Unsupervised Person Re-Identification by Camera-Aware Similarity Consistency Learning

Ancong Wu[1] , Wei-Shi Zheng[2,3,4*] , and Jian-Huang Lai[2,5]

[1]School of Electronics and Information Technology, Sun Yat-sen University, China
[2]School of Data and Computer Science, Sun Yat-sen University, China
[3]Peng Cheng Laboratory, Shenzhen 518005, China
[4]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
[5]Guangdong Province Key Laboratory of Information Security, China
wuancong@gmail.com, wszheng@ieee.org, stsljh@mail.sysu.edu.cn

Abstract

For matching pedestrians across disjoint camera views in surveillance, person re-identification (Re-ID) has made great progress in supervised learning. However, it is infeasible to label data in a number of new scenes when extending a Re-ID system. Thus, studying unsupervised learning for Re-ID is important for saving labelling cost. Yet, cross-camera scene variation is a key challenge for unsupervised Re-ID, such as illumination, background and viewpoint variations, which cause domain shift in the feature space and result in inconsistent pairwise similarity distributions that degrade matching performance. To alleviate the effect of cross-camera scene variation, we propose a Camera-Aware Similarity Consistency Loss to learn consistent pairwise similarity distributions for intra-camera matching and cross-camera matching. To avoid learning ineffective knowledge in consistency learning, we preserve the prior common knowledge of intra-camera matching in the pretrained model as reliable guiding information, which does not suffer from cross-camera scene variation as cross-camera matching. To learn similarity consistency more effectively, we further develop a coarse-to-fine consistency learning scheme to learn consistency globally and locally in two steps. Experiments show that our method outperformed the state-of-the-art unsupervised Re-ID methods.

1. Introduction

In recent years, person re-identification (Re-ID) has drawn much attention in surveillance applications. Many works focus on supervised learning [14, 45, 18, 6, 16, 1, 29] and have made great progress. However, in practise, manually labelling data for training is costly when developing Re-ID system for a large number of new scenes. To reduce labelling amount and exploit unlabelled data in a new scene, some previous works attempt to study unsupervised and transfer learning for Re-ID [26, 13, 40, 32, 8, 34, 7, 49], in which some recent advanced methods [40, 34, 7, 49] focus on scene variation between cameras. Cross-camera scene variation is a key challenge for unsupervised Re-ID [1], since illumination, background and viewpoint vary from camera to camera and cause domain shift in feature space.

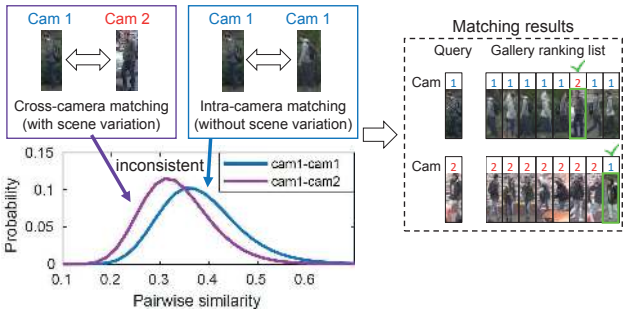


Figure 1. Illustration of the camera-aware similarity inconsistency problem. We match samples in two cameras (denoted by cam 1 and cam 2) on the DukeMTMC dataset [48] using a ResNet-50 model [12] pretrained on the MSMT17 dataset [34]. The pairwise similarities are computed between each pair of samples in intra-camera matching or cross-camera matching and the distributions are shown on the left. Top-8 matchings retrieved by cosine similarity are shown on the right with the correct matchings indicated by green bounding boxes. The cross-camera scene variation leads to domain shift in the feature space and results in inconsistent pairwise similarity distributions that degrade matching performance.

To show the effect of cross-camera scene variation, we visualize pairwise similarity distributions and some matching examples in Figure 1. We match samples in two cameras on the DukeMTMC dataset [48]. The pairwise similarities are computed between each pair of samples in intra-camera matching or cross-camera matching by a ResNet-50 model

*Corresponding author

[1]In this paper, unsupervised person re-identification is studied in the unsupervised domain adaptation setting, which learns a model for the target domain given labelled source data and unlabelled target data.

[12] pretrained on the MSMT17 dataset [34]. As shown in the distribution figure, the two pairwise similarity distributions are inconsistent that the average pairwise similarity of cross-camera matching is smaller than that of intra-camera matching, because the pairwise similarity is negatively related to the degree of scene variations of the camera pair, such as illumination, background and viewpoint variations. Matching based on pairwise similarities of inconsistent distributions leads to failure of retrieving the correct cross-camera sample in the top ranking list, as shown on the right of Figure 1. We call it the *camera-aware similarity inconsistency problem*, which is caused by cross-camera scene variation, a serious problem for unsupervised Re-ID.

To alleviate the effect of cross-camera scene variation, we solve the camera-aware similarity inconsistency problem by learning consistent pairwise similarity distributions for intra-camera and cross-camera matching. To avoid learning ineffective knowledge in consistency learning, we exploit the prior common knowledge of Re-ID (e.g., pretrained model) as guiding information for regularization. We preserve the prior common knowledge of intra-camera matching to guide learning cross-camera matching, since intra-camera matching does not suffer from cross-camera scene variation as cross-camera matching and thus is relatively more reliable. To achieve this, we propose a Camera-Aware Similarity Consistency Loss, which jointly learns intra-/cross-camera similarity consistency and preserves the common knowledge in intra-camera pairwise similarities.

To learn similarity consistency more effectively, we model pairwise similarities not only in the global feature space (i.e., all sample pairs) but also in the local neighbourhood of the feature space (i.e., the sample pairs of top-ranked nearest neighbours), since retrieving correct top-ranked samples for Re-ID relies on the nearest neighbours. Hence, we further develop a coarse-to-fine consistency learning scheme to learn consistency globally and locally.

Compared with advanced unsupervised Re-ID methods [40, 34, 7, 49] that handle cross-camera scene variation by camera-to-camera alignment, we explore the relation of pairwise similarity between intra-camera and cross-camera matching, so that cross-camera matching can benefit from the relatively reliable knowledge in intra-camera matching.

In summary, the contributions of this paper are: (1) We propose a Camera-Aware Similarity Consistency Loss to alleviate cross-camera scene variation for unsupervised Re-ID, which explores the relation of pairwise similarity between intra-camera and cross-camera matching; (2) We further develop a coarse-to-fine consistency learning scheme to learn consistency more effectively with our loss.

2. Related Work

Supervised Person Re-Identification. Person re-identification has witnessed a fast growing development recently, from feature design [10, 9, 20, 18, 22, 47, 38] to distance metric learning [35, 10, 27, 14, 45, 23, 25, 17, 38, 21, 24, 18, 6, 46, 39, 41, 33, 3] and end-to-end deep learning [16, 1, 36, 37, 31, 19, 42, 43, 48, 29]. With abundant labelled data, the supervised models achieve high performance, but heavy labelling cost hinders the scalability.

Unsupervised Person Re-Identification. Recently, reducing labelling cost for person re-identification has drawn more attention, since it is infeasible to label a large number of identities for each new scene. Most works study unsupervised learning [26, 13, 40, 32, 8, 34, 7, 4, 50, 49, 15] to learn from unlabelled data for Re-ID. Among the advanced unsupervised methods, most of them rely on source data of other scenes for transfer learning or learning prior knowledge of Re-ID. [40] and [8] use source data for pretraining and learn from unlabelled target data by clustering and finetuning. [34, 7, 4, 50, 49] learn to transfer knowledge by image-to-image transformation from source images to target images. [32] learns to transfer knowledge from attribute labels. [15] learns from associating tracklets in videos across cameras.

These methods exploit unlabelled data in different ways and most of them alleviate cross-camera scene variation explicitly or implicitly. They handle cross-camera scene variation by camera-to-camera alignment either at feature level [40] or at image level [34, 7, 4, 50, 49]. Our method also focus on alleviating cross-camera scene variation, which is significant for unsupervised Re-ID. Rather than camera-to-camera alignment, we further explore the relation of pairwise similarity between intra-camera matching and cross-camera matching, which is ignored in existing methods. We aim to learn consistent pairwise similarity distributions for intra-camera and cross-camera matching with the guidance of prior common knowledge of intra-camera matching.

Domain Adaptation. For alignment between cameras for alleviating the effect of cross-camera scene variation, domain adaptation techniques are closely related. For example, MMD [11], CORAL [28] and ADDA [30] are representative domain adaptation methods. MMD [11] minimizes the difference between the means of two domains. CORAL [28] minimizes the difference between the covariance matrices of two domains. ADDA [30] aligns two domains by adversarial learning. They are for domain alignment in the feature space, while our method learns consistency of pairwise similarity distributions of intra-camera matching and cross-camera matching in the similarity space and our method can benefit from the relation of pairwise similarity between intra-camera and cross-camera matching.

3. Camera-Aware Similarity Consistency

To study unsupervised Re-ID, we first formulate this problem as follows. In a new scene with N_{cam} cameras, a set of unlabelled pedestrian images $\{\mathbf{I}_{c,i}\}$ can be obtained,

in which $\mathbf{I}_{c,i}$ is the i-th person image in camera c. We aim to learn a model H from the unlabelled data $\{\mathbf{I}_{c,i}\}$ to compute the similarities between samples for retrieval.

3.1. Similarity Inconsistency Problem

As mentioned in Section 1, cross-camera scene variation is a serious problem for unsupervised Re-ID and it causes domain shift in the feature space and inconsistent pairwise similarity distributions. For visualization, we randomly select three cameras in the DukeMTMC [48] dataset denoted by cam1, cam2, cam3 to analyse the pairwise similarity distributions for all camera pairs. We apply a ResNet-50 [12] model pretrained on the MSMT17 [34] dataset for computing similarities. The pairwise similarity distributions are shown in the first distribution figure in Figure 2.

As shown in the first distribution figure, the three distributions of cross-camera matching are similar and so are those of intra-camera matching. The average pairwise similarity of cross-camera matching is smaller than that of intra-camera matching as pairwise similarity is negatively related to the degree of cross-camera scene variation. Inconsistency of pairwise similarity distributions caused by cross-camera scene variation degrades matching performance for unsupervised Re-ID, as shown in Figure 1 in Section 1. We call this the *camera-aware similarity inconsistency problem*.

To alleviate cross-camera scene variation, we propose camera-aware similarity consistency learning, which aims at learning consistent distributions of intra-camera similarity and cross-camera similarity, as shown in Figure 2.

3.2. Similarity Consistency Learning

3.2.1 Intra-/Cross-Camera Similarity Consistency

To address the camera-aware similarity inconsistency problem, we aim to minimize the difference between pairwise similarity distributions of intra-camera matching and cross-camera matching. Let $H(\cdot;\mathbf{\Theta}_1)$ denote a learnable feature extractor parameterized by $\mathbf{\Theta}_1$ and $\mathbf{x}_{p,i} = H(\mathbf{I}_{p,i};\mathbf{\Theta}_1) \in \mathbb{R}^d$ denote the feature of image $\mathbf{I}_{p,i}$. In our case, the feature $\mathbf{x}_{p,i}$ is normalized by $\ell 2$-norm, so that the inner product of two features $\mathbf{x}_{p,i}^\top \mathbf{x}_{p,j}$ is cosine similarity. Let $\mathbf{X}_p = [\mathbf{x}_{p,1}, \mathbf{x}_{p,2}, ..., \mathbf{x}_{p,N_p}] \in \mathbb{R}^{d \times N_p}$ denote the feature matrix extracted by model $H(\cdot;\mathbf{\Theta}_1)$ during training.

To learn consistent pairwise similarity distributions, we compute the pairwise similarities for intra-camera matching and cross-camera matching. For camera p, the intra-camera similarity matrix is $\mathbf{X}_p^\top \mathbf{X}_p$, in which the element $\mathbf{x}_{p,i}^\top \mathbf{x}_{p,j}$ in the i-th row and the j-th column is the similarity between samples $\mathbf{I}_{p,i}$ and $\mathbf{I}_{p,j}$. Likewise, for two cameras p and q, the cross-camera similarity matrix is $\mathbf{X}_p^\top \mathbf{X}_q$.

Then, to minimize the difference between pairwise similarity distributions of intra-camera matching and cross-camera matching, we minimize the difference of the means and standard deviations of all matrix elements between intra-camera similarity matrix $\mathbf{X}_p^\top \mathbf{X}_p$ and cross-camera similarity matrix $\mathbf{X}_p^\top \mathbf{X}_q$ as follow:

$$\min_{\mathbf{\Theta}_1} L_{con} = \sum_{p \neq q} (\text{mean}(\mathbf{X}_p^\top \mathbf{X}_p) - \text{mean}(\mathbf{X}_p^\top \mathbf{X}_q))^2 + (\text{std}(\mathbf{X}_p^\top \mathbf{X}_p) - \text{std}(\mathbf{X}_p^\top \mathbf{X}_q))^2, \quad (1)$$

where $\text{mean}(\cdot)$ and $\text{std}(\cdot)$ denote the functions of computing mean and standard deviation for all elements in the input matrix, respectively. We call L_{con} the *intra-/cross-camera similarity consistency loss*.

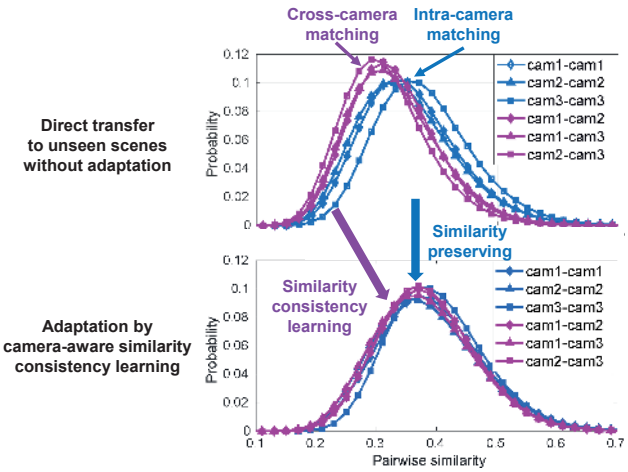


Figure 2. Illustration of camera-aware similarity consistency learning. The key idea is learning consistent pairwise similarity distributions for intra-camera and cross-camera matching with the relatively reliable common knowledge of intra-camera matching preserved as guidance. We show some pairwise similarity distributions of different camera pairs in three cameras in DukeMTMC [48] (denoted by cam 1, cam2 and cam3). The first distribution figure shows the case of directly applying a ResNet-50 model [12] (pretrained on MSMT17 [34]) and the second distribution figure shows the case after camera-aware similarity consistency learning. The pairwise similarity distributions of cross-camera matching becomes consistent with intra-camera matching after learning.

3.2.2 Intra-Camera Similarity Preservation

To avoid learning ineffective knowledge by similarity consistency learning, we exploit prior knowledge of Re-ID as guiding information for regularization, which can be learned by labelled source data in other scenes. To obtain prior common knowledge of Re-ID, we pretrain the model H on the currently largest Re-ID benchmark dataset MSMT17 [34]. Let $\mathbf{\Theta}_{pre}$ denote the fixed pretrained parameters of H. To benefit from the common knowledge, the parameters $\mathbf{\Theta}_1$ of model H are initialized by $\mathbf{\Theta}_{pre}$.

The knowledge of Re-ID is embedded in the pairwise similarities between features. To extract the common knowledge of Re-ID for a new scene of camera p, we extract

the feature $\mathbf{f}_{p,i} = H(\mathbf{I}_{p,i}; \mathbf{\Theta}_{pre}) \in \mathbb{R}^d$ for image $\mathbf{I}_{p,i}$ by the pretrained model, which is fixed during training for computing pairwise similarity. Let $\mathbf{F}_p = [\mathbf{f}_{p,1}, \mathbf{f}_{p,2}, ..., \mathbf{f}_{p,N_p}] \in \mathbb{R}^{d \times N_p}$ denote the feature matrix of images $\{\mathbf{I}_{p,i}\}$. To exploit reliable knowledge as guiding information, we choose to preserve the common knowledge of intra-camera matching, since there is no cross-camera scene variation when matching samples in the same camera, and thus it is relatively more reliable as compared to cross-camera matching that suffers from cross-camera scene variation.

To preserve the common knowledge of intra-camera matching, the pairwise similarities should be preserved as those of the pretrained model. For camera p, we minimize the distance between the intra-camera similarity matrix $\mathbf{X}_p^\top \mathbf{X}_p$ of the model $H(\cdot; \mathbf{\Theta}_1)$ and the intra-camera similarity matrix $\mathbf{F}_p^\top \mathbf{F}_p$ of the fixed pretrained model $H(\cdot; \mathbf{\Theta}_{pre})$ as follow:

$$\min_{\mathbf{\Theta}_1} L_{pre} = \sum_{p=1}^{N_{cam}} dist(\mathbf{X}_p^\top \mathbf{X}_p, \mathbf{F}_p^\top \mathbf{F}_p), \tag{2}$$

where $dist$ is a distance metric for matrices. We call L_{pre} the *intra-camera similarity preserving loss.*

In our case of mini-batch learning, the feature dimension d is larger than the batch size N_p. Generally, $\mathrm{rank}(\mathbf{X}_p) = \mathrm{rank}(\mathbf{F}_p) = N_p$ can be satisfied and the similarity matrices $\mathbf{X}_p^\top \mathbf{X}_p$ and $\mathbf{F}_p^\top \mathbf{F}_p$ have full rank, so that they are symmetric positive definite (SPD) matrices, which are intrinsically lying on a Riemannian manifold instead of a vector space, so we measure the distance using a Log-Euclidean Riemannian framework [2]. The intra-camera similarity preserving loss L_{pre} is reformulated by

$$\min_{\mathbf{\Theta}_1} L_{pre} = \sum_{p=1}^{N_{cam}} \left\| \log(\mathbf{X}_p^\top \mathbf{X}_p) - \log(\mathbf{F}_p^\top \mathbf{F}_p) \right\|_F^2, \tag{3}$$

where $\log(\mathbf{A})$ is the matrix logarithm of $\mathbf{A}$. For any SPD matrix $\mathbf{A}$, the logarithm of it is

$$\log(\mathbf{A}) = \mathbf{U}\mathrm{diag}(\log(\epsilon_1), \log(\epsilon_2), ..., \log(\epsilon_N))\mathbf{U}^\top, \tag{4}$$

where $\mathbf{U}$ is the orthonormal matrix of eigenvectors and ϵ_i is the eigenvalue, which are obtained from the eigendecomposition $\mathbf{A} = \mathbf{U}\mathrm{diag}(\epsilon_1, \epsilon_2, ..., \epsilon_N)\mathbf{U}^\top$.

3.2.3 Camera-Aware Similarity Consistency Loss

To analyse joint learning of L_{con} and L_{pre}, we first analyse the relation between them. In the intra-/cross-camera similarity consistency loss L_{con}, the intra-camera similarity matrix $\mathbf{X}_p^\top \mathbf{X}_p$ and the cross-camera similarity matrix $\mathbf{X}_p^\top \mathbf{X}_q$ are used to learn consistent pairwise similarity distributions. In the intra-camera similarity preserving loss L_{pre}, $\mathbf{X}_p^\top \mathbf{X}_p$ preserves the common knowledge of the pretrained model $H(\cdot; \mathbf{\Theta}_{pre})$. When L_{con} and L_{pre} are jointly learned, the intra-camera similarity matrix $\mathbf{X}_p^\top \mathbf{X}_p$ plays a role as a bridge between cross-camera matching and the reliable common knowledge of intra-camera matching in the pretrained model. Thus, L_{pre} provides prior common knowledge for regularizing consistency learning in L_{con}.

The objective function of camera-aware consistency learning is

$$\min_{\mathbf{\Theta}_1} L = L_{pre} + \lambda L_{con}, \tag{5}$$

where λ is a trade-off parameter. We call L the *Camera-Aware Similarity Consistency Loss*.

Analysis. We analyse the dependence of L_{pre} and L_{con}. When the intra-/cross-camera similarity consistency loss L_{con} is used individually, without preserving reliable common knowledge of intra-camera matching of the pretrained model as regularization, the incorrect knowledge of cross-camera matching hinders effective consistency learning. When the intra-camera similarity preservation loss L_{pre} is used individually, preserving the knowledge that already learned by the pretrained model cannot bring improvement. Thus, L_{con} and L_{pre} should be jointly learned.

To visualize the effect of our Camera-Aware Similarity Consistency Loss, we show the pairwise similarity distributions of different camera pairs after camera-aware similarity consistency learning in the second distribution figure in Figure 2. Compared to the distributions of direct transfer in the first distribution figure, the pairwise similarity distributions of all camera pairs become more consistent and the distributions of intra-camera matching are preserved.

4. Coarse-to-Fine Consistency Learning

In the last section, we introduce the Camera-Aware Similarity Consistency Loss, which learns consistent pairwise similarity distributions of intra-camera matching and cross-camera matching. To learn similarity consistency more effectively with the loss, we model pairwise similarities using not only all sample pairs in the global feature space but also sample pairs of top-ranked nearest samples in the local neighbourhood of the feature space, since retrieving correct top-ranked samples for Re-ID relies on the nearest neighbours. Learning similarity consistency in the global feature space can be regarded as learning coarse consistency. Then, based on coarse consistency, we aim to learn finer consistency in the local neighbourhood of the feature space.

We develop a coarse-to-fine consistency learning scheme as shown in Figure 3, which takes two steps:
(1) Coarse consistency learning in the global feature space;
(2) Fine consistency learning in the local neighbourhood of the feature space.

Step 1: Coarse Consistency Learning. To learn coarse consistency, we model pairwise similarities in the global feature space, that is, all sample pairs are used. For mini-batch learning, we use samples of two randomly sampled cameras p and q in each batch and the samples $\mathbf{I}_{p,i}, \mathbf{I}_{q,j}$

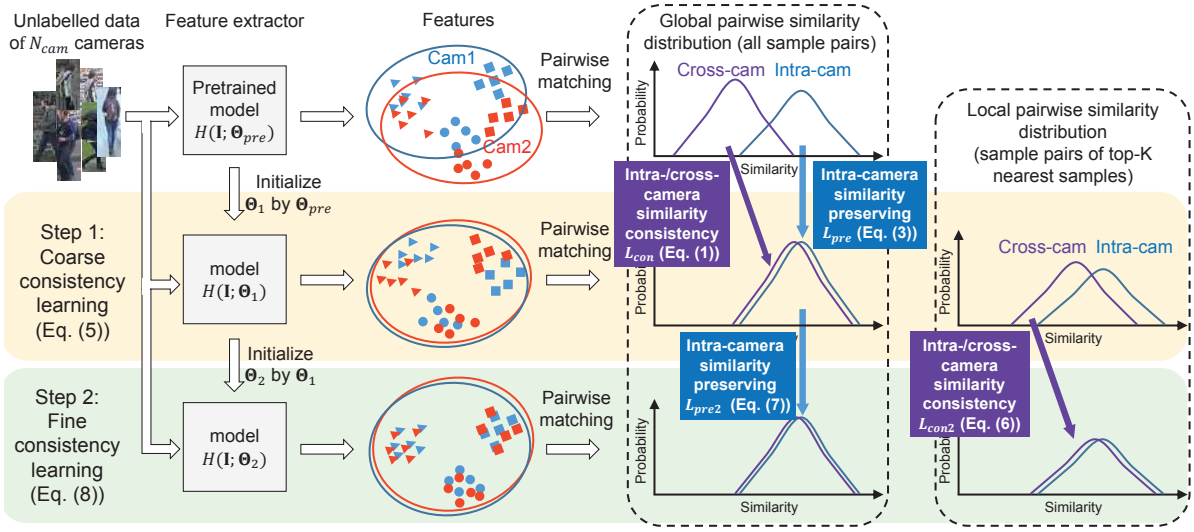


Figure 3. Illustration of the coarse-to-fine consistency learning scheme in Section 4. Given unlabelled data of N_{cam} cameras for learning, the feature extractor model H is trained in two steps. In coarse consistency learning (step 1), the model H is initialized by pretrained parameters $\mathbf{\Theta}_{pre}$, and then the Camera-Aware Similarity Consistency Loss L in Eq. (5) is optimized in the global feature space. In fine consistency learning (step 2), the model H is initialized by the parameters $\mathbf{\Theta}_1$ learned in step 1, and then the step 2 Camera-Aware Similarity Consistency Loss L_2 in Eq. (8) is optimized to further learn similarity consistency in local neighbourhood of the feature space. The scheme aims at learning consistent pairwise similarity distributions globally and locally from coarse to fine (best viewed in colour).

are randomly sampled from all samples. Then the Camera-Aware Similarity Consistency Loss (Eq. (5)) is applied to learn the model $H(\cdot;\mathbf{\Theta}_1)$ for step 1.

Step 2: Fine Consistency Learning. Fine consistency learning is based on coarse consistency learning. To learn the model $H(\cdot;\mathbf{\Theta}_2)$ parameterized by $\mathbf{\Theta}_2$ for step 2, we first initialize $H(\cdot;\mathbf{\Theta}_2)$ by the parameters $\mathbf{\Theta}_1$ learned in step 1. Then, we further model pairwise similarities in the local neighbourhood in the feature space, that is, the sample pairs of top-ranked nearest samples are used.

For mini-batch learning, we use samples of two randomly sampled cameras p and q. For camera $c \in \{p,q\}$, $\mathbf{X}_c$ is the feature matrix extracted by model $H(\cdot;\mathbf{\Theta}_2)$ of step 2 and let $\mathbf{F}_{c(s1)}$ denote the fixed feature matrix extracted by the model $H(\cdot;\mathbf{\Theta}_1)$ trained in step 1.

When forming a batch for computing intra-/cross-camera similarity consistency loss, we first randomly sample one sample $\mathbf{I}_{p,r}$ of camera p. Then, we search for the top-K nearest samples for $\mathbf{I}_{p,r}$ in both camera p and camera q to extract feature matrices $\mathbf{X}^{(K)}_{p,(r)}, \mathbf{X}^{(K)}_{q,(r)} \in \mathbb{R}^{d\times K}$. The cosine similarities for searching nearest neighbours are computed by fixed feature matrices $\mathbf{F}_{p(s1)}$ and $\mathbf{F}_{q(s1)}$.

The step 2 intra-/cross-camera similarity consistency loss L_{con2} is

$$\min_{\mathbf{\Theta}_2} L_{con2} = \sum_{r,p\neq q} (\mathrm{mean}(\mathbf{X}^{(K)\top}_{p,(r)}\mathbf{X}^{(K)}_{p,(r)}) - \mathrm{mean}(\mathbf{X}^{(K)\top}_{p,(r)}\mathbf{X}^{(K)}_{q,(r)}))^2 + (\mathrm{std}(\mathbf{X}^{(K)\top}_{p,(r)}\mathbf{X}^{(K)}_{p,(r)}) - \mathrm{std}(\mathbf{X}^{(K)\top}_{p,(r)}\mathbf{X}^{(K)}_{q,(r)}))^2, \quad (6)$$

where $\mathrm{mean}(\cdot)$ and $\mathrm{std}(\cdot)$ are defined as in Eq. (1). The terms $\mathrm{mean}(\mathbf{X}^{(K)\top}_{p,(r)}\mathbf{X}^{(K)}_{p,(r)})$ and $\mathrm{std}(\mathbf{X}^{(K)\top}_{p,(r)}\mathbf{X}^{(K)}_{p,(r)})$ are regarded as constants in optimization, since we expect to keep the learned reliable knowledge of intra-camera matching as much as possible in this finetuning process.

When forming $\mathbf{X}_c$ and $\mathbf{F}_{c(s1)}$ of camera $c \in \{p,q\}$ in a batch for computing intra-camera similarity preserving loss, the samples are randomly sampled. The step 2 intra-camera similarity preserving loss L_{pre2} is

$$\min_{\mathbf{\Theta}_2} L_{pre2} = \sum_{c\in\{p,q\}} \left\| \log(\mathbf{X}^\top_c\mathbf{X}_c) - \log(\mathbf{F}^\top_{c(s1)}\mathbf{F}_{c(s1)}) \right\|^2_F, \quad (7)$$

where $\log(\cdot)$ is defined as in Eq. (3).

The step 2 Camera-Aware Similarity Consistency Loss L_2 for fine consistency learning is

$$\min_{\mathbf{\Theta}_2} L_2 = L_{pre2} + \lambda_2 L_{con2}, \quad (8)$$

where λ_2 is a trade-off parameter.

In testing stage, cosine distance between features extracted by model $H(\cdot;\mathbf{\Theta}_2)$ trained by the coarse-to-fine consistency learning scheme is used for retrieval.

5. Experiments

We evaluated on two large person re-identification benchmark datasets Market-1501 [44] and DukeMTMC [48]. We compared our method with the state-of-the-art unsupervised person re-identification methods and further evaluated the key components and parameters in our method.

Experiment Settings and Datasets. The experiments were conducted on Market-1501 [44] and DukeMTMC [48] in

the unsupervised setting. Market-1501 [44] contains 32,217 images of 1,501 identities in 6 cameras. DukeMTMC-reID [48] consists of 36,411 images of 1,812 identities in 8 cameras. We followed the standard train/test split of Market-1501 [44] and DukeMTMC [48]. In training, we first pretrained our model on MSMT17 [34] to learn common knowledge of Re-ID. Then, we trained our model on the training set of Market-1501 or DukeMTMC without using identity labels. The performance metrics, cumulative matching characteristic (CMC) and mean Average Precision (mAP), were applied following the standard evaluation protocols in [44] and [48].

Implementation Details. For the feature extractor H, we adopted a ResNet-50 [12] model trained by the strategy of PCB [29]. We initialized the feature extractor H by pretraining on the training set of MSMT17 [34]. The input images were resized to 384×128. Our model was trained in two steps by coarse-to-fine consistency learning scheme (Section 4). In the coarse consistency learning step, we set $\lambda = 10.0$ (the weight of L_{con} in Eq. (5)). In the fine consistency learning step, we set $\lambda_2 = 1.0$ (the weight of L_{con2} in Eq. (8)) and set $K = 16$ (the number of top-ranked samples in Eq. (6)). For mini-batch learning, we used batch size of 64. In each batch, we randomly sampled two cameras and then sampled 32 samples for each camera, of which the sampling method is introduced in Section 4. When computing the terms L_{pre} in Eq. (3) and L_{pre2} in Eq. (7), they were divided by batch size to normalize the scales. For optimization, we used SGD optimizer [5] with momentum 0.9. We used 15 epochs for both coarse consistency learning and fine consistency learning. The learning rate was 0.1 in the first 10 epochs and was reduced to 0.01 in the last 5 epochs.

5.1. Comparison to Related Unsupervised Models

We compared with unsupervised Re-ID methods including unsupervised features LOMO [18], BOW [44] and unsupervised learning models UMDL [26], PTGAN [34], PUL [34], CAMEL [40], SPGAN [7], TJ-AIDL [32] and HHL [49]. The experiment results are shown in Table 1. We also evaluated using Market-1501 [44] or DukeMTMC [48] for pretraining and reported the results in Table 2.

Our method outperformed all compared unsupervised Re-ID methods. Among the competitive compared methods, CAMEL [40], PTGAN [34], SPGAN [7], HHL [49] also aim at alleviating the effect of cross-camera scene variation mentioned in Section 1. Compared to these methods that focus on camera-to-camera alignment, our method further explore the relation of pairwise similarity between intra-camera matching and cross-camera matching, so that cross-camera matching can benefit from the reliable prior common knowledge in intra-camera matching of the pretrained model, which is ignored in existing methods.

Table 1. Comparison with the state-of-the-art unsupervised Re-ID methods. Our model is pretrained using MSMT17 [34] as source dataset. "R-k" denotes rank-k accuracy (%). "mAP" denotes mean average precision (%). "-" denotes not reported.

Methods	Market-1501				DukeMTMC			
	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
LOMO [18]	27.2	41.6	49.1	8.0	12.3	21.3	26.6	4.8
BOW [44]	35.8	52.4	60.3	14.8	17.1	28.8	34.9	8.3
UMDL [26]	34.5	52.6	59.6	12.4	18.5	31.4	37.6	7.3
PTGAN [34]	45.5	60.7	66.7	20.5	30.0	43.4	48.5	16.4
PUL [8]	51.5	70.1	76.8	22.8	41.1	56.6	63.0	22.3
CAMEL [40]	54.5	-	-	26.3	-	-	-	-
SPGAN [7]	57.7	75.8	82.4	26.7	46.4	62.3	68.0	26.2
TJ-AIDL [32]	58.2	74.8	81.1	26.5	44.3	59.6	65.0	23.0
HHL [49]	62.2	78.8	84.0	31.4	46.9	61.0	66.7	27.2
Ours	**65.4**	**80.6**	**86.2**	**35.5**	**59.3**	**73.2**	**77.8**	**37.8**

Table 2. Comparison with unsupervised Re-ID methods using Market-1501 [44] or DukeMTMC [48] as source dataset for pretraining. The notations are the same as those in Table 1.

Source dataset	DukeMTMC				Market-1501			
Target dataset	Market-1501				DukeMTMC			
Methods	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
PUL [8]	51.5	70.1	76.8	22.8	41.1	56.6	63.0	22.3
TJ-AIDL [32]	58.2	74.8	81.1	26.5	44.3	59.6	65.0	23.0
HHL [49]	62.2	78.8	84.0	31.4	46.9	61.0	66.7	27.2
Ours	**64.7**	**80.2**	**85.6**	**35.6**	**51.5**	**66.7**	**71.7**	**30.5**

Table 3. Component-wise evaluation of our method. "Pretrained model" is the baseline. "L_{pre} (Eq. (3))" and "L_{con} (Eq. (1))" are the two terms in our Camera-Aware Similarity Consistency Loss L (Eq. (5)). "L_{pre} (w/o log)" denotes using Euclidean metric instead of Log-Euclidean metric in L_{pre}. "L_{pre} + L_{con} (step 1)" denotes our model trained by coarse consistency learning in step 1. "Full model (step 1 & 2)" denotes the full version of our method with two steps. The other notations are the same as those in Table 1.

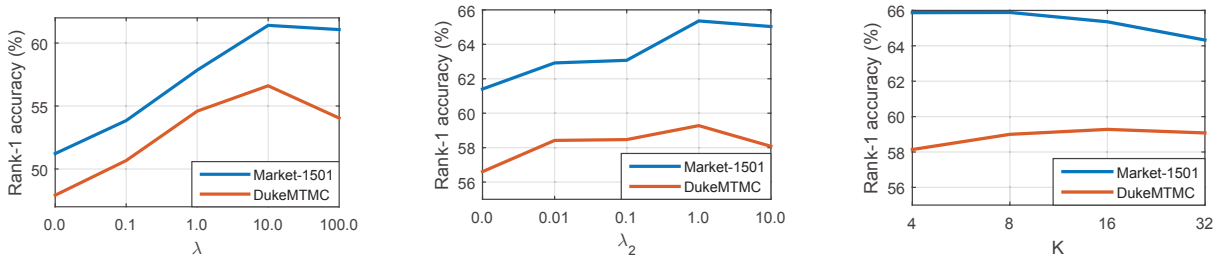
Methods	Market-1501				DukeMTMC			
	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
Pretrained model	51.5	67.2	73.7	24.9	47.6	64.2	70.4	30.6
L_{pre} (Eq. (3))	51.2	66.9	73.4	25.1	47.9	63.8	69.9	31.0
L_{con} (Eq. (1))	54.4	72.3	78.7	23.6	40.6	57.1	63.7	18.9
L_{pre} (w/o log) + L_{con}	59.2	75.4	81.4	31.4	55.4	70.9	76.2	36.2
L_{pre} + L_{con} (step 1)	61.4	78.0	83.8	32.1	56.6	71.8	76.9	35.8
Full model (step 1 & 2)	65.4	80.6	86.2	35.5	59.3	73.2	77.8	37.8

5.2. Further Evaluations

In this section, we further evaluate and analyse the components and parameters of our method.

Evaluation of Key Components. We verified the effectiveness of key components in our method, including the terms L_{con} and L_{pre} and two steps in coarse-to-fine consistency learning. The component-wise evaluations are as follows.

The pretrained model was regarded as the baseline model. In coarse-to-fine consistency learning in step 1, we applied the two terms in our Camera-Aware Similarity Consistency Loss L (Eq. (5)) individually, i.e., L_{pre} (Eq. (3)) and L_{con} (Eq. (1)), to show that they rely on each other. Then, we applied L by combining L_{pre} and L_{con} (denoted by "L_{pre} + L_{con} (step 1)"), i.e., coarse consistency learning



(a) Effect of λ (b) Effect of λ_2 (c) Effect of K

Figure 4. Effect of parameters λ, λ_2 and K. Parameter λ is the weight of intra-/cross-camera similarity consistency loss L_{con} in L in Eq. (5) in coarse consistency learning in step 1. Parameter λ_2 is the weight of intra-/cross-camera similarity consistency loss L_{con2} in L_2 in Eq. (8) in fine consistency learning in step 2. Parameter K is the number of top-ranked samples for computing L_{con2} in Eq. (6).

in step 1. To show the effectiveness of the Log-Euclidean metric in L_{pre}, we also compared with the case of using Euclidean metric (denoted by "L_{pre} (w/o log) + L_{con}"). Based on the model of step 1, we further applied fine consistency learning in step 2, which is the full version of our model (denoted by "Full model (step 1 & 2)"). The experiment results are shown in Table 3.

It can be observed that, using L_{pre} or L_{con} individually cannot bring improvement and the performance was even worse for L_{con} on DukeMTMC [48]. As analysed in Section 3.2.3, L_{con} is the leading role for consistency learning; while L_{pre} provides prior common knowledge as guiding information for regularizing consistency learning in L_{con} to avoid learning ineffective knowledge. Thus, they rely on each other. When L_{pre} and L_{con} are jointly learned, "L_{pre} + L_{con} (step 1)" achieved the best performance in step 1.

"L_{pre} + L_{con} (step 1)" is better than "L_{pre} (w/o log) + L_{con}" in most cases, since Log-Euclidean metric in L_{pre} can better preserve similarity than Euclidean metric because of the symmetric positive definite (SPD) property of intra-camera similarity matrices as explained in Section 3.2.3.

With fine consistency learning in step 2 in "Full model (step 1 & 2)", the performance was further improved as compared to "L_{pre} + L_{con} (step 1)", which shows the effectiveness of our coarse-to-fine consistency learning scheme.

Similarity Consistency Learning v.s. Feature Distribution Alignment. We propose to alleviate the effect of cross-camera scene variation by camera-aware similarity consistency learning. The cross-camera scene variation problem can also be regarded as feature distribution misalignment problem, thus domain adaptation methods for distribution alignment are closely related to this problem. We compared with two representative methods MMD [11] and CORAL [28]. When applied to Re-ID, MMD [11] minimizes the differences between the means of different cameras and CORAL minimizes the differences between the covariance matrices of different cameras. We also compared with the case of replacing our intra-/cross-camera similarity consistency loss L_{con} in L with MMD or CORAL to show the advantage of similarity consistency learning against feature distribution alignment. The results are reported in Table 4.

Table 4. Comparison with domain adaptation methods MMD [11] and CORAL [28] for feature distribution alignment. "Pretrained model" is the baseline. The other notations are as those in Table 3.

Methods	Market-1501				DukeMTMC			
	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
Pretrained model	51.5	67.2	73.7	24.9	47.6	64.2	70.4	30.6
MMD [11]	28.1	46.9	55.5	8.3	30.2	46.8	53.9	12.2
CORAL [28]	23.7	39.3	47.3	8.1	14.0	26.4	32.8	7.2
L_{pre} (Eq. (3)) + MMD	58.7	75.2	81.2	30.4	55.1	70.8	76.1	35.6
L_{pre} (Eq. (3)) + CORAL	58.5	75.4	81.7	29.7	54.6	70.8	76.0	34.8
L_{pre} + L_{con} (step 1)	61.4	78.0	83.8	32.1	56.6	71.8	76.9	35.8
Full model (step 1 & 2)	65.4	80.6	86.2	35.5	59.3	73.2	77.8	37.8

The results of MMD [11] and CORAL [28] are even much worse than the baseline pretrained model, because simply minimizing the differences between feature distributions without constraint degrades the common knowledge of Re-ID in the pretrained model and thus cannot avoid learning ineffective knowledge. When our intra-camera similarity preserving loss L_{pre} was applied with MMD and CORAL, feature distribution alignment with preserved common knowledge of intra-camera matching can bring improvement. This indicates that the common knowledge preserved in intra-camera similarity is reliable and significant.

Both our "L_{pre} + L_{con} (step 1)" and "Full model (step 1 & 2)" outperformed "L_{pre} (Eq. (3)) + MMD" and "L_{pre} (Eq. (3)) + CORAL", which shows the advantage of our camera-aware similarity consistency learning and coarse-to-fine consistency learning scheme. Compared with MMD and CORAL that align distribution in the feature space, our similarity consistency learning aligns the pairwise similarity distributions of intra-camera matching and cross-camera matching in the similarity space, which can benefit from the common knowledge preserved in similarities of intra-camera matching and thus is more robust. Moreover, MMD and CORAL cannot model samples in local neighbourhood in the feature space as our full model trained by the coarse-to-fine consistency learning scheme.

Parameter Evaluation. There are mainly three key parameters in our method, which are the weight λ of L_{con} in L

Figure 5. Some matching examples of direct transfer (applying the pretrained model), coarse consistency learning (step 1) and fine consistency learning (step 2) in testing on DukeMTMC [48] are shown. The correct matchings are indicated by green bounding boxes with ticks. In the matching results of direct transfer, incorrect samples in cameras with very similar background as compared to the query image are retrieved because of inconsistent pairwise similarity distributions of different camera pairs caused by cross-camera scene variation. Our method can improve matching results by coarse-to-fine consistency learning in two steps.

in Eq. (5), the weight λ_2 of L_{con2} in L_2 in Eq. (8) and the number of top-ranked samples K in L_{con2} in Eq. (6). We evaluated and analysed these parameters on Market-1501 [44] and DukeMTMC [48] as follows.

- Effect of Parameter λ. Parameter λ is the weight of L_{con} in Eq. (5) controlling the effect of intra-/cross-camera similarity consistency learning, which is used in coarse consistency learning in step 1. We varied λ from 0.0 to 100.0 and show the testing rank-1 accuracies of step 1 in Figure 4(a). With λ increasing in a wide range from 0.0 to 10.0, the improvement was increasingly significant. When λ was too large, L_{con} dominated L, so that the regularization of L_{pre} was weakened and cannot provide guiding information.

- Effect of Parameter λ_2. Parameter λ_2 is the weight of L_{con2} in Eq. (8) controlling the effect of intra-/cross-camera similarity consistency learning in the local neighbourhood of the feature space for fine consistency learning in step 2. As a step of further improving the model based on coarse consistency learning, we set λ_2 for fine consistency learning smaller than λ as the strategy of finetuning a model. We varied λ_2 from 0.0 to 10.0 and show the testing rank-1 accuracies of step 2 in Figure 4(b). The performance was improved when λ_2 was from 0.01 to 1.0.

- Effect of Parameter K. K is the number of top-ranked samples for computing the intra-/cross-camera similarity consistency loss L_{con2} in Eq. (6) for fine consistency learning in step 2. We varied K from 4 to 32 and show the testing rank-1 accuracies in Figure 4(c). It can be observed that fine consistency learning is rather not sensitive to K and the performance variation is lower than 2% when $K \in [4, 32]$.

Matching Examples. To have better visual understanding, we show some matching examples of direct transfer (applying the pretrained model), coarse consistency learning (step 1) and fine consistency learning (step 2) in testing on DukeMTMC [48] in Figure 5. The correct matchings are indicated by green bounding boxes with ticks.

In the failed cases of direct transfer, the pedestrian appearance and background of the retrieved gallery images are very similar to the query image, but they are incorrect matchings from the same or similar camera of the query image, while the correct matchings suffer from cross-camera scene variation and the pairwise similarity distributions are inconsistent for different camera pairs as illustrated in the camera-aware similarity inconsistency problem in Section 3.1. Our proposed method can alleviate this problem and improve the matching results by coarse-to-fine consistency learning in two steps.

6. Conclusion

In this paper, we study unsupervised person re-identification and focus on alleviating the effect of cross-camera scene variation (e.g., illumination, background and viewpoint), which is serious for unsupervised Re-ID. Cross-camera scene variation causes domain shift in the feature space and leads to inconsistent pairwise similarity distributions for different camera pairs and thus degrades the matching performance. We call it the camera-aware similarity inconsistency problem. To solve this problem, we propose a Camera-Aware Similarity Consistency Loss. Different from existing Re-ID methods that deal with cross-camera scene variation problem by camera-to-camera alignment, we further explore the relation of pairwise similarity between intra-camera matching and cross-camera matching. We can improve cross-camera matching by learning consistent pairwise similarity distributions for intra-camera and cross-camera matching with the guidance of the preserved reliable common knowledge of Re-ID in intra-camera matching. For more effective consistency learning, we further develop a coarse-to-fine consistency learning scheme to learn consistency globally and locally in two steps. The experiments show that our method outperformed the state-of-the-art unsupervised Re-ID methods.

Acknowledgement

This work was supported partially by NSFC (U1611461, U1811461, 61573387), Guangdong Province Science and Technology Innovation Leading Talents (2016TX03X157), Guangzhou Research Project (201902010037) and the Royal Society Newton Advanced Fellowship (NA150459).

References

[1] Ejaz Ahmed, Michael Jones, and Tim K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015. 1, 2

[2] Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Fast and simple calculus on tensors in the log-euclidean framework. In *MICCAI*, 2005. 4

[3] Slawomir Bak and Peter Carr. One-shot metric learning for person re-identification. In *CVPR*, 2017. 2

[4] Slawomir Bak, Peter Carr, and Jean-Francois Lalonde. Domain adaptation through synthesis for unsupervised person re-identification. In *ECCV*, 2018. 2

[5] Lon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT*, 2010. 6

[6] Ying-Cong Chen, Wei-Shi Zheng, Jian-Huang Lai, and Pong Yuen. An asymmetric distance model for cross-view feature mapping in person re-identification. *IEEE TCSVT*, 2015. 1, 2

[7] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification. In *CVPR*, 2018. 1, 2, 6

[8] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2018. 1, 2, 6

[9] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010. 2

[10] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*. 2008. 2

[11] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *JMLR*, 2012. 2, 7

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 2, 3, 6

[13] Elyor Kodirov, Tao Xiang, Zhenyong Fu, and Shaogang Gong. Person re-identification by unsupervised $\ell 1$ graph learning. In *ECCV*, 2016. 1, 2

[14] Martin Köstinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012. 1, 2

[15] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised person re-identification by deep learning tracklet association. In *ECCV*, 2018. 2

[16] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 1, 2

[17] Zhen Li, Shiyu Chang, Feng Liang, Thomas S Huang, Liangliang Cao, and John R Smith. Learning locally-adaptive decision functions for person verification. In *CVPR*, 2013. 2

[18] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015. 1, 2, 6

[19] Ji Lin, Liangliang Ren, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Consistent-aware deep learning for person re-identification in a camera network. In *CVPR*, 2017. 2

[20] Chunxiao Liu, Shaogang Gong, Chen Change Loy, and Xinggang Lin. Person re-identification: what features are important? In *ECCV Workshop*, 2012. 2

[21] Lianyang Ma, Xiaokang Yang, and Dacheng Tao. Person re-identification over camera networks using multi-task distance metric learning. *IEEE TIP*, 2014. 2

[22] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato. Hierarchical gaussian descriptor for person re-identification. In *CVPR*, 2016. 2

[23] Alexis Mignon and Frédéric Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *CVPR*, 2012. 2

[24] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton van den Hengel. Learning to rank in person re-identification with metric ensembles. In *CVPR*. 2015. 2

[25] Sateesh Pedagadi, James Orwell, Sergio Velastin, and Boghos Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *CVPR*, 2013. 2

[26] Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaogang Gong, Tiejun Huang, and Yonghong Tian. Unsupervised cross-dataset transfer learning for person re-identification. In *CVPR*, 2016. 1, 2, 6

[27] Bryan Prosser, Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by support vector ranking. In *BMVC*, 2010. 2

[28] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV Workshop*, 2016. 2, 7

[29] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling. In *ECCV*, 2018. 1, 2, 6

[30] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. *arXiv preprint arXiv:1702.05464*, 2017. 2

[31] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, 2016. 2

[32] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. *CVPR*, 2018. 1, 2, 6

[33] Xiaojuan Wang, Wei-Shi Zheng, Xiang Li, and Jianguo Zhang. Cross-scenario transfer person reidentification. *IEEE TCSVT*, 2016. 2

[34] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 2018. 1, 2, 3, 6

[35] Kilian Q Weinberger, John Blitzer, and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, 2005. 2

[36] Shangxuan Wu, Ying-Cong Chen, Xiang Li, Ancong Wu, Jinjie You, and Wei-Shi Zheng. An enhanced deep feature representation for person re-identification. In *WACV*, 2016. 2

[37] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, 2016. 2

[38] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*. 2014. 2

[39] Jinjie You, Ancong Wu, Xiang Li, and Wei-Shi Zheng. Top-push video-based person re-identification. In *CVPR*, 2016. 2

[40] Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *ICCV*, 2017. 1, 2, 6

[41] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a discriminative null space for person re-identification. In *CVPR*, 2016. 2

[42] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*, 2017. 2

[43] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, 2017. 2

[44] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 5, 6, 8

[45] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Reidentification by relative distance comparison. *IEEE TPAMI*, 2013. 1, 2

[46] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Towards open-world person re-identification by one-shot group-based verification. *IEEE TPAMI*, 2016. 2

[47] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shaogang Gong. Partial person re-identification. In *ICCV*, 2015. 2

[48] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017. 1, 2, 3, 5, 6, 7, 8

[49] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero-and homogeneously. In *ECCV*, 2018. 1, 2, 6

[50] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *CVPR*, 2018. 2