Person Re-Identification With Metric Learning Using Privileged Information

Xun Yang, Meng Wang, *Senior Member, IEEE*, and Dacheng Tao, *Fellow, IEEE*

***Abstract*— Despite the promising progress made in recent years, person re-identification remains a challenging task due to complex variations in human appearances from different camera views. This paper presents a logistic discriminant metric learning method for this challenging problem. Different with most existing metric learning algorithms, it exploits both original data and auxiliary data during training, which is motivated by the new machine learning paradigm—learning using privileged information. Such privileged information is a kind of auxiliary knowledge, which is only available during training. Our goal is to learn an optimal distance function by constructing a locally adaptive decision rule with the help of privileged information. We jointly learn two distance metrics by minimizing the empirical loss penalizing the difference between the distance in the original space and that in the privileged space. In our setting, the distance in the privileged space functions as a local decision threshold, which guides the decision making in the original space like a *teacher*. The metric learned from the original space is used to compute the distance between a probe image and a gallery image during testing. In addition, we extend the proposed approach to a multi-view setting which is able to explore the complementation of multiple feature representations. In the multi-view setting, multiple metrics corresponding to different original features are jointly learned, guided by the same privileged information. Besides, an effective iterative optimization scheme is introduced to simultaneously optimize the metrics and the assigned metric weights. Experiment results on several widely-used data sets demonstrate that the proposed approach is superior to global decision threshold-based methods and outperforms most state-of-the-art results.**

***Index Terms*— Person re-identification, learning using privileged information, metric learning, computer vision.**

I. Introduction

PERSON Re-identification (re-ID) [1] is a critical problem in video analytics applications such as security and surveillance and has attracted increasing attention in recent years. Although many approaches have been proposed, it remains a challenging problem since person's appearance usually undergoes dramatic changes across camera views due to changes in view angle, body pose, illumination and background clutter.

The fundamental problem is to compare a person of interest from a probe camera view to a gallery of candidates captured from a camera that does not overlap with the probe one. If a 'true' match to the probe exists in the gallery, it should have a high matching score, compared to incorrect candidates. Generally speaking, person re-ID involves two sub-problems: feature representation and metric learning. An effective feature representation [2], [3] is critical for person re-ID, which should be robust to complex variations in human appearances from different camera views. The general strategy is to concatenate multiple low-level visual features into a long feature vector. However, it inevitably brings massive redundant information which may degrade the ability of representation. Therefore, more efforts [4]–[12] have been made along the second direction. Some of them formulate re-ID as a subspace learning problem by learning a low-dimensional projection. Some others directly learn a Mahalanobis distance function parameterized by a positive semidefinite (PSD) matrix to separate positive person image pairs from negative pairs. This work follows the second approach, aiming to learn a suitable distance metric.

Despite the promising efforts have been made, most existing metric learning methods are limited in that they compare the distance between a pair of similar/dissimilar instances with a global threshold. Such global threshold based pairwise constraints may suffer from sub-optimal learning performance when coping with some real-world tasks with complex inter-class and intra-class variations, e.g., person re-ID. A natural solution to alleviate this limitation is to design a locally adaptive decision rule. Li *et al.* [10] proposed to learn a second-order local decision function in original feature space to replace the global threshold. Wang *et al.* [13] introduced an adaptive shrinkage-expansion rule to shrink/expand the Euclidean distance as an adaptive threshold. These two earlier works both leverage the information from original feature space to guide the decision making. However, the guidance from the original feature space might be relatively weak, since original feature is usually noisy and less discriminative. It is of great interest to design a solution that is able to exploit additional knowledge beyond the original data space.

It has been shown in a new learning paradigm - Learning Using Privileged Information (LUPI) [14] that a more reliable and effective model can be learned if some auxiliary

Manuscript received December 12, 2016; revised September 1, 2017 and October 6, 2017; accepted October 7, 2017. Date of publication October 23, 2017; date of current version November 22, 2017. This work was supported in part by the National 973 Program of China under Grant 2014CB347600 and in part by the National Nature Science Foundation of China under Grant 61432019. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Xiaochun Cao. *(Corresponding author: Meng Wang.)*

X. Yang and M. Wang are with the School of Computer and Information Engineering, Hefei University of Technology, Hefei 230009, China (e-mail: hfutyangxun@gmail.com; eric.mengwang@gmail.com).

D. Tao is with the UBTech Sydney Artificial Intelligence Institute, The University of Sydney, Darlington, NSW 2008, Australia, and also with the Faculty of Engineering and Information Technologies, School of Information Technologies, The University of Sydney, Darlington, NSW 2008, Australia (e-mail: dacheng.tao@sydney.edu.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2017.2765836

expert knowledge is exploited during training. Such auxiliary knowledge is called privileged information and is only available during training. It typically describes some important properties of the training instance, such as attributes, tags, textual descriptions or other high-level knowledge, etc. The LUPI paradigm is inspired by the human teaching-learning in which students will learn better if a teacher can provide some explanations, comments, comparison or other supervision.

Motivated by the LUPI paradigm, we present a logistic discriminant metric learning method for cross-view re-ID by exploiting both original data and auxiliary data to design a locally adaptive decision rule during training. In our setting, each training instance is represented with two forms of features: one is from the original space and the other is from the privileged space. We jointly learn two distance metrics with PSD constraints by minimizing the empirical loss penalizing the difference between the distance in the original space and the distance in the privileged space. During training, the distance in the privileged space functions as a local decision threshold to guide the metric learning in the original space like a *teacher*. The finally learned metric from the original space is used to compute the distance between a probe image and a gallery image during testing. Moreover, we extend the proposed algorithm from the single-view setting to a multi-view setting which is able to explore the complementation of multiple feature representations. In the multi-view setting, we simultaneously learn multiple distance metrics from different original feature spaces under the guidance of the same privileged knowledge. An effective iterative optimization algorithm is introduced to simultaneously optimize the metrics and the assigned metric weights.

Our main contributions can be summarized as follows:

(1) We present an effective logistic discriminant metric learning method by exploiting both original data and privileged information to design a locally adaptive decision rule during training. The proposed decision rule is different with the common way in most existing works that compares the distance between a pair of training instances with a global threshold to decide whether they are similar or dissimilar. In this work, such global threshold is replaced with the squared distance in the privileged space. (We term the proposed method as LDML+.)

(2) We extend the proposed method to a multi-view setting, which can explore the complementary information of multiple different features effectively. In this work, multiple distance metrics are learned simultaneously from different original feature spaces under the guidance of the same privileged knowledge, in which each metric is learned using a single feature. (We term the proposed multi-view approach as MVLDML+.)

(3) We conduct extensive evaluations on several widely-used datasets. Experimental results show that LDML+ is able to improve the performance of global decision threshold based metric learning methods with the help of privileged information and MVLDML+ can outperform most state-of-the-art results.

The proposed LDML+ method was first introduced in previous work [15]. In comparison with the preliminary version [15], we have improvements in the following aspects: (1) we regularize the privileged distance metric to control the complexity of model and avoid falling into local optimum; (2) we extend the single-view version in previous work to a multi-view setting in this work which can explore multiple original feature representations effectively; (3) we present extensive experimental evaluations and analyses to validate the effectiveness of the proposed LDML+ and MVLDML+ methods on several re-ID datasets.

II. Related Work

A. Metric Learning

During the past decades, many algorithms have been developed to learn a distance metric. In this subsection, we will briefly review some classical or related distance metric learning works.

Xing *et al.* [16] proposed to learn a distance metric by minimizing the distance between similar instances while keeping that between dissimilar instances larger than a predefined threshold. Globerson and Roweis [17] presented a metric learning algorithm by collapsing all examples in the same class to a single point and pushing examples in other classes infinitely far away. Schultz and Joachims [18] developed a method for learning a distance metric from relative comparison. Davis *et al.* [19] presented an information-theoretic metric learning approach, which formulates the problem as that of minimizing the differential relative entropy between two multivariate Gaussians under pairwise constraints on the distance function. Weinberger and Saul, [20] proposed to learn a Mahalanobis distance metric for kNN classification with the goal that k-nearest neighbors always belong to the same class while examples from different classes are separated by a large margin. Guillaumin *et al.* [21] designed a logistic discriminant approach to learn a distance metric. Bian and Tao [22] developed a risk minimization framework for metric learning. Mignon and Jurie [4] proposed to learn a distance metrics from sparse pairwise similarity/dissimilarity constraints in high dimensional input space.

Among the above-mentioned algorithms, our work is related to the methods [4], [21], [22] which explore a logistic discriminant approach for metric learning. However, Guillaumin's work [21] doesn't use any regularization term including the PSD constraint, which easily suffers from overfitting. Bian's work [22] relies on a strong assumption that the learned metric is bounded, which is too rigid. Besides, these works [4], [19], [21], [22] all adopt the global threshold based constraints, which easily suffer from sub-optimal learning performance. Compared with them, we learn a PSD metric by exploiting auxiliary information to construct a locally adaptive decision function, which is more robust and shows better performance.

Zha *et al.* [23] also presented a metric learning algorithm that exploits auxiliary knowledge during training. However, it's different with our work. Zha *et al.* [23] pre-trained several auxiliary metrics using several auxiliary datasets to assist the metric learning in the source dataset. Different with [23], we don't exploit any auxiliary datasets during training. In our work, we exploit auxiliary feature representation (privileged

information) of training instances during training, and the auxiliary feature representation is not available for testing.

B. Person Re-Identification

Person re-ID aims to retrieve a person of interest across spatially disjoint cameras. It can be seen as a image retrieval problem [24]–[26]. This paper focuses on tackling the person re-ID problem with the proposed metric learning scheme. In this subsection, we will briefly review some related works. Generally speaking, mainstream re-ID works can be roughly categorized into the following groups.

The first group of methods focus on designing discriminative and invariant features for re-ID [2], [3], [27]–[31]. Recently, some new proposed feature descriptors have gained good performance, i.e., local maximal occurrence (LOMO) feature [3], weighted histograms of overlapping stripes [31], and Gaussian of Gaussian (GOG) descriptor [2]. The general trend is that the dimensions of feature descriptors are getting higher by concatenating multiple low-level visual features, which may result in the so-called curse of dimensionality. To alleviate this problem, our work provides an effective way to simultaneously explore multiple feature representations.

The second group of methods aim to design discriminative distance functions for recognizing people from disjoint camera views [3]–[9], [11], [12], [32], [33]. In this group, some works aim to learn a Mahalanobis-like distance metric [5], [15], [34], while some other methods focus on seeking a discriminative subspace [7], [11], [35], [36]. These two subgroups actually are closely related. We briefly introduce some well-known works as follows. Zheng *et al.* [37] formulated re-ID as a relative distance comparison learning problem by maximizing the probability that relevant samples have smaller distance than the irrelevant ones. Liao and Li [5] proposed a logistic discriminant metric learning approach with PSD constraints and asymmetric sample weight strategy. Köstinger *et al.* [8] developed a simple and effective metric learning method by computing the difference between the intra-class and inter-class covariance matrix. As an improvement, Liao *et al.* [3] proposed a cross-view quadratic discriminant analysis (XQDA) method by learning a more discriminative distance metric and a low-dimensional subspace simultaneously. Pedagadi *et al.* [32] applied the local fisher discriminant analysis algorithm to match person images by maximizing the inter-class separability while preserving the multiclass modality, whose kernel version was presented for re-ID in [11]. Zhang *et al.* [7] proposed to overcome the small-sample-size problem in re-ID by learning a discriminative null space, where the within-class scatter is minimized to zero while maximizing the relative between-class separation simultaneously.

Different with them, we propose a novel metric learning method for re-ID, which incorporates auxiliary knowledge to guide the metric learning in original feature space. In addition, we also present a multi-view extension which can explore the complementation of multiple feature representations by simultaneously learning multiple metrics. Some existing works [9], [38] have investigated the effects of distance fusion approach for re-ID by a two-stage strategy. They first pretrain several base metrics using different descriptors or different metric learning algorithms, and then combine those base distance functions to obtain the final distance function. Different with them, we present a unified multi-metric learning scheme which can simultaneously learn base metrics and metric weights.

C. Learning Using Privileged Information

Our work is motivated by the LUPI [14] paradigm. We will briefly introduce this learning paradigm in this subsection.

LUPI is a new learning paradigm which was first incorporated into SVM in the form of SVM+ by Vapnik and Vashist [39], which uses the additional (privileged) information as a proxy for predicting the slack variables. It is equivalent to learning an oracle that tells which sample is easy or hard to be predicted. This paradigm has been used for multiple tasks, such as hashing [40], action and event recognition [41], information bottleneck learning [42], learning to rank [43], image categorization [44], object localization [45], and active learning [46], etc.

Recently, two related works [47], [48] are proposed to learn a metric using privileged information. Fouad *et al.* [47] proposed a two-stage strategy to exploit privileged information for metric learning using the information-theoretic approach [19]. They first learn a metric using privileged information to remove some outliers and then use the remaining pairs to learn a metric with original feature. Following [47], Xu *et al.* [48] proposed the ITML+ method, in which privileged information is used to design a slack function to replace the slack variables in ITML [19]. Different with these two ITML based methods [47], [48], we provide a new scheme to leverage privileged knowledge for metric learning under a general risk minimization framework. Moreover, we apply low rank selection for the learned metric in each iteration, which allows us to work directly with higher dimensional input data. While ITML based methods aim to learn a full matrix for the target metric that is in the square of the dimensionality, making it computationally unattractive for high dimensional data and prone to overfitting [4]. In addition, we present a multi-view extension which is able to simultaneously learn multiple metrics from different original feature spaces, which is also different with [47], [48].

III. A Generic Metric Learning Framework

In this section, we introduce a generic metric learning framework which doesn't exploit additional information during training. Suppose we have a pairwise constrained training set $\mathcal{Z} = \{(\mathbf{x}_i, \mathbf{z}_i, y_i) \,|i = 1, \cdots, n\}$, where $\mathbf{x}_i \in \mathbb{R}^d$, $\mathbf{z}_i \in \mathbb{R}^d$ are defined on the same original feature space, i is the index of the i-th pair of training instances, and y_i is the label of the pair $(\mathbf{x}_i, \mathbf{z}_i)$ defined by

$$y_i = \begin{cases} 1, & (\mathbf{x}_i, \mathbf{z}_i) \in \mathcal{S} \\ -1, & (\mathbf{x}_i, \mathbf{z}_i) \in \mathcal{D}, \end{cases} \tag{1}$$

where $\mathcal{S}$ denotes the set of similar pairs and $\mathcal{D}$ denotes the set of dissimilar pairs. The goal is to learn a Mahalanobis distance

metric defined by

$$d_{\mathbf{M}}(\mathbf{x}_i, \mathbf{z}_i) = \sqrt{(\mathbf{x}_i - \mathbf{z}_i)^T \mathbf{M}(\mathbf{x}_i - \mathbf{z}_i)}, \tag{2}$$

where $\mathbf{M} \succeq 0 \in \mathbb{R}^{d \times d}$ is a PSD distance metric. The learned distance $d_{\mathbf{M}}(\mathbf{x}_i, \mathbf{z}_i)$ is expected to be small if $\mathbf{x}_i$ and $\mathbf{z}_i$ are similar, or large if they are dissimilar.

Given a metric, how to determine whether two instances are similar or dissimilar? A common way [4], [5], [21], [22], [49] is to compare their distance with a global decision threshold σ. Hence, the decision function f can be defined by

$$f(\mathbf{x}_i, \mathbf{z}_i; \mathbf{M}) = \sigma - (\mathbf{x}_i - \mathbf{z}_i)^T \mathbf{M}(\mathbf{x}_i - \mathbf{z}_i). \tag{3}$$

If they are similar, the decision function $f > 0$, otherwise $f \leq 0$. Given the decision function, the problem of metric learning can be cast in a generic framework in which the metric is obtained by minimizing the empirical risk $J(\mathbf{M})$

$$\min_{\mathbf{M} \succeq 0} J(\mathbf{M}) = \sum_{i=1}^{n} w_i \mathcal{L}\big(y_i f(\mathbf{x}_i, \mathbf{z}_i; \mathbf{M})\big), \tag{4}$$

where $\mathcal{L}(\cdot)$ is a loss function that is decreasing monotonically, e.g., log loss and smooth hinge loss. w_i is a weight for the i-th pair. Existing works [4], [5], [21], [22], [49] are all under this framework.

IV. The Proposed Approach

A. Problem Formulation

As shown in section III, traditional pairwise constrained methods only exploit original data during training. They usually adopt the global threshold based decision function, which is too rough to obtain a reasonable metric. In this section, we aim to design a locally adaptive decision rule by exploiting additional knowledge.

Motivated by the LUPI paradigm [14], we exploit privileged information to design an adaptive decision function in the training stage. First, each training instance is represented with two forms of features: one is $\mathbf{x}_i \in \mathbb{R}^d$ in the original feature space; the other is $\mathbf{x}_i^* \in \mathbb{R}^{d^*}$ in the privileged space. The training set is reformulated as $\mathcal{Z} = \{(\mathbf{x}_i, \mathbf{x}_i^*, \mathbf{z}_i, \mathbf{z}_i^*, y_i) | i = 1, \ldots, n\}$. Then, we replace the global threshold σ in Eq. (3) using the squared distance $d_{\mathbf{P}}^2(\mathbf{x}_i^*, \mathbf{z}_i^*)$ in the privileged space, where $\mathbf{P} \in \mathbb{R}^{d^* \times d^*}$ is the distance metric corresponding to the privileged information. Here, $d_{\mathbf{P}}^2(\mathbf{x}_i^*, \mathbf{z}_i^*)$ functions as a local decision threshold for the i-th training pair. Our locally adaptive decision function is formulated by

$$\begin{aligned} &f(\mathbf{x}_i, \mathbf{z}_i; \mathbf{x}_i^*, \mathbf{z}_i^*; \mathbf{M}, \mathbf{P}) \\ &\quad = \beta d_{\mathbf{P}}^2(\mathbf{x}_i^*, \mathbf{z}_i^*) - d_{\mathbf{M}}^2(\mathbf{x}_i, \mathbf{z}_i) \\ &\quad = \beta\ (\mathbf{x}_i^* - \mathbf{z}_i^*)^T \mathbf{P}(\mathbf{x}_i^* - \mathbf{z}_i^*) - (\mathbf{x}_i - \mathbf{z}_i)^T \mathbf{M}(\mathbf{x}_i - \mathbf{z}_i), \end{aligned} \tag{5}$$

where $\beta > 0$ is a scale parameter. The idea behind Eq. (5) is that *teacher*'s concept of similarity between a pair of training instances is usually more credible. We expect the knowledge of *teacher* to be transferred from the privileged space to the original space where decision is made.

With the locally adaptive decision function, our problem can be formulated as

$$\langle \hat{\mathbf{M}}, \hat{\mathbf{P}} \rangle = \arg\min J(\mathbf{M}, \mathbf{P}) \quad s.t.\ \mathbf{M} \succeq 0; \mathbf{P} \succeq 0, \tag{6}$$

$$J(\mathbf{M}, \mathbf{P}) = \sum_{i=1}^{n} w_i \mathcal{L}\big(y_i f\big(\mathbf{x}_i, \mathbf{z}_i; \mathbf{x}_i^*, \mathbf{z}_i^*; \mathbf{M}, \mathbf{P}\big)\big) + \lambda \mathcal{R}(\mathbf{P}). \tag{7}$$

As seen from Eq. (7), the objective function $J(\mathbf{M}, \mathbf{P})$ is constituted by two terms: one is a loss term and the other is a regularization term on $\mathbf{P}$. $\lambda > 0$ is a regularization parameter which makes a trade-off between the two terms.

In this work, we denote $\mathcal{L}(\cdot)$ with the log loss

$$\mathcal{L}(u) = \ln(1 + \exp(-u)) \tag{8}$$

which has shown good performance in some existing works. The weight w_i is defined as $\frac{1}{|\mathcal{S}|}$ if $y_i = 1$, or $\frac{1}{|\mathcal{D}|}$ if $y_i = -1$, in which $|\mathcal{S}|$ and $|\mathcal{D}|$ denote the number of similar training pairs and the number of dissimilar training pairs, respectively. The regularization term $\mathcal{R}(\mathbf{P})$ is defined by

$$\mathcal{R}(\mathbf{P}) = \|\mathbf{P}\|_F^2 / d^*. \tag{9}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Here, the regularization term is added to control the model complexity. It should be noted that we only regularize the metric $\mathbf{P}$, while give the metric $\mathbf{M}$ more freedom. Since once $\mathbf{P}$ has higher degree of freedom than $\mathbf{M}$, the former may be shifted to the latter during training. It should be ensured that $\mathbf{M}$ is guided (supervised) by the privileged information. That is, *student*'s concept of similarity between training instances should be under the control of *teacher*.

In this subsection, a logistic discriminant metric learning method is presented that exploits auxiliary knowledge to build a locally adaptive decision rule during training. We term this method as **LDML+** for simplicity.

B. Multi-View Extension

The proposed LDML+ method only considers a single original features. In this subsection, we extend LDML+ from the single-view setting [1] to a multi-view setting to exploit the complementation of multiple original features.

In the multi-view setting, each training instance is represented by $\mathcal{M}$ original features $\{\mathbf{x}_i^m \in \mathbb{R}^{d^m}\}_{m=1}^{\mathcal{M}}$ and a single privileged feature $\mathbf{x}_i^* \in \mathbb{R}^{d^*}$. During training, we aim to simultaneously learn multiple metrics $\mathbf{M}^1, \cdots, \mathbf{M}^{\mathcal{M}}$ from different original spaces and a single metric $\mathbf{P}$ from the privileged space.

Our objective function that is being minimized can be reformulated as

$$\begin{aligned} &J\left(\mathbf{M}^1, \cdots, \mathbf{M}^{\mathcal{M}}, \mathbf{P}, \mathbf{a}\right) \\ &= \sum_{m=1}^{\mathcal{M}} a_m^r \left\{ \sum_{i=1}^{n} w_i \mathcal{L}\Big(y_i f\left(\mathbf{x}_i^m, \mathbf{z}_i^m; \mathbf{x}_i^*, \mathbf{z}_i^*; \mathbf{M}^m, \mathbf{P}\right)\Big) + \lambda \mathcal{R}(\mathbf{P}) \right\}, \\ &\quad s.t.\ \ \mathbf{M}^m \succeq 0; \mathbf{P} \succeq 0; a_m > 0; \sum_{m=1}^{\mathcal{M}} a_m = 1 \end{aligned} \tag{10}$$

[1] Although LDML+ uses two features for training, only the original feature is utilized during testing. Therefore, we categorize it as a single-view method.

where $r > 1$ and $\mathbf{a} = (a_1, \cdots, a_m, \cdots, a_{\mathcal{M}})$ consists of $\mathcal{M}$ weights for $\mathcal{M}$ metrics from original spaces. In Eq. (10), we adopt a trick utilized in [50] and [51] that uses a_m^r ($r > 1$) instead of a_m. It ensures that each view has a particular contribution to the final distance.

In this subsection, we simultaneously learn multiple view-specific metrics with the help of privileged information. We term the proposed multi-view method as **MVLDML+**.

C. Solution of Eq. (10)

In this subsection, we adopt an alternating optimization strategy to solve the minimization problem in Eq. (10) with the loss function in Eq. (8) and the regularizer in Eq. (9). More specifically, we alternatively update $\mathbf{M}^m (m = 1, \cdots, \mathcal{M})$, $\mathbf{P}$, and $\mathbf{a}$ to optimize the objective.

1) Optimization of $\mathbf{M}^m$*:* To optimize $\mathbf{M}^m$, we first fix $\mathbf{P}$, $\mathbf{a}$, and $\mathbf{M}^1, \cdots, \mathbf{M}^{m-1}, \mathbf{M}^{m+1}, \cdots, \mathbf{M}^{\mathcal{M}}$. Then we derive the derivative of J with respect to $\mathbf{M}^m$ as

$$\frac{\partial}{\partial \mathbf{M}^m} J(\mathbf{M}^1, \cdots, \mathbf{M}^{\mathcal{M}}, \mathbf{P}, \mathbf{a}) = \sum_{i=1}^{n} \frac{a_m^r w_i y_i \left(\mathbf{x}_i^m - \mathbf{z}_i^m\right)\left(\mathbf{x}_i^m - \mathbf{z}_i^m\right)^T}{1+\exp\left(y_i\left(\beta^m d_{\mathbf{P}}^2\left(\mathbf{x}_i^*, \mathbf{z}_i^*\right) - d_{\mathbf{M}^m}^2\left(\mathbf{x}_i^m, \mathbf{z}_i^m\right)\right)\right)}. \quad (11)$$

Based on the derivative $\frac{\partial J}{\partial \mathbf{M}^m}$, we adopt a general gradient descent to update the metric $\mathbf{M}^m$:

$$\mathbf{M}_t^m = \mathbf{M}_{t-1}^m - \eta_t^m \frac{\partial J}{\partial \mathbf{M}^m}|_{\mathbf{M}^m = \mathbf{M}_{t-1}^m}, \quad (12)$$

where $\mathbf{M}_{t-1}^m$ denotes the value of $\mathbf{M}^m$ in the $(t-1)$-th iteration. η_t^m denotes the step-size for $\mathbf{M}^m$ in the t-th iteration. In addition, since $\mathbf{M}^m$ is constrained to be PSD ($\mathbf{M}^m \succeq 0$), the output at Eq. (12) should further be projected into the PSD cone using *singular value decomposition* (SVD)

$$\mathbf{M}_t^m = \mathbf{SVD}\left(\mathbf{M}_{t-1}^m - \eta_t^m \frac{\partial J}{\partial \mathbf{M}^m}|_{\mathbf{M}^m = \mathbf{M}_{t-1}^m}\right), \quad (13)$$

where only the eigenvectors corresponding to positive eigenvalues are retained in the solution. Therefore, the solution $\mathbf{M}_t^m$ has a low-rank structure. It can be factorized into $\mathbf{U}\mathbf{U}^T$ in which $\mathbf{U}$ can be used for dimension reduction.

During the optimization of $\mathbf{M}^m$, we dynamically adapt the step-size η^m to accelerate the optimization process, while guaranteeing the convergence. We first initialize the step-size with a large value η_0^m. At each iteration, we perform the dynamic step-size search strategy to find a suitable step-size η^m that satisfies the following condition

$$J(\mathbf{M}^1, \cdots, \mathbf{M}_t^m, \cdots, \mathbf{M}^{\mathcal{M}}, \mathbf{P}, \mathbf{a}) < J(\mathbf{M}^1, \cdots, \mathbf{M}_{t-1}^m, \cdots, \mathbf{M}^{\mathcal{M}}, \mathbf{P}, \mathbf{a}). \quad (14)$$

If the condition Eq. (14) is not satisfied, we shrink the step-size η_t^m to $\eta_t^m/2$ and repeat the operation in Eq. (13) until the condition is satisfied [5]. When the condition is satisfied, we double the step-size as $\eta_{t+1}^m = 2\eta_t^m$ for next iteration. The effectiveness of enlarging the step-size in gradient-descent based metric learning has been shown in [52]. Besides, to avoid the step-size to be enlarged unboundedly, we bound the step-size $\eta_t^m < s\eta_1^m (s > 1)$ to make the optimization more stable. Here, η_1^m is the found step-size which satisfies the condition Eq. (14) at the first iteration. The parameter s controls the upper-boundary of the step-size.

2) Optimization of $\mathbf{P}$*:* Now we consider the optimization of $\mathbf{P}$. Considering $\mathbf{M}^m (m = 1, \cdots, \mathcal{M})$ and $\mathbf{a}$ are fixed, then we derive the derivative of J with respect to $\mathbf{P}$ as

$$\frac{\partial}{\partial \mathbf{P}} J(\mathbf{M}^1, \cdots, \mathbf{M}^{\mathcal{M}}, \mathbf{P}, \mathbf{a}) = \sum_{m=1}^{\mathcal{M}} \left\{ \sum_{i=1}^{n} \frac{-a_m^r w_i y_i \beta^m \left(\mathbf{x}_i^* - \mathbf{z}_i^*\right)\left(\mathbf{x}_i^* - \mathbf{z}_i^*\right)^T}{1+\exp\left(y_i\left(\beta^m d_{\mathbf{P}}^2\left(\mathbf{x}_i^*, \mathbf{z}_i^*\right) - d_{\mathbf{M}^m}^2\left(\mathbf{x}_i^m, \mathbf{z}_i^m\right)\right)\right)} + 2\frac{\lambda}{d^*} a_m^r \mathbf{P} \right\}. \quad (15)$$

Based on the derivative $\frac{\partial J}{\partial \mathbf{P}}$, we also adopt the gradient descent and PSD projection to update $\mathbf{P}$ as

$$\mathbf{P}_t = \mathbf{SVD}\left(\mathbf{P}_{t-1} - \eta_t^* \frac{\partial J}{\partial \mathbf{P}}|_{\mathbf{P}=\mathbf{P}_{t-1}}\right), \quad (16)$$

where $\mathbf{P}_{t-1}$ denotes the value of $\mathbf{P}$ in the $(t-1)$-th iteration. η_t^* is the step-size in the t-th iteration for $\mathbf{P}$. We also apply the dynamic step-size search strategy shown in subsection IV-C.1 for the optimization of $\mathbf{P}$. However, different with that for $\mathbf{M}^m$, the step-size η_t^* for $\mathbf{P}$ is not allowed to be enlarged during the optimization.

3) Optimization of $\mathbf{a}$*:* Considering $\mathbf{M}^m (m = 1, \cdots, \mathcal{M})$ and $\mathbf{P}$ are fixed, then the optimization problem at Eq. (10) can be transformed as

$$\min \sum_{m=1}^{\mathcal{M}} a_m^r F_m, \quad s.t. \quad a_m > 0; \sum_{m=1}^{\mathcal{M}} a_m = 1 \quad (17)$$

where $F_m = \sum_{i=1}^{n} w_i \mathcal{L}\left(y_i f\left(\mathbf{x}_i^m, \mathbf{z}_i^m; \mathbf{x}_i^*, \mathbf{z}_i^*; \mathbf{M}^m, \mathbf{P}\right)\right) + \lambda \mathcal{R}(\mathbf{P})$ which can be treated as a constant.

By using a Lagrange multiplier ξ to take $\sum_{m=1}^{\mathcal{M}} a_m = 1$ into consideration, we get the objective function as

$$Q(\mathbf{a}, \xi) = \sum_{m=1}^{\mathcal{M}} a_m^r F_m - \xi \left(\sum_{m=1}^{\mathcal{M}} a_m - 1 \right). \quad (18)$$

By setting the derivative of $Q(\mathbf{a}, \xi)$ with respect to $a_m (m = 1, \cdots, \mathcal{M})$ and ξ to zero

$$\begin{cases} \frac{\partial Q}{\partial a_1} = r a_1^{r-1} F_1 - \xi = 0 \\ \vdots \\ \frac{\partial Q}{\partial a_{\mathcal{M}}} = r a_{\mathcal{M}}^{r-1} F_{\mathcal{M}} - \xi = 0 \\ \frac{\partial Q}{\partial \xi} = \sum_{m=1}^{\mathcal{M}} a_m - 1 = 0 \end{cases}, \quad (19)$$

we can obtain the closed-form solution of $a_m (m = 1, \cdots, \mathcal{M})$

$$a_m = \frac{(1/F_m)^{\frac{1}{r-1}}}{\sum_{m=1}^{\mathcal{M}} (1/F_m)^{\frac{1}{r-1}}}, \quad (20)$$

Since F_m is positive, we have $a_m > 0$ naturally. When $\mathbf{M}^m (m = 1, \cdots, \mathcal{M})$ and $\mathbf{P}$ are fixed, Eq. (20) gives the global optimal $\mathbf{a}$.

According to Eq. (20), we have the following understanding for r in controlling a_m. If $r \to \infty$, a_m will close to each other, i.e., $a_m \to \frac{1}{\mathcal{M}}$ and each view has an equal contribution. If $r \to 1$, only a_m corresponding to the minimum J_m is close to 1; other weights will be close to zero. Therefore, the choice of r should be based on the complementary property of all views. Rich complementary prefers to a large r; otherwise, it should be small.

Note that we doesn't provide a separate solution for LDML+, since it can be seen as a special case of MVLDML+.

D. Person Re-ID

Once metrics $\mathbf{M}^m (m = 1, \cdots, \mathcal{M})$ and weights $\mathbf{a}$ have been learned after optimization, given a probe image $\mathbf{x}^p$ and a set of gallery images $\{\mathbf{x}_j^g\}_{j=1}^N$ during testing, we compute their squared distances as follows to perform person images matching.

$$d^2\left(\mathbf{x}^p, \mathbf{x}_j^g\right) = \sum_{m=1}^{\mathcal{M}} a_m (\mathbf{x}^p - \mathbf{x}_j^g)^T \mathbf{M}^m \ (\mathbf{x}^p - \mathbf{x}_j^g)$$
$$= \sum_{m=1}^{\mathcal{M}} a_m \|(\mathbf{U}^m)^T \mathbf{x}^p - (\mathbf{U}^m)^T \mathbf{x}_j^g\|^2, \quad (21)$$

where $\mathbf{U}^m$ is the low-dimensional projection of $\mathbf{M}^m$, obtained by SVD in Eq. (13). The gallery images can be ranked according to their distances to the probe image.

V. Experimental Results and Analysis

A. Datasets, Evaluation Protocol, and Setting

1) Datasets: The evaluation of the proposed methods is carried out on four benchmark person re-ID datasets: VIPeR [29], CUHK01 [53], PRID450S [34], and Market-1501 [54].

VIPeR [29] is the most commonly used person re-ID dataset containing 632 persons in which each person has a pair of images taken from widely differing views. It contains 1264 images in total. The large viewpoint change of 90 degrees or more as well as huge lighting variations in VIPeR make it one of the most challenging re-ID datasets. The evaluation protocol for VIPeR is to randomly split the dataset into half, 316 persons for training and 316 persons for testing.

CUHK01 [53] contains 971 persons from two disjoint camera views, where each person has two images in each camera view. It contains 3884 images in total. We randomly partition the CUHK01 dataset into 486 persons for training and 485 for testing.

PRID450S [34] is a commonly used dataset which contains 450 identities from two disjoint camera views, where each person has one image in each camera view. 225 persons are randomly selected for training and the rest for testing.

Market-1501 [54] is one of the largest person re-ID datasets, containing 32668 bounding boxes (cropped images) of 1501 identities. All the bounding boxes are detected by Deformable Part Model pedestrian detector [55]. Each identity has multiple images captured by at least two cameras and at most six cameras. We adopt the standard protocol [54] for Market-1501. Specifically, the training set contains 12936 bounding boxes of 750 identities. The testing set contains 19732 bounding boxes of 751 identities, where only one image of each identity is randomly selected as query image for each camera. In total, the testing set contains 3368 query images. There are 2793 images included as distractors in the original gallery set for testing.

TABLE I
The Characteristics of Four Person Re-ID Datasets

Datasets	# ID	# BBoxes	# Distra	# Cam
VIPeR [29]	632	1264	0	2
CUHK01 [53]	971	3884	0	2
PRID450S [34]	450	900	0	2
Market-1501 [54]	1501	32668	2793	6

Fig. 1. Sample images from four person re-identification datasets: (a) VIPeR; (b) CUHK01; (c) PRID450S; (d) Market-1501.

Table I provides a statistical summary of each dataset. In Table I, we indicate the number of identities (ID), bounding boxes (BBoxes), distractors (Distra), and cameras (Cam) in each dataset. Fig. 1 shows some image samples from these four datasets.

2) Evaluation Protocol: For the three small datasets: VIPeR, CUHK01, and PRID450S, we randomly divide each

dataset into training and testing sets containing half of the available individuals. As random selection is involved, we repeat the evaluation procedure for 10 times and report the mean results. Single-query (SQ) setting is adopted for all these three datasets. For the Market-1501 dataset, one of the largest re-ID datasets, we use its standard evaluation protocol [54]. Both the single-query and multi-query (SQ) matching results are reported on Market-1501. Two standard evaluation metrics are used in this work: Cumulative Matching Characteristics (CMC) curve and Mean Average Precision (mAP). The CMC curve provides a ranking for every image in the gallery with respect to the probe. It is used for all datasets. The mAP measure is only presented for Market-1501.

3) Features: The GOG descriptor [2] describes a local region in a person image via hierarchical Gaussian distribution in which both means and covariances are included in their parameters. It has four GOG features: {GOG_{RGB}, GOG_{Lab}, GOG_{HSV}, GOG_{nRnG}}, extracted respectively from four color channels {RGB, Lab, HSV, nRnG}. Here, the nRnG is the normalized RGB color space. In this work, we evaluate the effectiveness of LDML+ using the first three GOG features as original features respectively. All the GOG features are used to evaluate MVLDML+.

For the privileged information, it is better to be represented by some high-level features (e.g., attributes). However, up to now, it is hard to obtain ideal privileged features. To evaluate the effectiveness of the proposed method in this work, we consider an approximated setting for privileged information. For the VIPeR, CUHK01, and PRID450S datasets, we fuse multiple strong visual features (LOMO feature and FTCNN feature [56]) to approximate the privileged information. We apply the method in [3] to obtain a low-dimensional representation of the approximated privileged feature. For the Market-1501 dataset, we combine the predicted pedestrian attributes [57] and semantics-preserving deep embeddings [57] as the privileged information. Note that the privileged information is not available during testing.

4) Setting: For the scale parameter in Eq. (5), we set $\beta = \frac{mean(\mathbf{D_M})}{mean(\mathbf{D_P})}$ where $\mathbf{D_M}, \mathbf{D_P}$ denote the squared Euclidean distance matrices corresponding to the original features and the privileged features, respectively. The regularization parameter λ is empirically set to 0.0001 for Market-1501. For other datasets, λ is set to 0.001. We set the parameter $r = 3$ in Eq. (10) empirically. For the optimization of $\mathbf{M}$ and $\mathbf{P}$, we initialize the step-sizes with large values as $\eta_0^m = 2^{20}$ and $\eta_0^* = 2^{15}$ respectively. For the parameter s which controls the upper-boundary of η^m, we set it to $s = 2^5$. The maximal iteration number is set to 400 with a stopping criterion by $|\frac{J_t - J_{t-1}}{J_{t-1}}| \leq 10^{-4}$. PCA is applied for all datasets for dimension reduction but all energies are retained for the three small datasets. After PCA, the detailed dimensions of the original features on VIPeR, CUHK01, and PRID450S are 631, 1943, and 449 respectively. For the large dataset, Market-1501, we retain 99% energies. After PCA on Market-1501, the detailed dimensions of the GOG_{RGB}, GOG_{Lab}, GOG_{HSV} features are 4411, 4409, and 4468, respectively.

For the evaluation of LDML+, the following baseline methods are employed:

1) XQDA [3]. It is an efficient yet effective metric learning algorithm which learns a discriminative distance metric and a low-dimensional subspace simultaneously. It's a state-of-the-art method especially on small-size datasets. Default setting in [3] is applied for XQDA. Note that it doesn't apply PCA for dimension reduction. The input feature vectors will be projected into a low-dimensional subspace directly.
2) Global decision threshold based methods:
   - MLAPG [5]. It is a state-of-the-art logistic discriminant metric learning method under Eq. (4) in which the global decision threshold σ is set to the average squared Euclidean distance. Different with LDML+, it applies the Accelerated Proximal Gradient (APG) algorithm to optimize the metric. We use the default setting in [5] for MLAPG.
   - LDML. A logistic discriminant metric learning method under Eq. (4) in which the default setting of σ is the average squared Euclidean distance. LDML shares the same settings with LDML+ on the loss function $\mathcal{L}(\cdot)$, pair weights w, and optimization strategy of $\mathbf{M}$. We also provide the results of LDML with $\sigma = 1$ and term it as $\text{LDML}^{\sigma=1}$.

For the evaluation of MVLDML+, we build two baseline methods using XQDA [3] and MLAPG [5], respectively, based on a score-level fusion strategy. In detail, we learn an ensemble of distance functions, in which each base distance function is learned using a single feature descriptor. The final distance is calculated from a weighted sum of these distance functions with equal weights. The two baseline methods are termed as:

- Ensem-XQDA
- Ensem-MLAPG

B. Experiments on VIPeR, CUHK01, and PRID450S

In this subsection, we evaluate the effectiveness of LDML+ and MVLDML+ on the three small-size datasets: VIPeR, CUHK01, and PRID450S, respectively.

1) Evaluation of LDML+: We first evaluation the effectiveness of LDML+ using three GOG descriptors (GOG_{RGB}, GOG_{Lab}, and GOG_{HSV}), respectively. The evaluation results are shown in Table II.

It can be seen from Table II that the proposed LDML+ method performs better than the three global decision threshold based methods (MLAPG, $\text{LDML}^{\sigma=1}$, and LDML). Among them, LDML performs better than MLAPG and $\text{LDML}^{\sigma=1}$. On VIPeR, LDML+ surpasses its counterpart, LDML, by 1.71%, 1.3%, and 0.54% at rank-1 using three different GOG features, respectively. On CUHK01, LDML+ beats LDML by a large margin, 3.57%, 3.54%, 4.09% at rank-1, respectively. On PRID450S, the improvements are similar with those on VIPeR, which are 0.98%, 0.36%, and 1.91% at rank-1. It reveals that the locally adaptive decision rule can cope better with the complex intra-class and inter-class variations than the global threshold based decision rule, especially on the CUHK01 dataset that is larger than the VIPeR and PRID450S datasets.

TABLE II

TOP-RANKED AVERAGE RECOGNITION RATES (CMC@RANK-R, %) OF LDML+ AND FOUR BASELINE METHODS ON THE **VIPeR**, **CUHK01**, AND **PRID450S** DATASETS WITH THREE DIFFERENT VISUAL FEATURE DESCRIPTORS. A LARGER NUMBER INDICATES A BETTER RESULT. THE BEST RESULTS ARE SHOWN IN BOLDFACE

VIPeR		GOG_{RGB}				GOG_{Lab}				GOG_{HSV}			
		R=1	R=5	R=10	R=20	R=1	R=5	R=10	R=20	R=1	R=5	R=10	R=20
Baseline methods	XQDA [3]	43.77	74.81	84.34	93.32	44.24	74.91	85.13	92.94	39.30	68.39	79.59	89.68
	MLAPG [5]	42.66	74.40	85.47	93.83	44.30	75.38	85.66	93.51	39.59	69.08	80.82	89.94
	$LDML^{\sigma=1}$	43.26	74.84	85.54	93.86	44.08	75.51	85.82	**93.67**	39.46	69.56	80.57	90.19
	LDML	43.48	75.09	85.63	93.77	44.49	75.95	86.08	93.64	39.46	**69.68**	**80.89**	**90.51**
Ours	LDML+	**45.19**	**75.32**	**85.66**	**93.99**	**45.79**	**76.33**	**86.30**	93.16	**40.00**	69.40	80.79	89.94
CUHK01		GOG_{RGB}				GOG_{Lab}				GOG_{HSV}			
		R=1	R=5	R=10	R=20	R=1	R=5	R=10	R=20	R=1	R=5	R=10	R=20
Baseline methods	XQDA [3]	55.84	78.85	85.52	91.32	53.82	77.73	84.76	90.92	45.55	71.46	79.96	87.65
	MLAPG [5]	53.38	77.72	85.19	91.29	52.77	77.32	85.06	91.32	44.72	71.49	80.89	88.47
	$LDML^{\sigma=1}$	50.01	74.89	83.58	89.94	48.31	72.65	81.58	89.07	39.96	65.87	75.95	85.38
	LDML	54.61	78.49	85.88	91.53	53.34	77.40	85.20	91.42	45.47	71.51	81.16	88.67
Ours	LDML+	**58.18**	**80.97**	**87.45**	**92.87**	**56.88**	**80.25**	**87.08**	**92.47**	**49.56**	**75.35**	**83.60**	**90.94**
PRID450S		GOG_{RGB}				GOG_{Lab}				GOG_{HSV}			
		R=1	R=5	R=10	R=20	R=1	R=5	R=10	R=20	R=1	R=5	R=10	R=20
Baseline methods	XQDA [3]	**62.36**	**85.47**	**91.78**	**96.27**	**62.40**	**86.22**	92.53	**96.98**	56.00	80.93	89.02	94.71
	MLAPG [5]	59.56	83.47	90.89	95.96	58.71	84.22	92.13	96.40	53.82	79.69	88.09	94.31
	$LDML^{\sigma=1}$	58.62	83.11	91.20	95.60	55.96	81.29	90.94	95.51	48.62	75.42	85.07	91.64
	LDML	59.73	83.64	91.33	96.04	59.64	84.49	92.40	96.53	54.58	80.27	88.44	94.44
Ours	LDML+	60.71	84.58	91.73	96.18	60.00	85.38	**92.58**	96.58	**56.49**	**81.73**	**89.42**	**95.24**

TABLE III

TOP-RANKED AVERAGE RECOGNITION RATES (CMC@RANK-R, %) OF MVLDML+ AND TWO BASELINE METHODS ON THE **VIPeR**, **CUHK01**, AND **PRID450S** DATASETS. A LARGER NUMBER INDICATES A BETTER RESULT. THE BEST RESULTS ARE SHOWN IN BOLDFACE

GOG Features	Methods	VIPeR				CUHK01				PRID450S			
		R=1	R=5	R=10	R=20	R=1	R=5	R=10	R=20	R=1	R=5	R=10	R=20
RGB, Lab, HSV	Ensem-XQDA	47.72	76.93	86.99	94.75	57.59	79.86	86.47	92.42	64.53	87.24	93.02	96.98
	Ensem-MLAPG	47.47	78.01	87.59	**94.97**	56.95	80.33	87.11	92.77	62.76	85.60	92.40	96.62
	MVLDML+	**48.86**	**78.16**	**87.82**	94.53	**60.73**	**82.66**	**88.99**	**93.84**	**64.80**	**88.13**	**94.00**	**97.64**
RGB, Lab, HSV, nRnG	Ensem-XQDA	49.08	77.47	87.37	94.62	58.43	80.04	86.70	92.49	**68.00**	88.00	94.00	97.29
	Ensem-MLAPG	49.05	78.67	88.48	**94.97**	57.75	80.21	87.15	92.60	64.36	86.93	93.51	97.29
	MVLDML+	**50.03**	**79.21**	**88.54**	94.65	**61.37**	**82.74**	**88.88**	**93.85**	66.71	**88.80**	**94.44**	**97.51**

Our LDML+ method also outperforms the XQDA method on VIPeR and CUHK01 datasets, benefiting from the locally adaptive decision rule. On VIPeR, the improvements are 1.42%, 1.55%, and 0.7% at rank-1, respectively. On CUHK01, LDML+ yields significant improvements on XQDA at rank-1 by 2.34%, 3.06%, and 4.01%, respectively. XQDA performs better than our method using the GOG_{RGB} and GOG_{Lab} features on PRID450S. Since XQDA learns a discriminative low-dimensional subspace and a Mahalanobis distance metric simultaneously. Here, the discriminative subspace projection is a supervised technique for dimensionality reduction, which can retain more discriminative information than the PCA technique used in our method and other baselines. The advantage of a supervised dimensionality reduction technique will be more obvious when it is applied on a small dataset, e.g., PRID450S.

We note that LDML performs better than MLAPG on the three datasets although they both use the mean squared Euclidean distance as the global threshold, which can be owned to the effectiveness of the dynamical step-size adaption strategy applied in LDML+, $LDML^{\sigma=1}$, and LDML. MLAPG also applies a linear search strategy to find a suitable step-size. However, different with ours, it doesn't allow the step-size to be enlarged. Once a very small step-size is searched at the beginning of the optimization, the subsequent gradient descent would be much slower, thus resulting in suboptimal performance. In addition, as shown in Table II, compared with $LDML^{\sigma=1}$, LDML achieves better or comparable performances. It demonstrates that a data-dependent global threshold is better than a data-independent global threshold.

From the results shown in Table II, we can conclude that a more reliable metric can be learned by exploiting the privileged information to design a locally adaptive decision rule during training.

2) Evaluation of MVLDML+: By extending the LDML+ method to the multi-view setting in Eq. (10), we can simultaneously exploit multiple original features to learn an ensemble of base distance functions. The final distance is a weighted sum of these distance functions in Eq. (21). Table III shows the performance comparison of the proposed MVLDML+ method

TABLE IV

COMPARISONS OF TOP-RANKED AVERAGE RECOGNITION RATES (CMC@RANK-R, %) OF MVLDML+ WITH STATE-OF-THE-ART RESULTS ON THE **VIPeR** DATASET. A LARGER NUMBER INDICATES A BETTER RESULT. THE BEST RESULTS ARE SHOWN IN BOLDFACE

Methods	References	R=1	R=5	R=10	R=20
MVLDML+	Ours	**50.0**	79.2	88.5	94.7
GOG_{Fusion}+XQDA [2]	CVPR 2016	49.7	**79.7**	88.7	94.5
Cheng et al. [58]	CVPR 2016	47.8	74.7	84.8	91.1
GOG_{Fusion}+LDNS [7]	CVPR 2016	47.6	78.1	88.4	94.6
MetricEnsemble [9]	CVPR 2015	45.9	77.5	**88.9**	**95.8**
mFilter+LADF [38]	CVPR 2014	43.4	73.0	84.9	93.7
MirrorKMFA [59]	IJCAI 2015	43.0	75.8	87.3	94.8
LSSCDL [60]	CVPR 2016	42.7	-	84.3	91.9
LOMO+LDNS [7]	CVPR 2016	42.3	71.5	82.9	92.1
Su et al. [61]	ICCV 2015	42.3	72.2	81.6	89.6
Shi et al. [62]	CVPR 2015	41.6	71.9	86.2	95.1
LOMO+MLAPG [5]	ICCV 2015	40.7	-	82.3	92.4
LOMO+XQDA [3]	CVPR 2015	40.0	68.0	80.5	91.1
Xiao et al. [63]	CVPR 2016	38.6	-	-	-
Chen et al. [64]	TIP 2016	38.4	69.2	81.3	90.4
SCNCD [27]	ECCV 2014	37.8	68.5	81.2	90.4
Chen et al. [65]	CVPR 2015	36.8	70.4	83.7	91.7
Shen et al. [66]	ICCV 2015	34.8	68.7	82.3	91.8

TABLE V

COMPARISONS OF TOP-RANKED AVERAGE RECOGNITION RATES (CMC@RANK-R, %) OF MVLDML+ WITH STATE-OF-THE-ART RESULTS ON THE **CUHK01** DATASET. A LARGER NUMBER INDICATES A BETTER RESULT. THE BEST RESULTS ARE SHOWN IN BOLDFACE

Methods	References	R=1	R=5	R=10	R=20
MVLDML+	Ours	**61.4**	82.7	88.9	93.9
GOG_{Fusion}+LDNS [7]	CVPR 2016	60.8	81.7	88.4	93.5
CPDL [67]	IJCAI 2015	59.5	81.3	89.7	93.1
GOG_{Fusion}+XQDA [2]	CVPR 2016	57.8	79.1	86.2	92.1
Cheng et al. [58]	CVPR 2016	53.7	**84.3**	**91.0**	**96.3**
MetricEnsemble [9]	CVPR 2015	53.4	76.4	84.4	90.5
Chen et al. [64]	TIP 2016	50.4	75.9	84.0	91.3
LOMO+XQDA [3]	CVPR 2015	49.2	75.7	84.2	90.8
Ahmed et al. [68]	CVPR 2015	47.5	71.0	80.0	-
MirrorKMFA [59]	IJCAI 2015	40.4	64.6	75.3	84.1

TABLE VI

COMPARISONS OF TOP-RANKED AVERAGE RECOGNITION RATES (CMC@RANK-R, %) OF MVLDML+ WITH STATE-OF-THE-ART RESULTS ON THE **PRID450S** DATASET. A LARGER NUMBER INDICATES A BETTER RESULT. THE BEST RESULTS ARE SHOWN IN BOLDFACE

Methods	References	R=1	R=5	R=10	R=20
MVLDML+	Ours	66.8	**88.8**	**94.8**	97.7
GOG_{Fusion}+XQDA [2]	CVPR 2016	**68.4**	**88.8**	94.5	**97.8**
FFN [70]	WACV 2016	66.6	86.8	92.8	96.9
GOG_{Fusion}+LDNS [7]	CVPR 2016	64.8	88.1	94.0	97.6
LOMO+XQDA [3]	CVPR 2015	62.6	85.6	92.0	96.6
LSSCDL [60]	CVPR 2016	60.5	-	88.6	93.6
MirrorKMFA [59]	IJCAI 2015	55.4	79.3	87.8	93.9
MEDVL [70]	AAAI 2016	45.9	73.0	82.9	91.1
Shi et al. [62]	CVPR 2015	44.9	71.7	77.5	86.7
Shen et al. [66]	ICCV 2015	44.4	71.6	82.2	89.8
SCNCD [27]	ECCV 2014	41.6	68.9	79.4	87.8

with two score-level fusion based methods: Ensem-XQDA and Ensem-MLAPG.

By exploiting the first three visual GOG descriptors from the RGB, Lab, HSV color spaces, MVLDML+ achieves the best performance 48.86% at rank-1 on VIPeR. It outperforms Ensem-XQDA and Ensem-MLAPG by 1.14% and 1.39%, respectively, at rank-1. On CUHK01, the improvements are more obvious. MVLDML+ surpasses Ensem-XQDA and Ensem-MLAPG by over 3% at rank-1. On PRID450S, the improvements are 0.27% and 2.04% at rank-1 respectively. It indicates that MVLDML+ can explore the complementary of different visual features more effectively. By exploring the complementary of the three descriptors, MVLDML+ improves the rank-1 recognition rate of LDML+ with GOG_{RGB} by 3.67%, and that of LDML+ with GOG_{HSV} by 8.86%.

We also report the performance of MVLDML+ using all the four GOG features in Table III. By employing all the four descriptors, the rank-1 accuracy of MVLDML+ has been improved from 48.86% to 50.03% on VIPeR, 60.73% to 61.37% on CUHK01, and 64.80% to 66.71% on PRID450S. The results shown in Table III demonstrate the effectiveness of the MVLDML+ method.

3) Comparison With State-of-the-Art Results: In Table IV, we compare the performance of MVLDML+ with most recent state-of-the-art results on VIPeR at different ranks. By using four GOG descriptors, our method achieves the best rank-1 recognition rate 50.0%. Note that it is only slightly better than the result of GOG_{Fusion}+XQDA reported in [2]. That is because we use a different split of dataset with [2]. Our split is the same as that in MLAPG [5], where the dataset is randomly split with a fixed random seed (rng(0) on Matlab) for 10 times. If GOG_{Fusion}+XQDA is implemented using our split, it obtains 48.42% at rank-1, 78.23% at rank-5, and 87.63% at rank-10, which is clearly lower than our results.

We also report the results of LDNS [7] in Table IV, a more recent metric learning method, using the GOG_{Fusion} descriptor. It formulates a much stricter learning objective than the classic Fisher discriminative analysis (FDA) method. It aims to minimize the within-class scatter by collapsing all the images of the same person into a single point. This learning objective may be too rigorous to cope with the GOG_{Fusion} descriptor in which multiple different feature descriptors are fused, thus resulting inferior performance than MVLDML+ and XQDA.

The third best result in Table IV is obtained with a multi-channel parts-based convolutional neural network model [58] which has a very strong feature representation ability. Our result surpasses the third best result by 2.2% at rank-1.

Table V compares the top-ranked recognition rate of MVLDML+ with most recent state-of-the-art results on CUHK01. We can observe that, owing to the effectiveness of locally adaptive decision rule, MVLDML+ achieves a state-

TABLE VII

TOP-RANKED AVERAGE RECOGNITION RATES (CMC@RANK-R, %) OF LDML+ AND FOUR BASELINE METHODS ON THE **MARKET-1501** DATASET WITH THREE DIFFERENT VISUAL FEATURE DESCRIPTORS. A LARGER NUMBER INDICATES A BETTER RESULT. THE BEST RESULTS ARE SHOWN IN BOLDFACE

Market-1501 (SQ)		GOG_{RGB}				GOG_{Lab}				GOG_{HSV}			
		R=1	R=5	R=10	mAP	R=1	R=5	R=10	mAP	R=1	R=5	R=10	mAP
Baseline methods	XQDA [3]	43.29	65.38	73.81	23.26	43.32	65.65	74.26	23.30	35.24	56.84	68.29	17.55
	MLAPG [5]	46.26	68.56	77.05	24.13	47.48	69.15	77.94	24.96	40.26	64.43	73.63	19.79
	$LDML^{\sigma=1}$	41.24	65.29	73.78	21.52	35.30	57.69	66.48	16.83	34.62	60.15	69.83	16.65
	LDML	47.83	70.55	78.36	25.67	48.55	70.43	78.71	25.60	40.91	65.08	74.47	20.39
Ours	LDML+	**52.08**	**73.93**	**81.68**	**29.10**	**52.29**	**74.47**	**81.59**	**28.72**	**45.49**	**69.21**	**78.47**	**23.06**
Market-1501 (MQ)		GOG_{RGB}				GOG_{Lab}				GOG_{HSV}			
		R=1	R=5	R=10	mAP	R=1	R=5	R=10	mAP	R=1	R=5	R=10	mAP
Baseline methods	XQDA [3]	51.60	72.92	81.18	29.48	51.01	77.62	80.79	29.32	44.45	67.90	76.57	23.39
	MLAPG [5]	56.50	77.55	84.35	30.99	57.48	79.54	86.19	32.12	51.54	75.45	82.72	26.45
	$LDML^{\sigma=1}$	52.76	75.65	82.84	28.11	43.79	66.51	74.82	21.23	45.87	71.62	79.96	22.79
	LDML	58.61	79.63	86.19	32.97	58.28	80.40	86.67	33.00	52.08	76.37	83.73	27.34
Ours	LDML+	**63.63**	**83.05**	**88.33**	**37.16**	**62.53**	**83.28**	**88.48**	**37.05**	**57.60**	**79.16**	**85.39**	**31.05**

TABLE VIII

TOP-RANKED AVERAGE RECOGNITION RATES (CMC@RANK-R, %) OF MVLDML+ AND TWO BASELINE METHODS ON THE **MARKET-1501** DATASET. A LARGER NUMBER INDICATES A BETTER RESULT. THE BEST RESULTS ARE SHOWN IN BOLDFACE

Market-1501		SQ				MQ			
GOG Features	Methods	R=1	R=5	R=10	mAP	R=1	R=5	R=10	mAP
RGB,Lab,HSV	Ensem-XQDA	45.58	67.07	75.45	25.4	53.15	74.88	82.21	31.72
	Ensem-MLAPG	51.34	73.01	80.94	28.23	60.96	81.26	87.41	35.56
	MVLDML+	**55.94**	**77.94**	**84.35**	**32.19**	**66.81**	**85.33**	**90.23**	**40.63**
RGB,Lab,HSV,nRnG	Ensem-XQDA	47.3	68.14	75.67	26.36	55.20	75.50	82.72	32.65
	Ensem-MLAPG	51.99	73.57	81.03	28.62	61.25	81.15	87.02	35.92
	MVLDML+	**58.22**	**78.59**	**84.98**	**33.7**	**68.38**	**86.28**	**90.59**	**41.84**

of-the-art rank-1 accuracy 61.4%, improving the result of GOG_{Fusion}+XQDA by 3.6%.

Table VI compares the top-ranked recognition rates of MVLDML+ with several state-of-the-art results on PRID450S. It is observed that our method achieves the second best rank-1 recognition rate on PRID450S, which is comparable to deep network based result [69]. It can be mainly owed to the effectiveness of the locally adaptive decision rule and the multi-view learning strategy.

Note that, we only exploit four GOG color descriptors in this work. MVLDML+ can achieve a better performance by exploiting more visual features (e.g. WHOS [31]).

C. Experiments on Market-1501

In this subsection, we evaluate the effectiveness of the proposed LDML+ and MVLDML+ methods on the large dataset, Market-1501 [54]. The training set contains 12936 bounding boxes of 750 identities. The testing set contains 19732 bounding boxes of 751 identities. Compared to the above three small size datasets, Market-1501 has more complex intra-class and inter-class variations. To formulate a cross-camera setting, we split the training set into a probe set and a gallery set, where both the probe and gallery sets have 4914 training samples. 9828 training samples are finally used. 46518 positive pairs and over 20 million negative pairs are generated.

Table VII and Table VIII show the performance evaluation of the LDML+ and MVLDML+ methods respectively. Table IX compares the top-ranked recognition rates and mAP scores of MVLDML+ with several state-of-the-art results on Market-1501. Both the single-query and multi-query results are provided.

1) Evaluation of LDML+: As shown in Table VII, the proposed LDML+ method performs much better than the baseline methods on Market-1501, one of the largest datasets. In the single-query (SQ) setting, LDML+ obtains 52.08%, 52.29%, and 45.49% at rank-1, respectively, and 29.10%, 28.72% and 23.06% in terms of mAP, respectively, using the three GOG features. Our method beats the XQDA method, which performs very well on small size datasets, by a large margin. Using the GOG_{RGB} feature, LDML+ surpasses XQDA by 8.79% at rank-1 and 5.84% in terms of mAP. Similar improvements can also be obtained using the other two features. It is mainly due to that the Gaussian assumption of XQDA does not hold any more on such a large dataset with much more complex intra-class and inter-class variations.

In the SQ setting, using the three features respectively, LDML+ improves the rank-1 recognition rates of its counterpart, LDML, by 4.25%, 3.74%, and 4.58%, and the mAP scores by 3.43%, 3.12%, and 2.67%. It reflects that the locally adaptive decision rule performs much better than global decision rule at the large dataset. It can better characterize the similarity relationship between a pair of person samples and guide the gradient descending toward the right direction, thus resulting a more reliable metric.

In the multi-query (MQ) setting, using the three features respectively, the rank-1 recognition rates of LDML+ have been improved from 52.08% to 63.63%, 52.29% to 62.53%, and 45.49% to 57.60%. Significant improvements of LDML+ on the baseline methods can also be observed in the MQ setting.

Besides, we find that the rank of metric M in LDML+ drops faster than that in the two global decision threshold based methods, LDML and MLAPG, on Market-1501 dataset, thus resulting a lower-rank yet more discriminative metric.

The improvements shown in Table VII have clearly demonstrated the effectiveness of LDML+.

2) Evaluation of MVLDML+: As shown in Table VIII, by employing the three GOG features as input, MVLDML+ obtains 55.94% rank-1 recognition rate and 32.19% mAP score in SQ setting, and 66.81% rank-1 recognition rate and 40.63% mAP score in MQ setting. Comparing the results of MVLDML+ in Table VIII with the results of LDML in Table VII, we can find that MVLDML+ improves the performance of LDML+ by a large margin. It reveals that the complementary of multiple feature descriptors has been well exploited.

Table VIII compares the performance of MVLDML+ with Ensem-XQDA and Ensem-MLAPG in both the SQ and MQ settings. By only using three GOG features, MVLDML+ improves the performance of Ensem-XQDA by over 10% at rank-1 in SQ setting and 13% at rank-1 in MQ setting. It also beats Ensem-MLAPG by over 4% at rank-1 in SQ setting and 5% at rank-1 in MQ setting. Similar improvements can also be observed when all the four GOG features have been employed in MVLDML+. The results shown in Table VIII demonstrate the effectiveness of MVLDML+.

3) Comparison With State-of-the-Art Results: In Table IX, we compare the performance of MVLDML+ with most recent state-of-the-art results on Market-1501. Both the rank-1 recognition rate and mAP score are used as the evaluation metrics. We also apply a re-ranking strategy [79] to boost the performance of MVLDML+ in testing stage, which is termed as MVLDML+ (Re) in Table IX.

It should be noted that most recent results in Table IX are obtained by training deep neural networks on Market-1501, which can yield much stronger features. For example, Varior *et al.* [78] presents a gated Siamese architecture that yields the best rank-1 recognition rate in both SQ and MQ settings in Table IX. However, our proposed method uses the hand-crafted GOG color descriptors as input to learn a low-rank and discriminative metric, obtaining 58.22% rank-1 recognition rate in SQ setting and 68.38% rank-1 recognition rate in MQ setting. Although it is lower than the results in [78], it still performs better than most listed results in Table IX including multiple results of deep models [73], [75]–[77]. By employing the re-ranking technique, our results can be significantly boosted. MVLDML+ (Re) yields the best mAP scores (48.01% and 56.45%) in both SQ and MQ settings. It obtains 64.82% rank-1 recognition rate in SQ setting and 74.58% rank-1 recognition rate in MQ setting, which is comparable to the results of [78].

TABLE IX

PERFORMANCE COMPARISON (CMC@RANK-R AND MAP, %) ON THE **MARKET-1501** DATASET. A LARGER NUMBER INDICATES A BETTER RESULT. BOTH SINGLE-QUERY (SQ) AND MULTI-QUERY (MQ) EVALUATION RESULTS ARE PRESENTED RESPECTIVELY

Market 1501		SQ		MQ	
		R=1	mAP	R=1	mAP
MVLDML+	Ours	58.22	33.70	68.38	41.84
MVLDML+ (Re)	Ours	64.82	**48.01**	74.58	**56.45**
BoW+KISSME [54]	ICCV 2015	44.42	20.76	-	-
LOMO+XQDA [3]	CVPR 2015	43.80	22.20	-	-
WARCA [71]	ECCV 2016	45.16	-	-	-
LOMO+LDNS [7]	CVPR2016	55.43	29.87	67.96	41.89
SLSC [12]	CVPR2016	51.90	26.35	-	-
TMA [72]	ECCV 2016	47.92	22.31	-	-
AttentionNet [73]	TIP 2017	48.24	24.43	-	-
DeepAttribute [74]	ECCV 2016	39.40	19.60	49.00	25.80
MSTripletCNN [75]	MM 2016	45.10	-	55.40	-
DeepEmbedd [76]	NIPS 2016	59.47	-	-	-
SiameseLSTM [77]	ECCV 2016	-	-	61.60	35.30
ContrastiveLoss [78]	ECCV 2016	62.32	36.23	72.92	45.39
GatedSiamese [78]	ECCV 2016	**65.88**	39.55	**76.04**	48.45

D. Analysis of the Proposed Method

1) On the Parameter β: We introduce in Eq. (5) a scale parameter $\beta = \frac{mean(\mathbf{D_M})}{mean(\mathbf{D_P})}$ that is expected to globally bridge the gap between the privileged features $\{\mathbf{x}^*\}$ and original features $\{\mathbf{x}\}$. The reason for that is, in some case, $\{\mathbf{x}^*\}$ may have a significantly different distribution from $\{\mathbf{x}\}$, especially when the privileged information is partially or even totally wrong. Here, the scale parameter β can smooth the distance $\mathbf{D_P}$ in privileged space and is helpful for searching a suitable step-size at the beginning of the optimization. In this subsection, we compare the performance of LDML+ with and without β in two cases. One is representing $\{\mathbf{x}^*\}$ with the privileged features by the setting in Section V-A. The other case is replacing the privileged features with randomly-generated vectors. The performance comparison is shown in Fig. 2, where the VIPeR dataset is used as an example.

In the first case shown in Fig. 2(a), the improvements of LDML+ with β on that without β is not obvious. However, in the second case shown in Fig. 2(b), where the privileged information is totally wrong, LDML+ with β significantly surpasses LDML+ without β. It shows that in some extreme case where privileged information is totally wrong, the scale parameter β does help. With β, LDML+ can still yield satisfactory results, although lower than the results in Fig. 2(a). Without β, the gradient of $\mathbf{M}$ would descend toward a totally wrong direction, thus resulting a bad performance.

The results in Fig. 2 clearly demonstrates the effectiveness of the scale parameter β.

2) On the Parameter λ: The parameter λ modulates the effect of the regularization term $\mathcal{R}(\mathbf{P}) = \|\mathbf{P}\|_F^2/d^*$. If λ is too small, the metric $\mathbf{P}$ will have higher degree of freedom, which may result in slow convergence. While, a large λ may degrade the performance of our method because of premature convergence. In this subsection, we investigate the effects

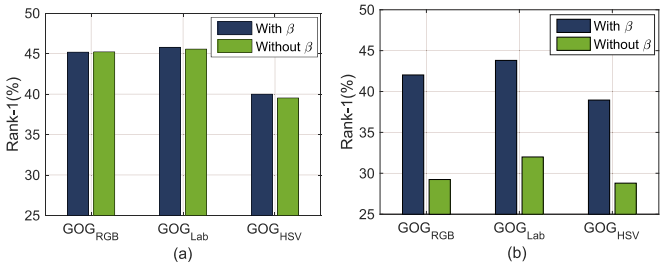

Fig. 2. Performance comparison of LDML+ with and without the parameter β on the VIPeR dataset. (a) $\{\mathbf{x}^*\}$ are represented by the privileged features; (b) $\{\mathbf{x}^*\}$ are represented by randomly-generated vectors.

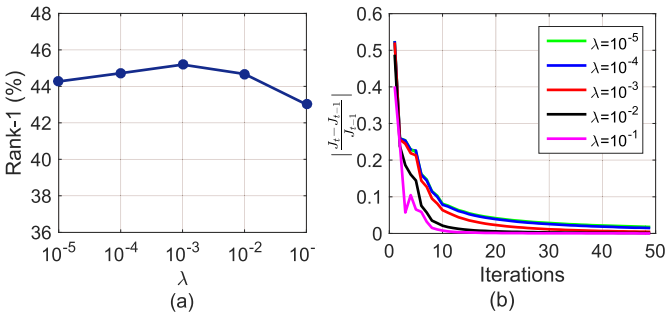

Fig. 3. Effects of the parameter λ on the performance and convergence of LDML+. (a) Rank-1 recognition rates (b) Convergence curves.

of λ on the performance and convergence of the proposed LDML+ method. Here, the VIPeR dataset is used as an example. We use the GOG_{RGB} descriptor as original feature representation. We illustrate the changes of rank-1 recognition rate of the LDML+ method in Fig. 3(a) by varying λ from 10^{-5} to 10^{-1}. To analyze of the effect of λ on the convergence, we use $|\frac{J_t - J_{t-1}}{J_{t-1}}|$ as the objective value and observe the changes of objective value with different settings of λ in Fig. 3(b).

It can be seen from Fig. 3(a) that the rank-1 recognition rate of LDML+ changes little when $10^{-5} \leq \lambda \leq 10^{-2}$. When $\lambda > 10^{-2}$, the performance drops fast. It shows that our method is sensitive to a large λ. As shown in Fig. 3(b), the larger the parameter λ is, the faster the algorithm converges. When $\lambda = 10^{-1}$, the algorithm converges in less than 20 iterations but with the lowest rank-1 recognition rate, since the algorithm has fallen in a bad local optima.

We empirically set $\lambda = 10^{-3}$ as a trade-off between the performance and the convergence speed on the three small size datasets. For the large dataset, Market-1501, we set λ to 10^{-4}.

3) Performance Comparison at Varying PCA Dimensions: PCA is applied for dimension reduction in this work but almost all energy is retained. In this subsection, we compare the performance of LDML+ with LDML and MLAPG at varying PCA dimensions. The VIPeR dataset is used as an example and the GOG_{RGB} feature is employed. Fig. 4 compares the rank-1 recognition rates of LDML+ with two baseline methods at varying PCA dimensions $d \in \{100, 200, 300, 400, 500, 631\}$. Here, 631 is the full PCA dimension on VIPeR.

As shown in Fig. 4, when d is high (e.g. 631) where more energy is retained, the improvement of LDML+ on the

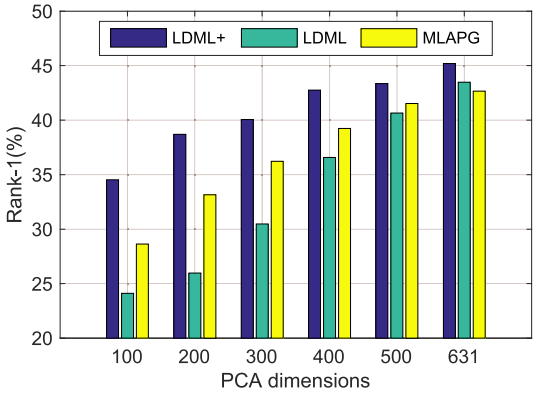

Fig. 4. Performance comparison of LDML+ with LDML and MLAPG at varying PCA dimensions.

two baseline methods is small. Nearly 2% improvement at rank-1 is observed when $d = 500$. With the decreasing of the dimension d, the advantage becomes more obvious. When the dimension is dropped to 100, LDML+ yields remarkable improvements on the two global decision threshold based methods, over 10% on LDML and 5% on MLAPG at rank-1. It is mainly because that given a much lower-dimensional representation, LDML and MLAPG may be more prone to overfitting on training set. While, benefiting from the locally adaptive decision rule built with privileged features, LDML+ can yield a metric with higher generalization capacity on testing set. Besides, high-dimension original features just like a *student* with strong learning ability. It can already learn a good metric without the help of *teacher* (privileged features). While, the low-dimensional features just like a *student* with poor learning ability, it can gain more help from the *teacher*.

4) The Effects of the Privileged Metric $\mathbf{P}$*:* In this subsection, we investigate the effects of the privileged metric $\mathbf{P}$ on the Market-1501 dataset in SQ setting. GOG_{RGB} feature is used as an example. Fig. 5 presents normalized distance histograms of positive training pairs and negative training pairs on both the original feature space and the privileged feature space before/after metric learning.

In this work, the metric $\mathbf{P}$ is always associated with the distance $d^2_{\mathbf{P}}(\mathbf{x}^*_i, \mathbf{z}^*_i)$ that functions as a local decision threshold to guide the learning of the target metric $\mathbf{M}$. During the training stage, given a positive pair, the distance $d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{z}_i)$ is expected to be smaller than $\beta d^2_{\mathbf{P}}(\mathbf{x}^*_i, \mathbf{z}^*_i)$; given a negative pair, the distance $d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{z}_i)$ is expected to be larger than $\beta d^2_{\mathbf{P}}(\mathbf{x}^*_i, \mathbf{z}^*_i)$. This expectation has been clearly illustrated by Fig. 5 (b) and (d). The reason of jointly learning metric $\mathbf{P}$ with $\mathbf{M}$ is that it may be too rigorous to directly use the Euclidean distances in the privileged feature space as the decision threshold for metric learning, due to the significant difference between the privileged feature distribution and the original feature distribution, which can be observed in Fig. 5 (a) and (c). Therefore, the privileged distance $d^2_{\mathbf{P}}(\mathbf{x}^*_i, \mathbf{z}^*_i)$ is constantly adapted for guiding the learning of $\mathbf{M}$ during the training stage. As shown in Fig. 5 (a) and (b), benefiting from jointly learning metric $\mathbf{P}$ with $\mathbf{M}$, almost all the positive pairs on the original feature space has been distinguished from the negative pairs.

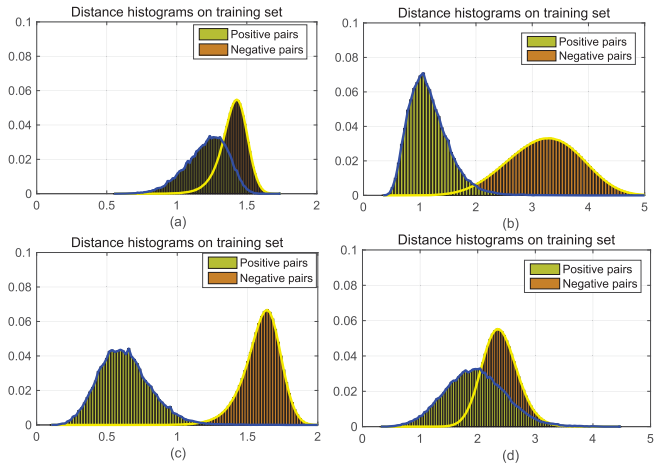

Fig. 5. Normalized distance histograms on the training set of Market-1501 before/after metric learning. (a) On the original feature space before learning **M**; (b) On the original feature space after learning **M**; (c) On the privileged feature space before learning **P**; (d) On the privileged feature space after learning **P**.

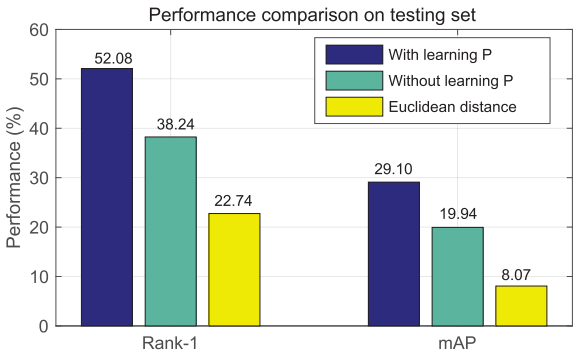

Fig. 6. Performance comparison of LDML+ with/without learning the privileged metric **P**.

Fig. 6 compares the performance of LDML+ with/without learning **P** in terms of rank-1 recognition rate and mAP score. Here, the performance of *LDML+ without learning* **P** is obtained by directly employing the Euclidean distance on privileged space as the decision threshold. We also use Euclidean distance (without training) of original features for person matching on testing set and employ its performance as a baseline in Fig. 6. As shown in Fig. 6, if directly using Euclidean distance for person matching, we can only achieve 22.74% rank-1 recognition rate and 8.07% mAP score. By learning a Mahalanobis distance on training set, the rank-1 recognition rate and mAP score have been significantly improved. By jointly learning **P** with **M**, we obtain 52.08% rank-1 recognition rate and 29.10% mAP score, which improves the performance of that without learning **P** by over 13% rank-1 recognition rate and 9% mAP score.

Here, the metric **P** bridges the original feature and privileged feature, enabling the knowledge of *teacher* to be smoothly transferred from privileged space to original space where *student* makes a decision.

VI. Conclusion

In this paper, we develop a logistic discriminant metric learning approach for cross-view person re-ID. It exploits privileged information to build a locally adaptive decision rule which can cope well with complex inter-class and intra-class variations. Besides, the proposed approach is extended to a multi-view setting, which explores the complementation of multiple different visual representations effectively. In addition, an effective iterative optimization strategy is introduced to solve the proposed method. Extensive experimental evaluations and analyses on multiple challenging datasets have demonstrated the effectiveness of the proposed work.

References

[1] L. Zheng, Y. Yang, and A. G. Hauptmann. (2016). "Person re-identification: Past, present and future." [Online]. Available: https://arxiv.org/abs/1610.02984

[2] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical Gaussian descriptor for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1363–1372.

[3] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. EEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2197–2206.

[4] A. Mignon and F. Jurie, "PCCA: A new approach for distance learning from sparse pairwise constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2666–2672.

[5] S. Liao and S. Z. Li, "Efficient psd constrained asymmetric metric learning for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3685–3693.

[6] L. Ma, X. Yang, and D. Tao, "Person re-identification over camera networks using multi-task distance metric learning," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3656–3670, Aug. 2014.

[7] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1239–1248.

[8] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2288–2295.

[9] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Learning to rank in person re-identification with metric ensembles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1846–1855.

[10] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3610–3617.

[11] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 1–16.

[12] D. Chen, Z. Yuan, B. Chen, and N. Zheng, "Similarity learning with spatial constraints for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1268–1277.

[13] Q. Wang, W. Zuo, L. Zhang, and P. Li, "Shrinkage expansion adaptive metric learning," in *Proc. ECCV*, 2014, pp. 456–471.

[14] V. Vapnik and R. Izmailov, "Learning using privileged information: Similarity control and knowledge transfer," *J. Mach. Learn. Res.*, vol. 16, pp. 2023–2049, Sep. 2015.

[15] X. Yang, M. Wang, L. Zhang, and D. Tao, "Empirical risk minimization for metric learning using privileged information," in *Proc. Int. Joint Conf. Artif. Intell.*, 2016, pp. 2266–2272.

[16] E. P. Xing, M. I. Jordan, S. Russell, and A. Y. Ng, "Distance metric learning, with application to clustering with side-information," in *Proc. NIPS*, 2002, pp. 505–512.

[17] A. Globerson and S. T. Roweis, "Metric learning by collapsing classes," in *Proc. NIPS*, 2005, pp. 451–458.

[18] M. Schultz and T. Joachims, "Learning a distance metric from relative comparisons," in *Proc. NIPS*, 2004, pp. 1–8.

[19] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proc. ICML*, 2007, pp. 209–216.

[20] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, Feb. 2009.

[21] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? Metric learning approaches for face identification," in *Proc. ICCV*, Sep./Oct. 2009, pp. 498–505.

[22] W. Bian and D. Tao, "Constrained empirical risk minimization framework for distance metric learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 8, pp. 1194–1205, Aug. 2012.

[23] Z.-J. Zha, T. Mei, M. Wang, Z. Wang, and X.-S. Hua, "Robust distance metric learning with auxiliary knowledge," in *Proc. Int. Joint Conf. Artif. Intell.*, 2009, pp. 1327–1332.

[24] L. Zheng, S. Wang, and Q. Tian, "Coupled binary embedding for large-scale image retrieval," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3368–3380, Aug. 2014.

[25] L. Zheng, Y. Yang, and Q. Tian, "SIFT meets CNN: A decade survey of instance retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2017.2709749.

[26] L. Zheng, S. Wang, and Q. Tian, "L_p-norm IDF for scalable image retrieval," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3604–3617, Aug. 2014.

[27] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 536–551.

[28] R. R. Varior, G. Wang, J. Lu, and T. Liu, "Learning invariant color features for person reidentification," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3395–3410, Jul. 2016.

[29] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 262–275.

[30] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3586–3593.

[31] G. Lisanti, I. Masi, A. D. Bagdanov, and A. Del Bimbo, "Person re-identification by iterative re-weighted sparse ranking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 8, pp. 1629–1642, Aug. 2015.

[32] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3318–3325.

[33] X. Yang, M. Wang, R. Hong, Q. Tian, and Y. Rui, "Enhancing person re-identification in a self-trained subspace," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 13, no. 3, pp. 27:1–27:23, 2017.

[34] P. M. Roth, M. Hirzer, M. Koestinger, C. Beleznai, and H. Bischof, *Mahalanobis Distance Learning for Person Re-Identification*. London, U.K.: Springer, 2014, pp. 247–267.

[35] L. An, S. Yang, and B. Bhanu, "Person re-identification by robust canonical correlation analysis," *IEEE Signal Process. Lett.*, vol. 22, no. 8, pp. 1103–1107, Aug. 2015.

[36] Y.-C. Chen, W.-S. Zheng, J.-H. Lai, and P. Yuen, "An asymmetric distance model for cross-view feature mapping in person reidentification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 8, pp. 1661–1675, Aug. 2016.

[37] W.-S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 653–668, Mar. 2013.

[38] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 144–151.

[39] V. Vapnik and A. Vashist, "A new learning paradigm: Learning using privileged information," *Neural Netw.*, vol. 22, nos. 5–6, pp. 544–557, 2009.

[40] J. T. Zhou, X. Xu, S. J. Pan, I. W. Tsang, Z. Qin, and R. S. M. Goh, "Transfer hashing with privileged information," in *Proc. Int. Joint Conf. Artif. Intell.*, 2016, pp. 2414–2420.

[41] L. Niu, W. Li, and D. Xu, "Exploiting privileged information from Web data for action and event recognition," *Int. J. Comput. Vis.*, vol. 118, no. 2, pp. 130–150, 2016.

[42] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, "Information bottleneck learning using privileged information for visual recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1496–1505.

[43] V. Sharmanska, N. Quadrianto, and C. H. Lampert, "Learning to rank using privileged information," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 825–832.

[44] W. Li, L. Niu, and D. Xu, "Exploiting privileged information from Web data for image categorization," in *Proc. ECCV*, 2014, pp. 437–452.

[45] J. Feyereisl, S. Kwak, J. Son, and B. Han, "Object localization based on structural SVM using privileged information," in *Proc. NIPS*, 2014, pp. 208–216.

[46] Y. Yan, F. Nie, W. Li, C. Gao, Y. Yang, and D. Xu, "Image classification by cross-media active learning with privileged information," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2494–2502, Dec. 2016.

[47] S. Fouad, P. Tino, S. Raychaudhury, and P. Schneider, "Incorporating privileged information through metric learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 7, pp. 1086–1098, Jul. 2013.

[48] X. Xu, W. Li, and D. Xu, "Distance metric learning using privileged information for face verification and person re-identification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 12, pp. 3150–3162, Dec. 2015.

[49] J. Hu, J. Lu, and Y.-P. Tan, "Discriminative deep metric learning for face verification in the wild," in *Proc. CVPR*, Jun. 2014, pp. 1875–1882.

[50] M. Wang, X.-S. Hua, R. Hong, J. Tang, G.-J. Qi, and Y. Song, "Unified video annotation via multigraph learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 5, pp. 733–746, May 2009.

[51] T. Xia, D. Tao, T. Mei, and Y. Zhang, "Multiview spectral embedding," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 6, pp. 1438–1446, Dec. 2010.

[52] J. Yu, M. Wang, and D. Tao, "Semisupervised multiview distance metric learning for cartoon synthesis," *IEEE Trans. Image Process.*, vol. 21, no. 11, pp. 4636–4648, Nov. 2012.

[53] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *Proc. Asian Conf. Comput. Vis.*, 2012, pp. 31–44.

[54] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1116–1124.

[55] H. Cho, P. E. Rybski, A. Bar-Hillel, and W. Zhang, "Real-time pedestrian detection with deformable part models," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2012, pp. 1035–1042.

[56] T. Matsukawa and E. Suzuki, "Person re-identification using cnn features learned from combination of attributes," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 2428–2433.

[57] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang. (2017). "Improving person re-identification by attribute and identity learning." [Online]. Available: https://arxiv.org/abs/1703.07220

[58] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1335–1344.

[59] Y.-C. Chen, W.-S. Zheng, and J. Lai, "Mirror representation for modeling view-specific transform in person re-identification," in *Proc. Int. Joint Conf. Artif. Intell.*, 2015, pp. 3402–3408.

[60] Y. Zhang, B. Li, H. Lu, A. Irie, and X. Ruan, "Sample-specific SVM learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1278–1287.

[61] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao, "Multi-task learning with low rank attribute embedding for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3739–3747.

[62] Z. Shi, T. M. Hospedales, and T. Xiang, "Transferring a semantic representation for person re-identification and search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4184–4193.

[63] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proc. CVPR*, Jun. 2016, pp. 1249–1258.

[64] S.-Z. Chen, C.-C. Guo, and J.-H. Lai, "Deep ranking for person re-identification via joint representation learning," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2353–2367, May 2016.

[65] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang, "Similarity learning on an explicit polynomial kernel feature map for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1565–1573.

[66] Y. Shen, W. Lin, J. Yan, M. Xu, J. Wu, and J. Wang, "Person re-identification with correspondence structure learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3200–3208.

[67] S. Li, M. Shao, and Y. Fu, "Cross-view projective dictionary learning for person re-identification," in *Proc. Int. Joint Conf. Artif. Intell.*, 2015, pp. 2155–2161.

[68] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3908–3916.

[69] S. Wu, Y.-C. Chen, X. Li, A.-C. Wu, J.-J. You, and W.-S. Zheng, "An enhanced deep feature representation for person re-identification," in *Proc. WACV*, Mar. 2016, pp. 1–8.

[70] Y. Yang, Z. Lei, S. Zhang, H. Shi, and S. Z. Li, "Metric embedded discriminative vocabulary learning for high-level person representation," in *Proc. Int. Joint Conf. Artif. Intell.*, 2016, pp. 3648–3654.

[71] C. Jose and F. Fleuret, "Scalable metric learning via weighted approximate rank component analysis," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 875–890.
[72] N. Martinel, A. Das, C. Micheloni, and A. K. Roy-Chowdhury, "Temporal model adaptation for person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 858–877.
[73] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3492–3506, Jul. 2017.
[74] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *Proc. ECCV*, 2016, pp. 475–491.
[75] J. Liu *et al.*, "Multi-scale triplet cnn for person re-identification," in *Proc. ACM Multimedia Conf.*, 2016, pp. 192–196.
[76] E. Ustinova and V. Lempitsky, "Learning deep embeddings with histogram loss," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4170–4178.
[77] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A siamese long short-term memory architecture for human re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 135–153.
[78] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 791–808.
[79] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1318–1327.

Xun Yang is currently pursuing the Ph.D. degree with the School of Computer and Information Engineering, Hefei University of Technology, China. He was a Visiting Research Student with the Centre for Quantum Computation and Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology Sydney, from 2015 to 2017. His research interests include person re-identification, multimedia content analysis, computer vision, and pattern recognition.

Meng Wang (SM'17) received the B.E. and Ph.D. degrees in the special class for the gifted young from the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, China, in 2003 and 2008, respectively. He is currently a Professor with the Hefei University of Technology, China. He has authored over 200 book chapters, journal and conference papers in his research areas. His current research interests include multimedia content analysis, computer vision, and pattern recognition. He was a recipient of the ACM SIGMM Rising Star Award 2014. He is an Associate Editor of the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.

Dacheng Tao (F'15) is currently a Professor of computer science with the School of Information Technologies, The University of Sydney. He mainly applies statistics and mathematics to artificial intelligence and data science. His research interests spread across computer vision, data science, image processing, machine learning, and video surveillance. His research results have expounded in one monograph and over 200 publications at prestigious journals and prominent conferences, such as the IEEE T-PAMI, T-NNLS, T-IP, JMLR, IJCV, NIPS, ICML, CVPR, ICCV, ECCV, AISTATS, ICDM, and ACM SIGKDD, with several best paper awards, such as the Best Theory/Algorithm Paper Runner Up Award at the IEEE ICDM07, the Best Student Paper Award at the IEEE ICDM13, and the 2014 ICDM 10-Year Highest Impact Paper Award. He is a fellow of OSA, IAPR, and SPIE. He received the 2015 Australian Scopus-Eureka Prize, the 2015 ACS Gold Disruptor Award, and the 2015 UTS Vice-Chancellors Medal for Exceptional Research.