Distilled Person Re-identification: Towards a More Scalable System

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

[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]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
[4]Guangdong Province Key Laboratory of Information Security, China
[5]YouTu Lab, Tencent

wuancong@gmail.com, wszheng@ieee.org, scorpioguo@tencent.com, stsljh@mail.sysu.edu.cn

Abstract

Person re-identification (Re-ID), for matching pedestrians across non-overlapping camera views, has made great progress in supervised learning with abundant labelled data. However, the scalability problem is the bottleneck for applications in large-scale systems. We consider the scalability problem of Re-ID from three aspects: (1) low labelling cost by reducing label amount, (2) low extension cost by reusing existing knowledge and (3) low testing computation cost by using lightweight models. The requirements render scalable Re-ID a challenging problem. To solve these problems in a unified system, we propose a Multi-teacher Adaptive Similarity Distillation Framework, which requires only a few labelled identities of target domain to transfer knowledge from multiple teacher models to a user-specified lightweight student model without accessing source domain data. We propose the Log-Euclidean Similarity Distillation Loss for Re-ID and further integrate the Adaptive Knowledge Aggregator to select effective teacher models to transfer target-adaptive knowledge. Extensive evaluations show that our method can extend with high scalability and the performance is comparable to the state-of-the-art unsupervised and semi-supervised Re-ID methods.

1. Introduction

With the development of surveillance systems, person re-identification (Re-ID) has drawn much attention in recent years. Most researches focus on supervised learning [25, 67, 31, 28, 1, 47] and have made great progress. However, system extension is still a significant obstacle for applying Re-ID in large-scale surveillance systems because of the scalability problem, which is still under-explored. Some previous works attempt to improve scalability from different aspects, such as unsupervised and transfer learning [42, 24, 62, 52, 51, 11] for reducing label amount and fast retrieval [56, 73] for large-scale applications. To build a more scalable Re-ID system, it is expected that these scalability problems can be addressed in a unified system.

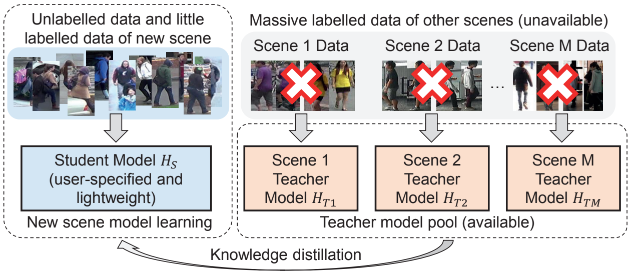
Unlabelled data and little labelled data of new scene
Massive labelled data of other scenes (unavailable)
Scene 1 Data
Scene 2 Data
...
Scene M Data
Student Model H_S (user-specified and lightweight)
New scene model learning
Scene 1 Teacher Model H_{T1}
Scene 2 Teacher Model H_{T2}
Scene M Teacher Model H_{TM}
Teacher model pool (available)
Knowledge distillation

Figure 1. A scalable adaptation Re-ID system. We propose to store the knowledge in teacher models trained in existing scenes. When extending to a new scene, we can flexibly transfer the knowledge from teacher models to a user-specified lightweight student model by using unlabelled data and little labelled data of the new scene, without using source domain data which may not be accessible.

We consider the requirements and challenges of the scalability problem of Re-ID mainly from three aspects:
(1) Low labelling cost. Existing supervised Re-ID methods require abundant labelled identities, which is unrealistic for a large-scale system. A scalable Re-ID system should be able to learn from unlabelled data and limited labelled data, which provide very limited information for learning.
(2) Low extension cost. When extending to a new scene, most existing Re-ID methods apply transfer learning, which requires auxiliary source domain data for pretraining or joint training. In some cases, source domain data of other scenes may not be accessible because of privacy problem or transfer problem. Even if source domain data is available, joint training increases computation costs. Moreover, pretrained models may not be applicable due to different user-specified requirements of model architecture and capacity. Thus, a scalable Re-ID system should be able to extend to a new scene flexibly with low cost. Knowledge transfer without accessing source domain data is a challenge.
(3) Low testing computation cost. Existing state-of-the-art

[*] Corresponding author

methods for Re-ID are based on large neural network models, e.g. ResNet-50 [20], which cannot meet the requirement of the camera hardware development trend of front-end processing on the chips. Thus, a scalable Re-ID system should be able to learn lightweight models.

To solve the above problems, we propose a scalable adaptation Re-ID system based on knowledge reuse as shown in Fig. 1. The system can extend with unlabelled data and only a few labelled identities in the target domain, without using source domain data. Instead of storing raw data, we store the knowledge of Re-ID for an existing scene i in a teacher model H_{Ti}. When extending to a new scene, knowledge in a teacher model pool $\{H_{Ti}\}_{i=1}^{M}$ is aggregated and transferred to a new user-specified lightweight student model H_S for the target domain. In this system, knowledge transfer and knowledge aggregation are two key problems and we address them as follows.

Knowledge transfer in our system is challenging, because limited labelled target data and absence of source data lead to little information for learning and knowledge need be compressed in a lightweight model. To solve these problems, we need to imitate and distill the knowledge in the teacher models, which is knowledge distillation [21]. For Re-ID, we exploit knowledge embedded in similarity and propose the Log-Euclidean Similarity Distillation Loss for imitating the sample pairwise similarity matrix of the teacher. Most knowledge distillation methods [21, 37, 61] are designed for closed-set classification and they convey knowledge by soft labels. They are not suitable for Re-ID, because Re-ID is an open-set identification problem, in which the identities are non-overlapping in training and testing.

Furthermore, to effectively aggregate knowledge from multiple teachers, we propose the Adaptive Knowledge Aggregator for adjusting the contributions of multiple teachers in the distillation loss dynamically, in order to select effective teachers and aggregate effective knowledge for the target domain. Only a few labelled identities (e.g. 10) are needed as validation data for computing empirical risk to guide knowledge aggregation. We further integrate the Log-Euclidean Similarity Distillation Loss and the Adaptive Knowledge Aggregator into a Multi-teacher Adaptive Similarity Distillation Framework as shown in Fig. 2.

In summary, the contributions of this paper are: (1) We propose the Log-Euclidean Similarity Distillation Loss for knowledge distillation for Re-ID, which is an open-set identification problem. (2) We propose the Adaptive Knowledge Aggregator for aggregating effective knowledge from multiple teacher models for learning a lightweight student model. (3) We further integrate them in a Multi-teacher Adaptive Similarity Distillation Framework for scalable person re-identification, which can simultaneously reduce labelling cost, extension cost and testing computation cost.

2. Related Work

Supervised Person Re-identification. Person Re-identification has witnessed a fast growing development in recent years, from feature design [19, 15, 32, 31, 35, 69, 58] to distance metric learning [54, 19, 43, 25, 67, 36, 41, 30, 58, 33, 38, 31, 9, 68, 60, 63, 69, 29, 52, 6] and end-to-end deep learning [28, 1, 55, 57, 49, 22, 64, 65, 70, 47]. Most existing works rely on abundant labelled data. Although high performance can be achieved by deep models, heavy labelling cost hinders the scalability of these methods.

Scalable Person Re-identification. Recently, scalable person re-identification has drawn more attention for reducing costs of system extension. Unsupervised learning [42, 24, 62, 51, 14, 53, 13, 7, 27, 71], transfer learning [72, 53, 13, 11], small sample learning [63, 2] and active learning [34, 50, 45] are for reducing labelling cost by minimizing the requirement of labelled data or selecting specific data for labelling. Fast adaptation [5, 39] is for reducing computation cost in training. Fast retrieval by lightweight model [56] and binary representation [8, 73] are for reducing computation cost in testing. These methods only address one of the problems of labelling cost, extension cost or testing computation cost, while our method address these problems simultaneously in a unified framework.

Knowledge Transfer/Distillation. Knowledge transfer/distillation [21] is for transferring knowledge from a large teacher model to a smaller student model by imitation. A vast majority of knowledge distillation methods such as [21, 37, 61, 16] are designed for closed-set classification problems by using soft labels of the teacher model to guide learning the student model. However, person re-identification is an open-set identification problem, in which the identities in training and testing are non-overlapping, and thus the soft-label-based distillation methods are not so suitable for Re-ID. Some methods also consider using information other than soft labels as knowledge. Fitnets [44] and FSP [59] exploit feature maps and PKT [40] exploits the probability distribution of data, which are not directly related to measuring similarity for matching in Re-ID. To more effectively represent and convey knowledge, we distill the knowledge embedded in sample similarity by imitating the teacher similarity matrix. As for distilling from multiple teacher models as in our proposed method, [16, 61] exploit the ensemble of multiple teachers, but they are for closed-set classification and the contributions of different teachers cannot be adaptively adjusted as in our method. Semi-supervised teacher-student frameworks [48, 17] are for closed-set classification and cannot solve our problem.

Hypothesis transfer learning (HTL) [26] studies learning from source models without source data. Existing HTL methods [3, 12] are for closed-set domain adaptation, which cannot solve the open-set identification problem for Re-ID.

3. Similarity Knowledge Distillation

As designed in the scalable adaptation Re-ID system (Fig. 1) in Section 1, to learn a model for the target domain with only a few labelled data, we can transfer knowledge from other domains. To transfer knowledge without accessing source domain data, we regard the model as a student, which learns to imitate the teacher models by knowledge distillation. In most knowledge distillation methods [21, 37, 16], soft labels are utilized. However, it is not so suitable to convey knowledge by identity soft labels for Re-ID, because Re-ID is an open-set identification problem in which there is no overlap of identity in training and testing.

To overcome this problem, we exploit the knowledge embedded in similarity and make the student model imitate the pairwise similarities of the teacher model. For N unlabelled image samples $\{\mathbf{I}_i\}_{i=1}^{N}$, let $\mathbf{A}$ denote the pairwise similarity matrix, where the element $a_{i,j}$ in the i-th row and the j-th column of $\mathbf{A}$ is the similarity between samples $\mathbf{I}_i$ and $\mathbf{I}_j$ determined by model H. Let H_S denote the student model to be learned and H_T denote the teacher model that is fixed. The pairwise similarity matrices of the student model H_S and the teacher model H_T are denoted by $\mathbf{A}_S$ and $\mathbf{A}_T$, respectively. To transfer knowledge from teacher to student, we minimize the distance between the student similarity matrix $\mathbf{A}_S$ and the teacher similarity matrix $\mathbf{A}_T$ as follow:

$$\min dist(\mathbf{A}_S, \mathbf{A}_T), \tag{1}$$

where $dist(\cdot)$ is a distance metric for similarity matrices. Note that $\mathbf{A}_T$ is fixed as the target for learning $\mathbf{A}_S$.

3.1. Construction of Similarity Matrices

Student Similarity Matrix $\mathbf{A}_S$. To obtain the pairwise similarity matrix $\mathbf{A}_S$ for student model H_S, we use cosine similarity, which is commonly used in neural networks for Re-ID [47]. For a sample $\mathbf{I}_i$, the student model H_S extracts a d-dimensional feature vector $H_S(\mathbf{I}_i; \mathbf{\Theta}_S)$, where $\mathbf{\Theta}_S$ is the parameter of H_S. Let $\mathbf{x}_{S,i} \in \mathbb{R}^d$ denote the non-negative normalized unit feature vector formulated by

$$\mathbf{x}_{S,i} = \mathrm{ReLU}(H_S(\mathbf{I}_i; \mathbf{\Theta}_S)) / \left\| \mathrm{ReLU}(H_S(\mathbf{I}_i; \mathbf{\Theta}_S)) \right\|, \tag{2}$$

where $\mathrm{ReLU}(x) = \max(0, x)$ is an activation function applied after the feature layer of H_S.

Let $\mathbf{X}_S = [\mathbf{x}_{S,1}, \mathbf{x}_{S,2}, ..., \mathbf{x}_{S,N}] \in \mathbb{R}^{d \times N}$ denote the feature matrix for samples $\{\mathbf{I}_i\}_{i=1}^{N}$. The student similarity matrix $\mathbf{A}_S$ is computed by

$$\mathbf{A}_S = \mathbf{X}_S^\top \mathbf{X}_S, \tag{3}$$

where $a_{S,i,j} = \mathbf{x}_{S,i}^\top \mathbf{x}_{S,j}$ in $\mathbf{A}_S$ is the cosine similarity between samples $\mathbf{I}_i$ and $\mathbf{I}_j$.

Properties of Student Similarity Matrix. We derive the properties of the student similarity matrix $\mathbf{A}_S$ as follows:
(1) The range of similarities in $\mathbf{A}_S$ is $[0, 1]$. The cosine similarity between non-negative unit feature vectors extracted by Eq. (2) is between 0 and 1.
(2) $\mathbf{A}_S$ is a symmetric positive semi-definite matrix. In Eq. (3), $\mathbf{X}_S^\top \mathbf{X}_S$ is symmetric positive semi-definite. In our case of mini-batch learning, the feature dimension d is larger than the batch size N. Generally, $\mathbf{X}_S \in \mathbb{R}^{d \times N}$ satisfies $\mathrm{rank}(\mathbf{X}_S) = N$ and has full rank, so that $\mathbf{A}_S \in \mathbb{R}^{N \times N}$ is a symmetric positive definite (SPD) matrix in this case.

Teacher Similarity Matrix $\mathbf{A}_T$. The teacher similarity matrix $\mathbf{A}_T$, as the target of student similarity matrix $\mathbf{A}_S$, should satisfy the properties of student similarity matrix.

Generally, when using neural network as teacher model, the teacher similarity matrix $\mathbf{A}_T$ can be computed as the student model in Eq. (3) by using the feature matrix $\mathbf{X}_T = [\mathbf{x}_{T,1}, \mathbf{x}_{T,2}, ..., \mathbf{x}_{T,N}]$ extracted by the teacher model H_T.

In other cases, the teacher similarity matrix $\mathbf{A}_T$ may not satisfy the two properties of the student similarity matrix. For example, when pairwise verification neural network is used as teacher model without constraint, the similarity matrix may not be SPD. Some simple transformations can be applied to make the teacher similarity matrix valid. First, if the range of similarities in $\mathbf{A}_T$ is not $[0, 1]$, it can be mapped to $[0, 1]$ by normalization. Second, if $\mathbf{A}_T$ is not symmetric positive definite, we can project it onto the cone of all positive semi-definite matrices as in [54]. More details are provided in the supplementary due to space limitation.

3.2. Log-Euclidean Similarity Distillation

After constructing the similarity matrices for student and teacher, we aim to distill the knowledge by minimizing the distance between the similarity matrices in Eq. (1). As analyzed above, the student and teacher similarity matrices are symmetric positive definite (SPD) matrices, which are intrinsically lying on a Riemannian manifold [4] instead of a vector space. Hence, when measuring the distance between $\mathbf{A}_S$ and $\mathbf{A}_T$, we take this property into consideration and measure the distance in a log-Euclidean Riemannian framework [4] instead of using Euclidean metric as follow:

$$dist(\mathbf{A}_S, \mathbf{A}_T) = \left\| \log(\mathbf{A}_S) - \log(\mathbf{A}_T) \right\|_F, \tag{4}$$

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

$$\log(\mathbf{A}) = \mathbf{U}\mathrm{diag}(\log(\lambda_1), \log(\lambda_2), ..., \log(\lambda_N))\mathbf{U}^\top, \tag{5}$$

where $\mathbf{U}$ is the orthonormal matrix of eigenvectors and λ_i is the eigenvalue, which are obtained from the eigendecomposition $\mathbf{A} = \mathbf{U}\mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_N)\mathbf{U}^\top$.

We distill the knowledge embedded in the similarity from teacher to student by minimizing the Log-Euclidean distance as follow:

$$\min_{\mathbf{\Theta}_S} L_T(\mathbf{X}_S) = \left\| \log(\mathbf{X}_S^\top \mathbf{X}_S) - \log(\mathbf{A}_T) \right\|_F^2, \tag{6}$$

where $\mathbf{\Theta}_S$ is the parameter of student model H_S, $\mathbf{X}_S$ is the feature matrix extracted by H_S and processed by Eq. (2),

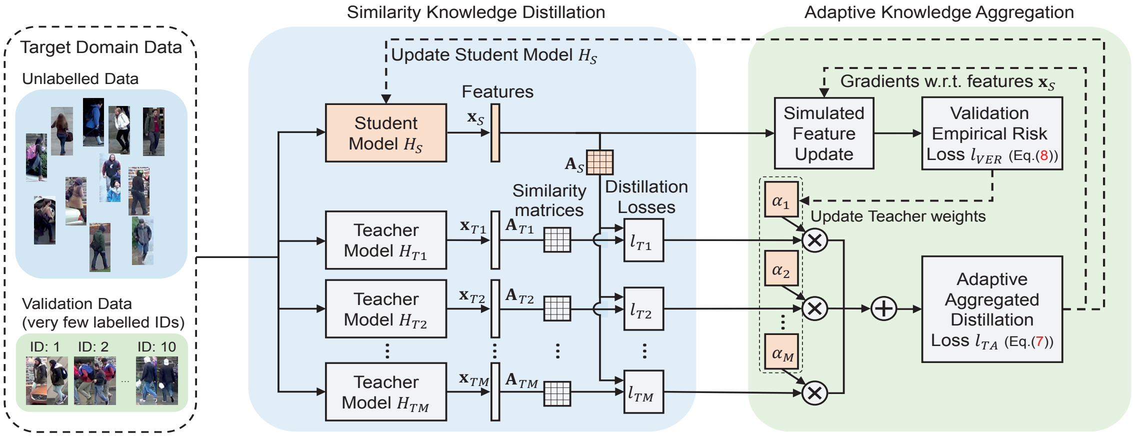

Figure 2. Overview of the Multi-teacher Adaptive Similarity Distillation Framework. Unlabelled data and a few labelled identities of target domain are required. The teacher models are fixed feature extractors for existing scenes in the system. The student model is a new user-specified lightweight model to be trained. The teacher weight α_i is for controlling the contribution of Log-Euclidean Similarity Distillation Loss L_{Ti} of teacher H_{Ti} and it is learned by minimizing validation empirical risk loss L_{VER}. The student model is trained by unlabelled data with the Adaptive Aggregated Distillation Loss L_{TA}. (← denotes forward propagation and ⇠ denotes backward propagation.)

$\mathbf{X}_S^\top \mathbf{X}_S$ is $\mathbf{A}_S$ as computed in Eq. (3) and $\mathbf{A}_T$ is fixed similarity matrix provided by teacher model H_T. We call L_T the *Log-Euclidean Similarity Distillation Loss* for H_T. The effectiveness of the Log-Euclidean metric is further validated in our experiments as compared to the Euclidean metric.

4. Learning to learn from Multiple Teachers

We have illustrated learning from a single teacher model by similarity knowledge distillation. In this section, we study learning from multiple teacher models. Since not all teachers can provide effective and complementary knowledge due to variations between cameras (e.g. illumination and background), we need learning to adjust the contributions of multiple teachers, i.e. learning to learn. We propose the Adaptive Knowledge Aggregator to aggregate effective knowledge from multiple teachers for learning a student model, which requires only a few labelled identities for reducing labelling cost. It is integrated with similarity knowledge distillation to form the Multi-teacher Adaptive Similarity Distillation Framework as shown in Fig. 2.

4.1. Multi-teacher Adaptive Aggregated Distillation

To learn a student model H_S from multiple teacher models in a teacher model pool $\{H_{Ti}\}_{i=1}^M$ simultaneously, the Log-Euclidean Similarity Distillation Loss L_T in Eq. (6) for a single teacher H_T is generalized as follow:

$$\min_{\mathbf{\Theta}_S} L_{TA}(\mathbf{X}_S; \{\alpha_i\}_{i=1}^M) = \sum_{i=1}^M \alpha_i L_{Ti}(\mathbf{X}_S), \tag{7}$$

where L_{Ti} is the Log-Euclidean Similarity Distillation Loss for teacher model H_{Ti}, α_i is the teacher weight for controlling the contribution of L_{Ti}. The teacher weight α_i should satisfy $\sum_{i=1}^M \alpha_i = 1$ and $\alpha_i \geq 0$. $\mathbf{X}_S$ is the feature matrix extracted by student model H_S parameterized by $\mathbf{\Theta}_S$.

The constrain $\sum_{i=1}^M \alpha_i = 1$ is for normalizing the scale of α_i. It can be simply satisfied by $\alpha_i = |\tilde{\alpha}_i| / \sum_{j=1}^M |\tilde{\alpha}_j|$, where $\tilde{\alpha}_i$ is an unconstrained real number parameter.

For unsupervised learning without prior knowledge, α_i can be set equally as $1/M$. However, in practice, there may be some ineffective teacher models that provide wrong knowledge. Hence, instead of regarding α_i as fixed hyper-parameter which needs tuning, we aim to learn α_i dynamically to make the loss L_{TA} adaptive to the target domain. We call L_{TA} the *Adaptive Aggregated Distillation Loss*.

4.2. Adaptive Knowledge Aggregation

To learn the teacher weights $\{\alpha_i\}_{i=1}^M$, guiding information is required. Generally, for validating whether a Re-ID system is working normally, it is necessary and feasible for human operator to label a small amount of identities (e.g. $\leq$ 10). Although the small amount of data is far from enough for training an effective model from scratch due to overfitting, we can compute the validation empirical risk on it to provide guiding information for aggregating knowledge.

Validation Empirical Risk. For the target domain, we have a large amount of unlabelled data $\mathcal{D}_U = \{\mathbf{I}_i^U\}_{i=1}^N$. Additionally, we label data of a few identities (not in $\mathcal{D}_U$) to form a small validation set $\mathcal{D}_L = \{(\mathbf{I}_i^L, y_i)\}_{i=1}^{N_v}$, where $y_i = 1, 2..., C_v$ is the label ($C_v = 10$ in our case). To indicate whether the student is learning correct knowledge from teachers, the empirical risk on validation data $\mathcal{D}_L$ can be computed. Let $\mathbf{x}_{S,k}^U$ and $\mathbf{x}_{S,i}^L$ denote the features of unlabelled sample $\mathbf{I}_k^U$ and labelled sample $\mathbf{I}_i^L$, respectively. Let $\mathcal{P}_L = \{(i,j) | y_i = y_j\}$ denote the set of index pair (i,j) of positive sample pair $(\mathbf{I}_i^L, \mathbf{I}_j^L)$ in the validation set $\mathcal{D}_L$. As there is no overlap identity in $\mathcal{D}_L$ and $\mathcal{D}_U$, $(\mathbf{I}_i^L, \mathbf{I}_k^U)$ is a negative sample pair. We apply a Softmax cross entropy loss for computing the validation empirical risk by

$$L_{VER}(\mathbf{X}_S^U, \mathbf{X}_S^L) = \sum_{(i,j)\in\mathcal{P}_L} -\log\frac{\exp(\mathbf{x}_{S,i}^{L\top}\mathbf{x}_{S,j}^L)}{\exp(\mathbf{x}_{S,i}^{L\top}\mathbf{x}_{S,j}^L) + \sum_{k=1}^N \exp(\mathbf{x}_{S,i}^{L\top}\mathbf{x}_{S,k}^U)}. \tag{8}$$

We call L_{VER} the *validation empirical risk loss*. When the similarity of positive pair $\mathbf{x}_{S,i}^{L\top}\mathbf{x}_{S,j}^L$ becomes larger and the similarity of negative pair $\mathbf{x}_{S,i}^{L\top}\mathbf{x}_{S,k}^U$ becomes smaller, the loss L_{VER} becomes smaller. Thus, it can indicate the empirical risk effectively.

Adaptive Knowledge Aggregator. To learn teacher weights $\{\alpha_i\}_{i=1}^M$ that can minimize the validation empirical risk L_{VER}, we propose the Adaptive Knowledge Aggregator for optimizing $\{\alpha_i\}_{i=1}^M$ by gradient descent.

In the learning process of student model H_S, at each step, we have the feature matrices $\mathbf{X}_S^U$ and $\mathbf{X}_S^L$ for unlabelled and labelled data, respectively. Feature learning is guided by gradient descent of the Adaptive Aggregated Distillation Loss L_{TA}. To evaluate whether the current teacher weights $\{\alpha_i\}_{i=1}^M$ are effective, it is expected that the features updated by L_{TA} parameterized by $\{\alpha_i\}_{i=1}^M$ can decrease the validation empirical risk loss L_{VER} to the most extent. We simulate one step update of the features $\mathbf{X}_S^U$ and $\mathbf{X}_S^L$ by using gradients of $L_{TA}(\mathbf{X}_S^U; \{\alpha_i\}_{i=1}^M)$ and $L_{TA}(\mathbf{X}_S^L; \{\alpha_i\}_{i=1}^M)$ with respect to $\mathbf{X}_S^U$ and $\mathbf{X}_S^L$ by

$$\begin{aligned} \mathbf{X}_S^{U'} &= \mathbf{X}_S^U - \beta\frac{\partial L_{TA}(\mathbf{X}_S^U; \{\alpha_i\}_{i=1}^M)}{\partial \mathbf{X}_S^U}, \\ \mathbf{X}_S^{L'} &= \mathbf{X}_S^L - \beta\frac{\partial L_{TA}(\mathbf{X}_S^L; \{\alpha_i\}_{i=1}^M)}{\partial \mathbf{X}_S^L}, \end{aligned} \tag{9}$$

where $\mathbf{X}_S^{U'}$ and $\mathbf{X}_S^{L'}$ are the simulated updated features and β is the step size of the simulated updating.

Then, we compute the validation empirical risk $L_{VER}(\mathbf{X}_S^{U'}, \mathbf{X}_S^{L'})$ of the updated features $\mathbf{X}_S^{U'}$ and $\mathbf{X}_S^{L'}$. Note that, the updated features $\mathbf{X}_S^{U'}$ and $\mathbf{X}_S^{L'}$ are related to the teacher weights α_i because the gradients $\frac{\partial L_{TA}(\mathbf{X}_S^U;\{\alpha_i\}_{i=1}^M)}{\partial \mathbf{X}_S^U}$ and $\frac{\partial L_{TA}(\mathbf{X}_S^L;\{\alpha_i\}_{i=1}^M)}{\partial \mathbf{X}_S^L}$ contain α_i. Thus, to minimize the validation empirical risk loss L_{VER}, the gradient $\frac{\partial L_{VER}(\mathbf{X}_S^{U'},\mathbf{X}_S^{L'})}{\partial \alpha_i}$ can be computed and the teacher weights α_i can be learned by gradient descent as follow:

$$\alpha_i' = \alpha_i - \gamma_\alpha\frac{\partial L_{VER}(\mathbf{X}_S^{U'}, \mathbf{X}_S^{L'})}{\partial \alpha_i}, \tag{10}$$

where α_i' is the updated value of teacher weight α_i by using learning rate γ_α.

With the objective of minimizing the validation empirical risk during training, the learning target $L_{TA}(\mathbf{X}_S; \{\alpha_i\}_{i=1}^M)$ parameterized by teacher weights $\{\alpha_i\}_{i=1}^M$ can adaptively select effective teacher models by weighting to provide better guidance for student model.

Optimization. Before training, the teacher weights α_1, α_2,..., α_M are initialized as $1/M$. There are mainly three steps in training: (1) Feature extraction and similarity matrix construction as illustrated in Section 3.1; (2) Updating teacher weights $\{\alpha_i\}_{i=1}^M$ by Adaptive Knowledge Aggregator; (3) Updating student model H_S by distillation loss L_{TA} with updated teacher weights $\{\alpha_i\}_{i=1}^M$. The three steps are repeated for training student model H_S. This process is also shown in Algorithm 1 in the supplementary.

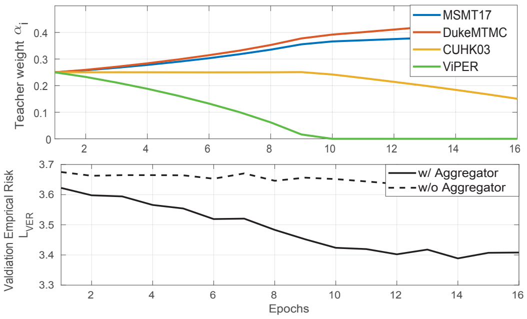

Figure 3. Change of teacher weights α_i (the upper figure) and change of validation empirical risk loss L_{VER} with and without using Adaptive Knowledge Aggregator (the lower figure) in the training process of Market-1501. The teacher weights can select more effective teacher models trained on larger datasets (MSMT17, DukeMTMC). The validation empirical risk loss L_{VER} can be further minimized with the learned teacher weights.

Visual Understanding. To better understand the effect of the adaptive teacher weights α_i, change of α_i and change of the validation empirical risk loss L_{VER} during training on Market-1501 are shown in Fig. 3. The experiment details are illustrated later in Section 5. It can be observed that, with the Adaptive Knowledge Aggregator, the learned teacher weights are larger for effective teachers (trained on large datasets MSMT17, DukeMTMC) and smaller for the ineffective teacher model (trained on small dataset ViPER). Compared with the case without using aggregator (i.e. using equal teacher weights), adaptive teacher weights can further minimize the validation empirical risk loss L_{VER}.

5. Experiments

We conducted extensive experiments on two large person re-identification benchmark datasets Market-1501 [66] and DukeMTMC [70]. Our Multi-teacher Adaptive Similarity Distillation Framework was evaluated and compared to knowledge distillation, unsupervised, semi-supervised and small sample learning methods. Further evaluations show the effectiveness of the Log-Euclidean Similarity Distillation Loss and the Adaptive Knowledge Aggregator.

Experiment Settings and Datasets. We used Market-1501 [66] and DukeMTMC [70] as target datasets. Source datasets were for training teacher models. To increase the diversity of effectiveness of teacher models for the target scene to simulate the practical situation, datasets of different scales collected in different scenes were used. We trained 5 teacher models $T1$, $T2$, $T3$, $T4$, $T5$ with labelled data in the training sets of MSMT17 [53], CUHK03 [28], ViPER [18], DukeMTMC [70] and Market-1501 [66], re-

Table 1. Basic information of datasets.

Teacher	Dataset	Images	Identities	Cameras	Scene
$T1$	MSMT17 [53]	126,441	4,101	15	outdoor, indoor
$T2$	CUHK03 [28]	28,192	1,467	2	indoor
$T3$	VIPeR [18]	1,264	632	2	outdoor
$T4$	DukeMTMC [70]	36,411	1,812	8	outdoor
$T5$	Market-1501 [66]	32,668	1,501	6	outdoor

spectively. Once a teacher model was trained, source data and extra training were not needed. Basic information of the datasets is in Table 1. Note that, when forming the teacher model pool, teacher model of the target dataset was excluded. For Market-1501, the teacher model pool with $M = 4$ teachers was $\{T1, T2, T3, T4\}$ (MSMT17, CUHK03, ViPER, DukeMTMC); while for DukeMTMC, the teacher model pool with $M = 4$ teachers was $\{T1, T2, T3, T5\}$ (MSMT17, CUHK03, ViPER, Market-1501).

The standard splits of training and testing IDs of Market-1501 and DukeMTMC were adopted as in [66] and [70]. Our experiments were conducted under two settings:
(1) Unsupervised setting: all data in training set was unlabelled. (2) Semi-supervised setting: Concerning labelling cost, $C_v = 10$ identities in training set were randomly selected to be labelled, and the remaining data was unlabelled.

Performance Metrics. In testing, similarities between query and gallery samples were determined by the student model. The performance metrics cumulative matching characteristic (CMC) and mean Average Precision (mAP) were applied following the standard evaluation protocol in [66] and [70]. Note that, scalability is also important in our evaluations, including training time, model size indicated by parameter number (#Para) and testing computation cost of the model indicated by floating-point operations (FLOPs).

Implementation Details. For teacher models of source scenes, an advanced Re-ID model PCB [47] was adopted. For student model of the target scene, a lightweight model MobileNetV2 [46] was adopted and a convolution layer was applied to reduce the last feature map channel number to 256. It was initialized by ImageNet pretraining, without training on any Re-ID dataset. The input images were resized to 384 × 128 and feature maps of the last convolution layer were extracted as feature vectors. In each batch, for computing validation empirical risk, we sampled two images for each identity from labelled data to guarantee positive pairs. More details are provided in the supplementary.

5.1. Comparison under Unsupervised Setting

Compared Methods. For unsupervised setting, we did not use labelled data for our method and set fixed equal teacher weights α_i as $1/M$ in the Adaptive Aggregated Distillation Loss L_{TA} in Eq. (7) without using the Adaptive Knowledge Aggregator. We compared with unsupervised Re-ID methods including unsupervised features LOMO [31], BOW [66] and unsupervised learning models UMDL [42], PTGAN [53], PUL [14], CAMEL [62], SPGAN [13], TJ-AIDL [51] and HHL [71]. Among them, the advanced deep models require source data for transfer learning or pretraining, while our method only requires teacher models. Moreover, we compared with a multi-teacher knowledge distillation method Fukuda et al. [16]. For evaluating scalability, training time was tested on a TITAN X GPU. In practice, since the model for target domain is a new user-specified model without learning from Re-ID data, the time of pretraining on source data was included in training time. The results as well as the parameter number (#Para) and testing computation costs (FLOPs) are reported in Table 2.

Table 2. Performance under unsupervised setting. "Ours (unsupervised)" denotes the unsupervised version of our method. "Backbones" denotes model architecture. "#Para" denotes the number of parameters. "FLOPs" denotes floating-point operations (testing computation cost). "Train" denotes training time. "R-1" denotes rank-1 accuracy (%). "mAP" denotes mean average precision (%).

Methods	Backbones	#Para (M)	FLOPs (G)	Market-1501			DukeMTMC		
				R-1	mAP	Train	R-1	mAP	Train
LOMO [31]	-	-	-	27.2	8.0	-	12.3	4.8	-
BOW [66]	-	-	-	35.8	14.8	-	17.1	8.3	-
UMDL [42]	-	-	-	34.5	12.4	-	18.5	7.3	-
PTGAN [53]	GoogleNet	6.8	1.5	38.6	-	-	27.4	-	-
PUL [14]	ResNet-50	25.6	4.1	45.5	20.5	-	30.0	16.4	-
CAMEL [62]	ResNet-56	0.9	6.2	54.5	26.3	13.7h	42.2	21.0	13.7h
SPGAN [13]	ResNet-50	25.6	4.1	57.7	26.7	-	46.4	26.2	-
TJ-AIDL [51]	MobileNet	4.2	0.6	58.2	26.5	-	44.3	23.0	-
HHL [71]	ResNet-50	25.6	4.1	**62.2**	31.4	21.0h	46.9	27.2	21.0h
Fukuda [16]	MobileNetV2	3.4	0.3	45.1	23.0	1.1h	27.6	18.9	1.4h
Ours (unsupervised)	MobileNetV2	3.4	0.3	61.5	**33.5**	1.1h	**48.4**	**29.4**	1.4h

Results and Analysis. Our method outperformed the compared unsupervised Re-ID and multi-teacher distillation methods, except that the rank-1 accuracy on Market-1501 is slightly lower than HHL [71]. Although our method does not require source data, the Log-Euclidean Similarity Distillation Loss can effectively transfer knowledge of source domains to the student model. The complementarity of knowledge of multiple teacher models can increase the generalization ability of the student model. The distillation method Fukuda et al. [16] also exploited multiple teachers, but it is based on soft labels for closed-set classification, which is not effective for the open-set Re-ID problem.

Scalability of the methods is compared as below. Training data required by our method contains only target data and is smaller than the other methods that require source data for pretraining or joint training. The computation cost indicated by FLOPs of our backbone model MobileNetV2 is much lower than the others. With smaller training set and lighter backbone model, the training time is much shorter than the methods with comparable performance such as CAMEL [62] and HHL [71]. Thus, our framework is more scalable than the compared methods.

5.2. Comparison under Semi-supervised Setting

Compared Methods. For semi-supervised setting, 10 labelled IDs were available, thus the full version of our method ("Ours (semi)") can be applied. For comparison, we chose two competitive recent advanced unsupervised Re-ID methods CAMEL [62] and HHL [71] to extend to semi-

Table 3. Performance (%) under semi-supervised setting with 10 labelled identities. Semi-supervised and small sample learning methods for Re-ID were compared (same notations as Table 2).

Methods	Backbones	#Para (M)	FLOPs (G)	Market-1501			DukeMTMC		
				R-1	mAP	Train	R-1	mAP	Train
DNS [63]	ResNet-50	25.6	4.1	11.8	5.3	3.2h	10.0	4.6	3.2h
CAMEL (semi)	ResNet-56	0.9	6.2	54.4	26.2	13.7h	42.1	21.1	13.7h
HHL (semi)	ResNet-50	25.6	4.1	**63.9**	34.4	21.0h	46.7	26.5	21.0h
CAMEL (semi)	MobileNetV2	3.4	0.3	13.2	5.0	1.4h	13.6	5.6	1.4h
HHL (semi)	MobileNetV2	3.4	0.3	56.7	27.7	21.0h	42.9	23.9	21.0h
Ours (semi)	MobileNetV2	3.4	0.3	63.7	**35.4**	1.3h	**57.4**	**36.7**	1.7h

supervised version ("CAMEL (semi)" and "HHL (semi)") by using the positive and negative sample pairs obtained from these labelled samples. We also tested using MobileNetV2 trained on MSMT17 (source dataset of the best teacher model) for CAMEL (semi) and using MobileNetV2 as backbone for HHL (semi). A small sample learning Re-ID method DNS [63] was also compared, for which we used a PCB [47] model trained on MSMT17 as the best teacher model $T1$. Training time was tested as the unsupervised setting in Section 5.1. The results are reported in Table 3.

Results and Analysis. Among the compared methods, the performance of our method is the best, except that rank-1 accuracy is slightly lower than HHL (semi) on Market-1501. Compared with the unsupervised results in Table 2, CAMEL (semi) and HHL (semi) benefited little from the 10 extra labelled identities. The small sample learning method DNS failed due to overfitting. Compared to our unsupervised version in Table 2, our method benefited more from the labelled identities especially on DukeMTMC (explained in Section 5.3). Although 10 labelled identities can provide little information for directly learning a model, they are sufficient for our method to compute the validation empirical risk to select better teacher models. Moreover, our method can learn MobileNetV2 more effectively than HHL and CAMEL. Scalability analysis is similar to Section 5.1.

5.3. Further Evaluations

In this section, we further evaluate and analyze the components and capabilities of our method.

Evaluation of Knowledge Distillation. Knowledge distillation is the key technique in our Multi-teacher Adaptive Similarity Distillation Framework. To fairly compare our Log-Euclidean Similarity Distillation Loss L_T in Eq. (6) with existing knowledge distillation losses, we evaluated learning from a single teacher model $T1$ (MSMT17). To evaluate the effectiveness of the Log-Euclidean metric in L_T, we also tested using Euclidean metic, which is denoted by "L_T w/o log". We compared with a representative soft-label-based distillation method Hinton et al. [21] and a recent advanced probability-distribution-based distillation method PKT [40]. The results are reported in Table 4.

The performance of our loss L_T is the best and very close to the teacher model, the upper bound for distillation methods. The performance of Hinton et al. [21] is lower than other methods, because it is designed for closed-set classification by using soft labels, which is not suitable for the open-set Re-ID problem. PKT [40] used probability distribution for knowledge distillation, which is not so effective as using similarity in our method for Re-ID. Comparisons between "L_T" and "L_T w/o log" show the effectiveness of the Log-Euclidean metric, which considers the symmetric positive definite (SPD) property of similarity matrices.

Table 4. Performance (%) of distilling a single teacher model $T1$. Our loss L_T without logarithm "L_T w/o log" and other knowledge distillation methods were compared. Please see text for details.

Methods	Market-1501		DukeMTMC	
	R-1	mAP	R-1	mAP
Teacher $T1$ (MSMT17)	51.5	24.9	47.6	30.6
Hinton et al. [21]	41.2	20.0	32.5	20.8
PKT [40]	46.1	22.3	44.7	28.6
L_T w/o log	47.7	23.0	45.0	29.8
L_T (Eq. (6))	49.7	24.6	47.6	31.1

Effect of the Learned Teacher Weights α_i. Another key component in our method is the Adaptive Knowledge Aggregator for learning teacher weights α_i to adjust the distillation loss L_{TA} in Eq. (7). We conducted some experiments as follows: (1) Evaluating the performance of all teachers individually. (2) Ensembles of all teachers by distance fusion with weights learned by RankSVM [43] on the validation set and with equal weights. Joint training a PCB model [47] with all source data. (3) Using multi-task weighting methods Uncertainty [23] and GradNorm [10] to learn teacher weights in our framework. (4) Training our framework without learning teacher weights α_i ("Ours (unsupervised)") and training our framework with subsets of teachers of top-k α_i. The backbone of our method was MobileNetV2 and the others were ResNet-50. The results and the teacher weights α_i after training are reported in Table 5.

- Individual Teacher. For Market-1501, teachers $T1$ (MSMT17), $T2$ (CUHK03) and $T4$ (DukeMTMC) are effective and comparable. For DukeMTMC, only teacher $T1$ (MSMT17) is effective, since DukeMTMC is more challenging with more camera views than Market-1501. For both datasets, $T3$ (ViPER) is the worst because its training set is small. $T3$ provides weak knowledge as interference to test the robustness of the system. Our method outperformed the best teacher by about 10% and is more lightweight.

- w/ and w/o Learning α_i. Teacher weights α_i learned by Adaptive Knowledge Aggregator can indicate the effectiveness of the teachers. The weights for the worst teacher $T3$ are nearly zero. When comparing "Ours (semi)" with "Ours (unsupervised)", for Market-1501, the selection by teacher weights brings limited improvement; whilst for DukeMTMC, the improvement is much more significant, because distance fusion with equal weights is already better than individual teachers for Market-1501 but it is not effective for DukeMTMC. Ensemble by distance fusion increases testing computation costs and is not as scalable as our method.

Furthermore, we ranked α_i in descending order to select a subset of teachers for training. The teachers with large α_i

Table 5. Performance (%) of evaluating the Adaptive Knowledge Aggregator for selecting teachers. "α_i" is the teacher weight learned by the Adaptive Knowledge Aggregator. Ensembles of teachers, joint training and task weighting were compared.

Market-1501 (teacher model pool $\{T1, T2, T3, T4\}$)

Models	#Para	FLOPs	α_i	R-1	mAP	Train
Teacher $T1$ (MSMT17)	25.6	4.1	0.398	51.5	24.9	3.4h
Teacher $T2$ (CUHK03)	25.6	4.1	0.145	51.7	26.2	1.4h
Teacher $T3$ (ViPER)	25.6	4.1	0.000	28.5	12.0	0.1h
Teacher $T4$ (DukeMTMC)	25.6	4.1	0.458	49.4	24.2	1.7h
RankSVM weighted fusion (all teachers)	102.4	16.4	-	57.6	31.6	-
Distance fusion (all teachers)	102.4	16.4	-	56.5	30.7	-
Joint training (all source data)	25.6	4.1	-	62.8	35.6	6.6h
Uncertainty [23]	3.4	0.3	-	61.3	33.3	1.1h
GradNorm [10]	3.4	0.3	-	60.5	32.8	1.1h
Ours (unsupervised)	3.4	0.3	-	61.5	33.5	1.1h
Ours (Ti of top 1 α_i $\{T4\}$)	3.4	0.3	-	48.1	23.8	1.1h
Ours (Ti of top 2 α_i $\{T4, T1\}$)	3.4	0.3	-	63.0	34.6	1.3h
Ours (Ti of top 3 α_i $\{T4, T1, T2\}$)	3.4	0.3	-	63.5	35.5	1.3h
Ours (semi)	3.4	0.3	-	63.7	35.4	1.3h

DukeMTMC (teacher model pool $\{T1, T2, T3, T5\}$)

Models	#Para	FLOPs	α_i	R-1	mAP	Train
Teacher $T1$ (MSMT17)	25.6	4.1	0.581	47.6	30.6	3.4h
Teacher $T2$ (CUHK03)	25.6	4.1	0.071	25.3	14.8	1.4h
Teacher $T3$ (ViPER)	25.6	4.1	0.029	19.7	10.7	0.1h
Teacher $T5$ (Market-1501)	25.6	4.1	0.320	30.8	18.6	1.3h
RankSVM weighted fusion (all teachers)	102.4	16.4	-	41.8	28.3	-
Distance fusion (all teachers)	102.4	16.4	-	39.7	26.8	-
Joint training (all source data)	25.6	4.1	-	53.6	36.1	6.2h
Uncertainty [23]	3.4	0.3	-	51.0	30.6	1.4h
GradNorm [10]	3.4	0.3	-	44.8	26.5	1.4h
Ours (unsupervised)	3.4	0.3	-	48.4	29.4	1.4h
Ours (Ti of top 1 α_i $\{T1\}$)	3.4	0.3	-	47.6	31.1	1.4h
Ours (Ti of top 2 α_i $\{T1, T5\}$)	3.4	0.3	-	57.9	36.7	1.7h
Ours (Ti of top 3 α_i $\{T1, T5, T2\}$)	3.4	0.3	-	57.5	36.7	1.7h
Ours (semi)	3.4	0.3	-	57.4	36.7	1.7h

($> 1/4$) can bring significant improvement while those with small α_i ($< 1/4$) cannot bring improvement because they are weak or cannot provide complementary knowledge.

- Comparison with Ensemble and Task Weighting. Our method outperformed ensemble, joint training and task weighting [23, 10]. Moreover, our testing computation cost was much lower than ensemble and our training time was shorter than joint training. These show the advantages of our similarity knowledge distillation and Adaptive Knowledge Aggregator for aggregating knowledge of teachers.

The Number of Validation IDs. We tested using different numbers of validation identities from 0 to 50. As shown in Table 6, our method can achieve comparable performance with 5 to 50 identities. Since the validation identities are only for learning teacher weights α_i and not for training the student model parameters, there is no overfitting problem even with only 1 labelled identity. Comparing using 1 ID with using 0 ID, the performance dropped significantly especially on DukeMTMC, which indicates the importance of validation empirical risk. Thus, our Adaptive Knowledge Aggregator is robust with only a few labelled identities.

Different Student Model Architectures. To show the flexibility of our method, we used MobileNetV2 [46], ResNet-18 and ResNet-50 [20] as backbones for the student model, which are with different architectures and capacities. The results in Table 7 show that, our method achieved comparable performance for all three models. Thus, knowledge can be effectively distilled to models of different architectures.

Finetuning with Our Method as Initialization. When more labelled data is given, the MobileNetV2 [46] student model learned by our method can be used as initialization for finetuning. We finetuned it on 20% labelled identities and compared with the models finetuned with 20% and 100% labelled identities initialized by ImageNet pretraining. The results are shown in Table 8. The model initialized by our method finetuned on 20% IDs can achieve comparable performance with the model initialized by ImageNet pretraining finetuned on 100% IDs, while the performance of finetuning on 20% IDs with ImageNet pretraining was much lower. With prior knowledge of Re-ID, our method can generalize better with fewer target labelled samples.

Table 6. Performance (%) of using different numbers of labelled identities in validation set.

Validation set IDs		0	1	5	10	20	50
Market	R-1	61.5	62.6	63.2	63.7	63.9	64.2
	mAP	33.5	34.2	34.6	35.4	35.4	35.8
DukeMTMC	R-1	48.4	55.9	57.6	57.4	57.5	57.9
	mAP	29.4	35.0	36.6	36.7	36.8	37.1

Table 7. Performance (%) of different backbones of student model.

Backbones	Market-1501		DukeMTMC	
	R-1	mAP	R-1	mAP
ResNet-50	63.5	35.2	56.1	35.4
ResNet-18	63.1	34.9	56.7	36.4
MobileNetV2	63.7	35.4	57.4	36.7

Table 8. Performance (%) of finetuning MobileNetV2 [46] with our method as initialization on a small subset of 20% IDs.

Initialization	Labelled IDs	Market-1501		DukeMTMC	
		R-1	mAP	R-1	mAP
ImageNet	100%	78.3	55.5	64.7	45.4
ImageNet	20%	60.1	35.6	49.9	29.7
Ours	20%	77.1	57.7	66.7	46.0

6. Conclusion

In this paper, we aim to address the scalability of person re-identification from three aspects, including labelling cost, extension cost and testing computation cost. We propose a Multi-teacher Adaptive Similarity Distillation Framework, which can flexibly train a new user-specified lightweight model, with only a few labelled identities and without source data. The framework stores knowledge of Re-ID in a teacher model pool. When extending to a new scene, knowledge can be adaptively aggregated and distilled to a lightweight student model. For knowledge distillation for Re-ID, an open-set identification problem, we propose the Log-Euclidean Similarity Distillation Loss to imitate the sample pairwise similarity matrix of the teacher model. To effectively learn from multiple teachers, we propose the Adaptive Knowledge Aggregator to adjust the contribution of each teacher model by minimizing the validation empirical risk computed on a few labelled identities. Extensive evaluations show that our method is more scalable and can achieve performance comparable to state-of-the-art unsupervised and semi-supervised Re-ID methods.

Acknowledgement

This work was supported partially by the National Key Research and Development Program of China (2016YFB1001002), NSFC(61522115, 61573387, U1611461), Guangdong Province Science and Technology Innovation Leading Talents (2016TX03X157), 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] TM Feroz Ali and Subhasis Chaudhuri. Maximum margin metric learning over discriminative nullspace for person re-identification. In *CVPR*, 2018. 2

[3] Xiang Li Ao, Shuang and Charles X. Ling. Fast generalized distillation for semi-supervised domain adaptation. In *AAAI*, 2017. 2

[4] Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Fast and simple calculus on tensors in the log-euclidean framework. In *MICCAI*, 2005. 3

[5] Song Bai, Xiang Bai, and Qi Tian. Scalable person re-identification on supervised smoothed manifold. In *CVPR*, 2017. 2

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

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

[8] Jiaxin Chen, Yunhong Wang, Jie Qin, Li Liu, and Ling Shao. Fast person re-identification via cross-camera semantic binary transformation. In *CVPR*, 2017. 2

[9] 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. 2

[10] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *ICML*, 2018. 7, 8

[11] De Cheng, Yihong Gong, Zhihui Li, Dingwen Zhang, Weiwei Shi, and Xingjun Zhang. Cross-scenario transfer metric learning for person re-identification. *PR*, 2018. 1, 2

[12] Boris Chidlovskii, Stephane Clinchant, and Gabriela Csurka. Domain adaptation in the absence of source domain data. In *SIGKDD*, 2016. 2

[13] 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. 2, 6

[14] 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. 2, 6

[15] 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

[16] Takashi Fukuda, Masayuki Suzuki, Gakuto Kurata, Samuel Thomas, Jia Cui, and Bhuvana Ramabhadran. Efficient knowledge distillation from an ensemble of teachers. *Proc. Interspeech*, 2017. 2, 3, 6

[17] Chen Gong, Xiaojun Chang, Meng Fang, and Jian Yang. Teaching semi-supervised classifier via generalized distillation. In *IJCAI*, 2018. 2

[18] Douglas Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *PETS*, 2007. 5, 6

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

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2, 8

[21] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NeurIP workshop*, 2015. 2, 3, 7

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

[23] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, 2018. 7, 8

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

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

[26] Ilja Kuzborskij and Francesco Orabona. Stability and hypothesis transfer learning. In *ICML*, 2013. 2

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

[28] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 1, 2, 5, 6

[29] Xiang Li, Wei-Shi Zheng, Xiaojuan Wang, Tao Xiang, and Shaogang Gong. Multi-scale learning for low-resolution person re-identification. In *ICCV*, 2015. 2

[30] 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

[31] 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

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

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

[34] Niki Martinel, Abir Das, Christian Micheloni, and Amit K Roy-Chowdhury. Temporal model adaptation for person re-identification. In *ECCV*, 2016. 2

[35] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato. Hierarchical gaussian descriptor for person re-identification. In *CVPR*, 2016. 2

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

[37] Asit Mishra and Debbie Marr. Apprentice: Using knowledge distillation techniques to improve low-precision network accuracy. *arXiv preprint arXiv:1711.05852*, 2017. 2, 3

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

[39] Rameswar Panda, Amran Bhuiyan, Vittorio Murino, and Amit K Roy-Chowdhury. Unsupervised adaptive re-identification in open world dynamic camera networks. In *CVPR*, 2017. 2

[40] Nikolaos Passalis and Anastasios Tefas. Learning deep representations with probabilistic knowledge transfer. In *ECCV*, 2018. 2, 7

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

[42] 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

[43] Bryan Prosser, Wei-Shi Zheng, Shaogang Gong, Tao Xiang, and Q Mary. Person re-identification by support vector ranking. In *BMVC*, 2010. 2, 7

[44] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 2

[45] Sourya Roy, Sujoy Paul, Neal E Young, and Amit K Roy-Chowdhury. Exploiting transitivity for learning person re-identification models on a budget. In *CVPR*, 2018. 2

[46] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018. 6, 8

[47] 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, 3, 6, 7

[48] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIP*, 2017. 2

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

[50] Hanxiao Wang, Shaogang Gong, Xiatian Zhu, and Tao Xiang. Human-in-the-loop person re-identification. In *ECCV*, 2016. 2

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

[52] Xiaojuan Wang, Wei-Shi Zheng, Xiang Li, and Jianguo Zhang. Cross-scenario transfer person reidentification. *IEEE TCSVT*, 2016. 1, 2

[53] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 2018. 2, 5, 6

[54] Kilian Q Weinberger, John Blitzer, and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. In *NeurIP*, 2005. 2, 3

[55] 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

[56] Weiji Wu, Ancong Wu, and Wei-Shi Zheng. Light person re-identification by multi-cue tiny net. In *ICIP*, 2018. 1, 2

[57] 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

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

[59] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *CVPR*, 2017. 2

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

[61] Shan You, Chang Xu, Chao Xu, and Dacheng Tao. Learning from multiple teacher networks. In *SIGKDD*, 2017. 2

[62] 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

[63] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a discriminative null space for person re-identification. In *CVPR*, 2016. 2, 7

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

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

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

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

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

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

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

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

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

[73] Xiatian Zhu, Botong Wu, Dongcheng Huang, and Wei-Shi Zheng. Fast open-world person re-identification. *IEEE TIP*, 2017. 1, 2