Easy Identification from Better Constraints: Multi-Shot Person Re-Identification from Reference Constraints

Jiahuan Zhou[1], Bing Su[2], Ying Wu[1]
[1]Electrical Engineering and Computer Science, Northwestern University, US
[2]Institute of Software, Chinese Academy of Science, China
{jzt011, yingwu}@eecs.northwestern.edu, subingats@gmail.com

Abstract

Multi-shot person re-identification (MsP-RID) utilizes multiple images from the same person to facilitate identification. Considering the fact that motion information may not be discriminative nor reliable enough for MsP-RID, this paper is focused on handling the large variations in the visual appearances through learning discriminative visual metrics for identification. Existing metric learning-based methods usually exploit pair-wise or triple-wise similarity constraints, that generally demands intensive optimization in metric learning, or leads to degraded performances by using sub-optimal solutions. In addition, as the training data are significantly imbalanced, the learning can be largely dominated by the negative pairs and thus produces unstable and non-discriminative results. In this paper, we propose a novel type of similarity constraint. It assigns the sample points to a set of ***reference points*** *to produce a linear number of* ***reference constraints****. Several optimal transport-based schemes for reference constraint generation are proposed and studied. Based on those constraints, by utilizing a typical regressive metric learning model, the closed-form solution of the learned metric can be easily obtained. Extensive experiments and comparative studies on several public MsP-RID benchmarks have validated the effectiveness of our method and its significant superiority over the state-of-the-art MsP-RID methods in terms of both identification accuracy and running speed.*

1. Introduction

Person re-identification (P-RID) is a critical yet very challenging task in video surveillance [29]. It generally evaluates 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. Research efforts have been devoted to single-shot person re-identification (SsP-RID) [16, 18, 17, 36, 39, 32] 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 (MsP-RID) [10, 27, 31, 35, 22] that utilizes multiple images for the same person from the same camera-view, expecting to improve the performance.

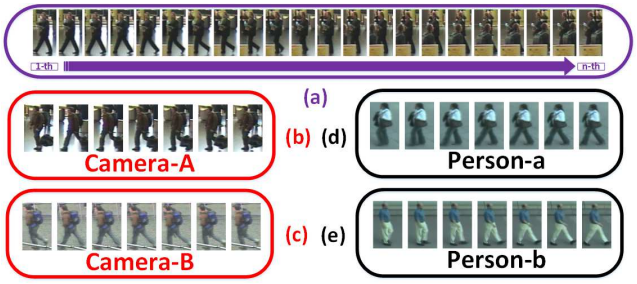

Figure 1. **(a)** The background occlusion completely conceals the motion information on the legs; **(b)** & **(c)** Even for the same person, the walking behavior can be very different; **(d)** & **(e)** For different persons, they may share very similar walking patterns.

One common approach in MsP-RID [31, 10, 35, 27] 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 [19] (Fig. 1(a)). Secondly, motion patterns may not be discriminative enough for identification since different persons may walk in the same walking pattern [31] (Fig. 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. 1(b) and (c), a person is walking with luggage cap-

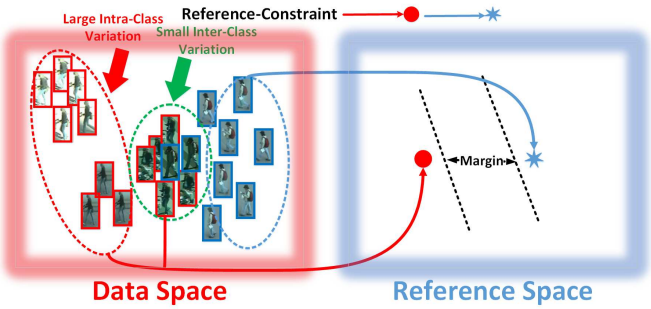

Figure 2. The proposed **reference constraint** correlates the original indiscriminative same class data to the common discriminative reference points (note: there can be multiple reference points to handle the multiple-mode distribution of same class data).

tured 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.

Another approach [13, 15] 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 [17, 33, 19], but there is still room for performance improvement. Recent methods have been focused on learning discriminative visual metrics to facilitate identification. Many such methods [17, 40, 31, 18] 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 [18] or triple-wise [40] 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 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 [26]. Moreover, in practice, the positive and negative samples are significantly imbalanced. As the learning can be largely dominated by the negative pairs [18], it leads to unstable and non-discriminative learning results.

To overcome these difficulties, in this paper, 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. 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 [6, 25] 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 [6] 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 [4], the P-RID 2011 [12], the iLIDS-VID [31] and the Market-1501 [39] datasets.

2. Easy Identification from Reference Constraints

2.1. 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. 2.2). 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.

2.2. Automatic Reference Constraint Generation

In this section, we will show how to automatically generate the reference constraints under a general optimal transport (OT) framework [30]. 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

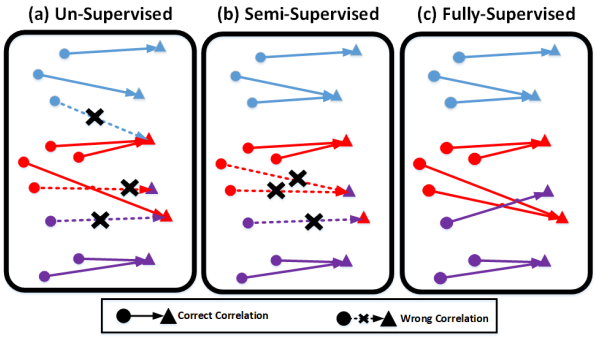

Figure 3. **(a)** is the result of an unsupervised OT method [8]; **(b)** is a semi-supervised OT method [7]; **(c)** is our proposed supervised OT method with cross-bin cost function Eqn. 2. Different colors (Red, Blue, Purple) represent different classes, and different shapes (●, ▲) mean different distributions.

be modeled as an optimal transport procedure [7, 24, 8]:

$$\arg\min_{\mathcal{T}} \quad \langle \mathcal{T}, \mathbf{C} \rangle_{\mathcal{F}} + \mathcal{G}(\mathcal{T}) \tag{1}$$

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 $\mathcal{G}(\mathcal{T})$.

2.2.1 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 person across a significant viewpoint change is difficult because of the visually spatial misalignment [26]. An intuitive idea to generate $\mathbf{R}$ 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 unsupervised since no class label information is used. Hence the correlations between two distributions are completely unconstrained (Fig. 3(a)) which will be problematic in the P-RID problem, where the identity label is given for each sample. In [7], 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. 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:

$$\mathbf{C}_A(i,j) = \|x_i^{\mathcal{A}} - x_j^{\mathcal{B}}\|_2 I(l_i^{\mathcal{A}} = l_j^{\mathcal{B}}) + \infty \cdot I(l_i^{\mathcal{A}} \neq l_j^{\mathcal{B}}) \tag{2}$$

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. 3:

$$\begin{aligned} \arg\min_{\mathcal{T}_A} \quad & \langle \mathcal{T}_A, \mathbf{C}_A \rangle_{\mathcal{F}} + \frac{1}{\lambda} \sum_{i,j} \mathcal{T}_A(i,j) \log \mathcal{T}_A(i,j) \\ & + \eta \sum_j \sum_c \|\mathcal{T}_A(l_i^{\mathcal{A}} = c, j)\|_q^p \end{aligned} \tag{3}$$

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 ([7, 24, 8]) as illustrated in Fig. 3(a) and (b) can be avoided, and thus a clean transport flow can be achieved (Fig. 3(c)). The objective Eqn. 3 can be efficiently solved via the alternation between the Sinkhorn-Knopp algorithm[8] and the Majoration-Minimization strategy[7]. 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 $\mathbf{R} = \mathbf{X}\mathcal{T}_A$

2.2.2 R from Class-based Discriminative Space

An efficient and straightforward idea is to explicitly determine the $\mathbf{R}$ 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 $\mathbf{X}$ to $\mathbf{R}$ can be modeled as optimizing:

$$\arg\min_{\mathcal{T}_C} \quad \langle \mathcal{T}_C, \mathbf{C}_C \rangle_{\mathcal{F}} \tag{4}$$

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

$$\mathcal{T}_C(x_i, l_i) = u_{\#l_i} \tag{5}$$

that all the samples in $\mathbf{X}$ from the same class $\#l_i$ will be transported into one single point $u_{\#l_i}$ in $\mathbf{R}$ 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 $\mathbf{X}$, $\mathcal{T}_C$ is equivalent to learn a lower-dimensional embedding where the samples drawn from different classes become much more discriminative.

Optimality of R from CDS: The similar idea of our CDS is shared by many existing works like the well-known

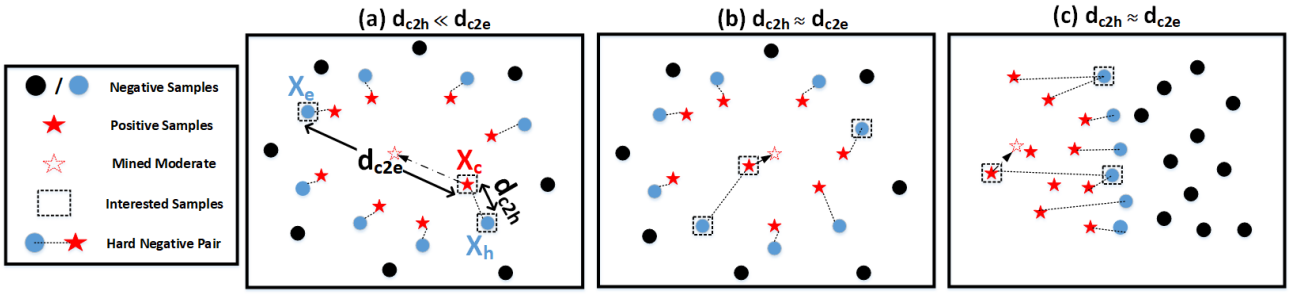

Figure 4. Moderate positive mining for a local unimodal data distribution.

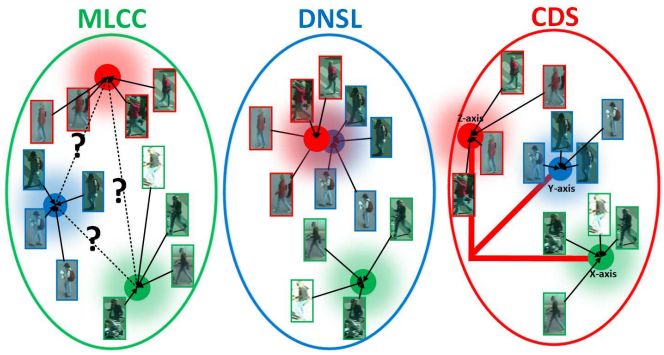

Figure 5. The comparison of three related algorithms: MLCC [11], DNSL [36] and our CDS method.

metric learning algorithm MLCC [11] and a recently state-of-the-art P-RID algorithm DNSL [36]. As illustrated by Fig. 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.

2.2.3 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 [2]. We propose to mine a set of "moderate" representations from $\mathbf{X}$ 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 $\mathbf{R}$ 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:

$$r^c = \frac{1}{2}(1 + \frac{d_{c2h}}{d_{c2e}})x^c + \frac{1}{2}(1 - \frac{d_{c2h}}{d_{c2e}})x_e^{\bar{c}} \tag{6}$$

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. 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. 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 $\mathbf{X}$ to $\mathbf{R}$ is:

$$\mathcal{T}_M(x_i, l_i) = r^{\#l_i} \tag{7}$$

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 [5] to $\mathcal{X}_c$ to divide $\mathcal{X}_c$ into several sub-class clusters, thus each cluster 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. Metric Learning from R via Regression

Once the reference set $\mathbf{R}$ 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 [6, 25]:

$$\mathbf{L}^* = \min_{\mathbf{L}} \frac{1}{n}\|\mathbf{X}\mathbf{L} - \mathbf{R}\|_{\mathcal{F}}^2 + \lambda\|\mathbf{L}\|_{\mathcal{F}}^2 \tag{8}$$

where the λ is a weighting parameter to balance the two terms. The closed-form solution to objective Eqn. 8 can be derived.

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

Theorem 1 *The optimal solution of objective Eqn. 8 has a closed form, as shown in the following two equivalent solutions:*

$$\boldsymbol{L} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda n\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{R} \tag{9}$$

$$\boldsymbol{L} = \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T + \lambda n\boldsymbol{I})^{-1}\boldsymbol{R} \tag{10}$$

Proof 1 *Compute the derivative of Eqn. 8:*

$$\frac{\partial f(\boldsymbol{L},\boldsymbol{X},\boldsymbol{R})}{\partial \boldsymbol{L}} = 2\left(\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}\right)\boldsymbol{L} - \frac{2}{n}\boldsymbol{X}^T\boldsymbol{R} \tag{11}$$

By setting this derivative to zero we can obtain:

$$\boldsymbol{L} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda n\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{R} \tag{12}$$

Theorem 2 *The optimal solutions Eqn. 9 and Eqn. 10 of objective Eqn. 8 are exactly equivalent.*

Proof 2 *For Eqn. 9, we perform Taylor expansion to the $(\boldsymbol{X}^T\boldsymbol{X} + \lambda n\boldsymbol{I})^{-1}\boldsymbol{X}^T$ part:*

$$\begin{aligned}
&(\boldsymbol{X}^T\boldsymbol{X} + \lambda n\boldsymbol{I})^{-1}\boldsymbol{X}^T = \frac{1}{\lambda n}(\boldsymbol{I} + \frac{1}{\lambda n}\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \\
&= \frac{1}{\lambda n}\sum_1^\infty(-1)^n(\frac{1}{\lambda n})^n(\boldsymbol{X}^T\boldsymbol{X})^n\boldsymbol{X}^T \\
&= \frac{\boldsymbol{X}^T}{\lambda n}\sum_1^\infty(-1)^n(\frac{1}{\lambda n})^n(\boldsymbol{X}\boldsymbol{X}^T)^{n-1}\boldsymbol{X}\boldsymbol{X}^T \\
&= \frac{\boldsymbol{X}^T}{\lambda n}\sum_1^\infty(-1)^n(\frac{1}{\lambda n})^n(\boldsymbol{X}\boldsymbol{X}^T)^n \\
&= \frac{\boldsymbol{X}^T}{\lambda n}(\boldsymbol{I} + \frac{1}{\lambda n}\boldsymbol{X}\boldsymbol{X}^T)^{-1} \\
&= \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T + \lambda n\boldsymbol{I})^{-1}
\end{aligned} \tag{13}$$

Therefore Eqn. 13 proves that Eqn. 9 and Eqn. 10 are exactly the same solution for the proposed objective Eqn. 8.

From Eqn. 9, we obtain the Mahalanobis metric $\mathbf{M}$:

$$\begin{aligned}
&\mathbf{M} = \mathbf{L}\mathbf{L}^T \\
&= (\mathbf{X}^T\mathbf{X} + \lambda n\mathbf{I})^{-1}\mathbf{X}^T\mathbf{R}\mathbf{R}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda n\mathbf{I})^{-1}
\end{aligned} \tag{14}$$

As we can see from Eqn. 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.

2.3.1 Non-Linear Extension by Kernelization

The linear model in Sec. 2.3 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. 10. Therefore the kernelized Mahalanobis metric $\mathbf{M}_K$ is written as:

$$\mathbf{M}_K = \phi(\mathbf{X})^T(K_{\mathbf{X}} + \lambda n\mathbf{I})^{-1}\mathbf{R}\mathbf{R}^T(K_{\mathbf{X}} + \lambda n\mathbf{I})^{-1}\phi(\mathbf{X}) \tag{15}$$

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

$$\begin{aligned}
&d^2_{\mathbf{M}_K}(x, x') = \\
&\phi(x)^T\mathbf{M}_K\phi(x) + \phi(x')^T\mathbf{M}_K\phi(x') - 2\phi(x)^T\mathbf{M}_K\phi(x')
\end{aligned}$$

that each term can be written as:

$$\begin{aligned}
&\phi(x)^T\mathbf{M}_K\phi(x) = \\
&K_{\mathbf{X}}(x)^T(K_{\mathbf{X}} + \lambda n\mathbf{I})^{-1}\mathbf{R}\mathbf{R}^T(K_{\mathbf{X}} + \lambda n\mathbf{I})^{-1}K_{\mathbf{X}}(x)
\end{aligned}$$

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.

2.3.2 Generalization Ability Analysis

For our objective Eqn. 8, the empirical error risk is $\mathcal{E}(\mathbf{L}, \mathcal{S}_r) = \frac{1}{n}\|\mathbf{X}\mathbf{L} - \mathbf{R}\|^2_{\mathcal{F}}$, which is to measure how close the projected samples $\mathbf{X}\mathbf{L}$ 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 [1].

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

$$\|\mathcal{E}(\boldsymbol{L}, \mathcal{D}_{\boldsymbol{R}}) - \mathcal{E}(\boldsymbol{L}, \mathcal{S}_r)\| \le \left(1 + \left(2n + \frac{\lambda n}{8B_x^2}\right)\sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right)\beta \tag{16}$$

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 $O(1/\sqrt{n})$.

3. Experiments

3.1. Experiment Details

Dataset. To evaluate our proposed method, we conduct thorough experiments on four widely-used multi-shot benchmarks: the CAVIAR [4], the PRID 2011 [12], the iLIDS-VID [31] and the Market-1501 [39] 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 [39] 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 [17] 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 [32, 3, 31, 35]. 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. 8 is chosen as $\lambda = 0.01$ for all the experiments, which empirically produces both small training errors and stable solutions.

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

State-of-the-art. For the comparison experiments, we select three state-of-the-art metric learners: MLAPG [18], XQDA [17], DNSL [36] 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.

3.2. 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. 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, the computational complexity of learning the metric $\mathbf{M}$ is quadratic to the training sample number n or data dimension d. Table. 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 [36] also has a closed-form solution, but it requires many times of SVD operation for the kernelized data matrix, which is indeed time-consuming.

3.3. 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

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

Table 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 3. The average empirical training error on CAVIAR.

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. 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.

3.4. 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 [3, 20, 32]. Due to the success of SsP-RID on CAVIAR, we would like to also report the state-of-the-art single-shot results, including SSCDL [20], MFA-χ^2 [32], EPKFM [3], PCCA-χ^2_{RBF} [32], LADF [16] and LFDA [23]. It can be observed from Table. 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. 4. As we can see, all of our proposed reference-based methods consistently outperform 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. 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. 6. The baseline [39] 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 [28] 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%.

4. Conclusion

In this paper, 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.

Scenario	Method	PRID 2011				iLIDS-VID				Reference
		R=1	R=5	R=10	R=20	R=1	R=5	R=10	R=20	
Multi-Shot	TDL	30.20	59.10	74.00	88.40	9.81	27.52	46.10	62.19	CVPR2016[35]
	MLAPG(lomo)	45.60	58.20	63.80	69.80	30.54	45.58	53.02	60.78	ICCV2015[18]
	XQDA(lomo)	47.50	60.20	66.20	72.00	30.66	44.48	51.84	59.53	CVPR2015[17]
	DNSL(lomo)	51.00	63.40	68.60	74.10	24.44	34.11	39.68	46.85	CVPR2016[36]
	DVDL	40.60	69.70	77.80	85.60	25.90	48.20	57.30	68.90	ICCV2015[13]
	Salience	25.80	43.60	52.60	62.00	10.20	24.80	35.50	52.90	CVPR2013[37]
	KISSME	28.54	59.78	72.13	83.26	10.67	28.33	39.80	57.00	CVPR2012[14]
	LFDA	26.40	56.07	69.89	81.12	7.80	23.93	36.47	50.80	CVPR2013[23]
	LADF	8.20	20.45	29.89	42.25	4.33	14.00	21.20	32.13	CVPR2013[16]
	LDA	27.64	58.09	69.66	82.47	10.27	27.40	39.80	55.27	AP2013[9]
Video-based	SMP	80.90	95.60	98.80	99.40	41.70	66.30	74.10	80.70	ICCV2017[21]
	DGM+IDE	56.40	81.30	88.00	96.40	36.20	62.80	73.60	82.70	ICCV2017[34]
	CNN+KISS	69.90	90.60	-	98.20	48.80	75.60	-	92.60	ECCV2016[38]
	TDL	56.74	80.00	87.64	93.59	56.33	87.60	95.60	98.27	CVPR2016[35]
	Co&LBP+DVR	37.60	63.90	75.30	88.30	34.50	56.70	67.50	77.50	ECCV2014[31]
	KISSME	34.38	61.68	72.13	81.01	36.53	67.80	78.80	87.07	CVPR2012[14]
	LFDA	43.70	72.80	81.69	90.89	32.93	68.47	82.20	92.60	CVPR2013[23]
	LADF	47.30	75.50	82.69	91.12	39.00	76.80	89.00	96.80	CVPR2013[16]
	LDA	15.84	41.46	55.51	70.67	42.06	79.13	89.40	94.47	AP2013[9]
Linear	$\mathcal{T}_C$-**L**	**70.10**	**79.10**	**83.30**	**87.10**	**44.67**	**57.33**	**63.33**	**68.67**	Proposed
	$\mathcal{T}_M$-**L**	**64.80**	**77.00**	**80.20**	**84.30**	**38.67**	**56.67**	**61.67**	**70.67**	Proposed
	$\mathcal{T}_A$-**L**	**70.40**	**80.90**	**85.60**	**88.40**	**42.67**	**58.67**	**63.33**	**72.07**	Proposed
Kernel	$\mathcal{T}_C$-**K**	**66.90**	**77.10**	**80.80**	**84.60**	**37.33**	**47.73**	**54.53**	**60.67**	Proposed
	$\mathcal{T}_M$-**K**	**65.10**	**77.30**	**78.70**	**85.30**	**39.33**	**56.00**	**59.33**	**65.74**	Proposed
	$\mathcal{T}_A$-**K**	**70.90**	**78.70**	**82.70**	**87.30**	**42.00**	**52.67**	**60.03**	**66.67**	Proposed

Table 4. Comparison results on PRID 2011 and iLIDS-VID under the **multi-shot** and **video-based** matching settings.

Method	R=1	R=5	R=10	R=20
MLAPG(lomo)[18]	50.00	71.85	84.25	93.11
XQDA(lomo)[17]	51.18	75.59	90.33	96.86
DNSL(lomo)[36]	53.54	77.17	86.61	94.69
SSCDL-S[20]	49.10	80.20	93.50	97.90
MLAPG(lomo)-S[18]	40.60	71.70	83.30	95.70
XQDA(lomo)-S[17]	42.20	69.90	82.50	95.50
DNSL(lomo)-S[36]	47.60	75.66	87.37	96.20
MFA-χ^2-S[32]	40.20	70.20	83.90	95.10
EPKFM-S[3]	40.10	65.60	78.00	90.50
PCCA-χ^2_{RBF}-S[32]	33.20	65.90	81.90	95.20
LFDA-S[23]	32.00	56.30	70.70	87.40
LADF-S[16]	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 5. Comparison results on CAVIAR under the **multi-shot** matching setting. **'-S'** means the single-shot result.

Method	Sing-Q R=1	Multi-Q R=1	Reference
Baseline	35.84	44.36	ICCV15[39]
MLAPG(lomo)	38.80	61.33	ICCV15[18]
XQDA(lomo)	44.80	55.82	CVPR15[17]
DNSL(lomo)	51.73	57.70	CVPR16[36]
KISSME(lomo)	40.50	N/A	ICCV15[39]
MFA-χ^2(lomo)	45.67	N/A	ECCV14[32]
kLFDA(lomo)	51.37	52.67	ECCV14[32]
Hist-Loss	59.47	N/A	NIPS16[28]
$\mathcal{T}_C$-**L**	**57.73**	**68.27**	Proposed
$\mathcal{T}_M$-**L**	**54.67**	**64.53**	Proposed
$\mathcal{T}_A$-**L**	**51.07**	**72.40**	Proposed
$\mathcal{T}_C$-**K**	**63.20**	**73.87**	Proposed
$\mathcal{T}_M$-**K**	**60.93**	**70.40**	Proposed
$\mathcal{T}_A$-**K**	**56.03**	**68.93**	Proposed

Table 6. Comparison results on Market-1501.

Acknowledgements

This work was supported in part by National Science Foundation grant IIS-1217302, IIS-1619078, the Army Research Ofice ARO W911NF-16-1-0138 and the National Natural Science Foundation of China grant No.61603373.

References

[1] O. Bousquet and A. Elisseeff. Stability and generalization. *JMLR*, 2002. 5

[2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *JAIR*, 2002. 4

[3] 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 *CVPR*, 2015. 6, 7, 8

[4] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *BMVC*, 2011. 2, 6

[5] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE TPAMI*, 2002. 4

[6] C. Cortes, M. Mohri, and J. Weston. A general regression framework for learning string-to-string mappings. *Predicting Structured Data*, 2007. 2, 4

[7] N. Courty, R. Flamary, and D. Tuia. Domain adaptation with regularized optimal transport. In *ECML PKDD*, 2014. 3

[8] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NIPS*, 2013. 3

[9] K. Fukunaga. *Introduction to statistical pattern recognition*. 2013. 8

[10] N. Gheissari, T. B. Sebastian, and R. Hartley. Person reidentification using spatio temporal appearance. In *CVPR*, 2006. 1

[11] A. Globerson and S. Roweis. Metric learning by collapsing classes. In *NIPS*, 2005. 4

[12] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *SCIA*, 2011. 2, 6

[13] S. Karanam, Y. Li, and R. J. Radke. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *ICCV*, 2015. 2, 8

[14] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012. 8

[15] X. Li, W.-S. Zheng, X. Wang, T. Xiang, and S. Gong. Multi-scale learning for low-resolution person re-identification. In *ICCV*, 2015. 2

[16] Z. Li, S. Chang, F. Liang, T. Huang, L. Cao, and J. Smith. Learning locally-adaptive decision functions for person verification. In *CVPR*, 2013. 1, 7, 8

[17] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015. 1, 2, 6, 8

[18] S. Liao and S. Z. Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *ICCV*, 2015. 1, 2, 6, 8

[19] K. Liu, B. Ma, W. Zhang, and R. Huang. A spatio-temporal appearance representation for viceo-based pedestrian re-identification. In *ICCV*, 2015. 1, 2

[20] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu. Semi-supervised coupled dictionary learning for person re-identification. In *CVPR*, 2014. 7, 8

[21] Z. Liu, D. Wang, and H. Lu. Stepwise metric promotion for unsupervised video person re-identification. In *ICCV*, 2017. 8

[22] N. McLaughlin, J. Martinez del Rincon, and P. Miller. Recurrent convolutional network for video-based person re-identification. In *CVPR*, 2016. 1

[23] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *CVPR*, 2013. 7, 8

[24] M. Perrot, N. Courty, R. Flamary, and A. Habrard. Mapping estimation for discrete optimal transport. In *NIPS*, 2016. 3

[25] M. Perrot and A. Habrard. Regressive virtual metric learning. In *NIPS*, 2015. 2, 4

[26] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, and S. Z. Li. Embedding deep metric for person re-identification: A study against large variations. In *ECCV*, 2016. 2, 3

[27] D. Simonnet, M. Lewandowski, S. A. Velastin, J. Orwell, and E. Turkbeyler. Re-identification of pedestrians in crowds using dynamic time warping. In *ECCV*, 2012. 1

[28] E. Ustinova and V. Lempitsky. Learning deep embeddings with histogram loss. In *NIPS*, 2016. 7, 8

[29] R. Vezzani, D. Baltieri, and R. Cucchiara. People reidentification in surveillance and forensics: A survey. *ACM CSUR*, 2013. 1

[30] C. Villani. *Optimal transport: old and new*. 2008. 2

[31] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *ECCV*. 2014. 1, 2, 6, 8

[32] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*. 2014. 1, 6, 7, 8

[33] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li. Salient color names for person re-identification. In *ECCV*. 2014. 2

[34] M. Ye, A. J. Ma, L. Zheng, J. Li, and P. C. Yuen. Dynamic label graph matching for unsupervised video re-identification. *ICCV*, 2017. 8

[35] J. You, A. Wu, X. Li, and W.-S. Zheng. Top-push video-based person re-identification. In *CVPR*, 2016. 1, 6, 8

[36] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *CVPR*, 2016. 1, 4, 6, 8

[37] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013. 8

[38] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, 2016. 8

[39] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 1, 2, 6, 7, 8

[40] W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR*, 2011. 2