Unsupervised Cross-dataset Person Re-identification by Transfer Learning of Spatial-Temporal Patterns

Jianming Lv[1], Weihang Chen[1], Qing Li[2], and Can Yang[1]

[1]South China University of Technology
[2]City University of Hongkongy
{jmlv,cscyang}@scut.edu.cn,csscut@mail.scut.edu.cn,qing.li@cityu.edu.hk

Abstract

Most of the proposed person re-identification algorithms conduct supervised training and testing on single labeled datasets with small size, so directly deploying these trained models to a large-scale real-world camera network may lead to poor performance due to underfitting. It is challenging to incrementally optimize the models by using the abundant unlabeled data collected from the target domain. To address this challenge, we propose an unsupervised incremental learning algorithm, TFusion*, which is aided by the transfer learning of the pedestrians' spatio-temporal patterns in the target domain. Specifically, the algorithm firstly transfers the visual classifier trained from small labeled source dataset to the unlabeled target dataset so as to learn the pedestrians' spatial-temporal patterns. Secondly, a Bayesian fusion model is proposed to combine the learned spatio-temporal patterns with visual features to achieve a significantly improved classifier. Finally, we propose a learning-to-rank based mutual promotion procedure to incrementally optimize the classifiers based on the unlabeled data in the target domain. Comprehensive experiments based on multiple real surveillance datasets are conducted, and the results show that our algorithm gains significant improvement compared with the state-of-art cross-dataset unsupervised person re-identification algorithms.*

1. Introduction

As one of the most challenging and well studied problem in the field of surveillance video analysis, person re-identification (Re-ID) aims to match the image frames which contain the same pedestrian in surveillance videos. The core of these algorithms is to learn the pedestrian features and the similarity measurements, which are view invariant and robust to the change of cameras.
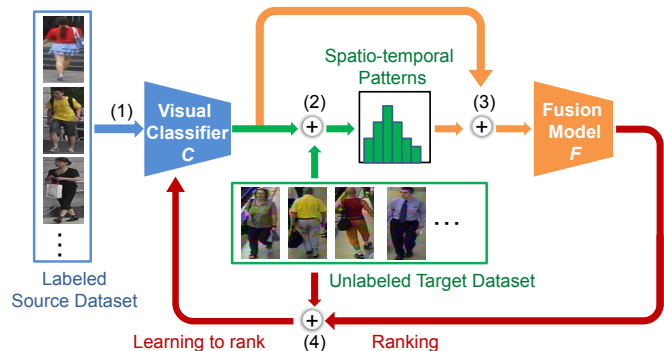

Figure 1: The *TFusion* model consists of 4 steps: (1) Train the visual classifier $\mathcal{C}$ in the labeled source dataset (Section 4.2); (2) Using $\mathcal{C}$ to learn the pedestrians' spatio-temporal patterns in the unlabeled target dataset (Section 4.3); (3) Construct the fusion model $\mathcal{F}$ (Section 4.4); (4) Incrementally optimize $\mathcal{C}$ by using the ranking results of $\mathcal{F}$ in the unlabeled target dataset (Section 4.6).

Most of the proposed algorithms [1][3] [30][14] [20] [24] conduct supervised learning on the labeled datasets with small size. Directly deploying these trained models to the real-world environment with large-scale camera networks can lead to poor performance, because the target domain may be significantly different from the small training dataset. Thus the incremental optimization in real-world deployment is critical to improve the performance of the Re-ID algorithms. However, it is usually expensive and impractical to label the massive online surveillance videos to support supervised learning. How to leverage the abundant unlabeled data is a practical and extremely challenging problem.

To address this problem, some unsupervised algorithms [13] [18] [27] are proposed to extract view invariant features and to measure the similarity of different images in unlabeled datasets. Without powerful supervised tuning and optimiza-

tion, the performance of above unsupervised algorithms is typically poor. Besides these unsupervised methods applied in a single dataset, a cross-dataset unsupervised transfer learning algorithm[21] is proposed recently, which transfers the view-invariant representation of a person's appearance from a source labeled dataset to another unlabeled target dataset by a dictionary learning mechanism, and gains much better performance. However, the performance of the above mentioned algorithms are still much weaker than the supervised learning algorithms. For example, in the CUHK01 [28] dataset, the unsupervised transfer learning algorithm [21] achieves 27.1% rank-1 accuracy, while the accuracy of the state-of-art supervised algorithm [25] can reach to 67% .

In this paper, we propose a novel unsupervised transfer learning algorithm, named *TFusion*, to enable high performance Re-ID in unlabeled target datasets. Different from the above algorithms which are only based on visual features, we try to learn and integrate with the pedestrians' spatio-temporal patterns in the steps shown in Fig. 1. Firstly, we transfer the visual classifier $\mathcal{C}$, which is trained from a small labeled source dataset, to learn the pedestrians' spatio-temporal patterns in the unlabeled target dataset. Secondly, a Bayesian fusion model is proposed to combine the learned spatio-temporal patterns with visual features to achieve a significantly improved fusion classifier $\mathcal{F}$ for Re-ID in the target dataset. Finally, a learning-to-rank scheme is proposed to further optimize the classifiers based on the unlabeled data. During the iterative optimization procedure, both of the visual classifier $\mathcal{C}$ and the fusion classifier $\mathcal{F}$ are updated in a mutual promotion way.

The comprehensive experiments based on real datasets (VIPeR [6], GRID [2], CUHK01 [28] and Market1501 [36]) show that *TFusion* outperforms the state-of-art cross-dataset unsupervised transfer algorithm [21] by a big margin, and can achieve comparable or even better performance than the state-of-art supervised algorithms using the same datasets.

This paper includes the following contributions:

- We present a novel method to learn pedestrians' spatio-temporal patterns in unlabeled target datsets by transferring the visual classifier from the source dataset. The algorithm does not require any prior knowledge about the spatial distribution of cameras nor any assumption about how people move in the target environment.

- We propose a Bayesian fusion model, which combines the spatio-temporal patterns learned and the visual features to achieve high performance of person Re-ID in the unlabeled target datasets.

- We propose a learning-to-rank based mutual promotion procedure, which uses the fusion classifier to teach the weaker visual classifier by the ranking results on unlabeled dataset. This mutual learning mechanism can be applied to many domain adaptation problems.

2. Related Work

Supervised Learning: Most existing person Re-ID models are supervised, and based on either invariant feature learning [7] [14] [35] [31] , metric learning [11][20] [24] [15] or deep learning [1] [3] [30] . However, in the practical deployment of Re-ID algorithms in large-scale camera networks, it is usually costly and unpractical to label the massive online surveillance videos to support supervised learning as mentioned in [21].

Unsupervised Learning: In order to improve the effectiveness of the Re-ID algorithms towards large-scale unlabeled datasets, some unsupervised Re-ID methods [34][26] [13] [18] [27] are proposed to learn cross-view identity-specific information from unlabeled datasets. However, due to the lack of the knowledge about identity labels, these unsupervised approaches usually yield much weaker performance compared to supervised learning approaches.

Transfer Learning: Recently, some cross-dataset transfer learning algorithms[17] [16][21][12] are proposed to leverage the Re-ID models pre-trained in other labeled datasets to improve the performance on target dataset. This type of Re-ID algorithms can be classified further into two categories: supervised transfer learning and unsupervised transfer learning according to whether the label information of target dataset is given or not. Specifically, in the **supervised transfer learning** algorithms [12] [17] [16], both of the source and target datasets are labeled or have weak labels. [12] is based on a SVM multi-kernel learning transfer strategy, and [16] is based on cross-domain ranking SVMs. [17] adopts multi-task metric learning models. On the other hand, the recently proposed cross-dataset **unsupervised transfer learning** algorithm for Re-ID, UMDL[21], is totally different from above algorithms, and closer to real-world deployment environment where the target dataset is totally unlabeled. UMDL[21] transfers the view-invariant representation of a person's appearance from the source labeled dataset to the unlabeled target dataset by dictionary learning mechanisms, and gains much better performance. Although this kind of cross-dataset transfering algorithms are proved to outperform the purely unsupervised algorithms, they still have a long way to catch up the performance of the supervised algorithms, e.g. in the CUHK01[28] dataset, UMDL [21] can achieve 27.1% rank-1 accuracy, while the accuracy of the state-of-art supervised algorithms [25] can reach 67%.

Besides the person Re-ID algorithms only based on visual features, some recent research works focus on **using the spatio-temporal constraint** in camera networks to improve the Re-ID precision. [9] considers the distance of cameras and filters the candidates with less possibility. [19] models the connection of any pair of cameras by measuring the average similarity score of the images from different cameras, and applies the relationship of cameras to filter the candidates with low probability. [10] makes statistics about

the temporal distribution of pedestrians' transferring among different cameras. All of these algorithms are designed on one single labeled dataset, while our model is adaptive to a cross-dataset transferring learning scenario where the target dataset is totally unlabeled. On the other hand, in above algorithms, the spatio-temporal patterns are learned independent of the visual classifier, and keep fixed at the initialization step. In this paper, we address that the visual classifier and the spatio-temporal patterns can be linked together to conduct an iterative co-train procedure to promote each other.

3. Preliminaries

3.1. Problem Definition of Person Re-ID

Given a surveillance image containing a target pedestrian, the design goal of a person Re-ID algorithms is to retrieve the surveillance videos for the image frames which contain the same person. For clarity of the problem definition, some notations describing Re-ID are introduced in this section.

Each surveillance image containing a pedestrian is denoted as S_i, which is cropped from an image frame of a surveillance video. The time when S_i is taken is denoted by t_i, and the ID of the corresponding camera is denoted by c_i. The ID of the pedestrian in S_i is denoted as $\Upsilon(S_i)$. Given any surveillance image S_i, the person Re-ID problem is to retrieve the images $\{S_j | \Upsilon(S_j) = \Upsilon(S_i)\}$, which contain the same person $\Upsilon(S_i)$.

The traditional strategy of person Re-ID is to train a classifier $\mathcal{C}$ based on visual features to judge whether two given images contain the same person. Given two images S_i and S_j, if $\mathcal{C}$ judges that S_i and S_j contain a same person, it is denoted as $S_i \Vdash_{\mathcal{C}} S_j$. Otherwise, it is denoted as $S_i \nVdash_{\mathcal{C}} S_j$.

The false positive error rate of the classifier $\mathcal{C}$ is given by:

$$E_p = Pr(\Upsilon(S_i) \neq \Upsilon(S_j) | S_i \Vdash_{\mathcal{C}} S_j) \qquad (1)$$

The false negative error rate of C_s is given by:

$$E_n = Pr(\Upsilon(S_i) = \Upsilon(S_j) | S_i \nVdash_{\mathcal{C}} S_j) \qquad (2)$$

3.2. Cross-Dataset Person Re-ID

Like most of the traditional person Re-ID algorithms [14] [35], we can conduct supervised learning on some public labeled dataset (denoted as Ω_s below), which is usually of small size, to train a classifier $\mathcal{C}$. While directly deploying the trained $\mathcal{C}$ to a real-world unlabeled target dataset Ω_t collected from a large-scale camera network, it tends to have poor performance, due to the significant difference between Ω_s and Ω_t.

How to effectively transfer the classifier trained in a labeled source dataset to another unlabeled target datset is the fundamental challenging problem addressed in this paper.

4. Model

4.1. Model overview

Because most of the time people move with definite purposes, their trajectories usually follow some non-random patterns, which can be utilized as important clues besides visual features to discriminate different persons. Motivated by this observation, we propose a novel algorithm to transfer the classifier, which is trained in a small source dataset, to learn the spatio-temporal patterns of pedestrians in the unlabeled target dataset. Then we combine the patterns with the visual features to build a more precise fusion classifier. Furthermore, we adopt a learning-to-rank scheme to incrementally optimize the classifier by using the unlabeled data in the target dataset. The architecture of the model is illustrated in Fig. 1, which contains the following main steps:

- **step (1): Supervised Learning in the Labeled Source Dataset**. In this warm-up initialization step, we adopt the supervised learning algorithm such as [37] to learn a visual classier $\mathcal{C}$ from an available small labeled source dataset. In the following steps, further optimization is needed for $\mathcal{C}$ to be applied in a large unlabeled target dataset. (Section 4.2)

- **step (2): Transfer Learning of the Spatio-temporal Pattern in the Unlabeled Target Dataset**. In this step, we transfer the classifier $\mathcal{C}$ to the unlabeled target dataset to learn pedestrians' spatio-temporal patterns in the target domain. (Section 4.3)

- **step (3): Fusion Model for the Target Dataset**. A Bayesian fusion model $\mathcal{F}$ is proposed to combine the visual classifier $\mathcal{C}$ and the newly learned spatio-temporal patterns for precise discrimination of pedestrian images. (Section 4.4)

- **step (4): Learning-to-rank Scheme for Incremental Optimization of Classifiers**. In this step, we leverage the fusion model $\mathcal{F}$ to further optimize the visual classifier $\mathcal{C}$ based on the learning-to-rank scheme. Firstly, given any surveillance image S_i, the fusion model $\mathcal{F}$ is applied to rank the images in the unlabeled target dataset according to the similarity with S_i. Secondly, the ranking results are fed back to incrementally train the visual classifier $\mathcal{C}$. (Section 4.6)

The model can be iteratively updated by repeating step (2) $\sim$ (4) until the number of iterations reaches a given threshold or the performance of the classifier converges. In this way, all of the visual classifier $\mathcal{C}$, the fusion model $\mathcal{F}$, and the spatio-temporal patterns can achieve collaborative optimization.

In the following sections, we will propose the detailed design and analysis of each key component of the model.
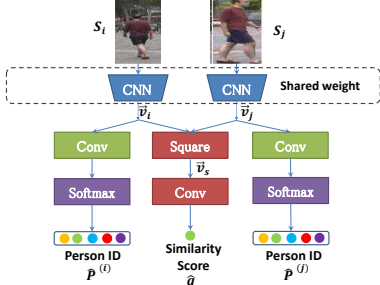
Figure 2: Visual classifier based on CNN.

4.2. Supervised Learning in Labeled Source Dataset

As shown in step (1) of Fig. 1, the supervised learning is conducted on the labeled source dataset to train the visual classifier $\mathcal{C}$, which measures the matching probability of the given two input images.

We select the recently proposed convolutional siamese network [37] as $\mathcal{C}$, which makes better use of the label information and has good performance in large-scale datasets such as Market1501[36]. The network architecture of $\mathcal{C}$ is shown in Fig. 2. The network adopts a siamese scheme including two ImageNet pre-trained CNN modules, which share same weight parameters and extract visual features from the input images S_i and S_j. The CNN module is achieved from the ResNet-50 network [8] by removing its final fully-connected (FC) layer. The outputs of the two CNN modules are flattened into two one-dimensional vectors: $\vec{v_i}$ and $\vec{v_j}$, which act as the embedding visual feature vectors of the input images. Finally, the model predicts the identities ($\hat{P}^{(i)}$ and $\hat{P}^{(j)}$) of the input images, and their similarity score $\hat{q}$. The cross entropy based verification loss and identification loss are adopted for training. Readers can refer to [37] or our appendix for the detail of the network.

While deploying this classifier to perform Re-ID, given two images S_i and S_j as input, the CNN modules extract their visual feature vectors $\vec{v_i}$ and $\vec{v_j}$ as shown in Fig. 2. The matching probability of S_i and S_j is measured as the cosine similarity of the two feature vectors:

$$Pr(S_i \Vdash_{\mathcal{C}} S_j | \vec{v_i}, \vec{v_j}) = \frac{\vec{v_i} \cdot \vec{v_j}}{\parallel \vec{v_i} \parallel_2 \parallel \vec{v_j} \parallel_2} \qquad (3)$$

If $Pr(S_i \Vdash_{\mathcal{C}} S_j | \vec{v_i}, \vec{v_j})$ is larger than a predefined threshold constant, S_i and S_j are judged to contain the same person. That is $S_i \Vdash_{\mathcal{C}} S_j$. Otherwise, they are judged as $S_i \nVdash_{\mathcal{C}} S_j$.

4.3. Spatio-temporal Pattern Learning

As reported in [10], due to the camera network topology, the time interval of pedestrians' transferring among different cameras usually follows specific patterns. These spatio-temporal patterns can provide non-visual clues for Re-ID.

Formally, the spatio-temporal pattern of pedestrians' transferring among different cameras can be defined as:

$$Pr(\triangle_{ij}, c_i, c_j | \Upsilon(S_i) = \Upsilon(S_j)). \qquad (4)$$

Here S_i is a surveillance image taken at the camera c_i at the time t_i, and S_j is another one at the camera c_j at the time t_j. $\triangle_{ij} = t_j - t_i$. Eq.(4) indicates the probability distribution of the time interval $\triangle_{ij}$ and camera IDs (c_i, c_j) of any pair of image frames S_i and S_j containing the same person ($\Upsilon(S_i) = \Upsilon(S_j)$).

To calculate the precise value of Eq.(4), it is needed to judge whether two images contain the same person firstly. However, this is impossible in unlabeled target datasets where person IDs are unknown. As shown in the step (2) of Fig. 1, we propose an approximation solution by transferring the visual classifier $\mathcal{C}$, which is trained in the labeled source dataset, to the unlabeled target dataset. With $\mathcal{C}$, we can make a rough judgment of any pair of images S_i and S_j to achieve the identification result $S_i \Vdash_{\mathcal{C}} S_j$ or $S_i \nVdash_{\mathcal{C}} S_j$. After applying $\mathcal{C}$ to every pair of images in the target dataset, we can obtain the statistics $Pr(\triangle_{ij}, c_i, c_j | S_i \Vdash_{\mathcal{C}} S_j)$, which indicates the probability distribution of the time interval and camera IDs of any pair of images which seem to contain the same person ($S_i \Vdash_{\mathcal{C}} S_j$). On the other hand, we can apply $\mathcal{C}$ to every pair of images in the target dataset to obtain the statistics $Pr(\triangle_{ij}, c_i, c_j | S_i \nVdash_{\mathcal{C}} S_j)$, which indicates the probability distribution of the time interval and camera IDs of any pair of images which seem to contain different persons. We can infer that:

$$\begin{aligned} &Pr(\triangle_{ij}, c_i, c_j | \Upsilon(S_i) = \Upsilon(S_j)) \\ =&(1 - E_n - E_p)^{-1}((1 - E_n) * Pr(\triangle_{ij}, c_i, c_j | S_i \Vdash_{\mathcal{C}} S_j) \\ &- E_p * Pr(\triangle_{ij}, c_i, c_j | S_i \nVdash_{\mathcal{C}} S_j)) \end{aligned} \qquad (5)$$

Thus, the spatio-temporal pattern $Pr(\triangle_{ij}, c_i, c_j | \Upsilon(S_i) = \Upsilon(S_j))$ can be expressed as a function of $Pr(\triangle_{ij}, c_i, c_j | S_i \Vdash_{\mathcal{C}} S_j)$ and $Pr(\triangle_{ij}, c_i, c_j | S_i \nVdash_{\mathcal{C}} S_j)$, both of which can be measured by the classier $\mathcal{C}$ in the following steps. We first calculate n, the number of the image pairs, which satisfy the conditions: 1) they are judged by $\mathcal{C}$ to contain the same person; 2) they are captured at the camera c_i and c_j, and 3) the time interval between them is in $[\triangle_{ij} - t, \triangle_{ij} + t]$. Here t is a small threshold. Then we can use n/N to estimate $Pr(\triangle_{ij}, c_i, c_j | S_i \Vdash_{\mathcal{C}} S_j)$, where N is the total number of testing image pairs. In a similar way, we can estimate $Pr(\triangle_{ij}, c_i, c_j | S_i \nVdash_{\mathcal{C}} S_j)$ through counting.

From Eq.(5), we can infer that while the error rates (E_p and E_n) are approaching 0, the estimated spatio-temporal pattern $Pr(\triangle_{ij}, c_i, c_j | S_i \Vdash_{\mathcal{C}} S_j)$ is approaching the ground-truth pattern $Pr(\triangle_{ij}, c_i, c_j | \Upsilon(S_i) = \Upsilon(S_j))$.

4.4. Bayesian Fusion model

As represented in the last section, the spatio-temporal pattern $Pr(\triangle_{ij}, c_i, c_j | \Upsilon(S_i) = \Upsilon(S_j))$, which is estimated from the visual classifier $\mathcal{C}$, provides a new perspective to discriminate surveillance images besides the visual features used in $\mathcal{C}$. This motivates us to propose a fusion model, which combines the visual features with the spatio-temporal
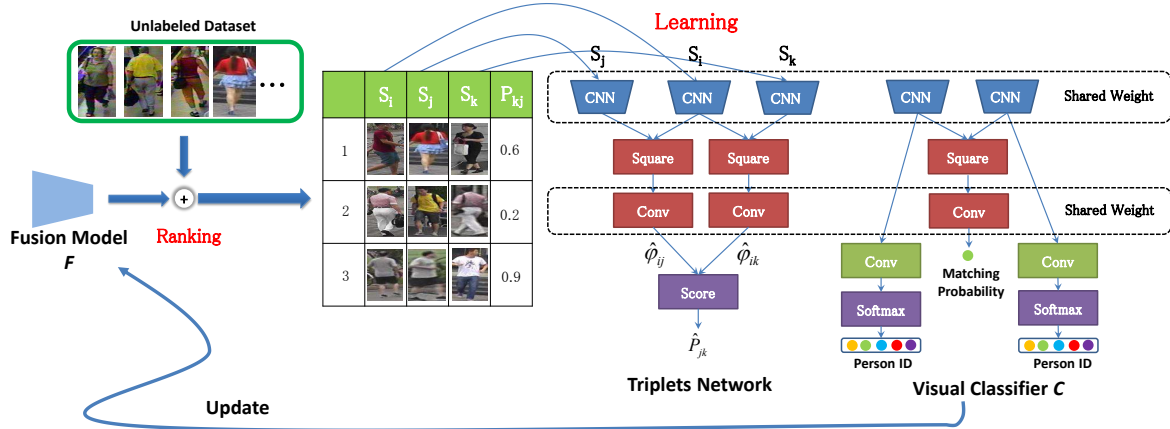
Figure 3: Incremental optimization by the learning-to-rank scheme.

pattern to achieve a composite similarity score of given pair of images, as shown in the step (3) of Fig. 1. Formally, the fusion model is based on the conditional probability:

$$Pr(\Upsilon(S_i) = \Upsilon(S_j)|\vec{v_i}, \vec{v_j}, \triangle_{ij}, c_i, c_j). \qquad (6)$$

Here S_i and S_j are any pair of surveillance images from the target dataset. S_i is taken at the camera c_i at the time t_i, and S_j is taken at the camera c_j at the time t_j. Their visual feature vectors are denoted as $\vec{v_i}$ and $\vec{v_j}$. The timing interval between them is $\triangle_{ij} = t_j - t_i$. Eq.(6) measures the probability of that S_i and S_j contain the same person conditional on their visual features and spatio-temporal information.

According to the Bayesian rule, we have:

$$
\begin{aligned}
&Pr(\Upsilon(S_i) = \Upsilon(S_j)|\vec{v_i}, \vec{v_j}, \triangle_{ij}, c_i, c_j) \\
&= \frac{Pr(\Upsilon(S_i) = \Upsilon(S_j)|\vec{v_i}, \vec{v_j}) * Pr(\triangle_{ij}, c_i, c_j|\Upsilon(S_i) = \Upsilon(S_j))}{Pr(\triangle_{ij}, c_i, c_j)}
\end{aligned}
\qquad (7)
$$

Here $Pr(\Upsilon(S_i) = \Upsilon(S_j)|\vec{v_i}, \vec{v_j})$ indicates the probability of that S_i and S_j contain the same person given their visual features. It can be derived from $Pr(S_i \Vdash_{\mathcal{C}} S_j|\vec{v_i}, \vec{v_j})$, which is the matching probability judged by the visual classifier $\mathcal{C}$:

$$
\begin{aligned}
&Pr(\Upsilon(S_i) = \Upsilon(S_j)|\vec{v_i}, \vec{v_j}) \\
=~& Pr(\Upsilon(S_i) = \Upsilon(S_j)|S_i \Vdash_{\mathcal{C}} S_j) * Pr(S_i \Vdash_{\mathcal{C}} S_j|\vec{v_i}, \vec{v_j}) + \\
& Pr(\Upsilon(S_i) = \Upsilon(S_j)|S_i \nVdash_{\mathcal{C}} S_j) * Pr(S_i \nVdash_{\mathcal{C}} S_j|\vec{v_i}, \vec{v_j}) \\
=~& (1 - E_p - E_n) * Pr(S_i \Vdash_{\mathcal{C}} S_j|\vec{v_i}, \vec{v_j}) + E_n
\end{aligned}
\qquad (8)
$$

On the other hand $Pr(\triangle_{ij}, c_i, c_j|\Upsilon(S_i) = \Upsilon(S_j))$ in Eq. (7) indicates the spatio-temporal pattern of pedestrains, and it can be calculated according to Eq.(5). By substituting Eq.(5) and (8) into Eq.(7), we have:

$$
\begin{aligned}
&Pr(\Upsilon(S_i) = \Upsilon(S_j)|\vec{v_i}, \vec{v_j}, \triangle_{ij}, c_i, c_j) \\
&= \frac{(M_1 + \frac{E_n}{1 - E_n - E_p})((1 - E_n)M_2 - E_p M_3)}{Pr(\triangle_{ij}, c_i, c_j)}
\end{aligned}
\qquad (9)
$$

Here, M_1,M_2, and M_3 are defined as follows:

$$
\begin{aligned}
M_1 &= Pr(S_i \Vdash_{\mathcal{C}} S_j|\vec{v_i}, \vec{v_j}) \\
M_2 &= Pr(\triangle_{ij}, c_i, c_j|S_i \Vdash_{\mathcal{C}} S_j) \\
M_3 &= Pr(\triangle_{ij}, c_i, c_j|S_i \nVdash_{\mathcal{C}} S_j)
\end{aligned}
\qquad (10)
$$

M_1 indicates the judgement of the classifier $\mathcal{C}$ based on the visual features, and it can be measured by $\mathcal{C}$ according to Eq. (3)). M_2 and M_3 represent the spatio-temporal patterns of the pedestrians moving in the camera network, and they can be calculated according to the steps mentioned in section 4.3. Based on Eq.(9), we can construct a fusion classifier $\mathcal{F}$, which takes the visual features and spatio-temporal information of two images as input, and outputs their matching probability. As E_q and E_n are unknown in the unlabeled target dataset, Eq.(9) can not be directly deployed. Thus we substitute E_q and E_n in Eq. (9) with two configurable parameters α and β to achieve a more general matching probability function of $\mathcal{F}$:

$$
Pr(S_i \Vdash_{\mathcal{F}} S_j|v_i, v_j, \triangle_{ij}, c_i, c_j) \qquad (11)
$$
$$
= \frac{(M_1 + \frac{\alpha}{1-\alpha-\beta})((1 - \alpha) * M_2 - \beta * M_3)}{Pr(\triangle_{ij}, c_i, c_j)}(0 \le \alpha, \beta \le 1)
$$

Here $S_i \Vdash_{\mathcal{F}} S_j$ means that the classifier $\mathcal{F}$ judges that S_i and S_j contain the same person.

In the person re-ID scenario, given any query image S_i, we can rank all the images $\{S_j\}$ in the database according to the matching probability $Pr(S_i \Vdash_{\mathcal{F}} S_j|v_i, v_j, \triangle_{ij}, c_i, c_j)$ defined in Eq.(11), and select out the images which have largest probability to contain the same person with S_i.

4.5. Precision Analysis of the Fusion Model

In this section, we will analyze the precision of the fusion model $\mathcal{F}$. Similar with Eq.(1) and (2), we define the false positive error rate E_p' of $\mathcal{F}$ as $Pr(\Upsilon(S_i) \ne \Upsilon(S_j)|S_i \Vdash_{\mathcal{F}} S_j)$ and the false negative error rate E_n' of $\mathcal{F}$ as $Pr(\Upsilon(S_i) =$

$\Upsilon(S_j)|S_i \nVdash_{\mathcal{F}} S_j)$. The following Theorem 1 shows the performance of the fusion model:

Theorem 1 : If $E_p + E_n < 1$ and $\alpha + \beta < 1$, we have $E_p' + E_n' < E_p + E_n$.

Theorem 1 means that the error rate of the fusion model $\mathcal{F}$ may be lower than the original visual classifier $\mathcal{C}$ under the conditions of $E_p + E_n < 1$ and $\alpha + \beta < 1$. It theoretically shows the effectiveness to fuse the spatio-temporal patterns with visual features. Due to the page limit, we put the proof of the theorem 1 in the appendix.

4.6. Incremental Optimization by Learning-to-rank

As shown in Fig. 1, the fusion model $\mathcal{F}$ is derived from the visual classifier $\mathcal{C}$ by integrating with spatio-temporal patterns. According to Theorem 1, $\mathcal{F}$ may perform better than $\mathcal{C}$ in the target dataset. That means, given a query image, when using the classifiers to rank the other images according to the matching probability, the ranking results of $\mathcal{F}$ may be more accurate than that of $\mathcal{C}$. Motivated by this, we propose a novel learning-to-rank based scheme to utilize $\mathcal{F}$ to optimize $\mathcal{C}$ by teaching it with the ranking results in the unlabeled target dataset. Subsequently, the improvement of $\mathcal{C}$ may also derive a better fusion model $\mathcal{F}$. In this mutual promotion procedure, both of the classifiers $\mathcal{C}$ and $\mathcal{F}$ can get incremental optimization in the unlabeled target dataset.

The detailed incremental optimization procedure is shown in the Fig. 3. In the first step, given any query image S_i, the fusion classifier $\mathcal{F}$ is applied to rank the other images in the unlabeled target dataset according to the matching probability defined in Eq.(11). Then we randomly select one image from the top $n(n > 0)$ results, and another one from the results, the rankings of which are in $(n, 2n]$. One of these two images is selected and denoted as S_j, and the other one is denoted as S_k. The matching probability between S_i and S_j measured by $\mathcal{F}$ is denoted as $\varphi_{i,j}$, and the matching probability between S_i and S_k is denoted as $\varphi_{i,k}$. The normalized ranking difference between S_j and S_k is defined as: $P_{j,k} = \frac{e^{\varphi_{i,j} - \varphi_{i,k}}}{1 + e^{\varphi_{i,j} - \varphi_{i,k}}}$.

In order to force $\mathcal{C}$ to learn the ranking difference judged by $\mathcal{F}$, we propose a triplets network based on $\mathcal{C}$ to predict the ranking difference. As shown in the Fig. 3, the triplets network takes the three images, $S_i, S_j,$ and S_k as input, and shares the CNN modules with $\mathcal{C}$ to extract visual features. The following square layer and convolutional layer, which are also shared with $\mathcal{C}$, are used to calculate the similarity scores of the image pairs (S_i, S_j) and (S_i, S_k). Their corresponding similarity scores are $\hat{\varphi}_{i,j}$ and $\hat{\varphi}_{i,k}$. In the final score layer, the predicted ranking difference is calculated as $\hat{P}_{j,k} = \frac{e^{\hat{\varphi}_{i,j} - \hat{\varphi}_{i,k}}}{1 + e^{\hat{\varphi}_{i,j} - \hat{\varphi}_{i,k}}}$.

When training the triplets network, the loss function is defined as the cross entropy of the predicted score $\hat{P}_{j,k}$ and the ranking difference $P_{j,k}$ calculated by $\mathcal{F}$: $LOSS_r =$

$-\hat{P}_{j,k} * log(P_{j,k}) - (1 - \hat{P}_{j,k}) * log(1 - P_{j,k})$.

After training the triplets network, the CNN modules which are shared with the classifier $\mathcal{C}$ get incrementally optimized. In this way, by using the ranking results of the classifier $\mathcal{F}$, we can achieve a upgraded $\mathcal{C}$. Subsequently, the value of M_1, M_2 and M_3 can be updated based on the new $\mathcal{C}$ (Eq. (10)). With the new M_1, M_2 and M_3, we can update $\mathcal{F}$, the probability function of which is calculated according to Eq.(11). The mutual promotion of $\mathcal{C}$ and $\mathcal{F}$ can be conducted in multiple iterations to achieve persistent evolving in the unlabeled target dataset, until the change of the loss $LOSS_r$ among different iterations is less than a threshold.

5. Experiment

5.1. Dataset Setting

Four widely used benchmark datasets are chosen in our Experiments[1], including GRID [2], Market1501 [36], CUHK01 [28], and VIPeR [6]. As shown in Table. 1, we select one of above datsets as the source dataset and another one as the target dataset to test the performance of cross-dataset person Re-ID. As mentioned in section 4.4, the capturing time of each image frame is required to build the fusion model. Thus we choose 'Market1501' and 'GRID' as target datasets, for they provide the detailed frame numbers in the video sequences, which can be used as timestamps of image frames. The source dataset is chosen without any constraint, because only the image content is used to train the visual classifier $\mathcal{C}$ in the initial step of the model as Fig. 1. In this way, there are totally 6 cross-dataset pairs for experiments as Table. 1. In each source dataset, all labeled images are used for the pre-training of the visual classifier $\mathcal{C}$. On the other hand, the configurations of the target datasets 'Market1501' and 'Grid' follow the instructions of these datasets [2][36] to divide the training and testing set. Specifically, in the 'GRID' dataset, a 10-fold cross validation is conducted. In the 'Market1501' dataset, 12,936 'bounding-box-train' images are chosen for training and incremental optimization, while 3,368 query images and 19,732 'bounding-box-test' images for single query evaluation.

When adopting Eq. (11) in the fusion model, α and β are two tunable parameters. By default, we set $\alpha = 0$ and $\beta = 0$. The performance of different combinations of the parameters are also tested in the following Section 5.4.

5.2. Learned Spatio-temporal Patterns

As shown in Fig. 1, learning the spatio-temporal patterns in the unlabeled target dataset is a key step of our fusion model. As shown in Section 4.3, the learned spatio-temporal pattern is represented as the spatio-temporal distribution $Pr(\triangle_{ij}, c_i, c_j | \Upsilon(S_i) \Vdash_{\mathcal{C}} \Upsilon(S_j))$. Here S_i and S_j are any pair of images captured from the cameras C_i and C_j, and

---

[1]Source Code: https://github.com/ahangchen/TFusion

Table 1: Unsupervised transfer learning results.

| Source | Target | Transfer Learning Step | | | | | | Incremental Optimization Step | | | | | |
| | | Visual Classifier $\mathcal{C}$ | | | Fusion Model $\mathcal{F}$ | | | Visual Classifier $\mathcal{C}$ | | | Fusion Model $\mathcal{F}$ | | |
		rank-1	rank-5	rank-10	rank-1	rank-5	rank-10	rank-1	rank-5	rank-10	rank-1	rank-5	rank-10
CUHK01	GRID	10.70	20.20	23.80	30.90	63.70	79.10	17.40	33.90	41.10	50.90	78.60	88.30
VIPeR	GRID	9.70	17.40	21.50	28.40	65.60	80.40	18.50	31.40	40.50	52.70	81.70	89.20
Market1501	GRID	17.80	31.20	36.80	49.60	81.40	88.70	22.30	38.10	47.20	60.40	87.30	93.40
GRID	Market1501	20.72	35.39	42.99	51.16	65.08	70.04	22.38	39.25	48.07	58.22	72.33	76.84
VIPeR	Market1501	24.70	40.91	49.52	56.18	71.50	76.48	25.23	41.98	50.33	59.17	73.49	78.62
CUHK01	Market1501	29.39	45.46	52.55	56.53	70.22	74.64	30.58	47.09	54.60	60.75	74.44	79.25
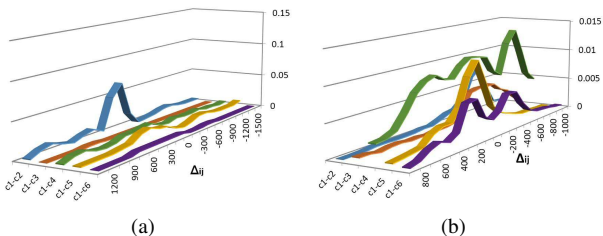
Figure 4: (a)Spatio-temporal pattern in the 'GRID' dataset. (b)Spatio-temporal pattern in the 'Market1501' dataset.

they are judged by the visual classifier $\mathcal{C}$ to contain the same person. $\triangle_{ij}$ is defined as : $\triangle_{ij} = t_i - t_j$, where t_i and t_j are the timestamps (frame number) of S_i and S_j. Fig. 4 shows the spatio-temporal distribution in the 'GRID' and 'Market1501' dataset. Due to the limit of pages, Fig. 4 only shows the distribution related to the first camera in the dataset. The full distribution is attached in the appendix. Fig. 4 shows clearly that the time interval of images from different pairs of cameras follows different non-random distribution, which indicates pedestrians' distinctive temporal patterns to transfer among different locations. This confirms that these spatio-temporal patterns can be used to filter out the matching results with less transferring probability to improve the precision of the person Re-ID system.

5.3. Re-ID Results

Table. 1 shows the performance of our model in each training step. Firstly, in the 'Transferring Learning Step', the 'Visual Classifier $\mathcal{C}$' column means to directly transfer the visual classifier $\mathcal{C}$ trained in the source dataset to the unlabeled target dataset without optimization. Not surprisingly, this kind of simple transferring method causes poor performance, due to the variation of data distribution in different datasets. The following 'Fusion Model $\mathcal{F}$' column shows that **the performance of the fusion model, which integrates with the spatio-temporal patterns, gains significant improvement compared with the original visual classifier** $\mathcal{C}$.

The 'Incremental Optimization step' in Table. 1 means the procedure to use the learning-to-rank scheme to further optimize the model as mentioned in Section 4.6. Table. 1 shows that, **with this incremental learning procedure, the visual classifier $\mathcal{C}$ achieves obvious improvement**. This proves the effectiveness of the learning-to-rank scheme to transfer knowledge from the fusion model $\mathcal{F}$ to the visual classifier $\mathcal{C}$

in the unlabeled target dataset. **Table. 1 also shows that the performance of the fusion model $\mathcal{F}$ achieves significant improvement after the incremental learning**. This is due to the mutual promotion of $\mathcal{F}$ and $\mathcal{C}$ as depicted in Fig. 1: a better $\mathcal{C}$ can derive a better $\mathcal{F}$, and a better $\mathcal{F}$ can train the $\mathcal{C}$ into a better one by the learning-to-rank procedure.

Table 2: Compare the precision of TFusion with the state-of-art unsupervised transfer learning methods.

| Method | Source | Target | Performance | | |
			rank-1	rank-5	rank-10
UMDL[21]	Market1501	GRID	3.77	7.76	9.71
	CUHK01	GRID	3.58	7.56	9.50
	VIPeR	GRID	3.97	8.14	10.73
	GRID	Market1501	30.46	45.07	52.38
	CUHK01	Market1501	29.69	44.33	51.40
	VIPeR	Market1501	30.34	44.92	52.14
TFusion-uns	Market1501	GRID	60.40	87.30	93.40
	CUHK01	GRID	50.90	78.60	88.30
	VIPeR	GRID	52.70	81.70	89.20
	GRID	Market1501	58.22	72.33	76.84
	VIPeR	Market1501	59.17	73.49	78.62
	CUHK01	Market1501	60.75	74.44	79.25

Table 3: Compare the precision of TFusion with the supervised methods on GRID.

| Method | Performance | | |
	rank-1	rank-5	rank-10
GOG + XQDA[25]	24.80	-	58.40
HIPHOP+LOMO+CRAFT[33]	26.00	50.60	62.50
SSM[23]	27.20	-	61.12
JLML[29]	37.5	61.4	69.4
TFusion-uns (Market1501->GRID)	60.40	87.30	93.40
TFusion-sup	64.10	91.90	96.50

Table 4: Compare the precision of TFusion with the supervised algorithms on Market1501.

| Method | Performance | | |
	rank-1	rank-5	rank-10
SLSC[4]	51.90	-	-
LDEHL[5]	59.47	80.73	86.94
S-CNN[22]	65.88	-	-
DLCE[37]	79.51	90.91	94.09
SVDNet[32]	82.3	-	-
JLML[29]	88.8	-	-
TFusion-uns (CUHK01->Market1501)	60.75	74.44	79.25
TFusion-sup	73.13	86.43	90.46

We also compare our model, named ***TFusion***, with the state-of-art unsupervised cross-dataset person Re-ID algorithm, *UMDL* [21]. *UMDL* addresses the similar problem with us, and aims to transfer the visual feature representation

from a labeled source dataset to another unlabeled target dataset. UMDL is based on the dictionary learning method and outperforms the state-of-art unsupervised learning algorithms as reported in [21]. We compare *TFusion* and *UMDL* under the same dataset configuration and show the results in Table. 2. **In all test cases, *TFusion* outperforms *UMDL* by a large margin.** Especially, for the cases where the target dataset is 'GRID', *TFusion* performs extremely well. This may be attributed to the distinct human motion pattern in the 'GRID' dataset, which is collected from a metro station. The fusion with pedestrians' spatio-temporal pattern can significantly improve the Re-ID performance.

To observe more clearly the strength of our algorithm to utilize the unlabeled data, we also compare its performance with the state-of-art supervised algorithms deployed on the labeled target datasets. Table. 3 shows the experimental results in the 'GRID' dataset. It is surprising to find that the *TFusion* model, which conducts unsupervised transferring from 'Market1501' to 'GRID' and does not use the label information of 'GRID', **outperforms the state-of-art supervised algorithms on 'GRID'.** This proves again the effectiveness of the fusion with spatio-temporal information. On the other hand, our model can be also run in a supervised mode (denoted as 'TFusion-sup' in Table. 3), where both the source dataset and the target dataset are the same. The performance of *TFusion-sup* is much better than the state-of-art supervised algorithms. It is also interesting to find that the performance of the unsupervised *TFusion* is very close to the supervised version *TFusion-sup*. This shows that the unlabeled data in the target dataset is utilized sufficiently by *TFusion* to achieve good performance. Similarly, Table. 4 compares *TFusion* with the state-of-art supervised algorithms on 'Market1501'. It also shows that **our unsupervised transferring model, *TFusion*, can achieve a comparable performance close to the supervised learning models.**

5.4. Parameter sensitivity

As mentioned in Eq. (11), α and β are two tunable parameters in the fusion model. Theorem 1 proves that when $\alpha + \beta < 1$, the fusion model $\mathcal{F}$ may have chance to perform better than the original visual classifier $\mathcal{C}$. Thus, we try different combinations of α and β, which satisfy $\alpha + \beta < 1$, and test the performance of the fusion model. Fig. 5(a) shows the rank-1 precision of the models with different α and β when transferring from 'Market1501' to 'GRID', and Fig. 5(b) shows the case from 'GRID' to 'Market1501'. It shows that the model with smaller α and β tends to have better performance. The combination that $\alpha = 0.25$ and $\beta = 0$ achieves a relatively good performance in both test cases.

As shown in Fig. 3, the incremental learning procedure consists of iterative learning-to-rank steps. In each iteration, the fusion model $\mathcal{F}$ is used to train the visual classifier $\mathcal{C}$, and

subsequently a more accurate $\mathcal{C}$ can derive a better $\mathcal{F}$. Fig. 6 shows how the number of learning-to-rank iterations affects the rank-1 precision of $\mathcal{F}$. It shows that the performance achieves big improvement in the first three iterations, and the precision tends to converge since then. This suggests us that the number of the learning-to-rank iterations can be configured as 3 in the real deployment of *TFusion*.
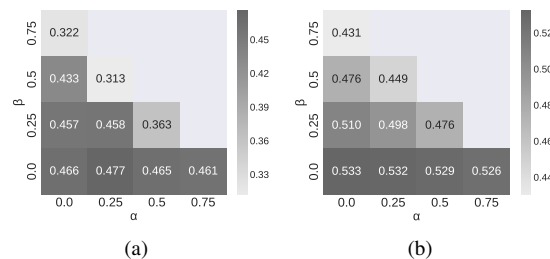

(a)   (b)

Figure 5: Performance under different setting of α and β (a) in the 'Grid' dataset; (b) in the 'Market1501' dataset.
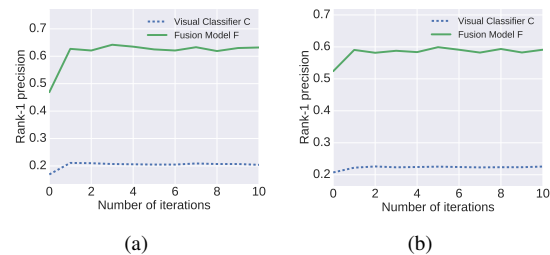

(a)   (b)

Figure 6: Performance vs. the number of iterations of the learning-to-rank optimization. (a) Performance in the 'Grid' dataset. (b) Performance in the 'Market1501' dataset.

6. Conclusions

In this paper, we have presented *TFusion* as a high-performance unsupervised cross-dataset person Re-ID algorithm. In particular, *TFusion* transfers the visual classifier trained in a small labeled source dataset to an unlabeled target dataset by integrating with the spatio-temporal patterns of pedestrians learned in an unsupervised way. Furthermore, an iterative learning-to-rank scheme is proposed to incrementally optimize the model based on the unlabeled data. Experiments show that *TFusion* outperforms the state-of-art unsupervised cross-dataset transferring algorithm by a big margin, and it also achieves a comparable or even better performance compared with the state-or-art supervised learning algorithms in multiple real datasets.

References

[1] E. Ahmed, M. J. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015. 1, 2

[2] C. Change, Loy, X. Tao, and G. Shaogang. Multi-camera activity correlation analysis. In *Computer Vision, IEEE International Conference on*, 2009. 2, 6

[3] S. Chen, C. Guo, and J. Lai. Deep ranking for person re-identification via joint representation learning. *IEEE Trans. Image Processing*, 25(5):2353–2367, 2016. 1, 2

[4] C. Dapeng, Y. Zejian, C. Badong, and Z. Nanning. Similarity learning with spatial constraints for person re-identification. *ECCV*, 2016. 7

[5] U. Evgeniya and L. Victor. Learning deep embeddings with histogram loss. *NIPS*, 2016. 7

[6] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. 01 2007. 2, 6

[7] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008. 2

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 4

[9] W. Huang, R. Hu, C. Liang, Y. Yu, Z. Wang, X. Zhong, and C. Zhang. Camera network based person re-identification by leveraging spatial-temporal constraint and multiple cameras relations. In *MMM*, 2016. 2

[10] O. Javed, K. Shafique, Z. Rasheed, and M. Shah. Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views. *Computer Vision and Image Understanding*, 109(2):146–162, 2008. 2, 4

[11] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*. IEEE Computer Society, 2012. 2

[12] R. Layne, T. M. Hospedales, and S. Gong. Domain transfer for person re-identification. In *ARTEMIS@ACM Multimedia 2013*, 2013. 2

[13] C. Liang, B. Huang, R. Hu, C. Zhang, X. Jing, and J. Xiao. A unsupervised person re-identification method using model based representation and ranking. In *MM*, 2015. 1, 2

[14] 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, 3

[15] G. Lisanti, I. Masi, A. D. Bagdanov, and A. D. Bimbo. Person re-identification by iterative re-weighted sparse ranking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(8):1629–1642, 2015. 2

[16] A. J. Ma, J. Li, P. C. Yuen, and P. Li. Cross-domain person reidentification using domain adaptation ranking svms. *IEEE Trans. Image Processing*, 24(5):1599–1613, 2015. 2

[17] L. Ma, X. Yang, and D. Tao. Person re-identification over camera networks using multi-task distance metric learning. *IEEE Trans. Image Processing*, 23(8):3656–3670, 2014. 2

[18] X. Ma, X. Zhu, S. Gong, X. Xie, J. Hu, K. Lam, and Y. Zhong. Person re-identification by unsupervised video matching. *Pattern Recognition*, 65:197–210, 2017. 1, 2

[19] N. Martinel, G. L. Foresti, and C. Micheloni. Person reidentification in a distributed camera network framework. *IEEE Trans. Cybernetics*, 47(11):3530–3541, 2017. 2

[20] S. Paisitkriangkrai, C. Shen, and A. van den Hengel. Learning to rank in person re-identification with metric ensembles. In *CVPR*, 2015. 1, 2

[21] P. Peixi, X. Tao, W. Yaowei, P. Massimiliano, G. Shaogang, H. Tiejun, and T. Yonghong. Unsupervised cross-dataset transfer learning for person re-identification. In *CVPR*, 2016. 2, 7, 8

[22] V. Rahul, Rama, H. Mrinal, and W. Gang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, 2016. 7

[23] B. Song, B. Xiang, and T. Qi. Scalable person re-identification on supervised smoothed manifold. In *CVPR*, 2017. 7

[24] D. Tao, Y. Guo, M. Song, Y. Li, Z. Yu, and Y. Y. Tang. Person re-identification by dual-regularized KISS metric learning. *IEEE Trans. Image Processing*, 25(6):2726–2738, 2016. 1, 2

[25] M. Tetsu, O. Takahiro, S. Einoshin, and S. Yoichi. Hierarchical gaussian descriptor for person re-identification. In *CVPR*, 2016. 2, 7

[26] H. Wang, S. Gong, and T. Xiang. Unsupervised learning of generative topic saliency for person re-identification. In *BMVC*, 2014. 2

[27] H. Wang, X. Zhu, T. Xiang, and S. Gong. Towards unsupervised open-set person re-identification. In *ICIP*, 2016. 1, 2

[28] L. Wei and W. Xiaogang. Glocally aligned feature transforms across views. In *Computer Vision, IEEE International Conference on*, 2013. 2, 6

[29] L. Wei, Z. Xiatian, and G. Shaogang. Person re-identification by deep joint learning of multi-loss classification. In *CVPR*, 2017. 7

[30] L. Wu, C. Shen, and A. van den Hengel. Deep linear discriminant analysis on fisher networks: A hybrid architecture for person re-identification. *Pattern Recognition*, 65:238–250, 2017. 1, 2

[31] 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

[32] S. Yifan, Z. Liang, D. Weijian, and W. Shengjin. Svdnet for pedestrian retrieval. In *ICCV*, 2017. 7

[33] C. Yingcong, Z. Xiatian, Z. Weishi, and L. Jianhuang. Person re-identification by camera correlation aware feature augmentation. In *CVPR*, 2017. 7

[34] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013. 2

[35] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *CVPR*, 2014. 2, 3

[36] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *Computer Vision, IEEE International Conference on*, 2015. 2, 4, 6

[37] Z. Zheng, L. Zheng, and Y. Yang. A discriminatively learned cnn embedding for person re-identification. *TOMM*, 2017. 3, 4, 7