Tensor Multi-task Learning for Person Re-identification

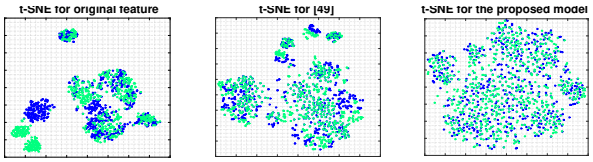
Zhizhong Zhang*, Yuan Xie*, *Member, IEEE,* Wensheng Zhang, Yongqiang Tang, Qi Tian, *Fellow, IEEE,*

***Abstract*—This paper presents a tensor multi-task model for person re-identification (Re-ID). Due to discrepancy among cameras, our approach regards Re-ID from multiple cameras as different but related classification tasks, each task corresponding to a specific camera. In each task, we distinguish the person identity as a one-vs-all linear classification problem, where one classifier is associated with a specific person. By constructing all classifiers into a task-specific projection matrix, the proposed method could utilize all the matrices to form a tensor structure, and jointly train all the tasks in a uniform tensor space. In this space, by assuming the features of the same person under different cameras are generated from a latent subspace, and different identities under the same perspective share similar patterns, the high-order correlations, not only across different tasks but also within a certain task, can be captured by utilizing a new type of low-rank tensor constraint. Therefore, the learned classifiers transform the original feature vector into the latent space, where feature distributions across cameras can be well-aligned. Moreover, this model can be incorporated into multiple visual features to boost the performance, and easily extended to the unsupervised setting. Extensive experiments and comparisons with recent Re-ID methods manifest the competitive performance of our method.**

***Index Terms*—Person Re-identification, Multi-task learning, Tensor optimization.**

I. Introduction

CAMERA networks are now ubiquitous in public infrastructure for surveillance. As one of the most fundamental tasks for surveillance systems, person re-identification (Re-ID) [1], [31], [42], [49] has received considerable academic attention recently, owing to its tremendous potential in security applications such as people tracking [25] and behavior analysis [2]. The essential objective of person Re-ID is to identify target person from a large amount of visual data, where the images of persons are captured from non-overlapping camera views. This makes person images be under various viewpoints, different illuminations and human poses, leading to a large intra-personal variation. Furthermore, the surveillance data captured from the cameras often has a low resolution, which also provides additional difficulties to match individuals.



(a) (b) (c)

Figure 1. The visual comparison of feature distributions alignment by using t-SNE. (a) feature distribution in the original feature space; (b) and (c) feature distributions in the projected common space produced by [44] and ours, respectively. Blue and green points denote the features captured by Camera 1 and Camera 2, respectively. The overlap ratio of the two feature distributions indicates the degree of alignment (the more, the better).

To deal with these issues, several methods have been investigated for decades. The recent success of person Re-ID usually stems from the discriminative metric learning process, which is commonly achieved by establishing a reliable view-generic or view-specific matching model upon the training data. Specifically, the view-generic methods [4], [31] utilize a uniform model for all the cameras to distinguish different people, while the view-specific model [13], [44] explicitly captures camera-wise discrepancy. Since feature distributions across different cameras can not be well-aligned by the view-generic model (see Fig. 1(a) for more details), it is expected that the view-specific model might be much more preferable.

A major problem for the view-specific model is the insufficiency of training samples, particularly in the condition of dramatic view-specific changes. Due to the difficulties of collecting matched pairs, previous methods usually resort to matrix regularization [13], [28], [32], [44], learning asymmetric distance to describe such view-wise discrepancy. One key assumption in this sort is that all view-specific projection matrices are correlated via a certain structure, which, for example, includes low-rank regularizer [44], Bregman discrepancy [32] [28]. Nevertheless, one obvious drawback of these methods is that they only focus on pairwise asymmetric metric learning, where the knowledge is merely shared across the cameras, *ignoring the high-order correlation among cameras and persons*. Furthermore, the complexity increases in proportion to the scale of the camera network.

In this paper, by considering each camera as one task, we propose a new multi-task learning model via tensor regularization for person Re-ID. Concretely, in each task (camera)[1], we recognize person identities via a standard one-vs-all multi-class classification formulation, where one classifier is used to

Z. Zhang, Y. Tang, and W. Zhang are with the Research Center of Precision Sensing and Control, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China, and also with the University of Chinese Academy of Sciences, Beijing, 101408, China. E-mail: {zhangzhizhong2014, tangyongqiang2014}@ia.ac.cn, zhangwenshengia@hotmail.com

Y. Xie is with the School of Computer Science and Technology, East China Normal University, Shanghai, China; E-mail: yxie@cs.ecnu.edu.cn

Q. Tian is with the Huawei Noah's Ark Lab, on leave from the Department of Computer Science, University of Texas at San Antonio, San Antonio, TX 78249-1604 USA. E-mail: tian.qi1@huawei.com and qitian@cs.utsa.edu

[1]We do not distinguish between cameras and tasks throughout the paper.

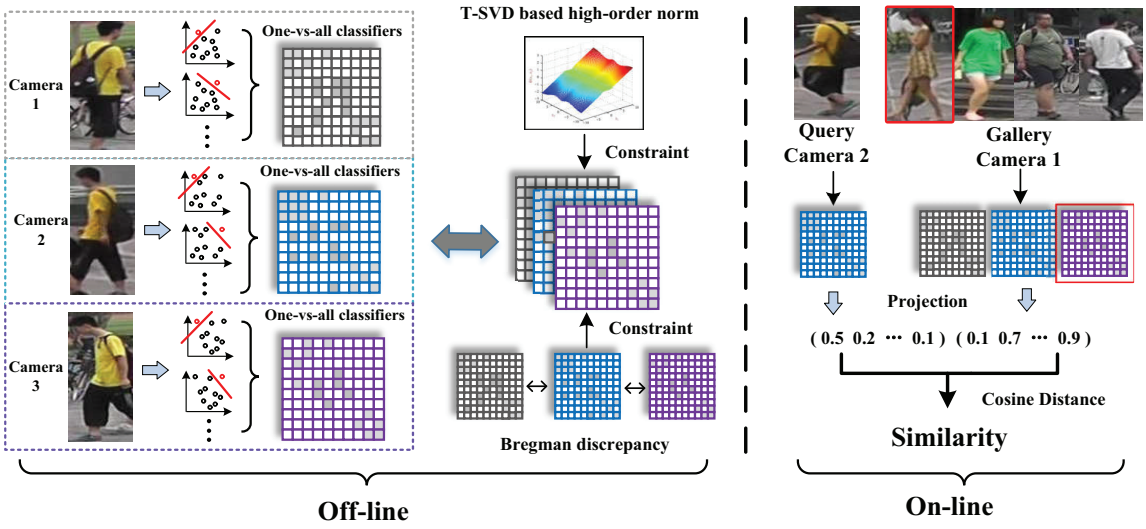


Fig. 2. The flowchart of the proposed approach. In the off-line stage, suppose there are three different cameras (tasks), and we formulate the person Re-ID under a certain camera (task) as a standard one-vs-all multi-class classification problem, where one classifier corresponds to a specific person. All these linear classifiers of a certain task are then constructed into a task-specific projection matrix (*e.g.*, grey matrix, blue matrix and purple matrix). By stacking these projection matrices to a tensor structure and imposing regularizer on it, all linear classifiers are optimized simultaneously to capture the high-order correlations. Optimal projection matrices are finally obtained in an alternative way. In on-line stage, given the probe and its Camera ID, we transform original features of probe and gallery images with corresponding projection matrix to a common space, which is produced by the outputs of classifiers. We compute cosine distance to rank all gallery images.

distinguish a specific person, and all these binary classifiers are stitched into a task-specific projection matrix. To achieve better generalization ability, among different tasks, we assume that the features of the same person under different cameras are generated from a latent/common subspace. Within a certain task, we assume that different identities under the same perspective should share similar patterns. By utilizing a new type of low-rank tensor constraint, the correlation among all the projection matrices can be captured, such that the feature distributions from different tasks can be well-aligned in the projected common space (see Fig. 1 for more details). Thanks to this well-founded tensor norm, namely tensor-Singular Value Decomposition (t-SVD) based nuclear norm [48], its circulant algebra could provide the relationship not only along third dimension (task-specific) but also among different columns (person-specific). This indicates the high-order correlationship not only across different tasks but also through the classifiers under a certain camera (within a certain task) , which is the *major motivation* of the proposed tensor multi-task model.

Inherited the merit from the multi-task learning, *i.e.,* the knowledge obtained from each task being reused by the others leads to a better generalization ability, our proposed model can deal with the situation that little or even no training labels are given. The derived model can be further extended by incorporating various visual representations in a direct but elegant way. Our experiments show the promising performance of the person Re-ID on both supervised and unsupervised settings. Fig. 2 illustrates the pipeline of our proposed scheme.

The main contributions of this paper are summarized as follows:

- We propose a new multi-task learning model, in both supervised and unsupervised manners, by taking advantage of a new tensor regularization to effectively handle person Re-ID, where the correlationship can be captured not only across different tasks but also within the task itself.
- We present an efficient optimization algorithm to solve the objective function of the proposed model, with relatively low computational complexity and theoretical convergence guarantee.
- The proposed model can be easily incorporated multiple visual features in a flexible way. Interestingly, no matter how we construct the multi-task tensor, the consistency and complementary among different visual features could be effectively exploited through high-order low-rank regularization.
- We conduct extensive evaluations of our method on several benchmark datasets, which manifests the competitive performance of our method.

The rest of this paper is organized as follows. Section II introduces related works. Section III gives the notations used throughout this paper. In Section IV, we firstly motivate the proposed model in detail, present it formally, then give an optimization algorithm to solve it, and extend the proposed method at last. Experimental analysis and completion results are shown in Section V to verify our method. Some analyses and discussions are also provided in this section. Finally, we conclude the proposed method in Section VI.

II. Related Work

Most person Re-ID methods fall in the scope of robust feature design and supervised/unsupervised distance metric learning. Their strengths and limitations are briefly reviewed below.

A. Robust Feature Design

Feature representation methods [4], [7] target on designing robust features against viewpoint, illumination and human pose changes. In the early stage, many hand-crafted features have been proposed, including texture descriptors [45], color information [7], gradient [49]. These local features do not need training process but are designed with human knowledge, which can be directly applied to any data since it doesn't need any domain knowledge. In this framework, Liao *et al.* [4] adopted the Retinex algorithm to conquer illumination variations and maximized the occurrence to deal with viewpoint changes. Matsukawa *et al.* [7] proposed to utilize hierarchical Gaussian distribution to preserve discriminative information for RE-ID, achieving promising performance. However, high precision needs to be accompanied with discriminative learning process [4], [31], [45].

More recently, deep learning is introduced into the Re-ID community and significantly promotes the performance. But its success partly stems from annotating a large number of labels. For sufficient training labels, SVDnet [46], MTDnet [1], CAN [12] and some other convolutional neural networks [3] have been proposed to produce more discriminative features. These methods differ significantly in network structure [18], training strategy [11] and loss function [40], leading different performance for Re-ID. On the contrary, there are also some methods aiming at designing domain-free feature extractors, such as JSTL [39], deep transfer model [30] and HIPHOP [5]. Moreover, to take full advantage of the strengths of different features, a lot of works [21] have already begun to combine different visual features to boost Re-ID performance. MHJLw [27] was proposed to explore the correlation among the probe and gallery data with various visual representations. MMF [8] proposed a multi-index fusion procedure to fuse multiple visual features.

B. Supervised Metric Learning

Supervised distance learning is another important issue for person Re-ID. Most works resort to metric learning [36], [47], rank learning [29], subspace learning [4] and deep learning [3] to give more reliable and stable similarities to identify persons. Theses methods can be further divided into view-generic model and view-specific model. Notable view-generic models include Metric Ensembles [29], SCSP [9], Null Space [16], XQDA [4], KISSME [31], MFA [45] and so on, which don't consider camera information and only build one model for all cameras. Hence, view-generic model is intrinsically limited, since the domain gap between cameras widely exists. But their computational complexity is usually low, making them scalable for large scale person Re-ID.

On the contrary, view-specific model either learns a matching model for each pair of cameras, or trains different projections for each camera. The former's representative method is MtMCML [13], which designs multiple Mahalanobis distance metrics to associate with the camera network. However, the complexity of these methods increases in proportion to the scale of the camera network, making them not feasible for real-world application. Instead, the projection based framework is becoming popular. In this framework, Su *et. al.* [44] proposed a multi-task classification framework to train view-specific classifiers simultaneously. Chen *et. al.* [32] generalized Mahalanobis distance to asymmetric distance to describe the view-wise discrepancy. These methods hold a common assumption that all view-specific models are different but correlated , such that information can be shared among all the models. However, previous works only focus on the relationship between cameras, ignoring the high-order information among cameras and persons, leading a suboptimal solution for person Re-ID.

C. Unsupervised Metric Learning

Learning a Re-ID model without training labels is a more challenging task but has broader application scenarios. To achieve this goal, a lot of transfer learning methods [6], [30], [39] have been proposed, which transfer the knowledge obtained from auxiliary datasets. Due to the domain gap across datasets, Peng *et al.* [6] developed a cross-dataset transfer model via multi-task dictionary learning. Geng *et al.* [30] addressed the data sparsity problem by a well-designed transfer deep structure and a loss-specific dropout strategy. Xiao *et al.* [39] also proposed a novel domain guided dropout to train images from all the datasets in a uniform deep structure, which can be regarded as an excellent deep feature extractor. Fan *et al.* [11] proposed a progressive training strategy by iteratively performing clustering and fine-tuning, indicating that the clustering results can also teach our model to improve identification.

Apart from transfer learning, several other methods focus on taking advantage of unlabeled training data to improving unsupervised Re-ID performance. Kodirov *et al.* [37] introduced l_1-norm graph Laplacian term into dictionary learning framework, jointly learning representation and discriminative information. Lisanti *et al.* [38] also explored additional information carried by neighboring individuals and proposed a solution for group Re-ID. To alleviate view-specific bias, Yu *et al.* [28] followed the idea [11] and proposed an unsupervised asymmetric metric via the clustering process. But this is achieved by performing clustering multiple times, which is time-consuming.

Summary: Our work is significantly differenced with previous works in a lot of aspects. First, the proposed method extends traditional metric learning methods [4], [31] to asymmetrical metric by exploiting camera information. Second, our model is based on the assumption that the MTL framework for Reid problem exists high-order correlations that not only across tasks but also within a certain task, while existing Re-ID model only focuses on exploring the relationship across tasks [32], [44]. Third, we show that the well-founded tensor structure can be flexibility incorporated into multiple visual features, and is easily generalized to the unsupervised setting.

III. Notation

The notations used throughout the paper will be introduced in this section. Specifically, we use lower case letters $x(i, j)$ to denote entries of matrix, bold lower case letters $\boldsymbol{x}$ to denote vector and bold upper case letters $\mathbf{X}$ to denote matrix. The

notation $\|\mathbf{X}\|_F := (\sum_{i,j} |x_{ij}|^2)^{\frac{1}{2}}$ is the Frobenius norm. And $\|\mathbf{X}\|_* := \sum_i \sigma_i(\mathbf{X})$ is the matrix nuclear norm, where $\sigma_i(\mathbf{X})$ denotes the i-th largest singular value of a matrix. The bold calligraphy letters are denoted for tensors (*i.e.,* $\boldsymbol{\mathcal{Z}} \in \mathcal{R}^{n_1 \times n_2 \times n_3}$ is a three-order tensor, where order means the number of ways of the tensor and is fixed at 3 in this paper). For a three-order tensor $\boldsymbol{\mathcal{X}}$, the 2D section $\boldsymbol{\mathcal{X}}(i,:,:)$, $\boldsymbol{\mathcal{X}}(:,i,:)$ and $\boldsymbol{\mathcal{X}}(:,:,i)$ (Matlab notation is used for better understanding) denote the ith horizontal, lateral and frontal slices, respectively. Analogously, the 1D section $\boldsymbol{\mathcal{X}}(i,j,:)$, $\boldsymbol{\mathcal{X}}(i,:,j)$ and $\boldsymbol{\mathcal{X}}(:,i,j)$ are the mode-1, mode-2 and mode-3 fibers of tensor. Specifically, $\boldsymbol{\mathcal{X}}^{(k)}$ is used to represent k-th Frontal Slice $\boldsymbol{\mathcal{X}}(:,:,k)$ for convenience. And $\boldsymbol{\mathcal{X}}_f$ denotes the tensor that we apply Fourier transform to $\boldsymbol{\mathcal{X}}$ along the third dimension.

IV. The Proposed Method

A. Motivation

Let $\mathbb{X} = \{\boldsymbol{x}_i \in \mathbb{R}^d\}_{i=1}^N$ be a set of N samples, where d is the dimension of the feature vector. Typical metric learning for person Re-ID aims to learn mahalanobis distance $\mathbf{M}$ to match individuals across multiple cameras, in which the distance between any two samples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is given by a symmetric model:

$$\begin{aligned} d(\boldsymbol{x}_i, \boldsymbol{x}_j) &= \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \mathbf{M} (\boldsymbol{x}_i - \boldsymbol{x}_j)} \\ &= ||\mathbf{U}^T \boldsymbol{x}_i - \mathbf{U}^T \boldsymbol{x}_j||_2, \end{aligned} \tag{1}$$

where $\mathbf{M} = \mathbf{U}\mathbf{U}^T$. By computing distance between feature vectors of probe and gallery images, we rank gallery images to complete re-identification. However, due to a unitary mapping for all cameras used in symmetric model [4], such a formulation fails to handle the situation where dramatic view-specific (*i.e.,* camera-specific) changes happen.

Hence, suppose a more general case that there are $V \geq 2$ cameras with significant camera-specific discrepancy. Let $\mathbf{X}^{(v)} = [\boldsymbol{x}_1^{(v)}, \boldsymbol{x}_2^{(v)}, ..., \boldsymbol{x}_{n_v}^{(v)}] \in \mathbb{R}^{d \times n_v}$ be the normalized feature vectors of pedestrian images captured by the v-th camera, where n_v is the number of images under the v-th camera. Note that we don't assume the numbers of training samples under different cameras are equal and this is suitable for most Re-ID applications. Like [28], we use the camera-specific projection matrix $\mathbf{U}^{(v)}, v = 1, 2, \ldots, V$ to transform each original feature vector to a latent space. In this latent space, each probe and gallery images can be represented by the projected feature. Thus, the distance between probe image $\boldsymbol{x}_i^{(v_1)}$ and gallery image $\boldsymbol{x}_j^{(v_2)}$ is reformulated by:

$$d(\boldsymbol{x}_i^{(v_1)}, \boldsymbol{x}_i^{(v_2)}) = ||\mathbf{U}^{(v_1)T} \boldsymbol{x}_i^{(v_1)} - \mathbf{U}^{(v_2)T} \boldsymbol{x}_i^{(v_2)}||_2. \tag{2}$$

In this way, distinct mapping matrices align feature distributions under different cameras, modeling the discrepancy among different cameras and generalizing the symmetric model (Eq. (1)) to asymmetric metric (Eq. (2)). As a result, the distance computed in this common space is more suitable for Re-ID on-line testing.

To learn suitable projection matrices, we formulate the person Re-ID as a one-vs-all classification problem under multi-task framework. Formally, we are given V sets of classifiers, in which one set corresponds to one specific camera (task). Let $\mathbb{U} = \{\mathbf{U} \in \mathbb{R}^{d \times C}\}_{i=1}^V$ be the sets of the classifiers, $\mathbf{U}^{(v)}$ denote the projection matrix for v-th camera, its column, *i.e.,* $\mathbf{U}_i^{(v)}$, represent the classifier for i-th identity under v-th camera, C is the number of classes (*i.e.,* persons). The relationship among classifiers, projection matrices, and tasks is illustrated in Fig. 3. Thus, given the training labels $\mathbf{Y}^{(v)} \in (0,1)^{C \times n_v}$

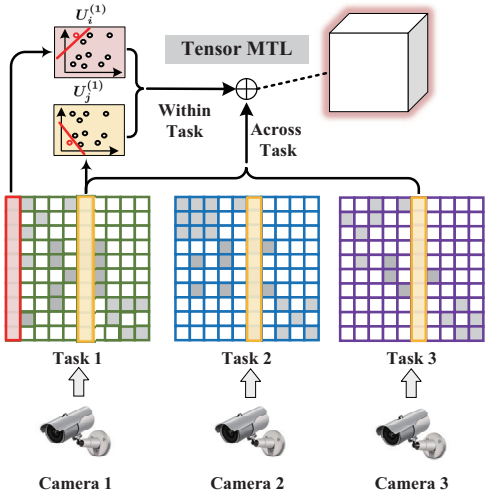


Fig. 3. The learning detail of the proposed approach. Each matrix represents the specific task learning problem. Each column of the matrix is a specific classifier. Across-task and within-task are incorporated in our model via the tensor structure.

whose c-th dimension is used to distinguish whether it belongs to the c-th identity, a general classification model can be described as:

$$\mathbf{U}^{(v)*} = \operatorname{argmin} \mathfrak{L}(\mathbf{U}^{(v)}, \mathbf{X}^{(v)}, \mathbf{Y}^{(v)}) + \lambda \mathfrak{R}(\mathbf{U}^{(v)}), \tag{3}$$

where $\mathfrak{L}(\mathbf{U}^{(v)}, \mathbf{X}^{(v)}, \mathbf{Y}^{(v)})$ indicates the classification term, and $\mathfrak{R}(\mathbf{U}^{(v)})$ denotes the regularizer term for $\mathbf{U}^{(v)}$, λ is the trade-off parameter to balance two terms.

However, due to the limited number of collected matched pairs, the learned classifiers are prone to be over-fitting. Therefore, the regularizer term $\mathfrak{R}(\mathbf{U}^{(v)})$ will play a critical role in learning process. Moreover, from the aspect of domain transfer, it is expected that the knowledge obtained from one camera can be re-used by others, which can further improve the discriminant and generalization ability. *This motivates us to adopt the multi-task (MTL) framework constrained by tensor based regularization to train the classifiers jointly.* It is worth noting that the typical MTL [44] only allows knowledge sharing across the tasks (*i.e.,* knowledge sharing is only across-tasks not within a task), while ignoring a critical issue that, the knowledge learned from the mapping to one output may be useful to the others within a certain task (*i.e.,* different person-specific classifiers under one camera). In other words, previous works mainly focus on pairwise asymmetric metric learning, ignoring the high-order information among cameras and persons. In the following, we will formally introduce our supervised/unsupervised Tensor-MTL (t-MTL) model, which extends to capture the high-order correlation across tasks and within a certain task, as well as the corresponding optimization algorithms and its attractive extension.

B. Supervised t-MTL Learning

One key assumption holds on MTL learning that all task-specific classifiers are correlated via a certain structure, so that the shared information can be transferred among tasks. To this end, we propose a new tensor structure for applications of person re-identification. Specifically, the proposed model is given as:

$$\min_{\mathbf{U}^{(v)}} \sum_{v} \mathfrak{L}(\mathbf{U}^{(v)}, \mathbf{X}^{(v)}, \mathbf{Y}^{(v)}) + \alpha \|\boldsymbol{\mathcal{U}}\|_{\circledast} + \beta \sum_{i \neq j} \|\mathbf{U}^{(i)} - \mathbf{U}^{(j)}\|_F^2 \tag{4}$$

where α and β denote the trade-off parameters, $\boldsymbol{\mathcal{U}} = \Phi(\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \ldots, \mathbf{U}^{(V)}) \in \mathbb{R}^{d \times C \times V}$ is a tensor by merging different $\mathbf{U}^{(v)}$ to a 3-order tensor along third dimension. That is to say, each frontal slice of $\boldsymbol{\mathcal{U}}$ is our task-specific projection matrix (*i.e.,* $\boldsymbol{\mathcal{U}}(:,:,v) = \mathbf{U}^{(v)}$). We emphasize here that $\|\cdot\|_{\circledast}$ denotes the t-svd nuclear norm [22], which constrains the tensor structure with a low-rank assumption. Here we give its formulation and the detailed mathematical derivation is introduced in supplementary material.

$$\begin{aligned} \|\boldsymbol{\mathcal{U}}\|_{\circledast} &= ||\mathrm{bcirc}(\boldsymbol{\mathcal{U}})||_* \\ &= || \begin{bmatrix} \mathbf{U}^{(1)} & \mathbf{U}^{(\mathrm{V})} & \cdots & \mathbf{U}^{(2)} \\ \mathbf{U}^{(2)} & \mathbf{U}^{(1)} & \cdots & \mathbf{U}^{(3)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{U}^{\mathrm{V}} & \mathbf{U}^{(\mathrm{V}-1)} & \cdots & \mathbf{U}^{(1)} \end{bmatrix} ||_*. \end{aligned} \tag{5}$$

This low-rank assumption allows us to capture the high order relationship by comparing every column in each task-specific classifiers (within task) and every frontal slice over the third dimension (across task). Specifically, by measuring every column of frontal slices (*i.e.,* $\mathbf{U}^{(i)}$), the classifiers, which are under different cameras but correspond to the specific identity, will be correlated. By measuring every row of frontal slices, the different identities, which are under the same perspective, would share similar patterns. It enables the model to achieve better performance and generation ability. Besides, the classification term $\mathfrak{L}(\mathbf{U}^{(v)}, \mathbf{X}^{(v)}, \mathbf{Y}^{(v)})$ can be any smooth and convex function measuring the discrepancy between groundtruth and predictions. Without loss of generality, we define the classification term as:

$$\mathfrak{L}(\mathbf{U}^{(v)}, \mathbf{X}^{(v)}, \mathbf{Y}^{(v)}) = ||\mathbf{U}^{(v)T}\mathbf{X}^{(v)} - \mathbf{Y}^{(v)}||_F^2. \tag{6}$$

In this way, each column of the $\mathbf{U}^v$ matrix corresponds to one identity in the training set and $\mathbf{U}^v$ transforms the feature vector into the label space rather than the distance space. Although the identities included in training and testing set are completely non-overlapping, it is trivial and beneficial to adopt this setting. The main advantage is that we can directly compute the classification loss in the transformed space without additional mapping, as well as knowing the number of identities in the testing set. From this perspective, each unseen identity can be represented by other training identities. This representation is also discriminative since it can be further mapped to binary training labels. Furthermore, there is no need to differentiate the intra-view and inter-view situations like [32], because of the uniform label space. Meanwhile, due to the clear physical meaning of $\mathbf{U}^v$, the consistent information both across and within tasks can be explored more thoroughly via the tensor structure. The Bregman discrepancy $\sum_{i \neq j} \|\mathbf{U}^{(i)} - \mathbf{U}^{(j)}\|_F^2$ is also used in our model to guarantee the discrepancy between transformations being controlled, leading to a more flexible way for the metric learning.

Algorithm 1: t-MTL

Input: Featre vector: $\mathbf{X}_v,\ v = 1, 2, \ldots, V$,
labels:$\mathbf{Y}_v,\ v = 1, 2, \ldots, V$, $\alpha > 0$, $\beta > 0$
Output: Classifiers $\mathbf{U}_v,\ v = 1, 2, \ldots, V$

Initialized: $\mathbf{U}_v = \mathbf{0}$; $\boldsymbol{\mathcal{G}} = \boldsymbol{\mathcal{W}} = \mathbf{0}$; $\rho = 10^{-5}$, $\rho_{\max} = 10^{10}$;
Construct: $\tilde{\mathbf{U}}$,$\tilde{\mathbf{X}}$,$\tilde{\mathbf{Y}}$,$\mathbf{M}$
while *not converge* **do**
  Update $\tilde{\mathbf{U}}$ by using (16);
  Obtain $\boldsymbol{\mathcal{U}}$ through $\tilde{\mathbf{U}}$;
  Update $\boldsymbol{\mathcal{G}}$ via Algorithm 2;
  Update Lagrange multipliers $\boldsymbol{\mathcal{W}}$ by using (21);
  Obtain $\tilde{\mathbf{G}}$ and $\tilde{\mathbf{W}}$ by $\boldsymbol{\mathcal{G}}$ and $\boldsymbol{\mathcal{W}}$;
  Update parameters ρ: $\rho = \min(\eta\rho, \rho_{\max})$;
end
Obtain $\mathbf{U}_v,\ v = 1, 2, \ldots, V$ via $\tilde{\mathbf{U}}$;
Return Classifiers $\mathbf{U}_v,\ v = 1, 2, \ldots, V$.

C. Optimization Procedure

The optimization problem (4) seems challenging to solve, not only because of the tensor low-rank norm on $\boldsymbol{\mathcal{U}}$, but also due to the Bregman discrepancy. We first rewrite Eq. (4) in a more compact form by constructing block matrices:

$$\tilde{\mathbf{U}} = [\mathbf{U}^{(1)}; \mathbf{U}^{(2)}; \ldots; \mathbf{U}^{(V)}], \tag{7}$$

$$\tilde{\mathbf{Y}} = [\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, \ldots, \mathbf{Y}^{(V)}]. \tag{8}$$

Thus the classification term of Eq. (6) can be reformulated as:

$$\sum_{v} ||\mathbf{U}^{(v)T}\mathbf{X}^{(v)} - \mathbf{Y}^{(v)}||_F^2 = ||\tilde{\mathbf{U}}^T\tilde{\mathbf{X}} - \tilde{\mathbf{Y}}||_F^2 \tag{9}$$

where

$$\tilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X}^{(1)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}^{(2)} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{X}^{(V)} \end{bmatrix}. \tag{10}$$

With identity matrix $\mathbf{I} \in \mathbb{R}^{d \times d}$, we define a block matrix $\mathbf{M}$ as:

$$\mathbf{M} = \begin{bmatrix} (V-1)\mathbf{I} & -\mathbf{I} & \cdots & -\mathbf{I} \\ -\mathbf{I} & (V-1)\mathbf{I} & \cdots & -\mathbf{I} \\ \vdots & \vdots & \ddots & \vdots \\ -\mathbf{I} & -\mathbf{I} & \cdots & (V-1)\mathbf{I} \end{bmatrix}, \tag{11}$$

and the Bregman discrepancy can be transferred as:

$$\sum_{i \neq j} \|\mathbf{U}^{(i)} - \mathbf{U}^{(j)}\|_F^2 = tr(\tilde{\mathbf{U}}^T\mathbf{M}\tilde{\mathbf{U}}). \tag{12}$$

Then, Eq. (4) can be rewritten as:

$$\min_{\tilde{\mathbf{U}}} ||\tilde{\mathbf{U}}^T\tilde{\mathbf{X}} - \tilde{\mathbf{Y}}||_F^2 + \alpha\|\boldsymbol{\mathcal{U}}\|_{\circledast} + \beta tr(\tilde{\mathbf{U}}^T\mathbf{M}\tilde{\mathbf{U}}) \tag{13}$$

The above optimization problem can be solved by using the Augmented Lagrange Multiplier (ALM) [23]. To adopt alternating direction minimizing strategy to problem Eq. (4), we need to make the objective function seperable. By introducing the auxiliary tensor variable $\boldsymbol{\mathcal{G}}$, the optimization problem can be transferred to minimize the following unconstrained problem:

$$\begin{aligned}\mathcal{L}(\tilde{\mathbf{U}};\boldsymbol{\mathcal{G}}) =&||\tilde{\mathbf{U}}^T\tilde{\mathbf{X}}-\tilde{\mathbf{Y}}||_F^2+\alpha\|\boldsymbol{\mathcal{G}}\|_{\circledast}+\beta tr(\tilde{\mathbf{U}}^T\mathbf{M}\tilde{\mathbf{U}}) \\ &+\langle\boldsymbol{\mathcal{W}},\boldsymbol{\mathcal{U}}-\boldsymbol{\mathcal{G}}\rangle+\frac{\rho}{2}||\boldsymbol{\mathcal{U}}-\boldsymbol{\mathcal{G}}||_F^2\end{aligned} \quad (14)$$

where the tensor $\boldsymbol{\mathcal{W}}$ represents Lagrange multiplier, ρ is actually the penalty parameter, which are adjusted by using adaptive updating strategy as suggested in [50]. We adopt an alternating scheme and partition the unconstrained problem into two steps alternatingly.

$\tilde{\mathbf{U}}$-subproblem: When tensor $\boldsymbol{\mathcal{G}}$ is fixed, since $\mathbf{G}^{(v)} = \Phi_v^{-1}(\boldsymbol{\mathcal{G}})$ and $\mathbf{W}^{(v)} = \Phi_v^{-1}(\boldsymbol{\mathcal{W}})$, where Φ_v^{-1} is the inverse operation w.r.t Φ by clipping v-th frontal slice of the tensor, our optimization task is transferred to solve the following subproblem for updating the projection matrix $\tilde{\mathbf{U}}$:

$$\begin{aligned}\tilde{\mathbf{U}}^* =&\underset{\tilde{\mathbf{U}}}{\operatorname{argmin}}\,||\tilde{\mathbf{U}}^T\tilde{\mathbf{X}}-\tilde{\mathbf{Y}}||_F^2+\beta tr(\tilde{\mathbf{U}}^T\mathbf{M}\tilde{\mathbf{U}}) \\ &+<\tilde{\mathbf{W}},\tilde{\mathbf{U}}-\tilde{\mathbf{G}}>+\frac{\rho}{2}||\tilde{\mathbf{U}}-\tilde{\mathbf{G}}||_F^2\end{aligned} \quad (15)$$

where $\tilde{\mathbf{W}}$ and $\tilde{\mathbf{G}}$ are the block matrices constructed by $[\mathbf{W}^{(1)};\mathbf{W}^{(2)};\ldots;\mathbf{W}^{(V)}]$ and $[\mathbf{G}^{(1)};\mathbf{G}^{(2)};\ldots;\mathbf{G}^{(V)}]$ like $\tilde{\mathbf{U}}$. It is easy to solve this optimization problem since it has a closed-form solution. We can obtain the solution by setting the derivative to zero:

$$\tilde{\mathbf{U}}^{T*} = (\tilde{\mathbf{Y}}\tilde{\mathbf{X}}^T+\rho\tilde{\mathbf{G}}^T-\tilde{\mathbf{W}}^T)(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T+\beta\mathbf{M}+\rho\tilde{\mathbf{I}})^{-1} \quad (16)$$

where $\tilde{\mathbf{I}}\in\mathbb{R}^{Vd\times Vd}$ denotes the identity matrix.

$\boldsymbol{\mathcal{G}}$-subproblem: When $\tilde{\mathbf{U}}$ is fixed, solving Eq. (14) is equal to minimize the following problem:

$$\boldsymbol{\mathcal{G}}^* = \underset{\boldsymbol{\mathcal{G}}}{\operatorname{argmin}}\,\alpha(||\boldsymbol{\mathcal{G}}||_{\circledast}+\frac{\rho}{2\alpha}||\boldsymbol{\mathcal{G}}-(\boldsymbol{\mathcal{U}}+\frac{1}{\rho}\boldsymbol{\mathcal{W}})||_F^2). \quad (17)$$

Algorithm 2: t-SVD based Tensor Nuclear Norm Minimization

Input: Observed tensor $\boldsymbol{\mathcal{F}}\in\mathbb{R}^{n_1\times n_2\times n_3}$, scalar $\tau>0$
Output: tensor $\boldsymbol{\mathcal{G}}$

$\boldsymbol{\mathcal{F}}_f = \text{fft}(\boldsymbol{\mathcal{F}},[\,],3)$, $\tau' = n_3\tau$;
for $j = 1 : n_3$ **do**
  $[\boldsymbol{\mathcal{U}}_f^{(j)},\boldsymbol{\mathcal{S}}_f^{(j)},\boldsymbol{\mathcal{V}}_f^{(j)}] = \text{SVD}(\boldsymbol{\mathcal{F}}_f^{(j)})$;
  $\boldsymbol{\mathcal{J}}_f^{(j)} = \text{diag}\{(1-\frac{\tau'}{\boldsymbol{\mathcal{S}}_f^{(j)}(i,i)})_+\}, \quad i = 1,\ldots,\min(n_1,n_2)$;
  $\boldsymbol{\mathcal{S}}_{f,\tau'}^{(j)} = \boldsymbol{\mathcal{S}}_f^{(j)}\boldsymbol{\mathcal{J}}_J^{(j)}$;
  $\boldsymbol{\mathcal{G}}_f^{(j)} = \boldsymbol{\mathcal{U}}_f^{(j)}\boldsymbol{\mathcal{S}}_{f,\tau'}^{(j)}\boldsymbol{\mathcal{V}}_f^{(j)^{\mathrm{T}}}$;
end
$\boldsymbol{\mathcal{G}} = \text{ifft}(\boldsymbol{\mathcal{G}}_f,[\,],3)$;
Return tensor $\boldsymbol{\mathcal{G}}$.

It can be reformulated in a compact way:

$$\min_{\boldsymbol{\mathcal{G}}}\ \tau||\boldsymbol{\mathcal{G}}||_{\circledast}+\frac{1}{2}||\boldsymbol{\mathcal{G}}-\boldsymbol{\mathcal{F}}||_F^2 \quad (18)$$

where $\tau=\frac{\alpha}{\rho}$ and $\boldsymbol{\mathcal{F}}=(\boldsymbol{\mathcal{U}}+\frac{1}{\rho}\boldsymbol{\mathcal{W}})$. The optimal solution of this problem is given by following theorem:

Theorem 1. *For $\tau>0$ and $\boldsymbol{\mathcal{G}},\boldsymbol{\mathcal{F}}\in\mathbb{R}^{n_1\times n_2\times n_3}$, the globally optimal solution to the following problem*

$$\min_{\boldsymbol{\mathcal{G}}}\ \tau||\boldsymbol{\mathcal{G}}||_{\circledast}+\frac{1}{2}||\boldsymbol{\mathcal{G}}-\boldsymbol{\mathcal{F}}||_F^2 \quad (19)$$

is given by the tensor tubal-shrinkage operator

$$\boldsymbol{\mathcal{G}} = \mathcal{C}_{n_3\tau}(\boldsymbol{\mathcal{F}}) = \boldsymbol{\mathcal{U}}*\mathcal{C}_{n_3\tau}(\boldsymbol{\mathcal{S}})*\boldsymbol{\mathcal{V}}^{\mathrm{T}}, \quad (20)$$

where $\boldsymbol{\mathcal{F}}=\sum_{i=1}^{\min(n_1,n_2)}\boldsymbol{\mathcal{U}}(:,i,:)*\boldsymbol{\mathcal{S}}(i,i,:)*\boldsymbol{\mathcal{V}}(:,i,:)^{\mathrm{T}}$ and $\mathcal{C}_{n_3\tau}(\boldsymbol{\mathcal{S}})=\boldsymbol{\mathcal{S}}*\boldsymbol{\mathcal{J}}$, herein, $\boldsymbol{\mathcal{J}}$ is an $n_1\times n_2\times n_3$ f-diagonal tensor whose diagonal element in the Fourier domain is $\boldsymbol{\mathcal{J}}_f(i,i,j)=(1-\frac{n_3\tau}{\boldsymbol{\mathcal{S}}_f^{(j)}(i,i)})_+$.

The proof is given by supplementary material. Additionally, the Lagrange multipliers $\boldsymbol{\mathcal{W}}$ need to be updated as follows

$$\boldsymbol{\mathcal{W}}^* = \boldsymbol{\mathcal{W}}+\rho(\boldsymbol{\mathcal{U}}-\boldsymbol{\mathcal{G}}) \quad (21)$$

The above two steps are repeated until the convergence condition is satisfied. Meanwhile based on [51], we have following theorem regarding the convergence of Algorithm 1.

Theorem 2. *The sequence ($\boldsymbol{\mathcal{G}}$, $\tilde{\mathbf{U}}$) generated by Algorithm 1 in each step converges to an accumulation point. Moreover, the accumulation point is an optimal solution of the optimization problem (13).*

Furthermore, the proposed method performs well and indeed converges fast in reality, which will be illustrated in Section V-C. In practice, we fix max iteration number to 30 for all datasets.

D. Unsupervised t-MTL Learning

In the previous sections, we have introduced the idea regarding the person Re-ID as an MTL classification problem. In this framework, the proposed supervised t-MTL learns suitable distance that matches the same individuals across multiple cameras. However, it is not always guaranteed that there are enough labels for training. Alternatively, a practical and intuitive solution is to make full use of cheap and valuable unlabeled data. But in the unsupervised setting, it becomes more challenging to train the model, as we have no labeled data to guide to distinguish similar appearance persons. Instead, motivated by [28], we can replace the label $\mathbf{Y}^{(v)}$ in classification term by a virtual label. Specifically, we first produce a virtual label for each training sample by clustering (*e.g.,* k-means clustering). Then, similar to our supervised t-MTL method, minimize the following objective function to obtain the projection matrices:

$$\min_{\mathbf{U}^{(v)}}\sum_v||\mathbf{U}^{(v)T}\mathbf{X}^{(v)}-\mathbf{P}||_F^2+\alpha\|\boldsymbol{\mathcal{U}}\|_{\circledast}+\beta\sum_{i\neq j}\|\mathbf{U}^{(i)}-\mathbf{U}^{(j)}\|_F^2 \quad (22)$$

where $\mathbf{P} \in \mathbb{R}^{C\times K}$ denotes the virtual label, and K is the number of cluster centers. Even though the virtual label is relatively undesirable w.r.t the ground truth, the tensor structure can benefit from the shared knowledge intra/inter-task to improve the generalization ability. Meanwhile, the proposed unsupervised model runs the clustering procedure only once, unlike [28], which must perform clustering many times until the algorithm converges.

E. Multiple Visual Features for Tensor-MTL

Existing person Re-ID methods [5], [16], [41], [43] usually utilize multiple types of visual features to boost performance by concatenating feature vectors. Nevertheless, such a kind of operation ignores the importance of different visual features, which can not exploit complementary information of multi-visual representations.

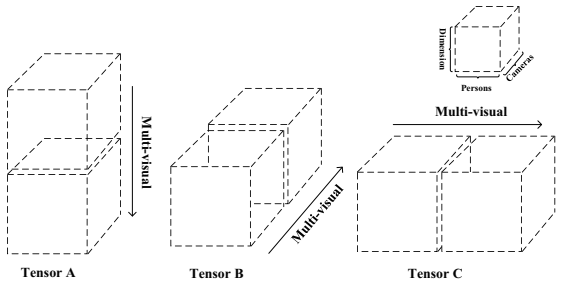


Fig. 4. The multi-visual tensor used in this paper. Each tensor is constituted by stacking the classifiers.

Built upon our multi-task learning problem, we can further explore the shared information contained in the multi-visual representations. Owing to the tensor structure, the proposed model can be easily incorporated with multiple visual features by stacking their corresponding projection matrices. Suppose we have L visual representations for each image, let $\mathbf{U}_{(l)}^{(v)}$ denote the projection matrix for l-th visual feature and v-th camera, $\{\mathbf{U}_{(l)}^{(v)}\}_{l=1}^{L}$ implicitly encode the shared information hidden in different visual representations. In practice, we consider three ways to construct our multi-task tensor as shown in Fig. 3, which are denoted as $\boldsymbol{\mathcal{U}}_A$, $\boldsymbol{\mathcal{U}}_B$ and $\boldsymbol{\mathcal{U}}_C$, respectively. Actually, these three ways are all equivalent due to the following Theorem.

Theorem 3. *The high-order nuclear norm on Tensor A, Tensor B and Tensor C is the same.*

The proof is provided in supplementary material. We implement our model by using the way as Tensor A (see Fig. 3) for all experiments. In contrast to concatenate the feature vectors directly, the proposal utilizes circulant algebra to compare every specific classifier to capture the consistence. Thus the complementary information is implicitly embedded in the projections, which leads a better performance. To sum up, no matter how we construct the multi-task tensor, the consistency and complementary among different visual features could be effectively exploited through high-order low-rank regularization.

F. Online Stage

In the online query stage, given the query image q with camera ID $v_q \in [1,\ldots,V]$, we first extract multiple visual features $\mathbf{x}(q) = [\mathbf{x}_{(1)}(q),\ldots,\mathbf{x}_{(L)}(q)]$. Notice that, for the Re-ID task, the classes in the training set can be the same as or different from those in the gallery and probe sets. That is to say, once the projection matrix $\mathbf{U}_{(l)}^{(v_q)}, l = 1,\ldots,L$ with respect to v_q-th camera are obtained in the off-line stage, we directly use it without any modifications to convert extracted features to the common space, no matter its label is in the training set or not.

$$\boldsymbol{x}'(q) = [\mathbf{U}_{(1)}^{(v_q)T}\boldsymbol{x}_{(1)}(q), \mathbf{U}_{(2)}^{(v_q)T}\boldsymbol{x}_{(2)}(q),\ldots,\mathbf{U}_{(L)}^{(v_q)T}\boldsymbol{x}_{(L)}(q)]. \tag{23}$$

This projected visual features need to be further normalized. For gallery image, we also transform their original feature vectors to the common space as done for query images. According to asymmetric model (Eq. (2)), we measure the similarity between these two features by cosine distance. Finally, we rank gallery images to complete Re-ID.

V. Experimental Results and Analysis

In this section, we perform experiments to present a comprehensive evaluation of the proposed method. All experiments are implemented on a workstation with Intel Xeon E5-2630 @ 2.30 GHz CPU, 128GB RAM, and TITANX GPU (12GB caches). To promote the culture of reproducible research, source codes and experimental results accompanying this paper will be released at https://www.researchgate.net/profile/Zhizhong_Zhang5.

To clearly illustrate our experimental strategies, we first introduce our experimental settings, including datasets, feature representations, and evaluation methodology. The main results and observations of the proposed methods are then presented, where both supervised and unsupervised settings are conducted. In both manners, the comparison is made to measure performance improvement on the baseline methods and some other state-of-the-art methods. Meanwhile, we also conduct experiments with variants of our approach and report results by integrating multiple visual features to further confirm the effectiveness of the extension of the proposal. At last, we analyze the characteristics of our t-MTL method, such as sensitivity, convergence, computational complexity and some insights for unsupervised t-MTL.

A. Experimental Settings

We evaluate the proposed algorithm on four public benchmark datasets, *i.e.*, ViPeR [33], CUHK01 [34], CUHK03 [35], and Market-1501 [49]. All of them are used to test our supervised model, while only Viper and Market-1501 are evaluated for our unsupervised t-MTL learning for simplicity. In the following, we will introduce some experimental details such as datasets, data representation, evaluation metric, and parameter setting.

Datasets: VIPeR [33] presents illumination variations and pose changes between pairs of views. We split the whole set of 632 image pairs randomly into two sets with equal size (316

pairs), one for training and the other one for testing. A single image from the probe set is selected and matched with all the images from the gallery set. CUHK01 [34] dataset is captured with two camera views in a campus environment. This dataset contains 971 persons, and each person has two images in each camera view. The person identities are split into 485 for training and 486 for the test. This dataset provides two evaluation modality: single-shot and multi-shot setting. CHUK03 [35] contains 13,164 images of 1,360 pedestrians captured from six surveillance camera views. Besides hand-cropped images, samples detected by a state-of-the-art pedestrian detector is provided. Market-1501 [49] is collected from six camera views in front of a supermarket in Tsinghua University. Overall, this dataset contains 32,668 annotated bounding boxes of 1,501 identities. There are 12,936 images used for training and other 19732 images for testing. **Data Representation for Supervised Model:** To obtain the image representations, we utilize two representative descriptors, (*i.e.,* Local Maximal Occurrence (LOMO) [4] and Gaussian Of Gaussian (GOG) [7]). In addition, we also use LOMO and GOG with the same weight to evaluate the performance of our multi-visual model. For market-1501, we follow the identification model proposed by [52] and train two baseline CNN networks, *i.e.,* CaffeNet [26] and ResNet50 [49] without any modifications .

Data Representation for Unsupervised Model: In our experiments, we use the deep learning based JSTL feature proposed in [39]. We implement it by using the convolutional layers, Inception modules, and fully connected layers [39], producing a 256-D feature. The original JSTL is adopted to our implementation to extract features on Market-1501. Note that the training set of the original JSTL contained VIPeR, violating the unsupervised setting. So we train a new JSTL model without using VIPeR training data to extract features.

Evaluation Measures: Three popular metrics are used to evaluate the performances, including cumulative match characteristic (CMC), Rank-1 accuracy (Rank-1), and mean average precision (mAP). In supervised setting, we evaluate the VIPeR, CUHK01 and CHUK03 by CMC as suggested in [4], [7], [27], [32], [44], while Rank-1 and mAP are adopted for Market-1501 [49]. In unsupervised setting, we follow [28] to report Rank-1 for VIPeR, Rank-1 and mAP for Market-1501. Note that in VIPeR and CUHK01, the reported final results on those metrics are measured by the average of 10 runs, while 20 runs for CUHK03.

Parameter setting: Only two parameters α and β need to be tuned. More details about the parameters will be discussed in Section V-C. The parameters in other competitors are set within ranges suggested by the original papers, and we tune those parameters so as to show the best results.

B. Experimental Results

1) Supervised t-MTL Results: **VIPeR**: For VIPeR dataset, we first compare the performance of the proposed method with some baseline models, which is shown in Table I. Our approach gets the Rank-1 accuracy of 44.7%, 50.6% and 56.1% for LOMO, GOG and multiple visual features, respectively. It is worth noting that our approach outperforms the existing asymmetric person Re-ID methods, such as CVDCA [32] and MTL-LORAE [44], by utilizing multiple visual features. Furthermore, compared to the symmetric metric learning method XQDA [4], our approach improves Rank-1 accuracy with an absolute gain of 4.7%, 0.9% and 2.8% for various features. But for Rank-5, Rank-10 and Rank-20 accuracy, our approach performs almost the same or slightly worse than XQDA. The reason for this may be that, our proposal is based on the classification model which is good at identifying similar appearance persons rather than concentrating on improving the ranking performance [1].

TABLE I
PERFORMANCES OF SUPERVISED T-MTL ON VIPER

Method	Feature	Viper			
		Rank-1	rank-5	Rank-10	rank-20
MTL-LORAE [44]	LBP+Attribute	42.3	72.2	81.6	89.6
CVDCA [32]	LOMO	43.7	74.1	84.8	91.9
XQDA [4]	LOMO	40.0	-	80.5	91.1
t-MTL($\alpha = 0$)	LOMO	31.2	64.5	79.1	89.6
t-MTL($\beta = 0$)	LOMO	32.6	58.0	69.7	80.2
t-MTL	LOMO	44.7	74.1	84.7	91.8
CVDCA [32]	GOG	50.4	78.8	88.0	94.5
XQDA [4]	GOG	49.7	-	88.6	94.5
t-MTL	GOG	50.6	78.4	87.3	93.3
CVDCA [32]	LOMO+GOG	49.5	78.6	87.7	94.1
XQDA [4]	LOMO+GOG	53.3	-	90.9	95.7
t-MTL (L1 loss)	LOMO+GOG	53.4	80.5	88.7	94.6
t-MTL (Concatenate)	LOMO+GOG	55.8	82.1	90.3	95.5
t-MTL (Tensor)	LOMO+GOG	56.1	82.1	90.3	95.5

To confirm the effectiveness of the proposed model, we also conduct experiments with variants of our approach, (*i.e.,* discard either low-rank regularizer or Bregman discrepancy constraint, and apply different tenor structures for multiple visual features to implement high-order norm). For LOMO feature, when we remove the tensor constraint or Bregman discrepancy, respectively, *i.e.* by setting $\alpha = 0$ (the fourth row in Table. I) or $\beta = 0$ (the fifth row in Table. I), it drops nearly 10% absolute reduction in term of Rank-1

TABLE II
PERFORMANCES OF SUPERVISED T-MTL ON VIPER WITH STATE-OF-THE ART METHODS

Method	Ref	Viper		
		Rank-1	Rank-10	rank-20
MTL-LORAE [44]	ICCV2015	42.3	81.6	89.6
XQDA [4]	CVPR2015	40.0	80.5	91.1
MetricEnsemple [29]	CVPR2015	45.9	88.9	95.8
SSDAL [10]	ECCV2016	43.5	81.5	89.0
NULL [16]	CVPR2016	51.2	90.5	95.9
GOG [7]	CVPR2016	49.7	88.6	94.5
KCVDCA [32]	TCSVT2017	43.3	83.5	92.2
MHJLw [27]	TNNLS2017	45.4	84.0	92.5
SSM [43]	CVPR2017	53.7	91.5	96.1
MTDnet [1]	AAAI2017	47.5	82.6	-
DictRW [41]	IJCAI2017	55.7	91.5	96.7
t-MTL (Tensor)	-	56.1	90.3	95.5

accuracy. Meanwhile, for multiple visual features, two variants with regard to different ways of tensor construction, *e.g.,* concatenate like Tensor A (the penultimate row in Table. I) and stack like Tensor B (the last row in Table. I), achieve almost the same performance. Notice that both of them are superior to XQDA, while their baselines behave similar to XQDA. This indicates that our t-MTL method could capture the complementary knowledge between the LOMO feature and GOG feature, and elevate the performance to a higher level. We also use L1 loss (*i.e.,* $||\mathbf{U}^{(v)T}\mathbf{X}^{(v)} - \mathbf{Y}^{(v)}||_1$) to replace L2 loss (*i.e.,* $||\mathbf{U}^{(v)T}\mathbf{X}^{(v)} - \mathbf{Y}^{(v)}||_2^2$) for our classification term. The detailed solution to this problem is shown in the supplementary material. The result is shown in Table I. Although we carefully tune the parameters, L2 loss still performs slightly better than L1 loss. It appears that a smooth classification term could be beneficial.

Furthermore, for ViPeR dataset, the comparison of visualization of distributions in the original feature spaces and the common spaces are provided by performing dimension reduction via PCA and t-SNE. As illustrated in Fig. 5 (a) and (b), the two feature distributions (blue and green points) obtained from two cameras have relatively low overlap ratio in original feature space, which indicates they are miss aligned for different tasks. Note that, under different cameras, the feature distributions in original feature space should be very different. But after the task-specific projections (see Fig. 5 (c) and (d)), the distributions are well aligned in common space, and data points are more separated than before.

TABLE III
PERFORMANCES OF SUPERVISED T-MTL. MEASURED BY RANK-1 ACCURACIES FOR MARKET-1501

Method	Single Query		Multiple Query	
	Rank-1	mAP	Rank-1	mAP
CaffeNet [49]	59.53%	32.85%	66.63%	41.25%
CaffeNet+ CVDCA [32]	59.80%	35.69%	-	-
CaffNet+XQDA [4]	62.00%	37.55%	70.28%	46.78%
CaffNet+KISSME [31]	61.02%	37.72%	69.86%	45.34%
CaffNet+t-MTL	62.35%	35.60%	71.38%	44.68%
ResNet50 [49]	75.62%	50.68%	81.26%	59.10%
ResNet-50+CVDCA [32]	74.82%	50.21%	-	-
ResNet50+XQDA [4]	76.01%	52.98%	81.12%	61.09%
ResNet50+KISSME [31]	77.52%	53.88%	82.16%	61.54%
ResNet50+t-MTL	78.33%	52.66%	84.14%	61.73%

Since enormous algorithms have reported results on VIPeR dataset, it is unrealistic to exhibit all of them. Hence, we only include those published in recent 3 years or have close relationships with our work. As demonstrated in Table II, the proposed approach achieves highly comparable (even better) results with the state-of-the-art methods, including domain transfer method [1], multi-task learning method [44], and also outperforms the multi-visual fusion method [27]. It is remarkable that SSM [43], as a postprocessing method, also provides comparable performance. It can be anticipated that SSM and t-MTL will benefit from each other, and lead a better performance. Meanwhile, DictRW [41] exceeds our proposal in terms of Rank-10 and Rank-20 accuracy, since they embed

TABLE IV
PERFORMANCES OF SUPERVISED T-MTL ON CUHK03

Method	Detected			
	Rank-1	rank-5	Rank-10	rank-20
DeepReID [35]	19.9	49.0	64.3	-
S-LSTM [17]	57.3	80.1	88.3	
Null [16]	54.7	84.8	94.8	95.2
S-CNN [18]	61.8	80.9	88.3	-
SSM [43]	72.7	92.4	96.1	-
XQDA+LOMO [4]	46.3	78.9	88.6	94.3
XQDA+GOG [7]	64.0	88.6	94.2	97.6
CVDCA+LOMO+GOG [32]	59.6	86.6	93.9	97.3
XQDA+LOMO+GOG	68.1	90.2	95.0	98.0
t-MTL+LOMO	50.5	78.5	86.3	92.2
t-MTL+GOG	59.3	84.5	91.5	96.2
t-MTL+LOMO+GOG	66.5	88.3	93.3	97.0

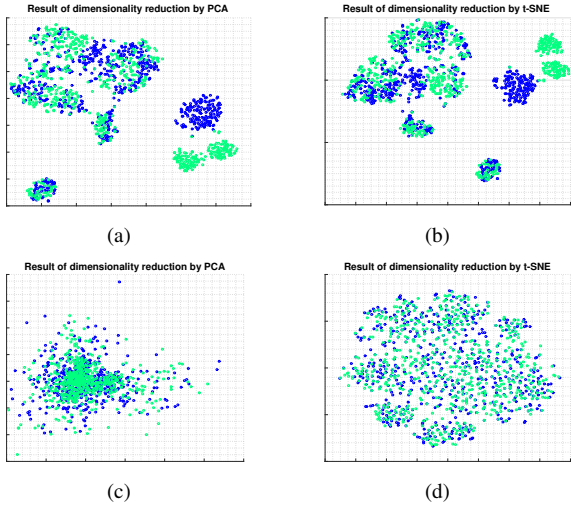


Figure 5. Result of dimensionality reduction for ViPeR dataset (LOMO feature). Figure. 5(a) and figure. 5(b) represent the original feature space, while figure. 5(c) and figure. 5(d) represent the projected common space. Blue points denote the features captured by Camera 1 and Green points denote the features captured by Camera 2.

triplet relationship into the model while we only focus on a classification model. By considering triplet loss, a slight improvement can be expected. However, this is out of the scope of this work. Besides, DictRW adopts deep feature to produce a strong baseline. By contrast, we only adopt handcrafted feature but achieve higher Rank-1 accuracy.

Market-1501: For Market-1501, since there are sufficient training samples and labels, deep learning based methods [49] achieve promising performance, as it is shown in Table III. With our proposal, we can find that, by exploring inter/intra-task correlations, the proposed method can further improve the performance accompanied with the same deep features. Concretely, for single query we outperform baseline models with a clear improvement in terms of Rank-1 accuracy and mAP. For CaffeNet, we achieve 2.82% and 2.70% absolute increase, and for ResNet-50, we obtain 2.71% and 1.98% absolute gain. As for multiple queries, similar results are also observed. Meanwhile, compared to typical metric learning

TABLE V
PERFORMANCES OF SUPERVISED T-MTL ON CUHK01

Method	single-shot				multiple-shot			
	Rank-1	rank-5	Rank-10	rank-20	Rank-1	rank-5	Rank-10	rank-20
MetricEnsemple [29]	53.4	76.4	84.4	90.5	-	-	-	-
KCVDCA [32]	47.8	74.2	83.4	89.9	-	-	-	-
CVDCA+LOMO+GOG [32]	73.0	89.1	93.9	95.5	-	-	-	-
MTDnet [1]	-	-	-	-	78.5	96.5	97.5	-
Null+LOMO [16]	-	-	-	-	65.0	85.0	89.9	94.4
MHJLw [27]	-	-	-	-	64.5	-	91.1	95.3
XQDA+LOMO [4]	48.7	73.0	81.3	88.2	63.2	83.9	90.0	94.2
XQDA+GOG [7]	57.8	79.1	86.2	92.1	67.3	86.9	91.8	95.9
XQDA+LOMO+GOG	68.5	87.3	92.4	96.3	76.9	91.5	95.4	97.9
t-MTL+LOMO	50.1	73.8	81.3	87.7	64.4	84.9	90.3	94.1
t-MTL+GOG	58.0	78.6	85.1	90.2	66.0	85.1	90.1	94.6
t-MTL+LOMO+GOG	74.1	89.1	92.9	95.9	80.3	92.4	95.1	96.9

methods (*i.e.,* KISSME [31] and XQDA [4]), our method is good at improving the top rank accuracy, but fails for the mAP, due to the aforementioned reason. However, when we fuse two types of visual features, the performance is slightly worse than the ResNet-50. The reasons behind such abnormal phenomena might be that, the CaffeNet and ResNet intrinsically belong to homogeneous feature, as well as have significant difference in performance, so they can't benefit from each other. It is worth mentioned that SVDnet [14] is also a similar tensor SVD projection in the CNN model, which shows nice improvements. But its motivation is very different from ours, since SVDnet is based on the observation that the last linear layers produce nonorthogonal projection. On this basis, it utilizes SVD decomposition to produce orthogonal layer to project features. Compared to SVDnet, we aim to find a tensor low-rank approximation to achieve better generalization ability, where t-SVD is adopted for the purpose of optimization solution. Re-ranking technique [8], [15] is also very relevant to our work, but they focus on embedding the relationship of the gallery images into the learned metric, where we aim to learn a suitable metric by exploring shared information between/within tasks.

CUHK01: The main results of CUHK01 dataset are shown in Table V. It is worth noting that similar experimental results are presented compared with VIPeR. Actually, for both single-shot and multi-shot, we perform better than the metric learning XQDA [4] by using LOMO feature, and achieve almost the same performance by using GOG feature. However, when we fuse two types of visual features, 6.5% absolute gain in terms of Rank-1 accuracy for single-shot has been achieved compared to XQDA, which beats most multi-shot metric methods. While for multi-shot, we also get a 3.4% improvement in terms of Rank-1 accuracy. The improvement is limited since our baseline for GOG feature is worse than XQDA. Compared with recent proposed methods [1], [16], [27], [29], [32] (the first group in Table V), the proposal achieves comparable performance for single visual feature, but outperforms them by a large margin for multiple visual features.

CUHK03: Since the scale of CUHK03 is relatively large, we apply principal component analysis to reduce the dimension of GOG and LOMO features to 1000D. All experimental results of CUHK03 dataset can be seen in Table IV where we compare our proposal with seven representative Re-ID methods [4], [7], [16], [17], [18], [35], [43]. XQDA [4] provides highly competent results, outperforming ours by a large margin with GOG feature. On this basis, SSM [43], as a matter of course, yields the best performance among all of the competitors by smoothing the learned metric. However, with LOMO feature, the proposed method performs slightly better than XQDA. Moreover, by fusing two kinds of visual features, our proposal obtains 66.5% Rank-1 accuracy, which is very close to 68.1% achieved by XQDA. It is worth noting that CVDCA [32] performs well on relatively small data sets, while for larger ones, its performance becomes poor. Due to the sufficient training samples, most involved deep methods, such as S-LSTM, S-CNN, are in the top performance group among all of the methods.

2) Unsupervised t-MTL Results: We compare our proposed unsupervised model with some other representative unsupervised methods on VIPeR and Market-1501.

ViPeR: In Table VI, we utilize three types of visual features (*i.e.,* LOMO, GOG and JSTL) to give a comprehensive evaluation for VIPeR dataset. Firstly, we evaluate our proposal with single feature. For the LOMO, GOG, and JSTL, we get the Rank-1 accuracy of 21.8%, 25.3% and 30.3%, respectively, which far exceeds Euclidean distance by using same visual feature. Furthermore, when we fuse the GOG and LOMO with the proposed tensor structure, a slight improvement is achieved. Specifically, we obtain 28.6% in term of Rank-1 accuracy, outperforming CAMEL [28] by 2.1%. For deep learning based feature JSTL, we also achieve highly comparable performance.

Market-1501: For Market-1501, as demonstrated in Table VII, we achieve the Rank-1 accuracy of 51.57% and mAP of 22.71% with single query, and 59.44%, and 30.75% for multiple queries. Here, we only use the training samples but without training labels, achieving an absolute improvement

TABLE VI
PERFORMANCES OF UNSUPERVISED T-MTL. MEASURED BY RANK-1 ACCURACIES FOR VIPER .

Method	Feature	Rank-1	Method	Feature	Rank-1
t-MTL	JSTL	31.8	l_2	JSTL	30.0
t-MTL	LOMO	21.8	l_2	LOMO	16.5
t-MTL	GOG	25.3	l_2	GOG	15.4
CAMEL [28]	LOMO	26.5	CAMEL [28]	JSTL	30.6
t-MTL	LOMO+GOG	28.6	t-MTL	JSTL	31.8

TABLE VII
PERFORMANCES OF UNSUPERVISED T-MTL. MEASURED BY RANK-1 ACCURACIES AND MAP FOR MARKET-1501

Market	Single Query		Multiple Query	
	Rank-1	mAP	Rank-1	mAP
JSTL [39]	43.0%	19.2%	52.9%	25.7%
PUL [11]	45.5%	20.5%	-	-
CAMEL [28]	-	-	54.5%	-
t-MTL	51.6%	22.7%	59.4%	30.8%

of 8.61% and 6.51% for Rank-1 accuracy compared with Euclidean distance [39]. We also compare the proposal with the most relevant work CAMEL [28], where almost identical experimental setting is adopted. Due to the well property possessed by tensor based multi-task regularization, we find that the proposal can benefit the shared information, leading a better performance than CAMEL [28]. Compared with the domain transfer method PUL [11], our method utilizes the unlabeled data from Markte-1501 and a joint learning feature extractor, outperforming PUL by a large margin. Fig. 6 shows some representative results produced by our unsupervised t-MTL model. Note that similar appearance persons are easier to be identified since only clustering results are used to guide the model.

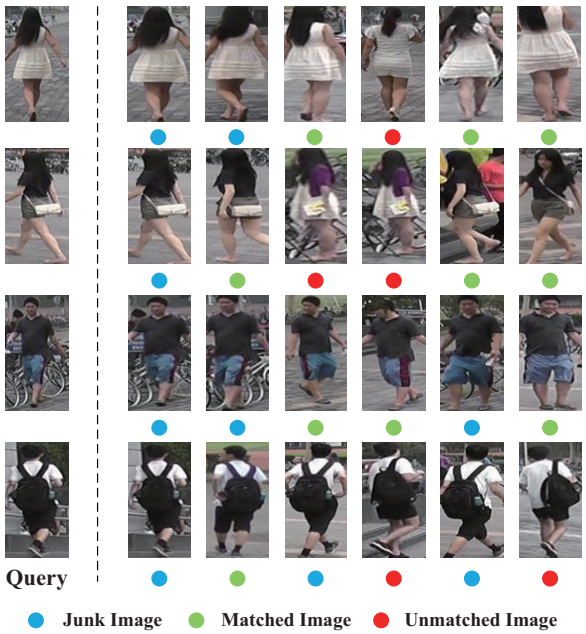


Fig. 6. Representative RE-ID results on the Market-1501 dataset produced by unsupervised t-MTL. Junk images indicate the distractors or the images that come from the same camera with the query, which is defined by [49].

C. Model Analysis

In this section, we conduct further analysis and experiments to better understand the characteristics of our t-MTL method.

1) Sensitive Analysis: First of all, we conduct a sensitivity analysis of the parameters. Two key parameters, *i.e.,* α and β, play an important role in our t-MTL approach. But most results are still much better than the baseline methods. Their values are set by cross-validation with training data. It is also worth noting that all the results are reported by random dataset spitting in avoid over-fitting.

For the supervised setting, we evaluate the impact of parameters by using different values of α and β. As shown in Fig. 7(c), the horizontal plane indicates the performance of the baseline model [7]. The values of the rank-1 accuracy of ViPeR dataset first climb, and then keeps relatively stable when α and β increase. But when they exceed the particular values, the performance begins to drop. The best performance is achieved by setting α within $[0.5, 1]$ and β within $[0.01, 0.05]$. Fig. 7(d) shows the sensitivity of parameters for CUHK01. Similar results are presented, and the plane also indicates the competitor [7]. As the scale of training identities increases, larger α is more suitable compared to ViPeR.

While for unsupervised setting, our t-MTL model fluctuates as the values of parameters changing on ViPeR, due to the random splitting. The best performance is achieved by setting $\alpha = 0.08$ and $\beta = 18$. It is noteworthy that there is no significant difference in performance as indicated in Fig 7(a). Furthermore, most values significantly outperform the baseline plane [28]. The main results of Market-1501 are shown in Fig 7(b). When the α increases, the Rank-1 accuracy firstly climbs to the peak point and then slowly decreases by fixing β. A similar observation can be found by fixing α. Also, there is a plane indicating the baseline model [39], and our proposal outperforms it by a clear margin. Empirically, α often

TABLE VIII
PERFORMANCES OF UNSUPERVISED T-MTL WHEN NUMBER OF CLUSTERS K VARIES. MEASURED BY RANK-1 ACCURACIES FOR MARKET-1501 AND VIPER .

K	500	800	1000	1500	2000
Market-1501	49.0	50.8	50.8	51.2	51.6
K	200	300	400	500	632
VIPeR	15.0	19.5	22.3	24.9	28.6

TABLE IX
PERFORMANCES OF UNSUPERVISED T-MTL WITH DIFFERENT CLUSTERING METHODS. MEASURED BY RANK-1 ACCURACIES FOR VIPER .

K	200	300	400	500	632
k-means	15.0	19.5	22.3	24.9	28.6
K	200	300	400	500	632
T-SVD-MSC	16.9	19.7	20.3	23.2	28.6

locates between 0.01 and 0.1, and $\beta = 10$ is suitable for most situations. Meanwhile, larger β shows clear physical meaning. Since there are no labels to guide the model in unsupervised setting, the symmetric model which preserves characteristics of visual features can lead a better result.

2) Deep Insights for Unsupervised t-MTL : For a comprehensive understanding of our unsupervised t-MTL, we further discuss the influence of the number of clusters w.r.t final performance, which is shown in Table VIII. It is remarked that, when the number of clusters increases, the Rank-1 accuracy of our proposal also slowly increases. In market-1501, the training set includes 750 identities and it appears that when the clusters increase to 750, the system performance keeps slightly improving and stable. The situation in VIPeR is quite different where the performance changes dramatically with the number of clusters increasing. This is a strange phenomenon, since it is expected that an adequate number of clusters will achieve better performance, but according to more recent papers, such as [53,54], their experiments show that instance discrimination is very effective in the unsupervised learning setting. They regard each sample as an independent category, and try to train a model to separate each of them. Their idea is simple that a typical discriminative learning method can learn visual data themselves rather than semantic annotations. In this perspective, under weak supervision, a finer division will benefit from the instance discrimination, and help model identify hard sample which is far away from the cluster center. Our experimental result is very consistent with theirs (*i.e.*, on ViPeR, taking 632 clusters, which is the number of training samples, achieves the best performance). Meanwhile, two reasonable regularizers, which can be considered as prior knowledge, help the model achieve better generalization ability. The same result is also presented in CAMEL [32]. Hence, it deduces that the more cluster centers will be beneficial.

More importantly, in CAMEL, an iterative generation of the virtual label is necessary for boosting the performance. But in our experiments, we find this process is nonsense for our model. As shown in Fig. 8, the Rank-1 accuracy reaches the peak at first several iterations, and slowly decreases until it is unchanged. We also conduct extra experiments on ViPeR dataset and it appears that with the increasing iterations, the performance first keeps stable and then slightly decreases regardless of the number of cluster centers. To in-depth analyze this abnormal phenomenon, we conduct another experiment on Market-1501. We select first 25 training samples from the same identity and observe the changes in their pseudo labels at each iteration. We find that after the first iteration, more samples are grouped in the same category, which indicates that the projected space becomes more discriminative. And with increasing iterations, the virtual labels change slightly and even give wrong supervision, where samples with the same virtual label are still assigned to the same cluster center with high probability at the next iteration. It deduces that the weakly supervised information can not further help our model improve the performance.

We also conduct experiments to present the influence of different kinds of clustering algorithms. To do that, two kinds of clustering methods, *i.e.*, k-means and t-SVD-MSC [21], are employed. Since the t-SVD-MSC, which achieves the promising performance for most clustering tasks, is designed for multiple visual features, we take GOG and LOMO as inputs. The results are shown in Table IX. With the same parameter setting, we find that the k-means performs as well as T-SVD-MSC for a various number of clusters K. It deduces that the clustering algorithm does not play a critical role in our proposal, which is attribute to the well-founded prior knowledge provided by two reasonable regularizers in our model.

3) Algorithm Complexity and Convergence: Although the optimization procedure seems complicated, as discussed above, the whole procedure only performs once at off-line training time. The bottleneck of our method is to solve the subproblem $\boldsymbol{\mathcal{G}}$, but it equals to calculate $(V-1)/2$ matrix SVD according to [48], whose dimension is $d \times C$. This special structure can be easily parallelized, and would be invested in our future work. In summary, it takes $\mathcal{O}(2dCV\log(V))$ for calculating the FFT and its inverse. And for each matrix, it takes $\mathcal{O}(\min(C^2d, d^2C))$ for calculating the SVD.

As a result, by considering V cameras, the complexity of subproblem $\boldsymbol{\mathcal{G}}$ is $\mathcal{O}(\min(VC^2d, Vd^2C))$ in each iteration. Since $min(d, C) \gg \log(V)$, the complexity of our t-MTL method is:

$$\mathcal{O}(\min(KVC^2d, KVd^2C))), \tag{24}$$

where K means the iteration number. In practice, K usually locates within $30 \sim 50$, and we set 30 for all the experiments, empirically. For real-world applications, it appears that the number of person identities plays a critical role in scalability. However, it is easy to handle tens of thousands identities, and a classifier in deep learning methods also encounters the same dilemma when scaling to hundreds of thousands. We also conduct experiments to present the execution time shown in Table X. Note that the off-line cost is increased

TABLE X
QUANTITATIVE ANALYSIS OF EXECUTION TIME FOR OUR SUPERVISED ALGORITHM.

Dataset	ViPeR	CUHK01	CUHK03	Market-1501
Time	7.2s	101.3s	52.4s	1267.3s

especially on larger datasets such as Market-1501, but we can still train it around 20 minutes. Meanwhile, the execution time of CUHK01 is almost twice than it of CUHK03 since we apply PCA on CUHK03.

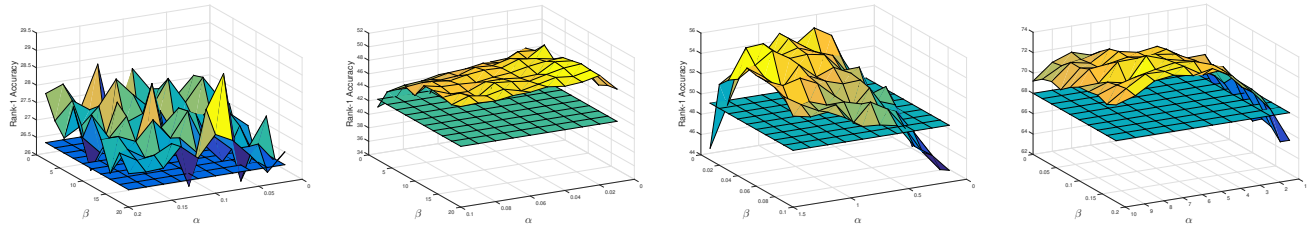
Additionally, the optimality gap produced in each iteration of our algorithm is monotonically decreasing and our subproblems are solved exactly. Thus we have:

$$||\boldsymbol{\mathcal{U}}_{t+1} - \boldsymbol{\mathcal{U}}_t||_F \to 0. \tag{25}$$

Hence, the convergence of our optimization can be indicated by the following criterion:

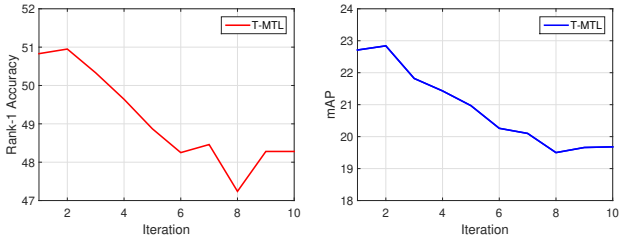
$$\text{Match Error} = ||\boldsymbol{\mathcal{U}} - \boldsymbol{\mathcal{G}}||_\infty. \tag{26}$$

Actually, our method converges fast in reality, as it is illustrated in Fig. 9, where the curve records the values of the Match Error (defined in Eq. (26)) in each iteration step.



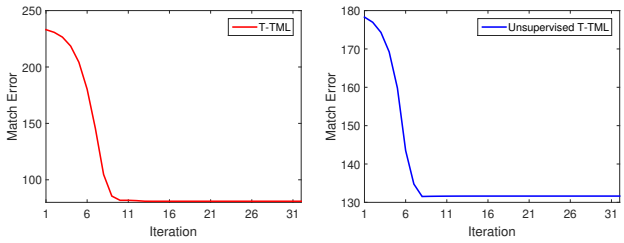
(a) (b) (c) (d)

Figure 7. Influence of parameter α and β in terms of Rank-1 accuracy. Fig. 7(a) and Fig. 7(b) show unsupervised results for ViPeR and Markte-1501. Fig. 7(c) and and Fig. 7(d) show supervised results for ViPeR and CUHK01.



(a) (b)

Figure 8. Influence of iteration number on Rank-1 accuracy and mAP for Markte-1501 by setting K=800



(a) (b)

Figure 9. Convergence Curve on ViPeR Dataset

4) Limitation: The main limitation of the proposed method is that the view-specific model is built upon a not very practical problem setup. This setup assumes that training and testing person images will be from the same cameras. However, in practice, an idea model trained from the camera network in location A should generalize to the camera network in location B. But for the moment, most existing successful models also do not have such transferability. In the future, we will further investigate this issue.

VI. Conclusion and Discussion

In this paper, to conquer the view-specific discrepancy problem, a tensor multi-task model is proposed to perform person Re-ID. To explore the high order correlations among cameras and persons, the proposal constrains the camera-specific classifiers through tensor multi-rank and Bregman discrepancy. Then, people identifying across cameras has been formulated in a unified multi-task classification framework, where an efficient algorithm is introduced to achieve the optimal solution. The proposed t-MTL also adopts a clustering procedure to assign a virtual label to each training samples, which significantly improves the baseline by using training samples but without labels. To extend our model to involving multiple visual features, the results show that t-MTL is very competitive with the recently proposed approach for both supervised and unsupervised setting.

There are still some issues in the proposed model that can be further improved. The most critical problem is the adaptive parameter learning, as it is discussed in Section V-C1. However, due to the domain gap and scale of datasets, the parameters need to be fine-tuned for individual datasets. Furthermore, we even can not obtain parameters via the cross-validation in the unsupervised setting. Hence, an adaptive approach is needed urgently. Another important issue is to utilize different deep structures to extend asymmetric distance learning in end-to-end training style. Moreover, future work will also include how to scale if data continuously arrives.

VII. Acknowledgments

The authors would like to thank editor and anonymous reviewers who gave valuable suggestion that has helped to improve the quality of the paper. The authors are also thankful for the financial support from the National Key R&D Program of China (No.2017YFC0803700), the National Natural Science Foundation of China (U1636220, 61432008, 61472423 and 61772524), and the Beijing Municipal Natural Science Foundation (4182067), and partly by the Fundamental Research Funds for the Central Universities associated with Shanghai Key Laboratory of Trustworthy Computing. Wensheng Zhang and Qi Tian are the corresponding authors.

References

[1] W. Chen, X. Chen, J. Zhang, K. Huang. A Multi-Task Deep Neural Networ for Person Re-identification. in *AAAI*, 2017.

[2] M. Li, F. Shen, J. Wang, C. Guan, J. Tang. Person Re-identification with Activity Prediction Based on Hierarchical Spatial-temporal Model. *Neurocomputing* , vol. 275, pp. 1200-1207, 2018

[3] H. Yao, S. Zhang, Y. zhang, J. Li, Q. Tian. Deep Representation learning with Part Loss for Person Re-identification. *arXiv preprint arXiv:1707.00798,* 2017.

[4] S. Liao, Y. Hu, X. Zhu, S. Z. Li. Person Re-identification by Local Maximal Occurrence Representation and Metric Learning. *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[5] Y. Chen, X. Zhu, W. Zheng, J. Lai. Person Re-Identification by Camera Correlation Aware Feature Augmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 40, no. 2, 2018.

[6] P. Peng, T. Xiao, Y. Wang, M. Pontil, S. Gang, T. Huang, Y. Tian. Unsupervised Cross-Dataset Transfer Learning for Person Re-identification. *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[7] T. Matsukawa, T. Okabe, E. Suzuki, Y. Sato. Hierarchical Gaussian Descriptor for Person Re-identification. *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[8] Z. Zhang, Y. Xie, W. Zhang, Q. Tian. Effective Image Retrieval via Multilinear Multi-index Fusion. *arXiv preprint arXiv:1709.09304,* 2017.

[9] D. Chen, Z. Yuan, B. Chen, and N. Zheng. Similarity learning with spatial constraints for person re-identification. *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[10] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian. Deep attributes driven multi-camera person re-identification. *European Conference on Computer Vision*, 2016.

[11] H. Fan, L. Zheng, Y. Yang. Unsupervised Person Re-identification: Clustering and Fine-tuning.*arXiv preprint arXiv:1705.10444*, 2017.

[12] H. Liu, J. Feng, M. Qi, J. Jiang, S. Yan. End-to-End Comparative Attention Networks for Person Re-Identification. *IEEE Trans. on Image Processing*, vol. 26, no. 7, pp. 3492-3506, 2017.

[13] L. Ma, X. Yang, D. Tao. Person Re-Identification Over Camera Networks Using Multi-Task Distance Metric Learning. *IEEE Trans. on Image Processing*, vol. 23, no. 8, pp. 3656-3670, 2014.

[14] Y. Sun, L. Zheng, W. Deng, S. Wang. Svdnet for Pedestrian Retrieval. *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[15] Z. Zhong, L. Zheng, D. Cao, S. Li. Re-ranking Person Re-identification with k-reciprocal Encoding. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[16] L. Zhang, T. Xiang, S. Gong. Learning a Discriminative Null Space for Person Re-identification. *IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

[17] R. R. Varior, B. Shuai, J. Lu, D. Xu, G. Wang. A Siamese Long Short-Term Memory Architecture for Human Re-identification. *European Conference on Computer Vision*, 2016.

[18] R. R. Varior, B. Shuai, M. Haloi, G. Wang. Gated Siamese Convolutional Neural Network Architecture for Human Reidentification. *European Conference on Computer Vision*, 2016.

[19] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian. Scalable Person Re-identification: A Benchmark. *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[20] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell. Caffe: Convolutional Architecture for Fast Feature Embedding. *Proceedings of ACM international conference on Multimedia*, 2014.

[21] Y. Xie, D. Tao, W. Zhang, L. Zhang. Multi-View Subspace Clustering via Relaxed L_1-Norm of Tensor Multi-Rank. *arXiv preprint arXiv:1610.07126*, 2016.

[22] Z. Zhang, G. Ely, S. Aeron, N. Hao, and M. Kilmer, Novel Methods for Multilinear Data Completion and De-noising based on Tensor-SVD. *Proceedings of IEEE Computer Vision and Pattern Recognition*, 2014.

[23] Z. Lin, M. Chen, Y. Ma. The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-rank Matrices. *Technical Report UILUENG-09-2215*, UIUC, 2009.

[24] O. Semerci, N. Hao, M. E. Kilmer, E. L. Miller. Tensor-based Formulation and Nuclear Norm Regularization for Multienergy Computed Tomography. *IEEE Trans. on Image Processing*, vol 24, no.4, pp.1678-1693, 2014.

[25] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, P. H. S. Torr. End-to-End Representation Learning for Correlation Filter Based Tracking. *Proceedings of IEEE Computer Vision and Pattern Recognition*, 2017.

[26] A. Krizhevsky, I. Sutskever, G. E. Hinton. Imagenet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 2012.

[27] X. Zhao, N. Wang, Y. Zhang, S. Du, Y. Gao, J. Sun. Beyond Pairwise Matching: Person Reidentification via High-Order Relevance Learning. *IEEE Trans. on Neural Networks and Learning System*, doi=10.1109/TNNLS.2017.2736640, 2017.

[28] H. Yu, A. Wu, W. Zheng. Cross-view Asymmetric Metric Learning for Unsupervised Person Re-identification. *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[29] S. Paisitkriangkrai, C. Shen, A. Hengel. Learning to Rank in Rerson Re-identification with Metric Ensembles. *Proceedings of IEEE Computer Vision and Pattern Recognition*, 2015.

[30] M. Geng, Y. Wang, T. Xiang, Y. Tian. Deep Transfer Learning for Person Re-identification. *arXiv preprint arXiv:1611.05244*, 2016.

[31] M. Hirzer, P. Wohlhart, P. M. Roth, H. Bischof. Large Scale Metric Learning from Equivalence Constraints. *Proceedings of IEEE Computer Vision and Pattern Recognition*, 2012.

[32] Y. Chen, W. Zheng, J. Lai, P. C. Yuen. An Asymmetric Distance Model for Cross-View Feature Mapping in Person Reidentification. *IEEE Trans. on Circuits and System for Video Technology*, vol. 27, No. 8, 2017.

[33] D. Gary, S. Brennan, H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. *International Workshop on Performance Evaluation for Tracking and Surveillance*, 2007.

[34] W. Li, R. Zhao, X. Wang. Human reidentification with transferred metric learning. *Asian Conference on Computer Vision*, 2012.

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

[36] S. Bak, P. Carr. One-Shot Metric Learning for Person Re-identification. *Proceedings of IEEE Computer Vision and Pattern Recognition*, 2017.

[37] E. Kodirov, T. Xiang, Z. Fu, S. Gang. Person Re-Identification by Unsupervised l_1 Graph Learning. *European Conference on Computer Vision*, 2016.

[38] G. Lisanti, N. Martinel, A. D. Bimbo, G. L. Foresti. Group Re-Identification via Unsupervised Transfer of Sparse Features Encoding. *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[39] T. Xiao, H. Li, W. Ouyang, X. Wang. Learning Deep Feature Representations with Domain Guided Dropout for Person Re-identification. *Proceedings of IEEE Computer Vision and Pattern Recognition*, 2016.

[40] A. Hermans, L. Beyer, B. Leibe. In Defense of the Triplet Loss for Person Re-Identification. *arXiv preprint arXiv:1703.07737*, 2017.

[41] D. Cheng, X. Chang, L. Liu. Dictionary Learning With Ranking Metric Embedded for Person Re-Identification. *Proceedings of International Joint Conference on Artificial Intelligence*, 2017.

[42] A. Wu, W. Zheng, J. Lai. Robust Depth-Based Person Re-Identification. *IEEE Trans. on Image Processing*, vol. 26, no. 6, 2017.

[43] S. Bai, X. Bai, Q. Tian. Scalable Person Re-identification on Supervised Smoothed Manifold. *Proceedings of IEEE Computer Vision and Pattern Recognition*, 2017.

[44] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, W. Gao. Multi-Task Learning with Low Rank Attribute Embedding for Person Re-Identification. *IEEE International Conference on Computer Vision*, 2015.

[45] F. Xiong, M. Gou, O. Camps, M. Sznaier. Person Re-Identification using Kernel-based Metric Learning Methods. *European Conference on Computer Vision*, 2014.

[46] Y. Sun, L. Zheng, W. Deng, S. Wang. SVDNet for Pedestrian Retrieval. *IEEE International Conference on Computer Vision*, 2017.

[47] F. Jurie, A. Mignon. PCCA: A New Approach for Distance Learning from Sparse Pairwise Constraints. *Proceedings of IEEE Computer Vision and Pattern Recognition*, 2012.

[48] M. E. Kilmer, K. Braman, N. Hao, R. C. Hoover. Third-order Tensors as Operators on Matrices: A Theoretical and Computational Framework with Applications in Imaging. *SIAM Journal on Matrix Analysis and Applications*, vol 34, no. 1, pp. 148-172, 2013.

[49] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian. Scalable Person Re-identification: A Benchmark. *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[50] Z. Lin, R. Liu, Z. Su. Linearized alternating direction method with adaptive penalty for low-rank representation. *In Advances in Neural Information Processing Systems*, pp. 612-620, 2011.

[51] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, vol 3, no. 1, pp. 1-122, 2011.

[52] L. Zheng, Y. Yang, A. Hauptmann. Person Re-identification: Past, Present and Future. *arXiv preprint arXiv:1610.02984*, 2016.

[53] Z. Wu, Y. Xiong, S. X. Yu, D. Lin. Unsupervised Feature Learning via Non-Parametric Instance Discrimination. *Proceedings of IEEE Computer Vision and Pattern Recognition*, 2018.

[54] Z. Zhong, L. Zheng, Z. Luo, S. Li, Y. Yang. Invariance Matters: Exemplar Memory for Domain Adaptive Person Re-identification. *Proceedings of IEEE Computer Vision and Pattern Recognition*, 2019.

Zhizhong Zhang is currently a Ph.D. candidate at the Institute of Automation, Chinese Academy of Sciences (CAS). He received his B.S. degree from the Department of Mathematics, Northeastern University, Shenyang, Liaoning, China, in 2014. His research interests include machine learning and computer vision.

Yuan Xie (M'12) received the Ph.D. degree in Pattern Recognition and Intelligent Systems from the Institute of Automation, Chinese Academy of Sciences (CAS), in 2013. He is currently a full professor with the School of Computer Science and Software Engineering, East China Normal University. His research interests include image processing, computer vision, machine learning and pattern recognition. He has published more than 35 papers in major international journals and conferences including the IJCV, IEEE TPAMI, TIP, TNNLS, TCYB, TCSVT, TGRS, TMM, and NIPS, CVPR, ECCV, etc. He also has served as a reviewer for more than 15 journals and conferences. Dr.Xie received the Hong Kong Scholar Award from the Society of Hong Kong Scholars and the China National Postdoctoral Council in 2014.

YongQiang Tang received the PhD degree from Institute of Automation, Chinese Academy of Sciences (CAS), in 2019. He is currently an assistant professor in the Institute of Automation, CAS. His research interests include computer vision, machine learning and data mining.

Wensheng Zhang received his Ph.D. degree in Pattern Recognition and Intelligent Systems from the Institute of Automation, Chinese Academy of Sciences (CAS), in 2000. He joined the Institute of Software, CAS, in 2001. He is a Professor of Machine Learning and Data Mining and the Director of Research and Development Department, Institute of Automation, CAS. His research interests include computer vision, pattern recognition and artificial intelligence.

Qi Tian is currently a Chief Scientist in Computer Vision at Huawei Noah's Ark Lab. He is also on faculty leave and a Full Professor in the Department of Computer Science, the University of Texas at San Antonio (UTSA). During 2008-2009, he took one-year Faculty Leave at Microsoft Research Asia (MSRA).

Dr. Tian received his Ph.D. in ECE from University of Illinois at Urbana-Champaign (UIUC) and received his B.E. in Electronic Engineering from Tsinghua University and M.S. in ECE from Drexel University, respectively. Dr. Tian's research interests include computer vision, multimedia information retrieval and machine learning and published over 440 refereed journal and conference papers. His Google citation is over 11500 with H-index 58. He was the co-author of best papers including ACM ICMR 2015, PCM 2013, MMM 2013, ACM ICIMCS 2012, a Top 10% Paper Award in MMSP 2011, a Student Contest Paper in ICASSP 2006, and co-author of a Best Paper/Student Paper Candidate in ICME 2015 and PCM 2007.

Dr. Tian research projects are funded by ARO, NSF, DHS, Google, FXPAL, NEC, SALSI, CIAS, Akiira Media Systems, HP, Blippar and UTSA. He received 2017 UTSA President's Distinguished Award for Research Achievement, 2016 UTSA Innovation Award, 2014 Research Achievement Awards from College of Science, UTSA, 2010 Google Faculty Award, and 2010 ACM Service Award. He is the associate editor of IEEE TMM, IEEE TCSVT, ACM TOMM, MMSJ, and in the Editorial Board of Journal of Multimedia (JMM) and Journal of MVA. Dr. Tian is the Guest Editor of IEEE TMM, Journal of CVIU, etc. Dr. Tian is a Fellow of IEEE.