IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 25, 2019, accepted October 11, 2019, date of publication October 29, 2019, date of current version November 7, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2950122

Distilled Camera-Aware Self Training for Semi-Supervised Person Re-Identification

ANCONG WU[1], WEI-SHI ZHENG[2], AND JIAN-HUANG LAI[2,3]
[1]School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510006, China
[2]School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China
[3]Guangdong Province Key Laboratory of Information Security, Guangzhou, China

Corresponding author: Wei-Shi Zheng (wszheng@ieee.org)

This work was supported in part by the NSFC under Grant U1811461, in part by the Guangdong Province Science and Technology Innovation Leading Talents under Grant 2016TX03X157, and in part by the Guangzhou Research Project under Grant 201902010037.

ABSTRACT Person re-identification (Re-ID), which is for matching pedestrians across disjoint camera views in surveillance, has made great progress in supervised learning. However, requirement of a large number of labelled identities leads to high cost for large-scale Re-ID systems. Consequently, it is significant to study learning Re-ID with unlabelled data and limited labelled data, that is, semi-supervised person re-identification. When labelled data is limited, the learned model tends to overfit the data and cannot generalize well. Moreover, the scene variations between cameras lead to domain shift in the feature space, which makes mining auxiliary supervision information from unlabelled data more difficult. To address these problems, we propose a Distilled Camera-Aware Self Training framework for semi-supervised person re-identification. To alleviate the overfitting problem for learning from limited labelled data, we propose a Multi-Teacher Selective Similarity Distillation Loss to selectively aggregate the knowledge of multiple weak teacher models trained with different subsets and distill a stronger student model. Then, we exploit the unlabelled data by learning pseudo labels by clustering based on the student model for self training. To alleviate the effect of scene variations between cameras, we propose a Camera-Aware Hierarchical Clustering (CAHC) algorithm to perform intra-camera clustering and cross-camera clustering hierarchically. Experiments show that our method outperformed the state-of-the-art semi-supervised person re-identification methods.

INDEX TERMS Person re-identification, semi-supervised learning, knowledge distillation, clustering.

I. INTRODUCTION

Person re-identification (Re-ID) has received much attention in recent years due to its significance in video surveillance applications. When abundant labelled data is given, many works [1]–[7] have made great progress in supervised learning. However, labelling cost should be considered in large-scale Re-ID system that consists of many cameras. To reduce labelling cost, studying semi-supervised learning to exploit unlabelled data and limited labelled data is a practical solution. Unsupervised person re-identification [8]–[15] has been studied to learn representation from unlabelled data, but how to effectively learn from limited labelled data is not considered in these methods. So far, semi-supervised person re-identification [16]–[20] is still under-explored.

The associate editor coordinating the review of this manuscript and approving it for publication was Hugo Proenca.

For semi-supervised Re-ID, exploiting unlabelled data and limited labelled data brings about some challenges. First, insufficient training data leads to overfitting for model learning and thus degrades generalization performance. Second, scene variations between cameras, such as illumination, background and viewpoint, cause domain shift in the feature space and create difficulty for mining auxiliary supervision information in unlabelled data to assist model training. The effect of scene variations is discussed in Section III-B later.

To address the challenges for semi-supervised Re-ID, we propose a Distilled Camera-Aware Self Training framework, as shown in Figure 1.

On the one hand, when training model with limited labelled data, motivated by ensemble learning and knowledge distillation, we aggregate the knowledge of multiple weak teacher models trained with different subsets and distill a stronger student model, in order to improve generalization performance without increasing model size. We propose a Multi-Teacher

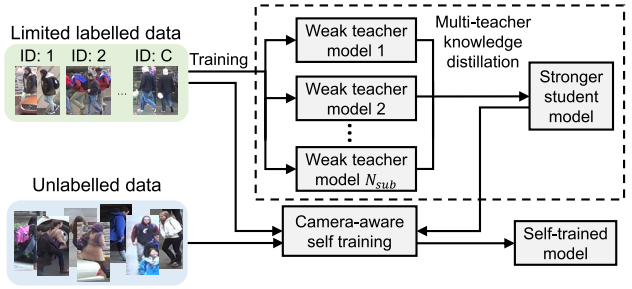


FIGURE 1. Overview of the Distilled Camera-Aware Self Training framework. First, it learns a stronger student model with limited labelled data by knowledge distillation from multiple weak teacher models. Then, it alleviates the effect of the scene variations between cameras by camera-aware self training to exploit the unlabelled data for training the distilled student model. Self-trained model is the output model for evaluation.

Selective Similarity Distillation Loss for selectively aggregate knowledge from multiple teacher models. The proposed loss benefits from the sparsity property of $\ell 2,1$-norm to selectively suppress noises in samples and teachers.

On the other hand, to alleviate the effect of scene variations between cameras for mining auxiliary supervision information from unlabelled data, we propose a Camera-Aware Hierarchical Clustering (CAHC) algorithm to learn pseudo labels for unlabelled data. As pedestrian data for Re-ID is intrinsically captured from different cameras and there exists domain shift between cameras, the scale of intra-camera similarity is generally larger than cross-camera similarity. Clustering based on these two types of similarities simultaneously results in confusion. Therefore, we separate the clustering process by performing intra-camera clustering and cross-camera clustering hierarchically, so that the influence of inconsistent similarity scales can be alleviated.

Our contributions are summarized as follows. First, we propose a Multi-Teacher Selective Similarity Distillation Loss for selectively aggregate and distill knowledge from multiple teachers to improve the generalization performance of our model trained with limited labelled data. Second, we propose a Camera-Aware Hierarchical Clustering (CAHC) algorithm to alleviate the effect of scene variations between cameras for learning pseudo labels for unlabelled data for self training. The above two processes cooperate to learn feature representation for semi-supervised person re-identification in a Distilled Camera-Aware Self Training framework.

II. RELATED WORK

A. SUPERVISED PERSON RE-IDENTIFICATION

In recent years, supervised person re-identification has undergone a fast development, from feature design [3], [21]–[23] to distance metric learning [1]–[4], [21], [24]–[29] and end-to-end deep learning [5]–[7], [30]–[34]. Among them, the most competitive methods are deep-learning-based models. Although high performance can be achieved, these methods rely on a large amount of labelled data for learning discriminative features and heavy labelling cost hinders the scalability of these methods.

B. UNSUPERVISED AND SEMI-SUPERVISED PERSON RE-IDENTIFICATION

Recently, reducing labelling cost for person re-identification has drawn more attention for developing scalable Re-ID system, since it is not feasible to label a large number of identities for each new scene. Unsupervised learning [8]–[15], [35]–[41] has been studied 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, and then learn from unlabelled data by pseudo label learning. In [10], [12], [41], the model was pretrained by source data and learned from unlabelled target data by clustering and fine-tuning. In [13]–[15], [35], [36], the knowledge was transferred from source domain to target domain by image-to-image transformation from source images to target images. Wang *et al.* [11] proposed transferring knowledge from attribute labels. Li et al. [37] learned from unlabelled data by associating tracklets in videos across cameras. Yu *et al.* [38] proposed to learn soft labels of source labelled data for target unlabelled data. Yang *et al.* [39] exploited image patches instead of whole images for more generalized unsupervised learning. Zhong *et al.* [40] proposed learning invariance from different aspects for unsupervised learning.

Compared with unsupervised learning, semi-supervised learning [12], [16]–[20], [42] is relatively under-explored for person re-identification. PUL [12] and MVC [16] were based on clustering for learning pseudo labels, which were similar to clustering-based unsupervised Re-ID methods. Liu *et al.* [19] proposed to learn representation by dictionary learning. Liu *et al.* [18] learned robust representation based on attribute learning. Li *et al.* [17] learned distance metric by exploring the neighbours. Ding *et al.* [20] aimed to improve the performance of learning from labelled data by extra unlabelled data generated by GAN.

The above methods mainly focused on mining the auxiliary supervision information in the unlabelled data, while the overfitting problem that degrades generalization performance caused by limited labelled data was seldom explored. Our method focuses on effectively exploiting the limited labelled data to improve generalization performance by dividing the labelled training set into several subsets and aggregating the shared and subset-specific knowledge of each subset in a single model by knowledge distillation. Moreover, in our Distilled Camera-Aware Self Training framework, we further consider the scene variations between cameras in the clustering algorithm, which is a significant problem in Re-ID but ignored in previous semi-supervised Re-ID methods.

C. ENSEMBLE LEARNING AND KNOWLEDGE DISTILLATION

As limited labelled training data easily incurs overfitting problem, ensemble learning has been a successful

solution [43], which reduces the variance in classifier decision and thereby improve the generalization performance. Generally, ensemble learning follows three steps, including data sampling, classifier training and classifier combination, in order to combine multiple weak classifiers to obtain a stronger classifier. Bootstrap aggregating (bagging) [44] and boosting [45] are representative ensemble learning algorithms.

Although ensemble learning is effective, it requires multiple models and increases model size and computation costs. Hinton *et al.* [46] proposed knowledge distillation for transferring knowledge from a large teacher model to a smaller student model by imitating the outputs. Likewise, knowledge distillation can also transfer knowledge from an ensemble system to a single model. Many existing knowledge distillation methods such as [46]–[49] are limited to closed-set classification, since they are based on soft labels of the teacher model to guide student model learning.

However, person re-identification is an open-set identification problem, in which the identities in training and testing are nonoverlapping, and thus the soft-label-based distillation methods are not suitable. Some methods also considered using information other than soft labels for imitation to convey knowledge. Fitnets [50] and FSP [51] exploited feature maps; PKT [52] exploited the probability distribution of data; RKD [53] exploited the relation information of data. The outputs that they used for imitation were not directly related to similarity measurement for Re-ID. To transfer knowledge more effectively, we distill knowledge embedded in similarities by imitating pairwise sample similarities.

As for the related methods of multi-teacher knowledge distillation, in [48], [49], their proposed methods exploited the ensemble of multiple teachers, but they were limited to closed-set classification and the knowledge of multiple teachers cannot be selectively aggregated as in our method.

III. APPROACH

To study semi-supervised person re-identification, we first formulate this problem as follows. From N_{cam} cameras, we obtain a set of labelled images $\{(\mathbf{I}_i, y_i^{cam}, y_i)\}_{i=1}^{N_L}$ and a set of unlabelled images $\{(\mathbf{I}_j, y_j^{cam})\}_{j=1}^{N_U}$, where $\mathbf{I}_i$, $\mathbf{I}_j$ are images, $y_i^{cam}, y_j^{cam} \in \{1, 2, \ldots, N_{cam}\}$ are the corresponding camera labels and $y_i \in \{1, 2, \ldots, C\}$ is the corresponding identity label of $\mathbf{I}_i$. N_U is the number of unlabelled samples. N_L is the number of labelled samples and C is the total number of labelled identities, which is limited for saving labelling cost. The identities of the labelled set and unlabelled set are nonoverlapping. Our objective is to learn a model F for computing similarities between samples for retrieval.

A. MULTI-TEACHER KNOWLEDGE DISTILLATION FOR LIMITED LABELLED DATA

To learn discriminative feature representation, we first train a model with the labelled data. As the number of labelled samples is limited, insufficient training data incurs overfitting for model learning. To alleviate the overfitting problem, ensemble learning [43] is commonly used, which can reduce the variance in classifier decision and thereby improve generalization performance by combining multiple models. However, exploiting multiple models significantly increases total model size and computation costs, which make ensemble learning not practical in a large-scale re-ID system.

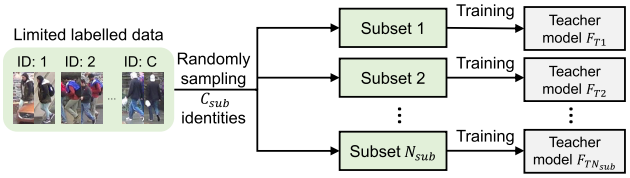


FIGURE 2. Training multiple teacher models by sampling multiple subsets from limited labelled data. Each subset consists of samples of randomly sampled C_{sub} identities. The teacher model F_{Tm} is trained by subset m.

Motivated by ensemble learning that aggregates the predictions of multiple models, we solve this problem by distilling the knowledge of multiple teacher models in a single student model, so that we can benefit from ensemble of models without increasing model size and computation costs. The process consists of two steps: (1) training multiple teacher models and (2) multi-teacher selective similarity distillation.

1) TRAINING MULTIPLE TEACHER MODELS

To train multiple teacher models that are complementary to each other for aggregation, in a similar way as bootstrap aggregating (bagging) [44], we train each model using a randomly drawn subset of the labelled training set. Each subset is formed by images of randomly selected C_{sub} identities from C identities. Random sampling can increase the diversity of data in different subsets. As shown in Figure 2, we draw N_{sub} subsets for training N_{sub} teacher models $\{F_{Tm}\}_{m=1}^{N_{sub}}$, respectively.

For person re-identification, the models can be trained by using Softmax cross entropy loss for classifying the identities. When extracting features for matching, we use the outputs before the last fully connected layer for classification.

2) MULTI-TEACHER SIMILARITY DISTILLATION

After training multiple teacher models, to combine the learned teacher models without increasing model size, we aggregate the knowledge of the teacher models $\{F_{Tm}\}_{m=1}^{N_{sub}}$ in a single student model F_S by knowledge distillation. The overview is shown in Figure 3.

We first consider the case of using one teacher model. Given a teacher model, knowledge distillation techniques transfer its knowledge to a student model by making the student model imitate the output of the teacher model. For general object classification problems, soft class label is commonly used for knowledge distillation [46]. However, different from object classification problem in a closed-set setting, the identities of training data and testing data are nonoverlapping for person re-identification, so that using soft

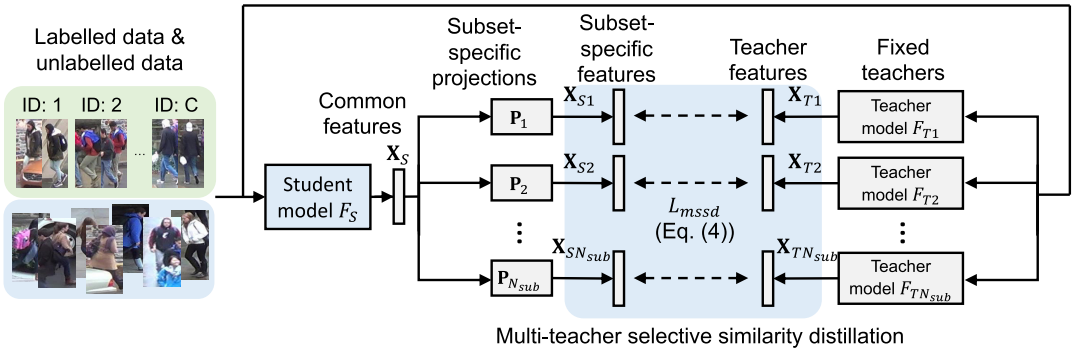


FIGURE 3. Overview of multi-teacher selective similarity distillation. The student model is expected to extract common features $\mathbf{X}_S$ that contain both shared and subset-specific knowledge. Subset-specific features are obtained by subset-specific projections. The knowledge of multiple teachers is aggregated in a student model by similarity distillation in different subset-specific subspaces by means of multi-task learning.

class label for knowledge distillation is not suitable. Instead of soft class label, the knowledge of matching for Re-ID is embedded in the similarities between samples. Motivated by this, we make the student model imitate the pairwise similarities of the teacher model for knowledge distillation for semi-supervised person re-identification. When the pairwise similarities of all samples are equal for the teacher model and the student model, the student model can obtain the same retrieval results as those of the teacher model, and thus the knowledge of the teacher model can be effectively transferred by means of imitating pairwise similarities.

To compute pairwise similarities, we apply all data $\{\mathbf{I}_k\}_{k=1}^{N} = \{\mathbf{I}_i\}_{i=1}^{N_L} \cup \{\mathbf{I}_j\}_{j=1}^{N_U}$ including both labelled and unlabelled data. The teacher model $F_{Tm}(\cdot\,;\Theta_{Tm})$ parameterized by Θ_{Tm} is learned from subset m and fixed. Given an image $\mathbf{I}_k$, we can extract its feature $\mathbf{x}_{Tm,k} = F_{Tm}(\mathbf{I}_k;\Theta_{Tm}) \in \mathbb{R}^{d_{Tm}}$. In our case, we normalize the feature vector by $\ell 2$-norm, so that the inner product of two feature vectors is cosine similarity. Let $\mathbf{X}_{Tm} = [\mathbf{x}_{Tm,1}, \mathbf{x}_{Tm,2}, \ldots, \mathbf{x}_{Tm,N}] \in \mathbb{R}^{d_{Tm}\times N}$ denote the feature matrix of all samples $\{\mathbf{I}_k\}_{k=1}^{N}$ extracted by teacher model F_{Tm}, which is constant for guiding the student model. The similarity matrix $\mathbf{A}_{Tm}$ of teacher model F_{Tm} is computed by

$$\mathbf{A}_{Tm} = \mathbf{X}_{Tm}^{\top}\mathbf{X}_{Tm}, \tag{1}$$

where the element in the p-th row and the q-th column of $\mathbf{A}_{Tm}$ is the similarity between samples $\mathbf{I}_p$ and $\mathbf{I}_q$.

To transfer knowledge of multiple teacher models to a student model $F_S(\cdot\,;\Theta_S)$, we aggregate the knowledge embedded in the similarity matrices $\{\mathbf{A}_{Tm}\}_{m=1}^{N_{sub}}$. Since the teacher models are trained by different subsets with both overlapping and nonoverlapping identities, both shared knowledge and subset-specific knowledge are learned in the teacher models. To explicitly and effectively model the subset-specific knowledge, we propose to distill knowledge of each subset in a subset-specific subspace.

Given samples $\{\mathbf{I}_k\}_{k=1}^{N}$, we expect that the student model F_S can extract common features $\mathbf{X}_S = [\mathbf{x}_{S,1}, \mathbf{x}_{S,2}, \ldots, \mathbf{x}_{S,N}] \in \mathbb{R}^{d_S\times N}$, which contains both shared knowledge and subset-specific knowledge. To achieve this, we aggregate and distill the knowledge of multiple teachers by means of multi-task learning and regard knowledge distillation for each teacher model as a task. For the distillation task of teacher model F_{Tm}, we introduce a subset-specific projection $\mathbf{P}_{Sm} \in \mathbb{R}^{d_S\times d_{Sm}}$ to map the features $\mathbf{X}_S$ to a subset-specific subspace by

$$\mathbf{X}_{Sm} = \mathbf{P}_{Sm}^{\top}\mathbf{X}_S, \tag{2}$$

where $\mathbf{X}_{Sm}$ is the subset-specific feature matrix for the distillation task of teacher model F_{Tm}.

To distill knowledge by imitating the pairwise similarities between samples, we minimize the distance between subset-specific similarity matrices of the student model and the similarity matrices of the teacher models by

$$\min_{\Theta_S,\{\mathbf{P}_{Sm}\}} L_{msd} = \sum_{m=1}^{N_{sub}} \|\mathbf{A}_{Sm} - \mathbf{A}_{Tm}\|_F^2, \tag{3}$$

where $\mathbf{A}_{Sm} = \mathbf{X}_{Sm}^{\top}\mathbf{X}_{Sm}$ is the subset-specific similarity matrix of the student model for the m-th distillation task. We call L_{msd} the *Multi-Teacher Similarity Distillation Loss*.

3) MULTI-TEACHER SELECTIVE SIMILARITY DISTILLATION

In the above Multi-Teacher Similarity Distillation Loss L_{msd} in Eq. (3), the errors of pairwise similarities between teacher models and the student model are measured by Frobenius norm, which is sensitive to outliers as illustrated in [54]. In practice, there exist noises in the similarities, since not all teacher models can provide correct and complementary knowledge for aggregation and some samples may be outliers. Therefore, the pairwise similarities should be selectively learned for the student model. We assume that the noises of similarities are in the minority and thereby the errors of similarities of these noises should be sparse. For more robust similarity distillation under the influence of noises, we take sparsity into account in the loss function and apply the $\ell 2$,1-norm for minimizing the errors of similarity matrices,

in order to benefit from the row sparsity property of $\ell 2$,1-norm for sample selection.

To enforce row sparsity for similarity matrices of all samples computed by all teacher models, we concatenate all subset-specific similarity matrices of the student model to obtain the joint student similarity matrix $\mathbf{A}_S = [\mathbf{A}_{S1}, \mathbf{A}_{S2}, \ldots, \mathbf{A}_{SN_{sub}}]^\top$ and concatenate the similarity matrices of all teacher models to obtain the joint teacher similarity matrix $\mathbf{A}_T = [\mathbf{A}_{T1}, \mathbf{A}_{T2}, \ldots, \mathbf{A}_{TN_{sub}}]^\top$. We minimize the distance between the joint similarity matrices by

$$\min_{\Theta_S, \{\mathbf{P}_{Sm}\}} L_{mssd} = \|\mathbf{A}_S - \mathbf{A}_T\|_{2,1}\,, \tag{4}$$

where the $\ell 2$,1-norm $\|\mathbf{M}\|_{2,1}$ is computed by $\|\mathbf{M}\|_{2,1} = \sum_{r=1}^{N_{row}} \|\mathbf{m}_r\|_2$. ($\mathbf{m}_r$ is the r-th row of $\mathbf{M}$ and N_{row} is the number of rows of $\mathbf{M}$.)

The r-th row of the similarity matrices $\mathbf{A}_S$ or $\mathbf{A}_T$ is a similarity vector that contains similarities between the r-th sample and all samples. The similarity vectors computed by different teacher models are in different rows. Therefore, the row sparsity property of $\ell 2$,1-norm can force the errors of similarity vectors of different samples computed by different teacher models to be sparse and selectively suppress the noises. We call L_{mssd} the *Multi-Teacher Selective Similarity Distillation Loss*.

In summary, the process of multi-teacher selective similarity distillation is shown in Figure 3. The teacher models trained with different subsets learn both shared and subset-specific knowledge for Re-ID. Joint learning of similarity distillation for multiple teacher models in different subset-specific subspaces can aggregate both shared and subset-specific knowledge in the student model in the way of multi-task learning.

B. CAMERA-AWARE HIERARCHICAL CLUSTERING FOR UNLABELLED DATA

For semi-supervised learning, besides learning from labelled data, mining auxiliary supervision information in the unlabelled data is another key problem. Pseudo label learning [55], [56] based on similarity measurement for the unlabelled data is commonly used for semi-supervised learning. For predicting pseudo labels for person re-identification, clustering algorithms [10], [12] are often used.

However, directly applying clustering algorithm for unlabelled data ignores the scene variations between cameras, which is a significant factor that degrades matching performance of person re-identification. Since the person images are captured in different cameras, different backgrounds and lighting conditions in different scenes cause domain shift in the feature space. As a result, the distributions of intra-camera similarity and cross-camera similarity are inconsistent and the scale of intra-camera similarity is generally larger than that of cross-camera similarity. We visualize the similarity distributions of intra-camera matching and cross-camera matching of two randomly selected cameras in DukeMTMC [33] in Figure 4. The similarities are computed

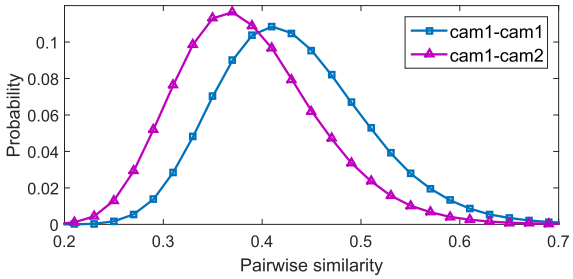


FIGURE 4. Comparison of intra-camera similarity distribution and cross-camera similarity distribution. "cam1" and "cam2" denote two randomly selected cameras in DukeMTMC [33] dataset. We compute pairwise similarities of samples of these two cameras by a ResNet-50 model [57] trained by limited training data (1/3 identities in the training set) and show the distributions of pairwise similarities of camera pairs "cam1-cam1" (intra-camera matching) and "cam1-cam2" (cross-camera matching). The scale of intra-camera similarity is generally larger than the scale of cross-camera similarity.

using a ResNet-50 model [57] trained by limited training data (1/3 identities of the training set). The similarity distributions of intra-camera matching and cross-camera matching are inconsistent as shown in in Figure 4. If the inconsistency of similarity distributions is ignored in clustering, the similarities with domain shift give rise to confusion when simultaneously associating intra-camera sample pairs and cross-camera sample pairs.

To address this problem, we divide the clustering process into two hierarchical steps: (1) intra-camera clustering and (2) cross-camera clustering, in order to handle intra-camera similarity and cross-camera similarity separately. In this way, clustering based on similarities of different scales can be avoided in each clustering step and thus the influence of domain shift can be alleviated.

We first introduce some notations for clustering. Let $\{(\mathbf{x}_j, y_j^{cam})\}_{j=1}^{N_U}$ denote the extracted features and camera labels of the unlabelled data $\{\mathbf{I}_j\}_{j=1}^{N_U}$. To perform clustering, we compute the pairwise distance matrix $\mathbf{D}$ for features $\{\mathbf{x}_j\}_{j=1}^{N_U}$. The element $d_{p,q}$ in the p-th row and the q-th column of $\mathbf{D}$ is the distance between features $\mathbf{x}_p$ and $\mathbf{x}_q$. Given the distance matrix $\mathbf{D}$, a clustering algorithm CLUSTER predicts pseudo labels $\mathbf{y}^p = \text{CLUSTER}(\mathbf{D})$ for $\{\mathbf{x}_j\}_{j=1}^{N_U}$.

1) BASIC CLUSTERING ALGORITHM

As the amount of unlabelled training data for person re-identification is uncertain, it is difficult to determine the number of clusters. Moreover, some samples are noises or have no corresponding positive sample of the same identity. Therefore, we apply a density-based clustering algorithm DBSCAN [58], which can determine the number of clusters and remove noise by searching for sample sets with high density to form clusters, as indicated in [41].

We briefly revisit the DBSCAN [58] algorithm. There are two parameters for DBSCAN: distance threshold ϵ and the minimum number of points N_{min} required to form a cluster. We start with an arbitrary point which is not visited. Then the

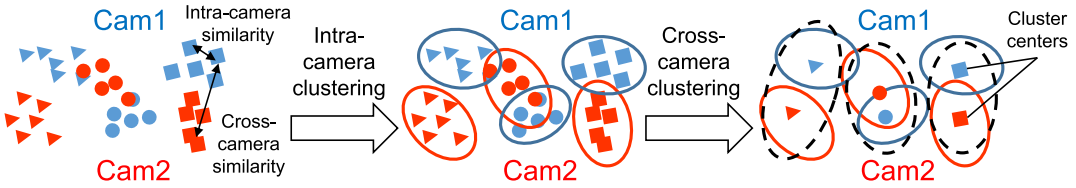


FIGURE 5. Overview of camera-aware hierarchical clustering (CAHC). Due to scene variations between cameras, the scales of intra-camera similarity and cross-camera similarity are different and lead to confusion in clustering. To address this problem, we separate the clustering process into two hierarchical steps: intra-camera clustering and cross-camera clustering. First, intra-camera sample pairs are clustered to predict intra-camera pseudo labels and the mean feature of each cluster is computed as cluster center. Second, based on the first step, cross-camera cluster centers are further associated to obtain the pseudo labels for each cluster center. Finally, the pseudo labels of cluster centers are assigned to each sample according to the intra-camera pseudo labels. Note that, in this figure, different colors denote different cameras and different shapes denote different identities. (Best viewed in color.).

samples of ϵ-neighbourhood are retrieved. If the number of retrieved samples is not fewer than N_{min}, a cluster is started; otherwise the point is regarded as noise. If a sample is a part of a cluster, its ϵ-neighbourhood also joins the cluster. After a cluster is completely found, we process new unvisited points following the above steps repeatedly.

Based on the basic clustering algorithm, we perform intra-camera clustering and cross-camera clustering hierarchically as follows. The overview of our Camera-Aware Hierarchical Clustering (CAHC) algorithm is shown in Figure 5.

2) STEP 1: INTRA-CAMERA CLUSTERING

Since associating intra-camera samples does not suffer from scene variations between cameras and is relatively less difficult than associating cross-camera samples, we perform intra-camera clustering in the first step. Intra-camera clustering only considers intra-camera sample pairs. We compute the pairwise distance matrix $\mathbf{D}$ and convert the distance $d_{p,q}$ in $\mathbf{D}$ to $d_{p,q}^{intra}$ to form an intra-camera distance matrix $\mathbf{D}^{intra}$ by

$$d_{p,q}^{intra} = \begin{cases} d_{p,q} & y_p^{cam} = y_q^{cam}, \\ d_{max} & y_p^{cam} \neq y_q^{cam}, \end{cases} \tag{5}$$

where $d_{p,q}^{intra}$ is the distance in the p-th row and the q-th column of intra-camera distance matrix $\mathbf{D}^{intra}$. d_{max} is the maximum value of distance. In our case, cosine distance is used and the maximum value is $d_{max} = 2$.

By replacing the cross-camera distances with the maximum distance, the cross-camera sample pairs are ignored and only intra-camera sample pairs can be assigned to the same cluster. Based on the intra-camera distance matrix $\mathbf{D}^{intra}$, we perform intra-camera clustering to obtain pseudo labels for all samples by

$$\mathbf{y}^{intra} = \text{DBSCAN}(\mathbf{D}^{intra}). \tag{6}$$

With the intra-camera pseudo labels $\mathbf{y}^{intra}$, we compute the mean feature of each cluster as center to represent a cluster of similar samples which are probably from the same identity. Let $\{(\mathbf{c}_l, y_l^{cam(c)})\}_{l=1}^{N_{cluster}}$ denote the set of cluster centers $\mathbf{c}_l$ and the corresponding camera labels $y_l^{cam(c)}$. $N_{cluster}$ is the number of clusters. Based on the cluster centers of intra-camera clustering, we further associate cross-camera cluster center pairs in the next step.

3) STEP 2: CROSS-CAMERA CLUSTERING

To associate cross-camera cluster center pairs, we first compute the center distance matrix $\mathbf{D}^{(c)}$ of cluster centers, in which $d_{p,q}^{(c)}$ in the p-th row and the q-th column is the distance between cluster centers $\mathbf{c}_p$ and $\mathbf{c}_q$. As intra-camera sample pairs have been associated in intra-camera clustering, we only associate cross-camera cluster center pairs in the second step. To ignore the intra-camera cluster pairs, we convert the distance $d_{p,q}^{(c)}$ in $\mathbf{D}^{(c)}$ to $d_{p,q}^{cross}$ to form a cross-camera center distance matrix $\mathbf{D}^{cross}$ by

$$d_{p,q}^{cross} = \begin{cases} d_{p,q}^{(c)} & y_p^{cam(c)} \neq y_q^{cam(c)}, \\ d_{max} & y_p^{cam(c)} = y_q^{cam(c)}, \end{cases} \tag{7}$$

where $d_{p,q}^{cross}$ is the distance in the p-th row and the q-th column of cross-camera distance matrix $\mathbf{D}^{cross}$. As in Eq. (5), $d_{max} = 2$ is the maximum value of cosine distance.

By replacing the intra-camera distances with the maximum distance, only cross-camera cluster centers can be associated. Based on the cross-camera distance matrix $\mathbf{D}^{cross}$, we perform cross-camera clustering to obtain pseudo labels for the cluster centers by

$$\mathbf{y}^{cross} = \text{DBSCAN}(\mathbf{D}^{cross}). \tag{8}$$

With the cross-camera cluster center pseudo labels $\mathbf{y}^{cross}$ and the intra-camera pseudo labels $\mathbf{y}^{intra}$, we assign $\mathbf{y}^{cross}$ to each cluster and then further assign the labels to each sample in that cluster as indicated by $\mathbf{y}^{intra}$. Then, we can obtain the pseudo labels of all samples $\mathbf{y}^{hier}$ after the two-step hierarchical clustering process. Note that, DBSCAN may annotate some samples as noises and we discard the noises in the subsequent processes.

Since intra-camera sample pairs and cross-camera sample pairs are clustered in a hierarchical way, the intra-camera similarity and cross-camera similarity with different scales are handled separately in two steps to avoid confusion in clustering caused by scene variations between cameras. We call this clustering algorithm the *Camera-Aware Hierarchical Clustering* (CAHC) algorithm.

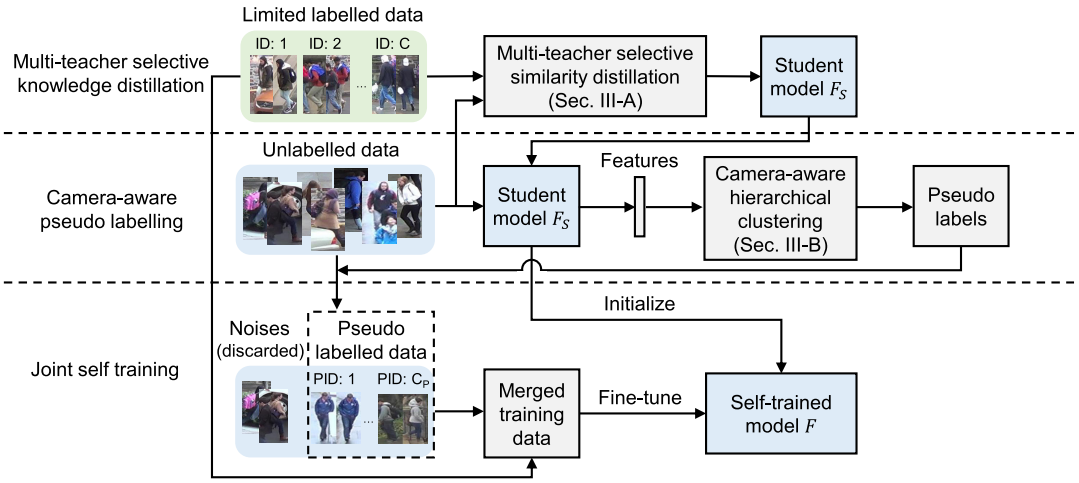


FIGURE 6. Overview of the distilled camera-aware self training framework. The framework consists of three steps: (1) multi-teacher selective knowledge distillation, (2) camera-aware pseudo labelling and (3) joint self training. First, multi-teacher selective knowledge distillation aggregates knowledge of multiple teacher models in a student model F_S by multi-teacher selective similarity distillation (Section III-A) to improve generalization performance for learning from limited labelled data. Then, camera-aware pseudo labelling assigns pseudo labels to unlabelled data by Camera-Aware Hierarchical Clustering (CAHC) (Section III-B) based on the features extracted by the student model F_S. Note that, DBSCAN [58] in our CAHC may annotate some samples as noises and we discard these samples. Finally, for joint self training, the pseudo labelled data and labelled data are merged for fine-tuning the student model F_S to obtain the self-trained model F.

C. DISTILLED CAMERA-AWARE SELF TRAINING

To jointly exploit unlabelled data and limited labelled data for semi-supervised learning, we develop a Distilled Camera-Aware Self Training framework, as shown in Figure 6. The framework consists of three steps: (1) multi-teacher selective knowledge distillation, (2) camera-aware pseudo labelling and (3) joint self training.

First, multi-teacher selective knowledge distillation aggregates knowledge of multiple teacher models trained with different subsets of training data and distills a student model F_S to improve generalization performance for learning from limited labelled data without increasing model size by multi-teacher selective similarity distillation (Section III-A). Then, based on the common features $\mathbf{X}_S$ extracted by the student model F_S, camera-aware pseudo labelling learns pseudo labels for unlabelled data by Camera-Aware Hierarchical Clustering (CAHC) (Section III-B), which can avoid the influence of domain shift caused by scene variations between cameras in clustering. Note that, DBSCAN [58] in our CAHC may annotate some samples as noises and we discard these samples. Finally, for joint self training, the pseudo labelled data and the labelled data are merged to form an extended training set. As there is no overlapping identity in labelled data and unlabelled data, they can be merged directly. After removing the subset-specific projections in the student model F_S, we fine-tune the student model F_S with the merged data to obtain the self-trained model F.

For testing, the self-trained model F is used for extracting features and cosine distances between samples are computed for retrieval.

IV. EXPERIMENTS

Our method was evaluated on two large-scale person re-identification benchmark datasets Market-1501 [59] and DukeMTMC [33]. We compared our method with the state-of-the-art semi-supervised and unsupervised person re-identification methods and evaluated the key components and parameters in our framework.

Experiment Settings and Datasets. The experiments were conducted on Market-1501 [59] and DukeMTMC [33]. Market-1501 [59] contains 32,217 images of 1,501 identities in 6 cameras. DukeMTMC [33] consists of 36,411 images of 1,812 identities in 8 cameras. We followed the standard train/test split of Market-1501 [59] and DukeMTMC [33]. For evaluation in the semi-supervised setting, we randomly selected data of 1/3 identities in the training set as labelled data, (i.e., 250 identities for Market-1501 [59] and 234 identities for DukeMTMC [33]) and the other data was unlabelled, following the setting in multi-view clustering (MVC) [16], a recent advanced semi-supervised Re-ID method. No source data of other pedestrian datasets was used for pretraining or transfer learning. The performance metrics, cumulative matching characteristic (CMC) and mean Average Precision (mAP), were applied following the standard evaluation protocols in [59] and [33].

Implementation Details. For training multiple teacher models for multi-teacher selective knowledge distillation, we first sampled $N_{sub} = 5$ subsets from labelled data by randomly selecting $C_{sub} = \lfloor (N_{sub} - 1)C/N_{sub} \rfloor$ identities to form each subset, where C is the number of identities of the labelled data and $\lfloor \cdot \rfloor$ is the floor rounding function.

For the self-trained model F, student model F_S and teacher models $\{F_{Tm}\}_{m=1}^{N_{sub}}$, we adopted ResNet-50 model [57] as backbone model. Both the teacher models $\{F_{Tm}\}_{m=1}^{N_{sub}}$ and student models F_S were initialized by ImageNet pretraining. The input images were resized to 384×128. In the teacher models $\{F_{Tm}\}_{m=1}^{N_{sub}}$, a convolution layer was applied to the feature maps before global average pooling of ResNet-50 to reduce the number of channels to 256 and then the feature maps were split into 6 stripes for computing Softmax cross-entropy loss individually, following the training strategy in [7]. For multi-teacher selective knowledge distillation, the network architecture of the student model F_S was the same with the teacher models $\{F_{Tm}\}_{m=1}^{N_{sub}}$ except that there were N_{sub} unshared subset-specific convolution layers for reducing the number of channels to 256, which were playing the roles of subset-specific projections $\mathbf{P}_m$ for multi-teacher selective similarity distillation (Section III-A). For camera-aware pseudo labelling, the common features $\mathbf{X}_S$ of the student model were extracted for clustering. Finally, for joint self training of the student model with both labelled data and pseudo labelled data, the layer of the common features of the student model were used for computing Softmax cross-entropy loss for classification.

In training, 30 epochs were used for training teacher models $\{F_{Tm}\}_{m=1}^{N_{sub}}$; 30 epochs were used for distilling the student model F_S; 45 epochs were used for fine-tuning the student model F_S with both labelled data and pseudo labelled data. For all training processes, the learning rate was 0.1 in the first 2/3 of all epochs and was reduced to 0.01 in the remaining epochs. For optimization, we used SGD optimizer [60] with momentum 0.9. The batch size was 64.

As for setting up the clustering algorithm DBSCAN [58], cosine distance between features extracted by the student model F_S was used for clustering; the number of minimum samples N_{min} in each cluster was 1 for intra-camera clustering and was 2 for cross-camera clustering; the threshold ϵ of distance for searching for neighbourhood was adaptively set by $\epsilon = 0.8 \cdot \bar{d}_{pos} + 0.2 \cdot \bar{d}_{neg}$, where $\bar{d}_{pos}$ is the mean distance of positive pairs and $\bar{d}_{neg}$ is the mean distance of negative pairs of the labelled data set. Detailed explanations of the parameters are introduced in "Basic Clustering Algorithm" in Section III-B.

A. COMPARISON TO RELATED SEMI-SUPERVISED AND UNSUPERVISED MODELS

We compared with recent advanced semi-supervised Re-ID methods multi-view clustering (MVC) [16] and PUL [12], which are closely related clustering-based methods. Our default setting was using data of 1/3 of identities in the training set as labelled data (denoted by "1/3 of IDs" in Table 1), which was the same with the setting of MVC [16]. For fair comparison with PUL [12], we also followed the setting in PUL [12] using data of 50 identities as labelled data (denoted by "50 IDs" in Table 1). Both MVC [16] and PUL [12] used ResNet-50 as backbone model, which was the same as ours.

TABLE 1. Comparison with related semi-supervised and unsupervised Re-ID methods. Data of 1/3 of identities in the training set was labelled for the semi-supervised setting. For semi-supervised Re-ID methods MVC [16] and PUL [12], our method was compared with them in the same setting, respectively. LOMO [3] and BOW [59] are unsupervised features. The others are unsupervised transfer learning methods that require source data. "R-*k*" denotes rank-*k* accuracy (%). "mAP" denotes mean average precision (%). "–" denotes not reported.

Methods	Market-1501			DukeMTMC		
	R-1	R-5	mAP	R-1	R-5	mAP
LOMO [3]	27.2	41.6	8.0	12.3	21.3	4.8
BOW [59]	35.8	52.4	14.8	17.1	28.8	8.3
PTGAN [13]	45.5	60.7	20.5	30.0	43.4	16.4
CAMEL [10]	54.5	-	26.3	-	-	-
SPGAN [14]	57.7	75.8	26.7	46.4	62.3	26.2
TJ-AIDL [11]	58.2	74.8	26.5	44.3	59.6	23.0
HHL [15]	62.2	78.8	31.4	46.9	61.0	27.2
MAR [38]	67.7	81.9	40.0	67.1	79.8	48.0
PAUL [39]	68.5	82.4	40.1	72.0	82.7	53.2
Zhong's [40]	75.1	87.6	43.0	63.3	75.8	40.4
PUL [12] (50 IDs)	50.9	66.5	24.8	36.5	52.6	21.5
Ours (50 IDs)	75.3	88.4	50.4	66.9	77.1	48.1
MVC [16] (1/3 of IDs)	75.2	-	52.6	57.6	-	37.8
Ours (1/3 of IDs)	**87.1**	**94.6**	**67.0**	**75.8**	**85.0**	**58.7**

Moreover, we also compared with unsupervised Re-ID methods including unsupervised features LOMO [3], BOW [59] and unsupervised transfer learning models PTGAN [13], CAMEL [10], SPGAN [14], TJ-AIDL [11], HHL [15], MAR [38], PAUL [39] and Zhong's [40], which required source data of other pedestrian datasets for training. The experimental results on Market-1501 [59] and DukeMTMC [33] datasets are shown in Table 1.

Our method outperformed the compared semi-supervised and unsupervised Re-ID methods. Compared with semi-supervised learning methods PUL [12] and MVC [16], our method significantly outperformed them using the same backbone model ResNet-50 in the same setting. Compared with the second best model MVC [16], the improvement on mAP of our method was about 15% on Market-1501 and 20% on DukeMTMC. PUL [12] and MVC [16] are based on clustering for learning pseudo labels as our framework. Compared with our Camera-Aware Hierarchical Clustering algorithm, they ignore the scene variations between cameras, which is a significant problem for person re-identification. Moreover, in order to improve the generalization performance of limited labelled data, compared with MVC [16] that learns multiple deep neural networks of different architectures for multi-view clustering, our method learns only one student model without increasing model size and computation costs, by means of distilling the shared and specific knowledge in different subsets of limited training data.

B. FURTHER EVALUATIONS

In this section, we further evaluated and analysed the components and parameters of our method. In our Distilled Camera-Aware Self Training framework shown in Figure 6, multi-teacher selective similarity distillation (Section III-A) and Camera-Aware Hierarchical Clustering (CAHC) (Section III-B) are the key techniques. We evaluated and analysed these two techniques in the following experiments.

TABLE 2. Evaluation of multi-teacher selective similarity distillation. Data of 1/3 of identities in the training set was labelled for the semi-supervised setting. "Supervised (1/3 of IDs)" is the baseline model. Please see Section IV-B.1 for details. The notations of performance are the same as those in table 1.

Methods	Market-1501			DukeMTMC		
	R-1	R-5	mAP	R-1	R-5	mAP
Supervised (1/3 of IDs)	81.8	92.3	57.9	69.7	81.6	49.7
Teacher model 1 (subset 1)	79.4	91.0	55.1	67.2	80.0	46.8
Teacher model 2 (subset 2)	79.4	91.1	54.1	67.8	80.3	47.3
Teacher model 3 (subset 3)	77.1	89.8	51.7	66.6	79.3	46.0
Teacher model 4 (subset 4)	78.5	90.6	53.0	68.4	80.7	48.0
Teacher model 5 (subset 5)	79.3	91.0	54.3	66.8	79.4	46.3
Teacher ensemble (5× model size)	83.6	93.0	60.5	**72.1**	82.6	53.3
L_{msd} (Eq. (3), Frobenius norm)	83.4	93.0	61.6	70.5	81.5	53.1
L_{mssd} (Eq. (4), $\ell 2$, 1-norm)	**84.6**	**93.8**	**62.8**	71.9	**82.7**	**55.0**

TABLE 3. Evaluation of the number of teachers N_{sub}. Data of 1/3 of identities in the training set was labelled for the semi-supervised setting. Our default parameter is $N_{sub} = 5$. The notations of performance are the same as those in table 1.

Number of teachers	Market-1501			DukeMTMC		
	R-1	R-5	mAP	R-1	R-5	mAP
2	78.9	90.6	55.1	66.7	79.0	48.5
3	82.7	92.6	59.9	70.1	81.2	52.6
4	84.0	93.2	61.4	71.5	82.3	54.0
5 (default)	84.6	93.8	62.8	71.9	82.7	55.0
6	84.4	93.8	62.7	72.5	83.1	55.3

1) EVALUATION OF MULTI-TEACHER SELECTIVE SIMILARITY DISTILLATION

For multi-teacher selective similarity distillation, we learned a student model with 1/3 of labelled identities. We evaluated training a ResNet-50 model using the same labelled data as a baseline model, denoted by "Supervised (1/3 of IDs)". For evaluating the effectiveness of our knowledge distillation method, we compared with ensemble of our five teacher models by similarity score fusion, denoted by "Teacher ensemble". For evaluating the effectiveness of $\ell 2$, 1-norm for multi-teacher selective similarity distillation by L_{mssd} (Eq. (4)), we compared with multi-teacher similarity distillation by L_{msd} (Eq. (3)) using Frobenius norm. These experimental results are shown in Table 2. Moreover, we analysed the number of teachers N_{sub}, a key parameter for multi-teacher selective similarity distillation. The parameter evaluation results are shown in Table 3.

Comparison with Teacher Ensemble. As shown in Table 2, the performances of the teacher models 1 to 5 are slightly lower than the baseline supervised model, since data of some identities was missing in each subset as compared with the whole training set. Both our method and ensemble of teacher models had notable improvement on the baseline supervised model, which indicates that different subsets of limited training data can provide complementary knowledge for improving generalization performance. It is an interesting observation that our method "L_{mssd}" even outperformed "Teacher ensemble" in most cases. Note that, ensemble of 5 teacher models requires 4 times more model size and computation costs than a single student model learned by distillation. Therefore, our multi-teacher selective similarity distillation method can effectively aggregate knowledge of multiple teachers without increasing model size.

TABLE 4. Evaluation of Camera-Aware Hierarchical Clustering. Data of 1/3 of identities in the training set was labelled for the semi-supervised setting. "Supervised (1/3 of IDs)" is the baseline model. "F_S (L_{mssd})" denotes the student model learned by our multi-teacher selective similarity distillation. "DBSCAN" is our basic clustering algorithm. "CAHC" denotes our Camera-Aware Hierarchical Clustering algorithm. "F_S (L_{mssd}) + CAHC" is our full model. "Fully supervised (all IDs)" is the ideal case of training with all ground-truth identities in the training set. The notations of performance are the same as those in Table 1.

Methods	Market-1501			DukeMTMC		
	R-1	R-5	mAP	R-1	R-5	mAP
Supervised (1/3 of IDs)	81.8	92.3	57.9	69.7	81.6	49.7
Supervised (1/3 of IDs) + CAHC	85.0	93.9	62.7	73.7	83.6	55.7
F_S (L_{mssd})	84.6	93.8	62.8	71.9	82.7	55.0
F_S (L_{mssd}) + DBSCAN [58]	86.0	94.2	65.9	73.7	83.4	56.2
F_S (L_{mssd}) + CAHC (full model)	**87.1**	**94.6**	**67.0**	**75.8**	**85.0**	**58.7**
Fully supervised (all IDs)	92.0	97.0	76.6	80.6	88.5	65.1

$\ell 2$, 1**-norm v.s. Frobenius Norm.** In Section III-A, we proposed a Multi-Teacher Similarity Distillation Loss L_{msd} using Frobenius norm to measure the errors of similarity matrices. We further proposed a Multi-Teacher Selective Similarity Distillation Loss L_{mssd} by applying $\ell 2$, 1-norm instead of Frobenius norm for selective distillation. Comparison between the results of L_{mssd} and L_{msd} shows that the performance can be improved by taking advantage of the row-sparsity property of $\ell 2$, 1-norm to suppress the noises in samples and teacher models.

Effect of the Number of Teachers N_{sub}. We sampled N_{sub} subsets in the training set for training N_{sub} teacher models, respectively. Each subset contained data of randomly sampled $C_{sub} = \lfloor (N_{sub} - 1)C/N_{sub} \rfloor$ identities, where C is the number of identities of the labelled data and $\lfloor \cdot \rfloor$ is the floor rounding function. Our default value was $N_{sub} = 5$. The results of using different numbers of teachers from $N_{sub} = 2$ to $N_{sub} = 6$ are reported in Table 3.

With N_{sub} increasing, the performance was improved and became stable after $N_{sub} = 4$, because using more teacher models can better reduce variance of the model predictions and achieve better generalization performance. Since more teacher models required more computation costs in training, we chose $N_{sub} = 5$ as default parameter.

2) EVALUATION OF CAMERA-AWARE HIERARCHICAL CLUSTERING

To verify the effectiveness of our Camera-Aware Hierarchical Clustering (CAHC) algorithm, we conducted component-wise evaluations. The results are reported in Table 4, in which "F_S (L_{mssd})" denotes the student model learned by our multi-teacher selective similarity distillation and "CAHC" denotes Camera-Aware Hierarchical Clustering algorithm. For comparison, we also evaluated the baseline model "Supervised (1/3 of IDs)" and the basic clustering algorithm DBSCAN [58].

Based on the models "Supervised (1/3 of IDs)" and "F_S (L_{mssd})", learning from unlabelled data by our clustering algorithm CAHC can further improve the performance with an increment of about 4% rank-1 accuracy and 4% mAP. The performance of our full model "F_S (L_{mssd}) + CAHC" is better than "Supervised (1/3 of IDs) + CAHC", with mAP increased by about 4% and 3% on Market-1501 and DukeMTMC, respectively. This indicates the significance of multi-teacher selective knowledge distillation for learning a better model for computing similarities for camera-aware pseudo labelling. To show the effectiveness of separating intra-camera clustering and cross-camera clustering in our CAHC, we compared with DBSCAN [58], the basic clustering algorithm of our CAHC. It can be observed that, CAHC is more effective than DBSCAN [58] and the improvement on DukeMTMC is more significant than that on Market-1501, since the scene variations between cameras are more severe on DukeMTMC. This indicates that, hierarchically performing intra-camera and cross-camera clustering in CAHC can effectively alleviate the influence of domain shift caused by scene variations between cameras for clustering.

As for the performance of our whole framework, compared with the ideal case of "Fully supervised (all IDs)" using the model trained with all ground-truth identities in the training set, our full model (using labelled data of 1/3 of identities) is close to the ideal case with a gap of about 5% rank-1 accuracy. Compared with the baseline model "Supervised (1/3 of IDs)", our full model can bring significant improvement of about 6% rank-1 accuracy and 9% mAP, which shows the effectiveness of our Distilled Camera-Aware Self Training framework.

V. CONCLUSION

In this paper, we study the semi-supervised person re-identification problem. Requirement of fewer labelled identities brings about better scalability of a person re-identification system, but lack of sufficient supervision information gives rise to challenges as well.

On the one hand, limited labelled data degrades the generalization performance of the learned model. To alleviate the overfitting problem, we propose a Multi-Teacher Selective Similarity Distillation Loss for selectively aggregating the knowledge of different subsets in a single student model to reduce variance of model prediction for improving generalization performance. Meanwhile, our model size can be kept unchanged just as using a single model. On the other hand, the scene variations between cameras cause domain shift in the feature space and make it difficult to mine auxiliary supervision information in the unlabelled data. To alleviate this effect when learning pseudo labels for unlabelled data, we propose a Camera-Aware Hierarchical Clustering algorithm to perform intra-camera clustering and cross-camera clustering hierarchically. To jointly exploit labelled and unlabelled data based on the above two techniques for semi-supervised learning, we develop a Distilled Camera-Aware Self Training framework. Experimental results show that our method outperformed the state-of-the-art unsupervised and semi-supervised person re-identification methods.

REFERENCES

[1] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 2288–2295.

[2] W.-S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 653–668, Mar. 2013.

[3] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2197–2206.

[4] Y.-C. Chen, W.-S. Zheng, J.-H. Lai, and P. C. Yuen, "An asymmetric distance model for cross-view feature mapping in person reidentification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 8, pp. 1661–1675, Aug. 2017.

[5] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 152–159.

[6] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3908–3916.

[7] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 480–496.

[8] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian, "Unsupervised cross-dataset transfer learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1306–1315.

[9] E. Kodirov, T. Xiang, Z. Fu, and S. Gong, "Person re-identification by unsupervised ℓ_1 graph learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2016, pp. 178–195.

[10] H.-X. Yu, A. Wu, and W.-S. Zheng, "Cross-view asymmetric metric learning for unsupervised person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 994–1002.

[11] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 2275–2284.

[12] H. Fan, L. Zheng, C. Yan, and Y. Yang, "Unsupervised person re-identification: Clustering and fine-tuning," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 14, no. 4, p. 83, Oct. 2018.

[13] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 79–88.

[14] W. Deng, L. Zheng, G. Kang, Y. Yang, Q. Ye, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 994–1003.

[15] Z. Zhong, L. Zheng, S. Li, and Y. Yang, "Generalizing a person retrieval model hetero- and homogeneously," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 172–188.

[16] X. Xin, J. Wang, R. Xie, S. Zhou, W. Huang, and N. Zheng, "Semi-supervised person re-identification using multi-view clustering," *Pattern Recognit.*, vol. 88, pp. 285–297, Apr. 2019.

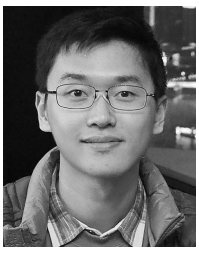
[17] J. Li, A. J. Ma, and P. C. Yuen, "Semi-supervised region metric learning for person re-identification," *Int. J. Comput. Vis.*, vol. 126, no. 8, pp. 855–874, 2018.

[18] W. Liu, X. Chang, L. Chen, and Y. Yang, "Semi-supervised Bayesian attribute learning for person re-identification," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, Feb. 2018, pp. 7162–7169.

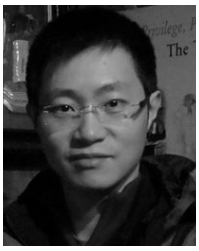
[19] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu, "Semi-supervised coupled dictionary learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 3550–3557.

[20] G. Ding, S. Zhang, S. Khan, Z. Tang, J. Zhang, and F. Porikli, "Feature affinity-based pseudo labeling for semi-supervised person re-identification," *IEEE Trans. Multimedia*, vol. 21, no. 11, pp. 2891–2902, Nov. 2019.

[21] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2008, pp. 262–275.
[22] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 2360–2367.
[23] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical Gaussian descriptor for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1363–1372.
[24] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary, "Person re-identification by support vector ranking," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Sep. 2010, pp. 1–6.
[25] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local Fisher discriminant analysis for pedestrian re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 3318–3325.
[26] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2014, pp. 1–16.
[27] W.-S. Zheng, S. Gong, and T. Xiang, "Towards open-world person re-identification by one-shot group-based verification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 591–606, Mar. 2016.
[28] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1239–1248.
[29] S. Bak and P. Carr, "One-shot metric learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1571–1580.
[30] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1249–1258.
[31] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 907–915.
[32] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3239–3248.
[33] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline *in vitro*," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3774–3782.
[34] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2138–2147.
[35] S. Bak, P. Carr, and J.-F. Lalonde, "Domain adaptation through synthesis for unsupervised person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 189–205.
[36] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 5157–5166.
[37] M. Li, X. Zhu, and S. Gong, "Unsupervised person re-identification by deep learning tracklet association," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 737–753.
[38] H.-X. Yu, W.-S. Zheng, A. Wu, X. Guo, S. Gong, and J.-H. Lai, "Unsupervised person re-identification by soft multilabel learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2148–2157.
[39] Q. Yang, H.-X. Yu, A. Wu, and W.-S. Zheng, "Patch-based discriminative feature learning for unsupervised person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3633–3642.
[40] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 598–607.
[41] L. Song, C. Wang, L. Zhang, B. Du, Q. Zhang, C. Huang, and X. Wang, "Unsupervised domain adaptive re-identification: Theory and practice," 2018, *arXiv:1807.11334*. [Online]. Available: https://arxiv.org/abs/1807.11334
[42] A. Wu, W.-S. Zheng, X. Guo, and J.-H. Lai, "Distilled person re-identification: Towards a more scalable system," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1187–1196.
[43] C. Zhang and Y. Ma, *Ensemble Machine Learning: Methods and Applications*. Berlin, Germany: Springer, 2012.
[44] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.
[45] Y. Freund, "Boosting a weak learning algorithm by majority," *Inf. Comput.*, vol. 121, no. 2, pp. 256–285, Sep. 1995.
[46] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *Comput. Sci.*, vol. 14, no. 7, pp. 38–39, 2015.
[47] A. Mishra and D. Marr, "Apprentice: Using knowledge distillation techniques to improve low-precision network accuracy," 2017, *arXiv:1711.05852*. [Online]. Available: https://arxiv.org/abs/1711.05852
[48] S. You, C. Xu, C. Xu, and D. Tao, "Learning from multiple teacher networks," in *Proc. ACM Int. Conf. Knowl. Discovery Data Mining (SIGKDD)*, Aug. 2017, pp. 1285–1294.
[49] T. Fukuda, M. Suzuki, G. Kurata, S. Thomas, J. Cui, and B. Ramabhadran, "Efficient knowledge distillation from an ensemble of teachers," in *Proc. Interspeech*, Aug. 2017, pp. 3697–3701.
[50] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for thin deep nets," 2014, *arXiv:1412.6550*. [Online]. Available: https://arxiv.org/abs/1412.6550
[51] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4133–4141.
[52] N. Passalis and A. Tefas, "Learning deep representations with probabilistic knowledge transfer," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 268–284.
[53] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3967–3976.
[54] Y. Pang, X. Li, and Y. Yuan, "Robust tensor analysis with L1-norm," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 2, pp. 172–178, Feb. 2010.
[55] H. Grabner, C. Leistner, and H. Bischof, "Semi-supervised on-line boosting for robust tracking," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, Oct. 2008, pp. 234–247.
[56] F. Roli and G. L. Marcialis, "Semi-supervised PCA-based face recognition using self-training," in *Proc. Joint IAPR Int. Workshops Stat. Techn. Pattern Recognit. Struct. Syntactic Pattern Recognit.* Berlin, Germany: Springer, Aug. 2006, pp. 560–568.
[57] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[58] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "Density-based spatial clustering of applications with noise," in *Proc. Int. Conf. Knowl. Discovery Data Mining (KDD)*, vol. 240, Aug. 1996, pp. 1–6.
[59] L. Zheng, L. Sheng, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1116–1124.
[60] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. 19th Int. Conf. Comput. Statist.*, Aug. 2010, pp. 177–186.



ANCONG WU received the bachelor's degree in intelligence science and technology from Sun Yat-Sen University, in 2015, where he is currently pursuing the Ph.D. degree with the School of Electronics and Information Technology. He is currently studying multimodality learning, metric learning, and knowledge distillation for person reidentification. His research interest includes computer vision.



WEI-SHI ZHENG received the Ph.D. degree in applied mathematics from Sun Yat-sen University, in 2008. He is currently a Professor with Sun Yat-Sen University. He has been a Postdoctoral Researcher on the EU FP7 SAMURAI Project with the Queen Mary University of London. His recent research interests include person reidentification, action/activity recognition, and large-scale machine learning algorithms. He joined the Microsoft Research Asia Young Faculty Visiting Programme. He has received the Outstanding Reviewer Award from ECCV 2016, the Excellent Young Scientists Fund from the National Natural Science Foundation of China, and the Royal Society-Newton Advanced Fellowship.

JIAN-HUANG LAI received the Ph.D. degree in mathematics from Sun Yat-sen University, in 1999. He is currently a Professor with the School of Data and Computer Science, Sun Yat-sen University. He has published over 100 scientific articles in international journals and conferences including the IEEE TPAMI, the IEEE TNN, the IEEE TIP, the IEEE TSMCB, PR, ICCV, CVPR, and ICDM. His current research interests include digital image processing, pattern recognition, multimedia communication, and wavelet and its applications.