		XQDA[**51**]	
Market-1501	CaffeNet[**43**]	64.7(41.9)→71.9(47.4)		63.6(36.3)→76.3(50.6)	
	VGG16[**80**]	63.9(42.5)→71.9(48.7)		64.0(38.4)→76.9(53.9)	
	GoogLeNet[**87**]	75.4(53.9)→82.9(60.9)		74.6(50.9)→85.7(65.4)	
	ResNet50 [**34**]	77.2(56.1)→82.7(61.8)		77.4(53.7)→85.0(64.4)	

Table 4.2. Comparison of Rank@1 performance with/without our method using different features under different baselines. For each result, the format is **baseline result w/o ours** → our result

covered by the learned local metrics under different k on different datasets. For the traditional individual learning methods [**114, 129**], n_p testing probes need n_p individually learned metrics to cover. However, our method needs much fewer jointly learned metrics. With the increase of k, fewer joint metrics are needed but more probes are covered since each joint metric will cover more visually similar query probes. The results of PRID 2011 (Fig. 4.6(a)) show even fewer joint metrics are used when k grows larger, better performance can be achieved due to the exhaustive utilization of the sharing information.

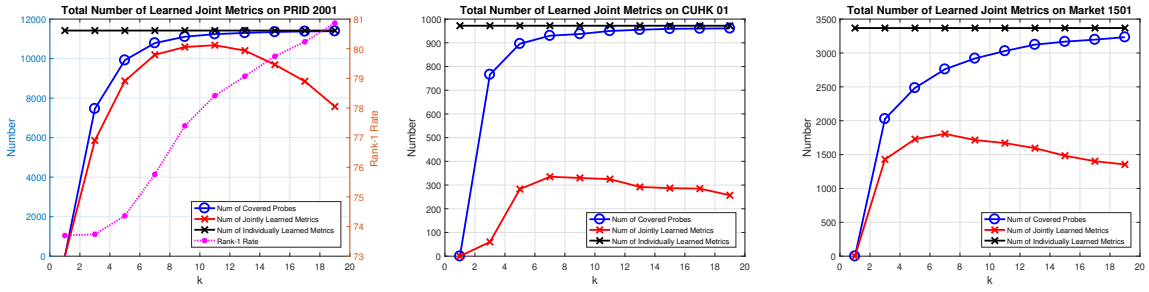


Figure 4.6. Computational cost comparison. The x-axis is the parameter k and the y-axis is the number of learned local metrics

4.5.6. Comparison with Re-Ranking Methods

We compare the aforementioned state-of-the-art recently methods in Sec. 4.4 with our method on different datasets. For a fair comparison, our method and [**129**] use the same NDB, and various parameters of [**127**] are tested to report the best performance. The Rank@1 performance is reported in Table. 4.3. If the original ranking result is pretty bad, [**127**] will further degrade the baseline performance. In contrast, [**129**] and ours will not harm the original result, while our method outperforms [**129**] by a large margin due to joint learning from sharing-subsets. Besides, our method can easily combine [**127**] to further boost the performance. Visual results are shown in Fig. 4.3, even for these "hard" queries that [**129**] cannot work well, our method is still able to handle them.

4.5.7. Comparison with the State-of-the-art

Experiments on VIPeR: The comparison results are reported in Table. 4.4. Our method achieves the second-best performance on all the ranks except for the recent SCSP[**114**]. But we can use the weak XQDA/MLAPG learners to beat the other strong learners by a large margin.

Method	**VIPeR**	**GRID**	**CUHK01**	**Market1501**
Euc	15.3	08.1	17.0	75.4(53.9)
Euc+[**127**]	11.8	09.5	16.8	78.2(66.2)
Euc+[**129**]	22.0	**21.5**	33.8	81.4(60.5)
Euc+Ours	**22.3**	**21.5**	**34.3**	**82.9(60.9)**
Euc+[**127**]+Ours	21.6	20.4	33.9	81.3(66.9)
XQDA	39.0	16.5	63.9	74.6(50.9)
XQDA+[**127**]	38.9	15.7	63.3	78.7(66.8)
XQDA+[**129**]	41.9	32.1	67.8	82.3(60.9)
XQDA+Ours	**44.3**	**33.0**	**68.9**	**85.7(65.4)**
XQDA+[**127**]+Ours	44.2	31.8	**68.9**	82.9(69.0)
DNSL	40.2	14.6	65.9	66.3(41.9)
DNSL+[**127**]	40.0	13.7	64.3	68.0(54.1)
DNSL+[**129**]	42.0	30.7	65.9	69.1(44.7)
DNSL+Ours	43.6	**31.0**	**69.5**	**71.1(47.2)**
DNSL+[**127**]+Ours	**43.9**	29.8	68.7	69.4(53.8)

Table 4.3. Comparison of different online re-ranking methods. **Rank@1** or **Rank@1(mAP)** performance is reported.

Experiments on Market-1501: Table. 4.4 shows the comparison results on Market-1501 by performing our method to different deep learning models. As can be seen, a significant improvement on Rank@1(mAP) can be achieved by performing our algorithm to it.

Experiments on DukeMTMC-reID: Table. 4.4 shows the poor performances of the weak XQDA/MLAPG learners can be boosted to the state-of-the-art level by utilizing our method.

VIPeR			Market-1501			DukeMTMC-reID		
Method	**R@1**	**R@20**	**Method**	**R@1**	**mAP**	**Method**	**R@1**	**mAP**
SCSP[**114**]	53.54	96.65	Baseline[**119**]	35.84	14.75	BoW+kissme[**119**]	25.1	12.2
LSSCDL[**114**]	42.66	91.93	kLFDA[**103**]	51.37	24.43	XQDA+lomo[**51**]	30.8	17.0
DNSL[**112**]	42.28	92.06	CAMEL[**110**]	55.00	27.10	Res50[**120**]	65.5	44.1
MLAPG[**52**]	40.73	92.37	Hist-Loss[**89**]	59.47	N/A	Res50+LSRO[**124**]	67.7	47.1
XQDA[**51**]	40.00	91.08	DNSL[**112**]	61.02	35.68	Res50+OIM[**101**]	68.1	-
TMA[**64**]	39.88	91.46	VGG-16[**80**]	63.93	42.50	Res50+MLAPG[**52**]	60.4	38.2
MFA[**103**]	38.67	89.02	CaffeNet[**43**]	64.70	41.90	Res50+XQDA[**51**]	64.7	44.4
kLFDA[**103**]	38.58	89.15	GoogLeNet[**87**]	75.39	53.90	Res50+OLMANS[**129**]	69.5	50.3
CAMEL[**110**]	30.90	-	Re-Ranking[**127**]	77.11	63.63	PAN[**125**]	71.6	51.5
DCMTL[**65**]	33.60	87.60	ResNet-50[**34**]	77.23	56.10	GAN[**124**]	67.7	47.1
OneSML[**4**]	34.30	-	**Ours+CaffeNet**	**71.85**	**47.40**	SVDNet[**85**]	67.6	45.8
LORAE[**83**]	42.30	89.60	**Ours+VGG16**	**71.85**	**48.70**	**Ours+Res50**	**73.0**	**55.8**
Ours+MLAPG	**44.56**	**92.94**	**Ours+GoogLeNet**	**82.93**	**60.90**	**Ours+Res50+MLAPG**	**66.3**	**47.9**
Ours+XQDA	**44.27**	**92.44**	**Ours+ResNet50**	**82.69**	**61.80**	**Ours+Res50+XQDA**	**71.1**	**53.7**

Table 4.4. State-of-the-art comparison results on VIPeR, Market-1501 and DukeMTMC-reID. **All the results are the best performances reported in their literatures**

4.6. Conclusion

In this chapter, unlike existing P-Rid re-ranking approaches which treat different query probes individually, we consider all the given testing queries as a *batch* query set. By utilizing their visual similarity sharing relationships, a novel joint multi-metric learning algorithm is proposed to simultaneously learn a set of sharing-based local Mahalanobis metrics for all the queries in the batch. Extensive experiments demonstrate that our method consistently and significantly outperforms the state-of-the-art methods.

No matter for our OL-MANS work in Chapter 3 and this joint multi-metric learning work in this chapter, both of them only focus on how to enhance the local discriminant of the probe-side samples in visual matching. However, for the counterpart gallery samples, no effort is paid to them but directly using them for visual matching. But the "hard" gallery samples which are still indistinguishable under the learned probe-specific local metrics will significantly influence the visual matching performance. In order to address these "hard" gallery distractors, the local discriminant of gallery samples still needs to be enhanced.

CHAPTER 5

Learning From Mixture Of Labeled and Unlabeled Samples: Online Bi-directional Local Discriminant Enhancement

5.1. Introduction

Person Re-Identification (P-Rid), focusing on retrieving the same identity images of a query probe from a gallery set, still remains a challenging task in computer vision. In order to address the challenges of large variations in human pose, camera viewpoint, illumination change, and background clutter, the existing P-Rid research mainly focuses on offline discriminative metric learning [**52, 51, 112, 4**] or feature embedding [**48, 56, 92**]. However, due to the critical distribution shifting issue of testing data, that the samples for testing are drawn from totally different distributions against the training data, the performance of the offline learned metrics and features is limited.

To narrow the gap between training and testing data distribution, various online rank refinement methods [**105, 28, 127, 129, 5**] are proposed for the sake. However, most online re-ranking methods [**105, 28, 127, 5**] simply treat different probes equally without considering the individual characteristics, so that their improvement performance is neither significant nor stable. To enhance the local discriminant of different probes, several algorithms [**114, 129**] are proposed to learn an instance-specific local metric for each query probe, while the involved gallery data is simply ignored in learning but only used for final retrieval. Even a discriminative local metric for one probe can be learned, the "hard"gallery samples with large intra-class variance and

Figure 5.1. For a query probe, the extreme challenging hard negative distractors in the gallery set (in blue box) will significantly influence the retrieval accuracy (1st row). Even using the state-of-the-art online rank refinement method [**129**] (2nd row), the ground-truth (in red box) still has a lower rank than the distractors. By taking advantage of our proposed bi-directional local discriminant enhancement method, the true-match is successfully re-ranked to the top (3rd row).

small inter-class variance will tremendously degrade the retrieval performance (Fig. 5.1) since such "hard" gallery samples are still indistinguishable under the learned probe-specific metric.

In this chapter, a novel online ranking refinement algorithm is proposed to fully utilize all the given probe and gallery data. By taking advantage of the instance-specific bi-directional local discriminant enhancement, a two-stage hierarchical local adaptation algorithm is designed as illustrated in Fig. 5.2. The local discriminant of a given probe is firstly enhanced via an extended probe-specific metric adaptation method to obtain a re-ranking list of the gallery samples. For the top-ranked gallery candidate, an instance-specific local discriminant enhancement algorithm is adopted to further refine its local similarity distribution. By performing a bi-directional

retrieval matching based on the bi-directional local discriminant enhancement result, a final re-ranking list is determined. Our main contributions of this work are three-fold: (1) To handle the severe shifted data distribution issue in P-Rid, we propose a novel instance-specific rank refinement algorithm by taking advantage of bi-directional local discriminant enhancement, which extends the state-of-the-art re-ranking scheme [**129**] to a more generalized and feasible model. (2) To fulfill the time-efficient requirement of re-ranking, a theoretical sound optimization solution is proposed for efficient learning which is theoretically proven to guarantee the improvement of baseline performance. (3) While we present our algorithm in the context of P-Rid, it can be potentially applied to any other general visual retrieval-based tasks. The efficiency and effectiveness of our proposed algorithm are verified by the extensive experiments on four large-scale P-Rid benchmarks (CUHK03, Market1501, DukeMTMC-reID and MSMT17).

The organization of the following sections are: in Sec. 5.2, some related works about online re-ranking of visual matching and the state-of-the-art methods are introduced. Our proposed method is presented in Sec. 5.3. Extensive experimental results of our proposed method are shown in Sec. 5.4 and conclusions are made in Sec. 5.5.

5.2. Related Work

5.2.1. CNN-based Feature Extraction For P-Rid

CNN-based feature extraction has achieved the state-of-the-art performance in P-Rid owning to a better spatial alignment of local image parts. A novel Harmonious Attention CNN (HA-CNN) proposed by Li *et al.* [**48**] tries to jointly learn attention selection and feature representation in a CNN by maximizing the complementary information of different levels of visual attention (soft attention and hard attention). Liu *et al.* [**56**] proposed a network called CAN which combines

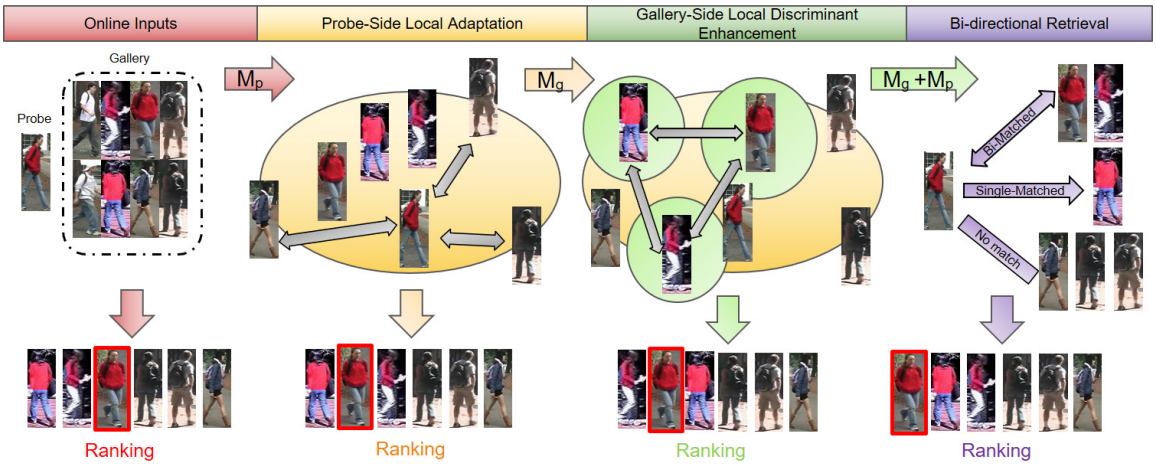


Figure 5.2. For online testing, a query probe and a gallery set is directly tested by a baseline method to achieve the initial ranking list. By performing probe-side local discriminant enhancement, a refined ranking list is obtained. For the top-N_g ranked gallery images, the gallery-side local discriminant enhancement is further performed to adjust the local similarity distributions of galleries. Therefore the final ranking list is obtained by a bi-directional retrieval matching.

attention methods with LSTM to obtain discriminative attention feature of the whole image. Wang *et al.* [**92**] proposed a novel deeply supervised fully attentional block that can be plugged into any CNNs to solve P-Rid problem, and a novel deep network called Mancs is designed to learn stable features for P-Rid. However, these well-trained networks are directly used for the testing data for deep feature extraction, no local adaptation is in the loop. The data shifting between training and testing samples definitely limits the performance of learned models. Therefore, our proposed method is suitable to any CNNs for sample-specific local adaptation in inference stage, which can handle the data shifting problem well and gain a further performance improvement.

5.2.2. Local Metric Learning For P-Rid

In recent years, more attention has been focused on learning local metrics to facilitate P-Rid. To tackle the multi-modal distributions of the person appearances, Zhang *et al.* [**111**] utilized the local distance comparison in P-Rid to obtain an accurate matching. Pedagadi *et al.* [**69**] proposed to combine the FDA and Locality Preserving Projections (LPP) together to exploit the local geometrical information of samples. Li *et al.* [**50**] combined a traditional distance metric with a local decision rule to form a Locally-Adaptive Decision Function (LADF) which aims to improve the local discriminant of given data. To handle the common over-fitting issue in P-Rid, a regularized local metric learning (RLML) method designed by Liong *et al.* [**53**] aims to utilize the merits of both the global and local metrics. A sample-specific SVM classifier is learned in Zhang *et al.* [**114**] for each training sample, then the weight parameters of a testing sample can be inferred. In order to relax the large-number labeled image pair requirement in P-Rid, a novel one-shot learning approach is proposed by Bak*et al.* [**4**] which only requires a single image from each camera for training, thus the learning result is specific to the only sample. However, these local metric learning methods still perform an offline global-learning procedure, a large number of labeled data are required. Their performance is indeed limited if testing data are from different distributions. To address this issue, our proposed method adopts an online local adaptation manner to adapt the offline learned baselines to each testing sample specifically.

5.2.3. Online Rank Refinement For P-Rid

Online rank refinement technique is widely adopted for further performance improvement in P-Rid. Ye *et al.* [**105**] revised the ranking list by considering the nearest neighbors of both the global and local features. An unsupervised re-ranking model proposed Garcia *et al.* [**28**] by

taking advantage of the content and context information in the ranking list. Zhong *et al.* [**127**] proposed a k-reciprocal encoding approach for re-ranking, which relies on a hypothesis that if a gallery image is similar to the probe in the k-reciprocal nearest neighbors, it is more likely to be a true match. Recently, Zhou *et al.* [**129**] proposed a novel online local metric adaptation algorithm to learn an instance-specific Mahalanobis metric for each query sample, and only the negative data are utilized for learning. Barman *et al.* [**5**] focused on how to make a consensus-based decision for retrieval by aggregating the ranking results from multiple algorithms, only the matching scores are needed. Unlike [**105, 28, 127, 5**] that simply treat different probes equally without considering the instance-specific characteristics, our method aims to handle different probes via different local discriminant enhancement. Compared with the probe-centric re-ranking algorithm [**129**], our method jointly utilizes both the probe and gallery data to achieve a better re-ranking performance.

5.3. Our Proposed Method

5.3.1. Problem Settings and Notations

Denote the three given disjoint **training set** $\mathcal{T}$, **probe set** $\mathcal{P}$ and **gallery set** $\mathcal{G}$, as

$$\begin{aligned} \mathcal{T} &= \left\{\left(t_i, l_i^t\right)\right\}_{i=1}^{n_t} \\ \mathcal{P} &= \{(p_i, l_i^p)\}_{i=1}^{n_p} \\ \mathcal{G} &= \{(g_i, l_i^g)\}_{i=1}^{n_g} \end{aligned} \tag{5.1}$$

that $t_i, p_i, g_i \in \mathbb{R}^d$ are the extracted feature representations from a baseline model, either hand-craft feature or learned deep feature via $\mathcal{T}$. $l_i^t \in \{1, 2, ..., c_t\}$ is the training sample label from c_t classes, and all the samples in $\mathcal{P}$ and $\mathcal{G}$ are drawn from the other c different classes which

have no overlap with the above c_t classes. The common-used *closed-set condition* is adopted that both the $\mathcal{P}$ and $\mathcal{G}$ contain samples from all the c classes respectively. Follow the setting in [**129**], an additional negative sample database (NDB), denoted by $\mathcal{Z} = \{z_i\}_{i=1}^{k}$ is provided.

5.3.2. Adapt $\mathcal{P}$ via Slacked-OLMANS

The goal of OLMANS [**129**] is to adaptively adjust the local similarities for all the samples in $\mathcal{P}$ by solely utilizing the NDB $\mathcal{Z}$. Revisit the objective in [**129**]:

$$\begin{aligned} &\mathbf{M}_{p_i} = \arg\min_{\mathbf{M}_{p_i}} \frac{1}{2}\|\mathbf{M}_{p_i}\|^2 \\ &w.r.t:\ \mathbf{M}_{p_i} \succeq 0 \\ &(p_i - z_j)^T \mathbf{M}_{p_i} (p_i - z_j) \geq 2,\ \forall 1 \leq j \leq k \end{aligned} \tag{5.2}$$

Eqn. 5.2 can be efficiently solved by a kernel-SVM solver in [**129**]. However, the original objective Eqn. 5.2 has strict margin requirement by the inequality constraints, which will result in over-fitting on non-separable distributions and performance degradation on the severe cases. In order to introduce more flexibility, we modify Eqn. 5.2 by adding the slack variables $\{\xi_j\}$ in the loop so that the very few extremely hard negatives is tolerated during learning.

$$\begin{aligned} &\mathbf{M}_{p_i} = \arg\min_{\mathbf{M}_{p_i}} \frac{1}{2}\|\mathbf{M}_{p_i}\|^2 + \Theta \sum \xi_j \\ &w.r.t:\ \mathbf{M}_{p_i} \succeq 0 \\ &(p_i - z_j)^T \mathbf{M}_{p_i} (p_i - z_j) \geq 2 - \xi_j,\ \forall 1 \leq j \leq k \\ &\xi_j \geq 0,\ \forall 1 \leq j \leq k \end{aligned} \tag{5.3}$$

Follow a similar proof in [**129**], by eliminating the PSD constraint in Eqn. 5.3, the above objective still can be solved via a kernel-SVM solver with slack variable constraints.

5.3.3. Adapt $\mathcal{G}$ via Local Discriminant Enhancement

Once $\mathbf{M}_{p_i}$ is learned, a re-ranking list $\mathcal{R}_{p_i} = \{g_{1,p_i}, g_{2,p_i}, ..., g_{N_g,p_i}\}$ containing the top-N_g retrieval results for p_i is obtained via:

$$D_{\mathbf{M}_{p_i}}(p_i, g_j) = \|p_i, g_j\|^2_{\mathbf{M}_{p_i}} = (p_i - g_j)^T \mathbf{M}_{p_i}(p_i - g_j) \tag{5.4}$$

However, as we present in Fig. 5.1, the re-ranking list $\mathcal{R}_{p_i}$ (2nd row) may suffer from ambiguous distractors. The similar gallery images from different identities will significantly degrade the discriminant of $\mathbf{M}_{p_i}$ since such distractors are still indistinguishable under $\mathbf{M}_{p_i}$. So the local discriminant of these indistinguishable gallery samples needs to be further enhanced. To facilitate the reading, we re-write $\mathcal{R}_{p_i} = \{g_1, g_2, ..., g_{N_g}\}$ for simplicity and their identity label is $\{l_{g_1}, l_{g_2}, ..., l_{g_{N_g}}\}$. Thus for each $g_j \in \mathcal{R}_{p_i}$, we optimize the following objective:

$$\begin{aligned}
&\mathbf{M}_{g_j} = \arg\min_{\mathbf{M}_{g_j}} \frac{1}{2}\left\|\mathbf{M}_{g_j}\right\|^2 + \Theta_+ \sum \xi_j^{pos} + \Theta_- \sum \xi_j^{neg} \\
&w.r.t: \ \mathbf{M}_{g_j} \succeq 0 \\
&\left(g_j - g_j^{pos}\right)^T \mathbf{M}_{g_j} \left(g_j - g_j^{pos}\right) \leq \xi_j^{pos}, \\
&\forall g_j, g_j^{pos} \in \mathcal{R}_{p_i}, l_{g_j} = l_{g_j^{pos}} \\
&\left(g_j - g_j^{neg}\right)^T \mathbf{M}_{g_j} \left(g_j - g_j^{neg}\right) \geq 2 - \xi_j^{neg}, \\
&\forall g_j, g_j^{neg} \in \mathcal{R}_{p_i}, l_{g_j} \neq l_{g_j^{neg}}
\end{aligned} \tag{5.5}$$

The Eqn. 5.5 proposes to collapse the same-identity galleries g_j^{pos} together to reduce inter-identity variation, meanwhile push the different-identity ones g_j^{neg} far away, so that the local discriminant of g_j is enhanced. By transforming Eqn. 5.5 via a spanning space projection, the optimization of Eqn. 5.5 is proved to be solving a projected version [**25**]:

Theorem 11. *Define the spanning space $\mathcal{S} = span\{g_j - g_j^{pos}\}_j$ and its orthogonal space $\mathcal{S}^{\perp}$, then project $g_j - g_j^{neg}$ on $\mathcal{S}^{\perp}$ to obtain a projected vector denoted as v_j. The solution of Eqn. 5.5 is equivalent to solve a re-formed objective Eqn. 5.6.*

$$
\begin{aligned}
&\boldsymbol{M}_{g_j} = \arg\min_{\boldsymbol{M}_{g_j}} \frac{1}{2}\left\|\boldsymbol{M}_{g_j}\right\|^2 + \Theta_- \sum \xi_j^{neg} \\
&w.r.t: \ \boldsymbol{M}_{g_j} \succeq 0 \\
&{v_j}^T \boldsymbol{M}_{g_j} v_j \geq 2 - \xi_j^{neg}, \ \forall g_j^{neg} \in \mathcal{R}_{p_i} \\
&\boldsymbol{M}_{g_j} s = 0, \ \forall s \in \mathcal{S}
\end{aligned}
\tag{5.6}
$$

PROOF. Since $\mathbf{M}_{g_j}$ is PSD, the constraint $\left(g_j - g_j^{pos}\right)^T \mathbf{M}_r \left(g_j - g_j^{pos}\right) \leq \xi_j^{pos}$ is equivalent to $\mathbf{M}_{g_j}\left(g_j - g_j^{pos}\right) = 0$ which means the $\mathbf{M}_{g_j} s = 0$ for all $s \in \mathcal{S}$. By projecting $g_j - g_j^{neg}$ on $\mathcal{S}$ and $\mathcal{S}^{\perp}$ to obtain two projected vectors denoted as s_j and v_j respectively, each $g_j - g_j^{neg}$ is equivalent to the summation of this two orthogonal bases, $g_j - g_j^{neg} = s_j + v_j$. So the negative constraints in Eqn 5.5 is re-written as:

$$
\begin{aligned}
&\left(g_j - g_j^{pos}\right)^T \mathbf{M}_{g_j} \left(g_j - g_j^{pos}\right) \\
&= \left(s_j + v_j\right)^T \mathbf{M}_{g_j} \left(s_j + v_j\right) \\
&= {v_j}^T \mathbf{M}_{g_j} v_j
\end{aligned}
\tag{5.7}
$$

Now we can re-form Eqn. 5.5 via replacing the negative constraints in Eqn. 5.5 by Eqn. 5.7, which gives us the equivalent object Eqn. 5.6. □

Revisit Eqn. 5.6, we prove that it is equivalent to a kernel slacked-SVM problem by relaxing its PSD constraint $\mathbf{M}_{g_r} \succeq 0$ and equality constraints, and the solution is still a PSD metric.

Theorem 12. *The solution to Eqn. 5.5 is actually equivalent by relaxing the equality and PSD constraints in Eqn. 5.6, since they are indeed off-the-shelf.*

PROOF. Eliminating the equality and PSD constraints in Eqn. 5.6 gives us:

$$\begin{aligned} &\mathbf{M}_{g_j} = \arg\min_{\mathbf{M}_{g_j}} \frac{1}{2}\left\|\mathbf{M}_{g_j}\right\|^2 + \Theta_- \sum \xi_j^{neg} \\ &w.r.t : {v_j}^T \mathbf{M}_{g_j} v_j \geq 2 - \xi_j^{neg},\ \forall g_j^{neg} \in \mathcal{R}_{p_i} \end{aligned} \tag{5.8}$$

Eqn. 5.8 is exactly the same form of the objective in [**129**] but with slack variables. So the positive semi-definiteness of $\mathbf{M}_{g_r}$ is guaranteed even if no PSD constraint is explicitly imposed since $\mathbf{M}_{g_r} = \sum \alpha_i \varphi(v_i) = \sum \alpha_i v_i \cdot v_i^T \succeq 0$. For the equality constraints in Eqn. 5.6, $\forall s \in \mathcal{S}$, we have:

$$\mathbf{M}_{g_j} s = \left(\sum \alpha_i v_i \cdot v_i^T\right) s = \sum \alpha_i v_i \cdot (v_i^T s) = 0 \tag{5.9}$$

so the equality constraints is satisfied as well. □

Figure 5.3. The visualization of rank improvement on CUHK03(top two cases) and Market1501(bottom two cases) based on HA-CNN. For each case, its top-10 (from lest to right) matches are presented and the true-match is labeled by the red box. The 1st row is the baseline result, the 2nd row is the result only using M_p and the 3rd row is the result using our full model.

5.3.4. Re-Ranking From Bi-Directional Retrieval

Finally, for a query probe image $p_i \in \mathcal{P}$ and an enrolled gallery image $g_j \in \mathcal{G}$, the final distance between them is:

$$\begin{aligned} & D(p_i, g_j)_{\mathbf{M}_{p_i}+\lambda\mathbf{M}_{g_j}} \\ & = \|p_i, g_j\|^2_{\mathbf{M}_{p_i}} + \lambda\|p_i, g_j\|^2_{\mathbf{M}_{g_j}} \\ & = (p_i - g_j)^T(\mathbf{M}_{p_i} + \lambda\mathbf{M}_{g_j})(p_i - g_j) \end{aligned} \tag{5.10}$$

where the λ is a weight parameter to balance the importance of $\mathbf{M}_{p_i}$ and $\mathbf{M}_{g_j}$. For each p_i and $\forall g_j \in \mathcal{R}_{p_i}$, based on $D(p_i, g_j)\mathbf{M}_{p_i} + \lambda\mathbf{M}_{g_j}$, a refined ranking list $\mathcal{R}^*_{p_i}$ is obtained as the final retrieval result. Previous proofs in [**129**] have guaranteed the improvement of rankings by $\mathbf{M}_{p_i}$. As proved by Theorem. 11 and Theorem. 12, the local discriminant enhancement via $\mathbf{M}_{g_j}$ also guarantees the improvement of rankings since the learning of $\mathbf{M}_{g_j}$ is equivalent to the learning of $\mathbf{M}_{p_i}$. Eqn. 5.10 is a late fusion of $\mathbf{M}_{p_i}$ and $\mathbf{M}_{g_j}$, the fused result is better than each single kernel which has been verified by previous multi-task and multi-kernel learning researches [**88, 81, 82**].

5.4. Experiments

5.4.1. Experimental Settings

Datasets. We mainly focus on four large-scale challenging P-Rid benchmarks: CUHK03 [**46**], Market1501 [**119**], DukeMTMC-reID [**124**] and MSMT17 [**97**]. The statistic details of the above datasets are summarized in Table. 5.1. For CUHK03 [1], the new splitting protocol proposed by [**127**] is adopted in our experiment so that 767 identities are used for training as well

[1] In our experiment, the CUHK detected dataset is utilized

as the left 700 identities are used for testing. As for the other three benchmarks, Market1501, DukeMTMC-reID and MSMT17, the pre-determined probe and gallery sets are directly utilized with no modification.

Dataset	**cuhk03**[**46**]	**market**[**119**]	**duke**[**124**]	**msmt17**[**97**]
#T-IDs	767	751	702	1040
#P-IDs	700	750	702	3060
#G-IDs	700	751	1110	3060
#cam	2	6	8	15
#images	28192	32668	36411	126441

Table 5.1. The statistics of CUHK03 [**46**], Market1501 [**119**], DukeMTMC-reID [**124**] and MSMT17 [**97**] benchmarks.

Baselines. Several most recent CNN-based P-Rid models are selected as our baselines to implement our proposed method to: ResNet50 [**34**], DenseNet121 [**37**] and HA-CNN [**48**]. Besides, the other state-of-the-art P-Rid methods [**38, 75, 113, 85, 86, 11, 84**] are further compared. Moreover, some **online rank refinement** methods including the OLMANS [**129**] and a recently proposed re-ranking approach [**127**] are compared with our algorithm. Various ablation studies of our proposed model are explored in Sec. 5.4.5.

Method	CUHK03			Market1501			DukeMTMC-reID			MSMT17		
	R@1	**R@20**	**mAP**	**R@1**	**R@20**	**mAP**	**R@1**	**R@20**	**mAP**	**R@1**	**R@20**	**mAP**
SqueezeNet[**38**]	26.9	N/A	25.8	72.2	N/A	47.9	58.8	N/A	37.8	30.6	N/A	13.0
MobileNetv2[**75**]	41.0	N/A	40.3	84.2	N/A	65.8	73.2	N/A	52.5	44.9	N/A	21.1
SuffleNet[**113**]	31.9	N/A	31.7	80.0	N/A	58.4	69.3	N/A	46.8	39.6	N/A	17.8
ResNet50[**34**]	47.9	84.8	46.8	88.5	98.3	71.3	77.7	92.9	58.8	63.4	86.1	34.2
DenseNet121[**37**]	41.0	76.8	40.1	88.2	97.9	69.2	78.6	93.0	58.5	66.0	86.6	34.6
HA-CNN[**48**]	48.0	85.4	47.6	90.6	98.3	75.3	80.7	94.3	64.4	61.8	85.8	34.6
Ours(ResNet50)	**66.9**	**87.4**	**60.7**	**95.4**	**98.8**	**82.6**	**84.7**	**94.9**	**68.5**	**72.8**	**88.6**	**55.0**
Ours(DenseNet121)	**61.6**	**81.6**	**54.4**	**95.3**	**98.6**	**81.2**	**84.9**	**95.1**	**68.0**	**75.5**	**89.9**	**43.1**
Ours(HA-CNN)	**69.8**	**88.8**	**63.5**	**96.3**	**98.9**	**85.2**	**87.1**	**95.8**	**72.2**	**68.0**	**87.8**	**37.8**

Table 5.2. Comparison results w/ and w/o our proposed algorithm on CUHK03, Market1501, DukeMTMC-reID and MSMT17.

Evaluation. For evaluation, we follow the same official evaluation protocols in [**119, 124, 46, 97**], the single-shot evaluation setting is adopted and all the results are shown in the form of Cumulated Matching Characteristic (CMC) at several selected ranks and mean Average Precision (mAP).

CUHK03			Market-1501			DukeMTMC-reID		
Method	**R@1**	**mAP**	**Method**	**R@1**	**mAP**	**Method**	**R@1**	**mAP**
ResNet50[**34**]	47.9	46.8	ResNet50[**34**]	88.5	71.3	ResNet50[**34**]	77.7	58.8
DenseNet121[**37**]	41.0	40.1	DenseNet121[**37**]	88.2	69.2	DenseNet121[**37**]	78.6	58.5
HA-CNN[**48**]	48.0	47.6	HA-CNN[**48**]	90.6	75.3	HA-CNN[**48**]	80.7	64.4
PCB[**86**]	63.7	67.5	PCB[**86**]	83.3	69.2	PCB[**86**]	83.3	69.2
SVDNet[**85**]	41.5	37.3	SVDNet[**85**]	82.3	62.1	SVDNet[**85**]	76.7	56.8
DPFL[**14**]	40.7	37.0	DNSL[**112**]	61.0	35.6	DuATM[**78**]	81.8	64.6
Mancs[**92**]	69.0	63.9	Mancs[**92**]	93.1	82.3	SPreID[**40**]	85.9	73.3
PAN[**125**]	36.3	34.0	Part-aligned[**84**]	91.7	79.6	Part-aligned[**84**]	84.4	69.3
MLFN[**11**]	52.8	47.8	PN-GAN[**72**]	77.1	63.6	PAN[**125**]	71.6	51.5
DaRe[**96**]	55.1	51.3	DeepCC[**73**]	89.5	75.7	GAN[**124**]	67.7	47.1
Ours(ResNet50)	**66.9**	**60.7**	**Ours(ResNet50)**	**95.4**	**82.6**	**Ours(ResNet50)**	**84.7**	**68.5**
Ours(DenseNet121)	**61.6**	**54.4**	**Ours(DenseNet121)**	**95.3**	**81.2**	**Ours(DenseNet121)**	**84.9**	**68.0**
Ours(HA-CNN)	**69.8**	**63.5**	**Ours(HA-CNN)**	**96.3**	**85.2**	**Ours(HA-CNN)**	**87.1**	**72.2**

Table 5.3. State-of-the-art comparison results on CUHK03, Market1501 and DukeMTMC-reID. **All the results are the best performances reported in their literatures**

Method	CUHK03	Market1501	DukeMTMC
HA-CNN[**48**]	48.0(47.6)	90.6(75.3)	80.7(64.4)
HA-CNN+RR [**127**]	54.8(55.7)	91.4(79.0)	82.5(69.9)
HA-CNN+OL [**129**]	62.3(56.5)	92.7(78.9)	83.7(67.8)
HA-CNN+Ours	**69.8**(63.5)	**96.3**(**85.2**)	**87.1**(72.2)
HA-CNN+RR [**127**]+Ours	67.2(**63.7**)	95.8(**85.2**)	86.4(**72.5**)
Dense121[**37**]	41.0(40.1)	88.2(69.2)	78.6(58.5)
Dense121+RR [**127**]	48.1(51.5)	90.2(85.0)	83.7(**76.9**)
Dense121+OL [**129**]	53.1(49.3)	90.4(74.0)	80.2(64.1)
Dense121+Ours	**61.6**(54.4)	**95.3**(81.2)	84.9(68.0)
Dense121+RR [**127**]+Ours	56.7(**56.0**)	94.2(**85.3**)	**86.3**(74.7)

Table 5.4. Comparison of different online rank refinement methods. **Rank@1(mAP)** performance is reported.

5.4.2. Implementation Details

All the experiments are conducted in the same experimental setting for a fair comparison. The NDB $\mathcal{Z}$ is simply determined as the unlabeled probe set $\mathcal{P}$ for $\mathbf{M}_p$ learning, so no extra data is utilized. For efficiency and effectiveness consideration, for each probe p_i, not all the samples in $\mathcal{Z}$ are used but we only choose the top-N_p hard negatives in $\mathcal{Z}$ for learning (the same N_p setting is also applied to [**129**]). So we set $N_g = N_p = 100$ for the learning of Eqn. 5.3 and Eqn. 5.5. A GPU-based SVM solver, ThunderSVM [**100**] is used for the optimization of our method, the slack variables $\xi_j, \xi_j^{pos}, \xi_j^{neg}$ are set to be 1 as default. The well-trained parameters for ResNet-50 [**34**], DenseNet121 [**37**] and HA-CNN [**48**] by a Pytorch implementation [2] are utilized which achieve the comparative results as in the papers. The weighting parameter λ in Eqn. 5.10 is set to be 1 for all the experiments. All the experiments are conducted on a remote server with 256G memory and four Titan X Pacal GPUs.

[2]https://github.com/KaiyangZhou/deep-person-reid

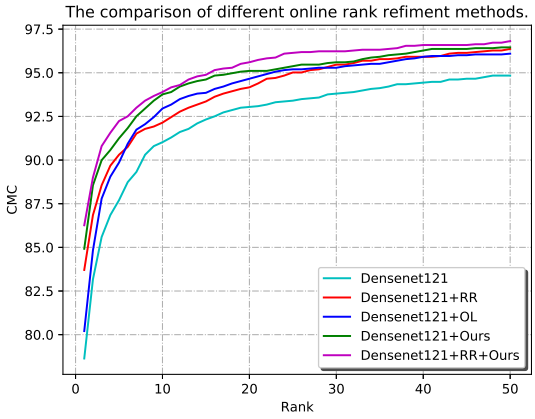
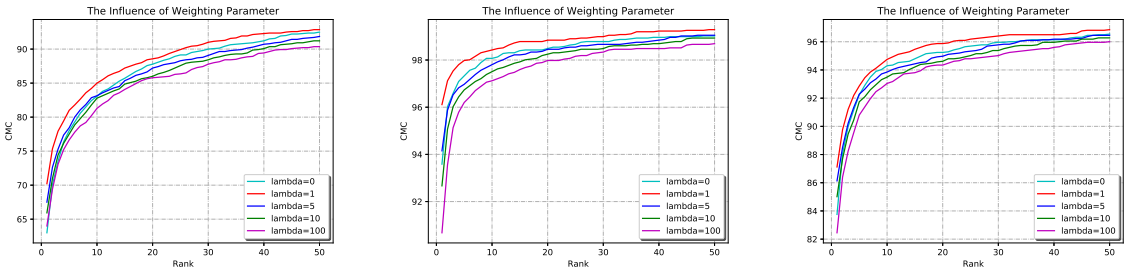


Figure 5.4. The full CMC plot of DenseNet121 on DukeMTMC-reID dataset.



(a) CUHK03 (b) Market1501 (c) DukeMTMC-reID

Figure 5.5. The influence of λ on (a) CUHK03, (b) Market1501 and (c) DukeMTMC-reID based on HA-CNN baseline.

5.4.3. Comparison with the State-of-the-arts

Evaluation on CUHK03: The experimental results under the novel 767/700 splitting protocol are presented in Table. 5.2 and Table. 5.3. Our method significantly boosts the baseline Rank@1(mAP) performance of ResNet50, DenseNet121 and HA-CNN, from 47.9%(46.8%), 41.0%(40.1%) and 48.0%(47.6%) to 66.9%(60.7%), 61.6%(54.4%) and 69.8%(63.5%), respectively with a 40.0%(29.7%), 50.2%(35.7%) and 45.4%(33.4%) relative improvement. Even compared with the state-of-the-art results in Table. 5.3, our model can still beat them by a large

Figure 5.6. The efficiency-effectiveness trade-off on CUHK03 based on HA-CNN baseline.

Method	**CUHK03**			**Market1501**		
	R@1	**R@20**	**mAP**	**R@1**	**R@20**	**mAP**
HA-CNN[**48**]	48.0	85.4	47.6	90.6	98.3	75.3
Our only w/ M_p w/o ξ_j	62.3	86.5	56.5	92.7	98.3	78.9
Our only w/ M_p	63.4	87.6	63.5	93.8	98.8	81.2
Our only w/ M_g	65.4	86.2	57.3	94.2	98.4	79.1
Our-Full	**69.8**	**88.8**	**63.5**	**96.3**	**98.9**	**85.2**
Method	**DukeMTMC-reID**			**MSMT17**		
	R@1	**R@20**	**mAP**	**R@1**	**R@20**	**mAP**
HA-CNN[**48**]	80.7	94.3	64.4	61.8	85.8	34.6
Our only w/ M_p w/o ξ_j	83.7	94.8	67.8	66.1	86.7	37.0
Our only w/ M_p	83.9	95.3	69.0	66.6	87.3	37.9
Our only w/ M_g	83.6	94.4	65.7	63.1	83.3	33.4
Our-Full	**87.1**	**95.8**	**72.2**	**68.0**	**87.8**	**37.8**

Table 5.5. The ablation study about the influence of each component in our algorithm.

margin. The visualization results of rank improvement shown in Fig. 5.3 also demonstrate the effectiveness of our method.

Evaluation on Market1501: Table. 5.2 and Table. 5.3 show the comparison results of our method on the baselines and against the state-of-the-art results. Although the state-of-the-art approaches have achieved a pretty high performance ($\geq 90\%$) on Market1501, the improvement

Setting	N_g=10	N_g=50	N_g=100	N_g=200
N_p=10	54.1(47.1)	65.3(58.2)	67.1(59.7)	67.4(60.3)
N_p=50	60.2(52.2)	69.1(62.3)	70.1(63.4)	70.1(63.8)
N_p=100	60.4(52.4)	70.0(62.6)	70.2(63.5)	70.1(63.8)
N_p=200	60.1(52.4)	69.4(62.5)	70.0(63.5)	69.9(63.8)

Table 5.6. The ablation study about the influence of the number of negative sample. The Rank@1(mAP) results of HA-CNN on CUHK03(767/700) are reported.

of our method is over 6%(10%) on Rank@1(mAP) for all the three baselines by handling the "hard" galley samples well (Fig. 5.3).

Evaluation on DukeMTMC-reID: DukeMTMC-reID is a recent benchmark proposed for P-Rid, but the lasted methods have obtained promising performances. As show in Table. 5.3, the recently published methods, SPreID [**40**], PCB [**86**] and Part-aligned [**84**], boost the state-of-the-art to 85.9%(73.3%). By implementing our proposed method on HA-CNN, the Rank@1 (mAP) result is boosted from 80.7%(64.4%) to 87.1%(72.2%), which beats SPreID by a large margin.

Evaluation on MSMT17: MSMT17 is the latest and largest benchmark proposed recently. The extreme large-scale identities and a large number of distractors make this dataset pretty challenging. We evaluate the performance of the baselines on the MSMT17 dataset with(w/) and without(w/o) our algorithm which are reported in Table. 5.2. Our method improve the DenseNet121 Rank@1(mAP) performance from 66.0% (34.6%) to a state-of-the-art 75.5% (43.1%).

5.4.4. Comparison with Related Works

Two state-of-the-art online rank refinement methods, OL [**129**] and RR [**127**], are compared with our algorithm. The comparison results in Table. 5.4 show that our proposed method performs better than the other two approaches by a large margin at both Rank@1 and mAP evaluation. OL [**129**] works better on improving Rank@1 performance but has little improvement on mAP due to the lack of gallery-specific local discriminant enhancement. In contrast, since RR [**127**] considers the k-reciprocal nearest neighbors of both probe and gallery data, it achieves a large improvement on mAP but with limited improvement on Rank@1 owing to the lack of instance-specific local adaptation. By integrating the RR [**127**] result with our method, its Rank@1(mAP) performance is further boosted which is also shown in Fig. 5.4.

5.4.5. Ablation Study

The Influence of Model Components: The final retrieval performance of Eqn. 5.10 relies on a bi-directional retrieval matching, so the performance of keeping each component is shown in Table. 5.5. As can be seen, the introduction of slack variable actually helps. By only keeping the probe-side adapted metric $\mathbf{M}_p$ or the gallery-side one $\mathbf{M}_g$, we still can achieve a significant improvement. While by performing bi-directional matching as a full-model, the performance is further boosted by a large margin.

The Influence of λ in Eqn. 5.10: The weighting parameter λ in Eqn. 5.10 will balance the importance of bi-directional local discriminant enhancement. The full CMC performances w.r.t λ of HA-CNN on CUHK03, Market1501 and DukeMTMC-reID are plotted in Fig. 5.5 respectively. As can be seen, setting $\lambda = 1$ gives the best performance since we perform

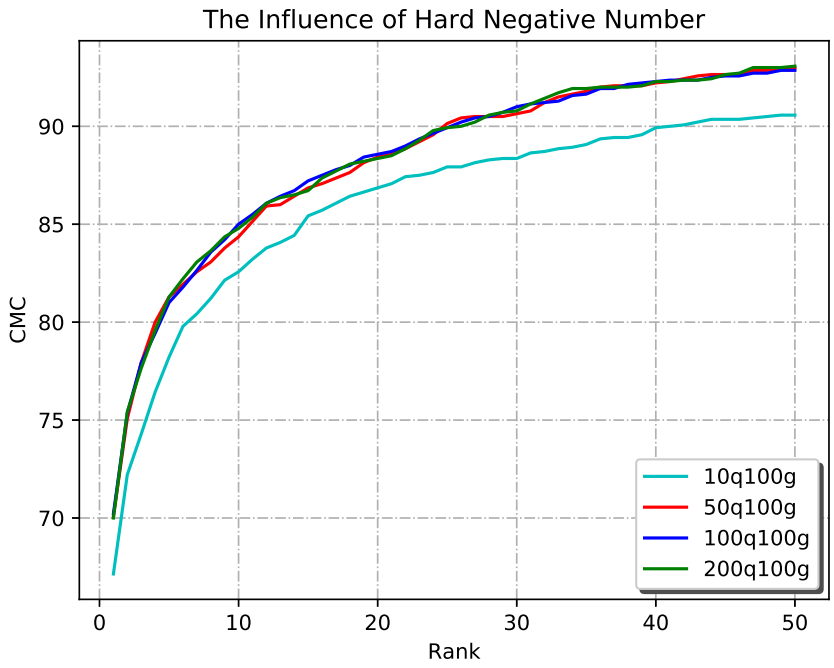


Figure 5.7. The influence of N_p and N_g on CUHK03 based on HA-CNN baseline.

$\max$ normalization to both $\mathbf{M}_p$ and $\mathbf{M}_g$ during learning, over-weighting either side is prone to suppress the other side's impact.

The Influence of N_p and N_g: As proved in [**129**], although massive negatives can be accessed, the performance of Eqn. 5.3 and Eqn. 5.5 rely on the small-amount hard negative samples, which is controlled by the parameters N_p and N_g. The influence of both N_p and N_g is explored in Table. 5.6. If N_p or N_g is too small, few useful negatives are utilized resulting in a poor learning performance. While if N_p or N_g is too large, the further performance improvement is trivial but causes an increase of computational burden.

The Efficiency-Effectiveness Trade-off: The extra testing time introduced by our proposed method includes the local discriminant enhancement of both probes and gallery data.

Such extra time is controlled by the values of N_p and N_g, meanwhile different negative sample numbers give different performances (As shown in Table. 5.6). Therefore we study the efficiency-effectiveness trade-off of our method based on HA-CNN baseline which is presented in Fig. 5.6. By increasing N_p and N_g gradually, the time cost still increases, but the Rank@1 performance is prone to be fixed. Therefore, by choosing $N_p = 100$ and $N_g = 100$, the trade-off between extra computation time and performance improvement is acceptable.

5.5. Discussion

Existing online rank refinement methods for P-Rid either handle different probes equally without considering the individual characteristics or only perform probe-centric learning by ignoring the gallery data. Therefore their re-ranking performance is neither significant nor stable. In this work, we propose a two-stage hierarchical local discriminant enhancement algorithm to simultaneously refine the local metric for each probe/gallery instance specifically. Our proposed method can be readily applied to any existing P-Rid baselines with the guarantee of performance improvement, and a theoretical sound optimization solution keeps a low online computational burden. Compared with the other state-of-the-art rank refinement approaches, our method achieves significant improvement on Rank@1(mAP) performance. Moreover, by implementing our method to the state-of-the-art baselines, their performance is further boosted by a large margin on four main large-scale P-Rid benchmarks.

CHAPTER 6

Chapter 6. Conclusion

The visual matching problem plays an important role in computer vision area. Many computer vision tasks rely on it as the core component, including the landmark-based face recognition [**24**], template-based object tracking [**126**], exemplar-based super resolution [**128**], image-based person re-identification [**121**], etc. The quality of visual matching directly determines the performances of these tasks. In order to obtain robust and discriminative visual matching, various approaches are proposed among which the metric learning ones proved to be a powerful tool for learning good quality visual matching. However, due to severe small-size sample problem of visual matching data, the learning ability of metric learning methods is significantly limited.

In order to address the critical small-size sample problem in visual matching, this dissertation proposes various metric learning methods based on small-size positive samples, including few-shot positives, only one positive, unlabeled samples and the mixture of labeled and unlabeled samples. All these small-size sample settings are the critical issues that must be addressed to facilitate the learning of visual matching as well as make the applications of visual matching in the intelligent video surveillance better. To summarize, the following contributions have been made in the dissertation:

- A novel learning constraint called reference constraint is proposed to facilitate the poor and difficult metric learning solution caused by the large-scale and imbalanced constraints used before. Our proposed reference constraint aims to associate the samples from the same class to one or multiple By utilizing our proposed reference constraint, a ridge-regression based global metric learning from few positives and no negatives is proposed to learn a discriminative metric. A closed-form solution can be obtained for our metric learning objective so that the learning is not only effective and also efficient compared with the related metric learning approaches [**52, 51, 112**].
- In order to address the two critical issues of the proposed global metric learning work, including the failure on one-shot positive scenario which is an extremely challenging small-size sample setting and the failure on handling the hard negative distractors, a novel online instance-specific local metric learning is proposed by using only one positive but a large number of negatives for learning. Our proposed online local metric adaptation algorithm can be applied to any offline learned baselines on any features, and an efficient optimization solution is proposed to our method which requires very trivial online learning cost. Three theoretical sound justifications guarantee the improvement of our method under both the asymptotic scenario and practical learning scenario.
- A novel online joint multi-metric learning algorithm is designed via learning from unlabeled samples. By considering the given matching samples as an unlabeled batch-shot query set, the intrinsic visual similarity sharing relationships among the samples can be utilized by mining different sharing-subsets. For each sharing-subset, the samples share the same visual similarity relationship are grouped from which a joint

Mahalanobis metric is learned to jointly adapt the local distribution of all the subset samples. Compared with our instance-specific local metric adaptation work, our joint multi-metric learning algorithm is not only more effective for matching performance improvement but also has lower online learning cost.

- A novel bidirectional local discriminant enhancement from the combination of few-shot labeled and unlabeled samples is proposed to perform a two-stage hierarchical local metric adaptation for both the probe and gallery samples in visual matching. Unlike the previous method which only focus on the local discriminant enhancement of the given matching probes, the local discriminant of the gallery samples is also enhanced by our method so that the "hard" gallery distractors which are indistinguishable will be well tackled by our method.

References

[1] AHMED, E., JONES, M., AND MARKS, T. K. An improved deep learning architecture for person re-identification. *Differences 5* (2015), 25.

[2] BAI, S., AND BAI, X. Sparse contextual activation for efficient visual re-ranking. *IEEE Transactions on Image Processing 25*, 3 (2016), 1056–1069.

[3] BAI, S., BAI, X., AND TIAN, Q. Scalable person re-identification on supervised smoothed manifold. In *CVPR* (2017).

[4] BAK, S., AND CARR, P. One-shot metric learning for person re-identification. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on* (2017), IEEE, pp. 1571–1580.

[5] BARMAN, A., AND SHAH, S. K. Shape: A novel graph theoretic algorithm for making consensus-based decisions in person re-identification systems. In *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), IEEE, pp. 1124–1133.

[6] BARTLETT, P. L., AND MENDELSON, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research 3*, Nov (2002), 463–482.

[7] BEDAGKAR-GALA, A., AND SHAH, S. K. A survey of approaches and trends in person re-identification. *Image and Vision Computing 32*, 4 (2014), 270–286.

[8] BOUSQUET, O., AND ELISSEEFF, A. Stability and generalization. *Journal of Machine Learning Research 2*, Mar (2002), 499–526.

[9] CARUNA, R. Multitask learning: A knowledge-based source of inductive bias. In *Machine Learning: Proceedings of the Tenth International Conference* (1993), pp. 41–48.

[10] CHANG, C.-C., AND LIN, C.-J. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST) 2*, 3 (2011), 27.

[11] CHANG, X., HOSPEDALES, T. M., AND XIANG, T. Multi-level factorisation net for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision* (2018), vol. 1, p. 2.

[12] CHAWLA, N. V., BOWYER, K. W., HALL, L. O., AND KEGELMEYER, W. P. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research 16* (2002), 321–357.

[13] CHEN, D., YUAN, Z., HUA, G., ZHENG, N., AND WANG, J. Similarity learning on an explicit polynomial kernel feature map for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 1565–1573.

[14] CHEN, Y., ZHU, X., AND GONG, S. Person re-identification by deep learning multi-scale representations. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 2590–2600.

[15] CHENG, D. S., CRISTANI, M., STOPPA, M., BAZZANI, L., AND MURINO, V. Custom pictorial structures for re-identification. In *BMVC* (2011), vol. 1, p. 6.

[16] CHUM, O., PHILBIN, J., SIVIC, J., ISARD, M., AND ZISSERMAN, A. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on* (2007), IEEE, pp. 1–8.

[17] COMANICIU, D., AND MEER, P. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence 24*, 5 (2002), 603–619.

[18] CORTES, C., MOHRI, M., AND WESTON, J. A general regression framework for learning string-to-string mappings. *Predicting Structured Data* (2007), 143–168.

[19] COURTY, N., FLAMARY, R., AND TUIA, D. Domain adaptation with regularized optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (2014), Springer, pp. 274–289.

[20] COVER, T., AND HART, P. Nearest neighbor pattern classification. *IEEE transactions on information theory 13*, 1 (1967), 21–27.

[21] CUTURI, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems* (2013), pp. 2292–2300.

[22] DAVIS, J. V., KULIS, B., JAIN, P., SRA, S., AND DHILLON, I. S. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning* (2007), ACM, pp. 209–216.

[23] DING, S., LIN, L., WANG, G., AND CHAO, H. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition* (2015).

[24] EDWARDS, G. J., COOTES, T. F., AND TAYLOR, C. J. Face recognition using active appearance models. In *European conference on computer vision* (1998), Springer, pp. 581–595.

[25] FETAYA, E., AND ULLMAN, S. Learning local invariant mahalanobis distances. *Proceedings of The 32st International Conference on Machine Learning* (2015).

[26] FUKUNAGA, K. *Introduction to statistical pattern recognition*. Academic press, 2013.

[27] GAO, H., YU, L., HUANG, Y., DONG, Y., AND CHAN, S. Multi-task learning for person re-identification. In *IScIDE* (2017).

[28] GARCIA, J., MARTINEL, N., MICHELONI, C., AND GARDEL, A. Person re-identification ranking optimisation by discriminant context information analysis. In *Proceedings of the IEEE International Conference on Computer Vision* (2015), pp. 1305–1313.

[29] GHEISSARI, N., SEBASTIAN, T. B., AND HARTLEY, R. Person reidentification using spatio temporal appearance. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on* (2006), vol. 2, IEEE, pp. 1528–1535.

[30] GLOBERSON, A., AND ROWEIS, S. Metric learning by collapsing classes. In *Nips* (2005), vol. 18, pp. 451–458.

[31] GRAY, D., BRENNAN, S., AND TAO, H. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)* (2007), vol. 3, Citeseer.

[32] GUILLAUMIN, M., VERBEEK, J., AND SCHMID, C. Is that you? metric learning approaches for face identification. In *Computer Vision, 2009 IEEE 12th international conference on* (2009), IEEE, pp. 498–505.

[33] HASTIE, T., AND TIBSHIRANI, R. Discriminant adaptive nearest neighbor classification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 18*, 6 (1996), 607–616.

[34] HE, K., ZHANG, X., REN, S., AND SUN, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 770–778.

[35] HIRZER, M., BELEZNAI, C., ROTH, P. M., AND BISCHOF, H. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis* (2011), Springer, pp. 91–102.

[36] HIRZER, M., ROTH, P. M., KÖSTINGER, M., AND BISCHOF, H. Relaxed pairwise learned metric for person re-identification. In *Computer Vision–ECCV 2012*. Springer, 2012, pp. 780–793.

[37] HUANG, G., LIU, Z., VAN DER MAATEN, L., AND WEINBERGER, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[38] IANDOLA, F. N., HAN, S., MOSKEWICZ, M. W., ASHRAF, K., DALLY, W. J., AND KEUTZER, K. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360* (2016).

[39] JIANG, N., LIU, W., AND WU, Y. Order determination and sparsity-regularized metric learning adaptive visual tracking. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (2012), IEEE, pp. 1956–1963.

[40] KALAYEH, M. M., BASARAN, E., GÖKMEN, M., KAMASAK, M. E., AND SHAH, M. Human semantic parsing for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 1062–1071.

[41] KARANAM, S., LI, Y., AND RADKE, R. J. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *Proceedings of the IEEE International Conference on Computer Vision* (2015), pp. 4516–4524.

[42] KOESTINGER, M., HIRZER, M., WOHLHART, P., ROTH, P. M., AND BISCHOF, H. Large scale metric learning from equivalence constraints. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (2012), IEEE, pp. 2288–2295.

[43] KRIZHEVSKY, A., SUTSKEVER, I., AND HINTON, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (2012), pp. 1097–1105.

[44] LENG, Q., HU, R., LIANG, C., WANG, Y., AND CHEN, J. Person re-identification with content and context re-ranking. *Multimedia Tools and Applications 74*, 17 (2015), 6989–7014.

[45] LI, W., WU, Y., MUKUNOKI, M., AND MINOH, M. Common-near-neighbor analysis for person re-identification. In *Image Processing (ICIP), 2012 19th IEEE International Conference on* (2012), IEEE, pp. 1621–1624.

[46] LI, W., ZHAO, R., AND WANG, X. Human reidentification with transferred metric learning. In *ACCV (1)* (2012), pp. 31–44.

[47] LI, W., ZHAO, R., XIAO, T., AND WANG, X. Deepreid: Deep filter pairing neural network for person re-identification. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on* (2014), IEEE, pp. 152–159.

[48] LI, W., ZHU, X., AND GONG, S. Harmonious attention network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[49] LI, X., ZHENG, W.-S., WANG, X., XIANG, T., AND GONG, S. Multi-scale learning for low-resolution person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision* (2015), pp. 3765–3773.

[50] LI, Z., CHANG, S., LIANG, F., HUANG, T., CAO, L., AND SMITH, J. Learning locally-adaptive decision functions for person verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2013), pp. 3610–3617.

[51] LIAO, S., HU, Y., ZHU, X., AND LI, S. Z. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 2197–2206.

[52] LIAO, S., AND LI, S. Z. Efficient psd constrained asymmetric metric learning for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision* (2015), pp. 3685–3693.

[53] LIONG, V. E., LU, J., AND GE, Y. Regularized local metric learning for person re-identification. *Pattern Recognition Letters 68* (2015), 288–296.

[54] LISANTI, G., MASI, I., AND DEL BIMBO, A. Matching people across camera views using kernel canonical correlation analysis. In *Proceedings of the International Conference on Distributed Smart Cameras* (2014), ACM, p. 10.

[55] LIU, C., CHANGE LOY, C., GONG, S., AND WANG, G. Pop: Person re-identification post-rank optimisation. In *Proceedings of the IEEE International Conference on Computer Vision* (2013), pp. 441–448.

[56] LIU, H., FENG, J., QI, M., JIANG, J., AND YAN, S. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing 26*, 7 (2017), 3492–3506.

[57] LIU, K., MA, B., ZHANG, W., AND HUANG, R. A spatio-temporal appearance representation for viceo-based pedestrian re-identification. In *Proceedings of the IEEE International Conference on Computer Vision* (2015), pp. 3810–3818.

[58] LIU, X., SONG, M., TAO, D., ZHOU, X., CHEN, C., AND BU, J. Semi-supervised coupled dictionary learning for person re-identification. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on* (2014), IEEE, pp. 3550–3557.

[59] LIU, Z., WANG, D., AND LU, H. Stepwise metric promotion for unsupervised video person re-identification. In *Computer Vision (ICCV), 2017 IEEE International Conference on* (2017), IEEE, pp. 2448–2457.

[60] LOY, C. C., XIANG, T., AND GONG, S. Multi-camera activity correlation analysis. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on* (2009), IEEE, pp. 1988–1995.

[61] MA, B., SU, Y., AND JURIE, F. Covariance descriptor based on bio-inspired features for person re-identification and face verification. *Image and Vision Computing 32*, 6 (2014), 379–390.

[62] MA, L., YANG, X., AND TAO, D. Person re-identification over camera networks using multi-task distance metric learning. *Image Processing, IEEE Transactions on 23*, 8 (2014), 3656–3670.

[63] MALISIEWICZ, T., GUPTA, A., EFROS, A., ET AL. Ensemble of exemplar-svms for object detection and beyond. In *Computer Vision (ICCV), 2011 IEEE International Conference on* (2011), IEEE, pp. 89–96.

[64] MARTINEL, N., DAS, A., MICHELONI, C., AND ROY-CHOWDHURY, A. K. Temporal model adaptation for person re-identification. In *European Conference on Computer Vision* (2016), Springer, pp. 858–877.

[65] MCLAUGHLIN, N., DEL RINCON, J. M., AND MILLER, P. C. Person reidentification using deep convnets with multitask learning. *IEEE TCSVT* (2017).

[66] MCLAUGHLIN, N., MARTINEZ DEL RINCON, J., AND MILLER, P. Recurrent convolutional network for video-based person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2016).

[67] NOH, Y.-K., ZHANG, B.-T., AND LEE, D. D. Generative local metric learning for nearest neighbor classification. In *Advances in Neural Information Processing Systems* (2010), pp. 1822–1830.

[68] PAISITKRIANGKRAI, S., SHEN, C., AND HENGEL, A. V. D. Learning to rank in person re-identification with metric ensembles. *arXiv preprint arXiv:1503.01543* (2015).

[69] PEDAGADI, S., ORWELL, J., VELASTIN, S., AND BOGHOSSIAN, B. Local fisher discriminant analysis for pedestrian re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2013), pp. 3318–3325.

[70] PERROT, M., COURTY, N., FLAMARY, R., AND HABRARD, A. Mapping estimation for discrete optimal transport. In *Advances in Neural Information Processing Systems* (2016), pp. 4197–4205.

[71] PERROT, M., AND HABRARD, A. Regressive virtual metric learning. In *Advances in Neural Information Processing Systems* (2015), pp. 1810–1818.

[72] QIAN, X., FU, Y., XIANG, T., WANG, W., QIU, J., WU, Y., JIANG, Y.-G., AND XUE, X. Pose-normalized image generation for person re-identification. In *The European Conference on Computer Vision (ECCV)* (September 2018).

[73] RISTANI, E., AND TOMASI, C. Features for multi-target multi-camera tracking and re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 6036–6046.

[74] ROTH, P. M., HIRZER, M., KÖSTINGER, M., BELEZNAI, C., AND BISCHOF, H. Mahalanobis distance learning for person re-identification. In *Person Re-Identification*. Springer, 2014, pp. 247–267.

[75] SANDLER, M., HOWARD, A., ZHU, M., ZHMOGINOV, A., AND CHEN, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 4510–4520.

[76] SHI, H., YANG, Y., ZHU, X., LIAO, S., LEI, Z., ZHENG, W., AND LI, S. Z. Embedding deep metric for person re-identification: A study against large variations. In *European Conference on Computer Vision* (2016), Springer, pp. 732–748.

[77] SHI, Z., HOSPEDALES, T. M., AND XIANG, T. Transferring a semantic representation for person re-identification and search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 4184–4193.

[78] SI, J., ZHANG, H., LI, C.-G., KUEN, J., KONG, X., KOT, A. C., AND WANG, G. Dual attention matching network for context-aware feature sequence based person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2018).

[79] SIMONNET, D., LEWANDOWSKI, M., VELASTIN, S. A., ORWELL, J., AND TURKBEYLER, E. Re-identification of pedestrians in crowds using dynamic time warping. In *European Conference on Computer Vision* (2012), Springer, pp. 423–432.

[80] SIMONYAN, K., AND ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[81] SONNENBURG, S., RÄTSCH, G., AND SCHÄFER, C. A general and efficient multiple kernel learning algorithm. In *Advances in neural information processing systems* (2006), pp. 1273–1280.

[82] SONNENBURG, S., RÄTSCH, G., SCHÄFER, C., AND SCHÖLKOPF, B. Large scale multiple kernel learning. *Journal of Machine Learning Research 7*, Jul (2006), 1531–1565.

[83] SU, C., YANG, F., ZHANG, S., TIAN, Q., DAVIS, L. S., AND GAO, W. Multi-task learning with low rank attribute embedding for person re-identification. In *ICCV* (2015).

[84] SUH, Y., WANG, J., TANG, S., MEI, T., AND MU LEE, K. Part-aligned bilinear representations for person re-identification. In *The European Conference on Computer Vision (ECCV)* (September 2018).

[85] SUN, Y., ZHENG, L., DENG, W., AND WANG, S. Svdnet for pedestrian retrieval. In *Computer Vision (ICCV), 2017 IEEE International Conference on* (2017), IEEE, pp. 3820–3828.

[86] SUN, Y., ZHENG, L., YANG, Y., TIAN, Q., AND WANG, S. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *The European Conference on Computer Vision (ECCV)* (September 2018).

[87] SZEGEDY, C., LIU, W., JIA, Y., SERMANET, P., REED, S., ANGUELOV, D., ERHAN, D., VANHOUCKE, V., RABINOVICH, A., ET AL. Going deeper with convolutions. Cvpr.

[88] TORRALBA, A., MURPHY, K. P., AND FREEMAN, W. T. Sharing visual features for multiclass and multiview object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence 29*, 5 (2007), 854–869.

[89] USTINOVA, E., AND LEMPITSKY, V. Learning deep embeddings with histogram loss. In *Advances in Neural Information Processing Systems* (2016), pp. 4170–4178.

[90] VEZZANI, R., BALTIERI, D., AND CUCCHIARA, R. People reidentification in surveillance and forensics: A survey. *ACM Computing Surveys (CSUR) 46*, 2 (2013), 29.

[91] VILLANI, C. *Optimal transport: old and new*, vol. 338. Springer Science & Business Media, 2008.

[92] WANG, C., ZHANG, Q., HUANG, C., LIU, W., AND WANG, X. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 365–381.

[93] WANG, H., GONG, S., ZHU, X., AND XIANG, T. Human-in-the-loop person re-identification. In *European Conference on Computer Vision* (2016), Springer, pp. 405–422.

[94] WANG, J., KALOUSIS, A., AND WOZNICA, A. Parametric local metric learning for nearest neighbor classification. In *Advances in Neural Information Processing Systems* (2012), pp. 1601–1609.

[95] WANG, T., GONG, S., ZHU, X., AND WANG, S. Person re-identification by video ranking. In *Computer Vision–ECCV 2014*. Springer, 2014, pp. 688–703.

[96] WANG, Y., WANG, L., YOU, Y., ZOU, X., CHEN, V., LI, S., HUANG, G., HARIHARAN, B., AND WEINBERGER, K. Q. Resource aware person re-identification across multiple resolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 8042–8051.

[97] WEI, L., ZHANG, S., GAO, W., AND TIAN, Q. Person transfer gan to bridge domain gap for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2018).

[98] WEINBERGER, K. Q., BLITZER, J., AND SAUL, L. K. Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems* (2005), pp. 1473–1480.

[99] WEINBERGER, K. Q., AND SAUL, L. K. Distance metric learning for large margin nearest neighbor classification. *The Journal of Machine Learning Research 10* (2009), 207–244.

[100] WEN, Z., SHI, J., LI, Q., HE, B., AND CHEN, J. ThunderSVM: A fast SVM library on GPUs and CPUs. *Journal of Machine Learning Research 19* (2018), 1–5.

[101] XIAO, T., LI, S., WANG, B., LIN, L., AND WANG, X. Joint detection and identification feature learning for person search. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on* (2017), IEEE, pp. 3376–3385.

[102] XING, E. P., NG, A. Y., JORDAN, M. I., AND RUSSELL, S. Distance metric learning with application to clustering with side-information. *Advances in neural information processing systems 15* (2003), 505–512.

[103] XIONG, F., GOU, M., CAMPS, O., AND SZNAIER, M. Person re-identification using kernel-based metric learning methods. In *Computer Vision–ECCV 2014*. Springer, 2014, pp. 1–16.

[104] YANG, Y., YANG, J., YAN, J., LIAO, S., YI, D., AND LI, S. Z. Salient color names for person re-identification. In *Computer Vision–ECCV 2014*. Springer, 2014, pp. 536–551.

[105] YE, M., CHEN, J., LENG, Q., LIANG, C., WANG, Z., AND SUN, K. Coupled-view based ranking optimization for person re-identification. In *International Conference on Multimedia Modeling* (2015), Springer, pp. 105–117.

[106] YE, M., LIANG, C., YU, Y., WANG, Z., LENG, Q., XIAO, C., CHEN, J., AND HU, R. Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing. *IEEE Transactions on Multimedia 18*, 12 (2016), 2553–2566.

[107] YE, M., MA, A. J., ZHENG, L., LI, J., AND YUEN, P. C. Dynamic label graph matching for unsupervised video re-identification. In *International conference on commuter vision* (2017).

[108] YI, D., LEI, Z., LIAO, S., AND LI, S. Z. Deep metric learning for person re-identification. In *Pattern Recognition (ICPR), 2014 22nd International Conference on* (2014), IEEE, pp. 34–39.

[109] YOU, J., WU, A., LI, X., AND ZHENG, W.-S. Top-push video-based person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2016).

[110] YU, H.-X., WU, A., AND ZHENG, W.-S. Cross-view asymmetric metric learning for unsupervised person re-identification. In *IEEE International Conference on Computer Vision* (2017).

[111] ZHANG, G., WANG, Y., KATO, J., MARUTANI, T., AND MASE, K. Local distance comparison for multiple-shot people re-identification. In *Computer Vision–ACCV 2012*. Springer, 2013, pp. 677–690.

[112] ZHANG, L., XIANG, T., AND GONG, S. Learning a discriminative null space for person re-identification. In *CVPR* (2016).

[113] ZHANG, X., ZHOU, X., LIN, M., AND SUN, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2018).

[114] ZHANG, Y., LI, B., LU, H., IRIE, A., AND RUAN, X. Sample-specific svm learning for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2016).

[115] ZHAO, R., OUYANG, W., AND WANG, X. Person re-identification by salience matching. In *Proceedings of the IEEE International Conference on Computer Vision* (2013), pp. 2528–2535.

[116] ZHAO, R., OUYANG, W., AND WANG, X. Unsupervised salience learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2013), pp. 3586–3593.

[117] ZHAO, R., OUYANG, W., AND WANG, X. Learning mid-level filters for person re-identification. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on* (2014), IEEE, pp. 144–151.

[118] ZHENG, L., BIE, Z., SUN, Y., WANG, J., SU, C., WANG, S., AND TIAN, Q. Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision* (2016), Springer, pp. 868–884.

[119] ZHENG, L., SHEN, L., TIAN, L., WANG, S., WANG, J., AND TIAN, Q. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision* (2015), pp. 1116–1124.

[120] ZHENG, L., YANG, Y., AND HAUPTMANN, A. G. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984* (2016).

[121] ZHENG, W.-S., GONG, S., AND XIANG, T. Associating groups of people. In *BMVC* (2009), vol. 2, p. 6.

[122] ZHENG, W.-S., GONG, S., AND XIANG, T. Person re-identification by probabilistic relative distance comparison. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on* (2011), IEEE, pp. 649–656.

[123] ZHENG, W.-S., GONG, S., AND XIANG, T. Reidentification by relative distance comparison. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 35*, 3 (2013), 653–668.

[124] ZHENG, Z., ZHENG, L., AND YANG, Y. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 3754–3762.

[125] ZHENG, Z., ZHENG, L., AND YANG, Y. Pedestrian alignment network for large-scale person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology* (2018).

[126] ZHONG, Y., JAIN, A. K., AND DUBUISSON-JOLLY, M.-P. Object tracking using deformable templates. *IEEE transactions on pattern analysis and machine intelligence 22*, 5 (2000), 544–549.

[127] ZHONG, Z., ZHENG, L., CAO, D., AND LI, S. Re-ranking person re-identification with k-reciprocal encoding. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on* (2017), IEEE, pp. 3652–3661.

[128] ZHOU, J., AND WU, Y. Finding the right exemplars for reconstructing single image super-resolution. In *Image Processing (ICIP), 2016 IEEE International Conference on* (2016), IEEE, pp. 1414–1418.

[129] ZHOU, J., YU, P., TANG, W., AND WU, Y. Efficient online local metric adaptation via negative samples for person reidentification. In *The IEEE International Conference on Computer Vision (ICCV)* (2017), vol. 2, p. 7.

APPENDIX A

Appendix A.

A.1. Theorem. 3 in Sec. 2.4

Once the reference point set $\mathbf{R}$ is obtained, we aim to learn a positive semi-definite (PSD) Mahalanobis metric $\mathbf{M} = \mathbf{L}\mathbf{L}^T$ by optimizing the following objective:

$$\mathbf{L}^* = \min_{\mathbf{L}} \frac{1}{n}\|\mathbf{X}\mathbf{L} - \mathbf{R}\|_{\mathcal{F}}^2 + \lambda\|\mathbf{L}\|_{\mathcal{F}}^2 \tag{A.1}$$

Theorem 13. *Assume $\|r\|_2 \leq B_r$ for any $r \in \mathcal{R}$, and $\|x\|_2 \leq B_x$ for any $x \in \mathcal{X}$. With probability $1-\delta$, for any matrix $\boldsymbol{L}$ which is the optimal solution of Eqn.A.1 with stability $\beta = \frac{8B_x^2B_r^2}{\lambda n}\left(1 + \frac{B_x}{\sqrt{\lambda}}\right)^2$, we have:*

$$\|\mathcal{E}(\boldsymbol{L}, \mathcal{D}_{\boldsymbol{R}}) - \mathcal{E}(\boldsymbol{L}, \mathcal{S}_r)\| \leq \left(1 + \left(2n + \frac{\lambda n}{8B_x^2}\right)\sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right)\beta \tag{A.2}$$

To prove Theorem. 13, we need to prove five important lemmas to facilitate our proof.

Lemma 1. *$\boldsymbol{L}$ is the optimal solution of Eqn. A.1, so that:*

$$\|\boldsymbol{L}\|_{\mathcal{F}} \leq \frac{B_r}{\sqrt{\lambda}} \tag{A.3}$$

PROOF. Since $\mathbf{L}$ is the optimal solution, so we have: $f(\mathbf{L}, \mathbf{X}, \mathbf{R}) \leq f(\mathbf{0}, \mathbf{X}, \mathbf{R})$

$$\begin{aligned}
&\Leftrightarrow \frac{1}{n}\|\mathbf{XL} - \mathbf{R}\|_{\mathcal{F}}^2 + \lambda\|\mathbf{L}\|_{\mathcal{F}}^2 \leq \frac{1}{n}\|\mathbf{X0} - \mathbf{R}\|_{\mathcal{F}}^2 + 0 \\
&\Leftrightarrow \lambda\|\mathbf{L}\|_{\mathcal{F}}^2 \leq \frac{1}{n}\|\mathbf{X0} - \mathbf{R}\|_{\mathcal{F}}^2 \\
&\Leftrightarrow \lambda\|\mathbf{L}\|_{\mathcal{F}}^2 \leq \frac{1}{n}\sum \|r\|_2^2 \\
&\Leftrightarrow \lambda\|\mathbf{L}\|_{\mathcal{F}}^2 \leq B_r^2 \\
&\Leftrightarrow \|\mathbf{L}\|_{\mathcal{F}} \leq \frac{B_r}{\sqrt{\lambda}}
\end{aligned}$$

□

Lemma 2. *Our loss* $l = \|x^T\boldsymbol{L} - r^T\|_2^2$ *has an upper bound* $B_l = B_r^2\left(1 + \frac{B_r}{\sqrt{\lambda}}\right)^2$

PROOF.

$$\begin{aligned}
\|x^T\mathbf{L} - r^T\|_2^2 &\leq (\|x^T\|_2\|\mathbf{L}\|_{\mathcal{F}} + \|r^T\|_2)^2 \\
&\leq \left(B_x\frac{B_r}{\sqrt{\lambda}} + B_r\right)^2 \\
&\leq B_r^2\left(1 + \frac{B_r}{\sqrt{\lambda}}\right)^2
\end{aligned}$$

□

Lemma 3. *Our loss* $l = \|x^T\boldsymbol{L} - r^T\|_2^2$ *is* σ*-admissible with* $\sigma = 2B_xB_r\left(1 + \frac{B_r}{\sqrt{\lambda}}\right)$

PROOF. A loss function l is σ-admissible **[8]** if it is convex with respect to its first argument and the following condition holds:

$$\forall \mathbf{L}, \mathbf{L}' \in \mathbb{R}^{d\times d'}, \forall (x, r), |l(\mathbf{L}, (x, r)) - l(\mathbf{L}', (x, r))| \leq \sigma\|\mathbf{L} - \mathbf{L}'\|_{\mathcal{F}}$$

So that

$$
\begin{aligned}
&|\|x^T\mathbf{L}' - r^T\|_2^2 - \|x^T\mathbf{L} - r^T\|_2^2| \\
&= |\|x^T\mathbf{L}' - r^T\|_2 - \|x^T\mathbf{L} - r^T\|_2|\|x^T\mathbf{L}' - r^T\|_2 + \|x^T\mathbf{L} - r^T\|_2| \\
&\leq \|x^T\mathbf{L}' - r^T - x^T\mathbf{L} + r^T\|_2\|x^T\mathbf{L}' - r^T\|_2 + \|x^T\mathbf{L} - r^T\|_2| \\
&\leq \|\mathbf{L}' - \mathbf{L}\|_{\mathcal{F}} 2B_x B_r \left(1 + \frac{B_r}{\sqrt{\lambda}}\right)
\end{aligned} \tag{A.4}
$$

□

Then we would like to prove that our algorithm is uniformly stable. For simplicity, in the following $\hat{E}(\mathbf{L})$ is the empirical risk over dataset $\mathcal{S}_r$, and $\hat{E}^i(\mathbf{L})$ is the empirical risk over a new dataset $\mathcal{S}_r^i$ obtained from $\mathcal{S}_r$ by replacing the i^{th} sample. f and f^i denote the functions to optimize in our objective using the sets of examples $\mathcal{S}_r^i$ and $\mathcal{S}_r$ respectively.

Lemma 4. *Let f and f^i be the optimization functions. $\boldsymbol{L}$ and $\boldsymbol{L}^i$ are the optimal solutions respectively. Let $\Delta\boldsymbol{L} = \boldsymbol{L} - \boldsymbol{L}'$, for any $t \in [0, 1]$ we have:*

$$
\|\boldsymbol{L}\|_{\mathcal{F}}^2 - \|\boldsymbol{L} - t\Delta\boldsymbol{L}\|_{\mathcal{F}}^2 + \|\boldsymbol{L}^i\|_{\mathcal{F}}^2 - \|\boldsymbol{L}^i + t\Delta\boldsymbol{L}\|_{\mathcal{F}}^2 \leq \frac{4tB_xB_r}{\lambda n}(1 + \frac{B_x}{\sqrt{\lambda}})\|\Delta\boldsymbol{L}\|_{\mathcal{F}} \tag{A.5}
$$

PROOF. Since $\hat{E}$ is a convex function, for any $t \in [0, 1]$, we have:

$$
\begin{aligned}
\hat{E}^i(\mathbf{L} - t\Delta\mathbf{L}) - \hat{E}^i(\mathbf{L}) &\leq t(\hat{E}^i(\mathbf{L}^i) - \hat{E}^i(\mathbf{L})) \\
\hat{E}^i(\mathbf{L}^i + t\Delta\mathbf{L}) - \hat{E}^i(\mathbf{L}^i) &\leq t(\hat{E}^i(\mathbf{L}) - \hat{E}^i(\mathbf{L}^i))
\end{aligned} \tag{A.6}
$$

Summing the above two inequalities gives:

$$
\hat{E}^i(\mathbf{L} - t\Delta\mathbf{L}) - \hat{E}^i(\mathbf{L}) + \hat{E}^i(\mathbf{L}^i + t\Delta\mathbf{L}) - \hat{E}^i(\mathbf{L}^i) \leq 0 \tag{A.7}
$$

And we also have:

$$\begin{aligned} f(\mathbf{L}) - f(\mathbf{L} - t\Delta\mathbf{L}) \leq 0 \\ f^i(\mathbf{L}^i) - f^i(\mathbf{L}^i + t\Delta\mathbf{L}) \leq 0 \end{aligned} \tag{A.8}$$

Summing Eqn. A.7 and Eqn. A.8, we have:

$$\begin{aligned} &\hat{E}^i(\mathbf{L} - t\Delta\mathbf{L}) - \hat{E}^i(\mathbf{L}) + \hat{E}(\mathbf{L}) - \hat{E}(\mathbf{L} - t\Delta\mathbf{L}) \\ &+ \lambda\|\mathbf{L}\|_{\mathcal{F}}^2 - \lambda\|\mathbf{L} - t\Delta\mathbf{L}\|_{\mathcal{F}}^2 + \lambda\|\mathbf{L}^i\|_{\mathcal{F}}^2 - \lambda\|\mathbf{L}^i + t\Delta\mathbf{L}\|_{\mathcal{F}}^2 \leq 0 \end{aligned} \tag{A.9}$$

So that we can write:

$$\lambda\|\mathbf{L}\|_{\mathcal{F}}^2 - \lambda\|\mathbf{L} - t\Delta\mathbf{L}\|_{\mathcal{F}}^2 + \lambda\|\mathbf{L}^i\|_{\mathcal{F}}^2 - \lambda\|\mathbf{L}^i + t\Delta\mathbf{L}\|_{\mathcal{F}}^2 \leq B_{const} \tag{A.10}$$

with $B_{const} = \hat{E}^i(\mathbf{L}) - \hat{E}^i(\mathbf{L} - t\Delta\mathbf{L}) + \hat{E}(\mathbf{L} - t\Delta\mathbf{L}) - \hat{E}(\mathbf{L})$. Using Lemma. 3 we have:

$$\begin{aligned} B_{const} &\leq |\hat{E}^i(\mathbf{L}) - \hat{E}^i(\mathbf{L} - t\Delta\mathbf{L}) + \hat{E}(\mathbf{L} - t\Delta\mathbf{L}) - \hat{E}(\mathbf{L})| \\ &\leq |\frac{1}{n}\sum_{(x,r)\in\mathcal{S}_r} l(\mathbf{L} - t\Delta\mathbf{L}, (x,r)) - \frac{1}{n}\sum_{(x,r)^i\in\mathcal{S}_r^i} l(\mathbf{L} - t\Delta\mathbf{L}, (x,r)^i) \\ &+ \frac{1}{n}\sum_{(x,r)^i\in\mathcal{S}_r^i} l(\mathbf{L}, (x,r)^i) - \frac{1}{n}\sum_{(x,r)\in\mathcal{S}_r} l(\mathbf{L}, (x,r))| \\ &\leq \frac{1}{n}|l(\mathbf{L} - t\Delta\mathbf{L}, (x_i,r_i)) - l(\mathbf{L} - t\Delta\mathbf{L}, (x_i,r_i)^i) + l(\mathbf{L}, (x_i,r_i)^i) - l(\mathbf{L}, (x_i,r_i))| \\ &\leq \frac{1}{n}|l(\mathbf{L} - t\Delta\mathbf{L}, (x_i,r_i)) - l(\mathbf{L}, (x_i,r_i))| \\ &+ \frac{1}{n}|l(\mathbf{L}, (x_i,r_i)^i) - l(\mathbf{L} - t\Delta\mathbf{L}, (x_i,r_i)^i)| \\ &\leq \frac{4tB_xB_r}{n}\left(1 + \frac{B_x}{\sqrt{\lambda}}\right)\|\Delta\mathbf{L}\|_{\mathcal{F}} \end{aligned} \tag{A.11}$$

□

Lemma 5. *Recall the definition of uniform stability in* [**8**]*, our algorithm has a uniform stability in* $\beta = \frac{8B_x^2 B_r^2}{\lambda n}\left(1 + \frac{B_x}{\sqrt{\lambda}}\right)^2$

PROOF. Recall the definition of uniform stability in [**8**] that an algorithm A has uniform stability β with respect to the loss function l if the following holds:

$$\forall \mathcal{S}_r, \forall i, \sup_{(x,r)\sim \mathbb{D}_r} |l(A_{\mathcal{S}_r}, (x,r)) - l(A_{\mathcal{S}_r^i}, (x,r))| \leq \beta$$

By setting $t = \frac{1}{2}$ in Lemma. 4, the LHS is:

$$\|\mathbf{L}\|_{\mathcal{F}}^2 - \|\mathbf{L} - \frac{1}{2}\Delta\mathbf{L}\|_{\mathcal{F}}^2 + \|\mathbf{L}^i\|_{\mathcal{F}}^2 - \|\mathbf{L}^i + \frac{1}{2}\Delta\mathbf{L}\|_{\mathcal{F}}^2 = \frac{1}{2}\|\Delta\mathbf{L}\|_{\mathcal{F}}^2$$

So that

$$\begin{aligned}
\frac{1}{2}\|\Delta\mathbf{L}\|_{\mathcal{F}}^2 &\leq \frac{2B_x B_r}{\lambda n}\left(1 + \frac{B_x}{\sqrt{\lambda}}\right)\|\Delta\mathbf{L}\|_{\mathcal{F}} \\
\Rightarrow \|\Delta\mathbf{L}\|_{\mathcal{F}} &\leq \frac{4B_x B_r}{\lambda n}\left(1 + \frac{B_x}{\sqrt{\lambda}}\right)
\end{aligned}$$

Consider Lemma. 3, we have:

$$\begin{aligned}
|l(\mathbf{L}, (x,r)) - l(\mathbf{L}^i, (x,r))| &\leq 2B_x B_r\left(1 + \frac{B_x}{\sqrt{\lambda}}\right)\|\Delta\mathbf{L}\|_{\mathcal{F}} \\
&\leq \frac{8B_x^2 B_r^2}{\lambda n}\left(1 + \frac{B_x}{\sqrt{\lambda}}\right)^2
\end{aligned}$$

□

Recall the **Theorem 12 in** [**8**]. Let A be an algorithm with uniform stability β w.r.t. a loss function l such that $0 \leq l(A_{\mathcal{S}_r}, (x,r)) \leq B_l$ for all (x,r) and all sets $\mathcal{S}_r$. For any $n \geq 1$ the

following bound holds with probability at least $1 - \delta$ over the random draw of the sample $\mathcal{S}_r$:

$$E(A_{\mathcal{S}_r}) \leq \hat{E}(A_{\mathcal{S}_r}) + \beta + (2n\beta + B_l)\sqrt{\frac{\ln \frac{1}{\delta}}{2n}}$$

We have already shown that our algorithm is uniformly stable and that our loss is bounded, hence we can directly apply this theorem to obtain our Theorem. 13 using the bound on the loss presented in Lemma. 2 and the uniform stability of our algorithm proven in Lemma. 5.

Vita

Education

- Ph.D, Aug. 2013 - Dec. 2018, Dept.of EECS, Northwestern University, Evanston, IL
- B.S Aug. 2009 - June. 2013, Dept.of Automation, Tsinghua University, Beijing, China

Selected Publications:

- Xinzhao Li, Yuehu Liu, Zeqi Chen, **Jiahuan Zhou** and Ying Wu. Fused Discriminative Metric Learning for Low Resolution Pedestrian Detection. in IEEE International Conference on Image Processing (ICIP'18), Athens, Greece, Oct. 2018.
- **Jiahuan Zhou**, Bing Su and Ying Wu. Easy Identification from Better Constraints: Multi-Shot Person Re-Identification from Reference Constraints. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'18), Salt Lake City, USA, June. 2018.
- **Jiahuan Zhou**, Pei Yu, Tang Wei and Ying Wu. Efficient Online Local Metric Adaptation via Negative Samples for Person Re-Identification. in Proceedings of International Conference on Computer Vision (ICCV'17), Venice, Italy, Oct. 2017.
- Wei Tang, Pei Yu, **Jiahuan Zhou**, and Ying Wu. Towards a Unified Compositional Model for Visual Pattern Modeling. in Proceedings of International Conference on Computer Vision (ICCV'17), Venice, Italy, Oct. 2017.

- Bing Su, **Jiahuan Zhou**, Xiaoqing Ding and Ying Wu, "Unsupervised Hierarchical Dynamic Parsing and Encoding for Action Recognition" IEEE Transactions on Image Processing, 26.12 (2017): 5784-5799.
- Bing Su, **Jiahuan Zhou**, Hao Wang and Ying Wu, "Hierarchical Dynamic Parsing and Encoding for Action Recognition", in Proc. European Conf. on Computer Vision (ECCV'16), Amsterdam, Netherlands, Oct. 2016.

- Pei Yu, **Jiahuan Zhou** and Ying Wu, "Learning Reconstruction-based Gaze Estimation", in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'16), Las Vegas, USA, June. 2016.

- **Jiahuan Zhou** and Ying Wu, Finding the Right Exemplars for Reconstructing Single Image Super-Resolution, in Proc. IEEE Intl Conf. on Image Processing (ICIP'16), Phoenix, USA, Sep. 2016. **(Oral)**

- Han Hu, **Jiahuan Zhou**, Jianjiang Feng and Jie Zhou. Multi-way Constrained Spectral Clustering via Nonnegative Restriction. Interntional Conference on Pattern Recognition (ICPR'12), Tsukuba, Japan, Nov. 2012. **(Oral)**

Awards and Honors:

- **The National Encouragement Scholarship**, Tsinghua University *2009 - 2010*

- **Academic Excellence Award**, Tsinghua University *2010 - 2011*

- **Outstanding Graduate Scholarship**, Tsinghua University *2012 - 2013*

- **The Murphy Fellowship**, Northwestern University *2013 - 2014*