IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 10, 2021, accepted May 3, 2021, date of publication May 6, 2021, date of current version May 17, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3077952

Asymmetric Mutual Mean-Teaching for Unsupervised Domain Adaptive Person Re-Identification

YACHAO DONG, HONGZHE LIU, AND CHENG XU, (Member, IEEE)
Beijing Key Laboratory of Information Service Engineering, Beijing Union University, Beijing 100101, China

Corresponding author: Hongzhe Liu (liuhongzhe@buu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61871039, Grant 61906017, and Grant 61802019; in part by the Beijing Municipal Commission of Education Project under Grant KM202111417001 and Grant KM201911417001; in part by the National Engineering Laboratory for Agri-Product Quality Traceability Project under Grant AQT-2020-YB2; in part by the Collaborative Innovation Center for Visual Intelligence under Grant CYXC2011; in part by the Academic Research Projects of Beijing Union University under Grant ZB10202003, Grant ZK80202001, Grant XP202015, and Grant BPHR2019AZ01; and in part by the Beijing Union University Graduate Research and Innovation Funding Project under Grant YZ2020K001.

ABSTRACT Unsupervised domain adaptive person re-identification aims to solve the problem of poor performance caused by transferring an unlabeled target domain from the labeled source domain in the re-identification task. The clustering pseudo-labels method in unsupervised learning is widely used in unsupervised adaptive person re-identification tasks, and it maintains state-of-the-art performance. However, pseudo-labels obtained through clustering often have much noise, and the use of a single network model structure and a single clustering algorithm can easily cause model learning to stagnate, making the model not generalizable. To solve this problem, this paper proposes an asymmetric mutual mean-teaching method for unsupervised adaptive person re-identification. In terms of feature extraction, two asymmetric network models with different structures are used for mutual mean-teaching on the target domain, making the features extracted by the network more robust. In terms of feature clustering, two clustering methods are used for mutual teaching to dynamically update the centroid of clusters to improve the confidence of clustering pseudo-labels. Finally, the triplet loss is improved based on the updated cluster centroid to improve the clustering effect. The proposed method is used to perform a large number of verification experiments on three public datasets. The experimental results show that the proposed method has better accuracy than other unsupervised person re-identification based on clustering pseudo-labels.

INDEX TERMS Deep learning, person re-identification, unsupervised domain adaptation, mutual mean-teaching.

I. INTRODUCTION

The purpose of person re-identification (re-ID) is to match the same person across cameras, which plays an important role in applications such as intelligent security, intelligent transportation, and video surveillance. An important step in the deep learning task is to label a large amount of data, and the data labeling in the re-ID task incurs greater work costs. Some unsupervised re-ID methods [46]–[48], [53] and one-shot learning methods [51], [52] are proposed to solve this problem. Although these unsupervised re-ID and one-shot learning methods have made great progress, in the face of a large number of labeled datasets, the new domain datasets will cause a significant decrease in the accuracy of re-ID due to the different domains. In addition, unsupervised learning without pre-training on the source domain datasets not only causes the model to converge slowly, but also the model performance is difficult to further improve. Unsupervised domain adaptive (UDA) re-ID can alleviate this situation, which aims to improve the accuracy of unsupervised learning by transferring the knowledge learned from the labeled source domain dataset to the unlabeled target domain. Due to the large domain shift and powerful supervision in source dataset, it is a popular approach for unsupervised re-ID without target

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

dataset labels [49]. Most such methods can be categorized as either image-based domain transfer methods, pseudo-label methods based on image and feature similarity, or cluster-based pseudo-label methods.

The image-based domain transfer method tries to narrow the background difference between the source and target domains and the style difference of the camera's perspective, so the network concentrates on the person feature map. Deng *et al*. [1] used generative adversarial networks (GANs) to convert the image style of the source domain to the style of the target domain while maintaining the identity of the original person and fine-tuned the model. Li *et al*. [2] transferred the pose of a person through a GAN, so that the model learned features irrelevant to the pose of the person. Jin *et al*. [3] used a plug-and-play-style normalization and stylization framework to normalize styles of different cameras, illuminations, and resolutions, reducing the difference between the source and target domains. The retrieval performance of these methods depends greatly on the ability of domain conversion, such asl the quality of images generated by the GAN. However, these conversion methods are not stable across a variety of complex environments.

The pseudo-label method based on image and feature similarity obtains some labels by calculating the distance, such as from reference images, features, and person attributes. It tries to align the feature distribution between the source and target domains. The reference may come from the source domain or a memory module. To obtain a good classifier and solve the gradient dispersion problem in traditional multi-label classification loss, MMCL [4] uses memory-based multi-label classification loss to align the features of the target domain and obtain some robust pseudo-labels. Yang *et al*. [5] explored the internal connection between the global and local features of the source and target domains and minimized the gap between domains for unsupervised person re-ID. Zhong *et al*. [6] introduced a sample memory to store the features of the target domain to adapt to sample, camera, and neighborhood invariance and assign soft labels using exemplary memory modules to store average features. Yu *et al*. [7] conducted multiple soft-label learning by comparing a reference person and an unlabeled person of the labeled auxiliary dataset and solved the problem of a lack of paired label guidance. However, it is difficult for reference images and features to be representative and generalized, and it is impossible to generate accurate labels for high-level performance.

To solve the problem of the image-based domain transfer method or GAN based method is not stable across a variety of complex environments, some scholars have proposed methods based on clustering pseudo-labels, which is widely used and has proved to be more effective than other methods. Lin *et al*. [8] proposed bottom-up clustering BUC to jointly optimize the relationship between the network and unlabeled samples. Yang *et al*. [9] proposed self-similarity grouping from global to local methods to mine the potential similarity of unlabeled samples, automatically establish multiple clusters from different viewpoints, and label these

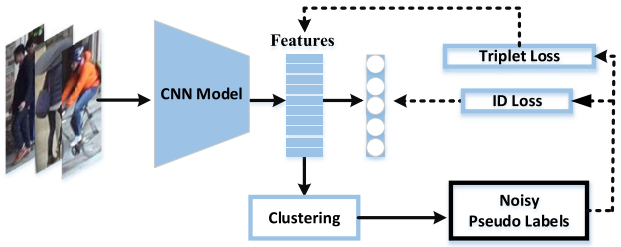

FIGURE 1. Baseline method of unsupervised learning based on clustering pseudo-labels. This figure illustrates that the pseudo-label output is used as its own supervision labels.

independent clusters as pseudo-labels to supervise training. Zhai *et al*. [10] used target domain sample generation to increase the clustering points, and hence the diversity of categories, and a feature encoder to minimize the distance between images within a class in the feature space to improve the accuracy of cross-domain person re-ID. Jin *et al*. [11] tried to distinguish the distributions of positive and negative samples using a momentum update strategy during training. Ge *et al*. [12] proposed a mutual mean-teaching framework (MMT) using two symmetric network models to supervise each other and designed a symmetric framework with hard pseudo-labels and refined soft labels, using more robust soft labels to optimize the pseudo-labels online.

The MMT currently maintains advanced accuracy and excellent inference speed, so this paper is based on the MMT method and improved it, and proposed an asymmetric mutual mean-teaching method for unsupervised adaptive person re-identification (AMMT). The difference between AMMT and MMT is: 1) The two-stage asymmetric network is used for mutual mean-teaching. The asymmetric network makes the process of mutual mean-teaching more diverse and complementary. The two-stage method enables the network model to dig deeper information. 2) MMT uses a single clustering method to obtain clustering pseudo-labels with low confidence, while AMMT uses two asymmetric clustering methods to guide each other to cluster, and the resulting pseudo-labels are of higher quality. 3) Since 1) and 2) make the noise of clustering pseudo-labels smaller, it is more reasonable to use the cluster centroid to distinguish positive and negative sample pairs for triplet loss.

The baseline method based on clustering pseudo-labels is shown in Fig. 1. It uses a single symmetrical network model and a single clustering algorithm. The pseudo-labels generated by clustering are directly used as their own supervised labels for supervised training, which can easily cause the features extracted by the network to be insufficiently robust, and the pseudo-labels obtained by clustering contain much noise.

A. MOTIVATIONS

- Pseudo-label noise generated by the clustering algorithm greatly hinders the training of the neural network and is ignored by most methods. This will bias the centroid of the clustering category, resulting in a poor clustering

effect when the class centroid is used as a part of the loss function, causing the network model to have no generalization. Therefore, the reliability of the false label generated by clustering is still not well resolved.

- With the two symmetric network models of mutual teaching or learning, extracted features will be relatively similar and single, without diversity and complementarity, resulting in stagnation in the two models' mutual mean-teaching. It remains to be clarified how to use multiple models and different clustering methods' mutual mean-teaching to improve the robustness of the network and the accuracy of clustering.

B. CONTRIBUTIONS

- In terms of feature extraction, the method in this paper learns and adapts to network model with two different architectures. It has two stages for mutual mean-teaching. The first stage learns the blind spot features that a single network cannot learn through mutual mean-teaching. In the second stage, in-depth mutual mean-teaching is carried out to dig out more in-depth information.
- To improve the confidence of clustering pseudo-labels, different networks are used to learn distinguishing features, two clustering algorithms learn from each other, and the strategy of seeking common ground while reserving differences extracts more reliable clustering instances and further reduces the noise generated by clustering.
- AMMT tries to change the design of the triplet loss according to the cluster centroids derived from clustering instances with higher confidence, so that the cluster centroids are closer to the same category instance, and different category instances are farther apart, making the clustering more discriminative. The final result is a single model, and no complicated multi-model integration is required. Experimental results show that the proposed method can effectively improve the effect of re-ID on the public datasets Market1501 [13], DukeMTMC-reID [14], and MSMT17 [15].

II. RELATED WORK

We introduce research of person re-ID in the UDA field and some work related to our proposed method, i.e., unsupervised domain adaptive person re-ID and mutual mean-teaching.

A. UNSUPERVISED DOMAIN ADAPTIVE PERSON RE-ID

UDA learning refers to learning labeled data features in the source domain, followed by training, testing, and application in the unlabeled target domain. It not only saves the cost of manual labeling but is indispensable in deep learning in actual situations. Among the three types of UDA re-ID methods, we use the better-performing method based on clustering pseudo-labels, whose steps are to 1) pre-train the model on the source domain, 2) cluster the last layer of classification features output by the pre-training model to generate pseudo-labels, and 3) use pseudo-labels to supervise the learning of the network on the target domain. The method loops through the second and third steps until the network is stable [16].

However, pseudo-labels clustered using this method have much noise. Zhang *et al*. [17] proposed a progressively enhanced self-training method, combining conservative and promoting stages to enhance the performance of the model in the target domain. The asymmetric collaborative teaching network [18] reduces the noise generated by clustering and uses some unlabeled external points to make the model perform better in the target domain, but the asymmetric network used is the same network type, and only the network parameters are different. Ding *et al*. [41] proposed an AE (Adaptive Exploration) method, which addressed the domain-shift problem of re-ID in an unsupervised manner by introducing the re-ID model in the target domain to maximize the distance between all person images and minimize the distance between similar person images. The mutual mean teaching framework [12] refines soft and hard pseudo-labels and extracts more reliable pseudo-labels. The same type of network is used, but the parameters are different.

Inspired by [12], we adopt a multi-network mutual mean-teaching strategy, because the strategy of mutual mean teaching aims to supervise each other through a variety of pseudo-labels to make the extracted features more robust and diverse, while the self-supervision of a single network does not have these advantages. In order to further expand this diversity of mutual supervision, we adopt different types of networks for mutual mean-teaching in terms of breadth, increasing the diversity of the network learned features, so that the two can complement each other. In terms of depth, we conduct deep-level mutual mean-teaching, i.e., the two models obtained from mutual mean-teaching are used as input training again, so each can fully learn the knowledge that the other cannot learn.

B. MUTUAL MEAN-TEACHING

Mutual mean-teaching, deep mutual learning, knowledge transfer, and knowledge distillation have the same ideas. Both networks guide each other or use teacher-student model teaching methods to learn better model knowledge. The method of extracting knowledge from a trained neural network and transferring it to another model network has been extensively studied in recent years [19]–[24]. The typical method of knowledge transfer is teacher-student model learning, which uses the soft output distribution of the teacher network to supervise the student network, so that the student model learns the ability to discriminate the teacher model. However, methods with a teacher-student mechanism are mainly designed to solve the problem of supervision or knowledge distillation. The labeled and unlabeled data share the same set of labels, so they cannot be directly used in the UDA re-ID task. The mean-teacher model [25] averages the model weights in different training iterations instead of the predictions of unlabeled samples. Each iteration

will update the teacher network, which will improve the performance of the teacher network. Deep mutual learning [26] uses a series of student models to guide and learn from each other by training under each other's supervision. Mutual mean teaching uses a symmetric frame with hard pseudo-labels and refined soft labels for UDA re-ID. It can be seen as a combination of mean-teacher model and deep mutual learning.

Most methods that adopt the teacher-student mechanism use symmetric frameworks, but this largely ignores the importance of mutual mean-teaching in asymmetric networks and mutual teaching in asymmetric clustering methods.

C. QUALITY OF PSEUDO LABELS

Methods based on clustering pseudo-labels rely heavily on the quality of pseudo-labels, so how to reduce the noise of pseudo-labels and improve the quality of pseudo-labels is a focus of research.

Co-teaching [40] and Co-mining [50] mainly adopted refining the training strategies to optimize the quality of pseudo-labels and reduce the noise of pseudo-labels. Zoph *et al*. [42] proposed a self-training method using the loss normalization technique to reduce the noise in the pseudo-label. Dong *et al*. [43] used multi-modal learning methods to utilize mutual information from multiple models to improve the semi-supervised performance. Dong *et al*. [44] proposed an interaction mechanism between a teacher and two students to generate more reliable pseudo labels for unlabeled data. The two students are instantiated as dual detectors, the teacher learns to judge the quality of the generated pseudo-labels. Before the retraining stage, the students filter out unqualified sample. In this way, the student model gets feedback from its teacher and retrains with the high-quality data generated by itself. Dong *et al*. [45] proposed an Isometric Propagation Network (IPN) method, which learned to generate the vision feature with semantic information for unlabeled data/unseen classes.

III. PROPOSED APPROACH

We use asymmetric mutual mean-teaching to study the problem of UDA re-ID from the source domain of labeled data to the target domain of unlabeled data. Given the labeled source domain dataset $D_S = \{X_S, Y_S\}$, the unlabeled target domain is represented as $D_T = \{X_T\}$, where X_S and X_T are the images of the source and target domain, respectively, and Y_S is the identity domain of the source domain image. Each sample $X_{S,j}$ in D_S corresponds to the character identity $Y_{S,j}$ in Y_S. Each sample $X_{T,j}$ in the target domain $D_T = \{X_T\}$ has no corresponding identity label. We aim to transfer the knowledge learned in the source domain D_S to the model of the target domain D_T without labels. We introduce the overall network structure and method implementation process of AMMT.

A. APPROACH OVERVIEW

Fig. 2 shows the overall framework of AMMT. First, pre-train the model on the labeled source domain, and then conduct asymmetric mutual mean teaching on the unlabeled target domain, so that the model can perform well in the unlabeled target domain. AMMT has three parts.

(1) Pre-training multiple models on the source domain. We use ResNet network models [27] with different structures, ResNet50-IBN-a [28] and ResNeSt50 [29], to perform supervised pre-training on the source domain dataset Market1501 and DukeMTMC, and mark these respective models as M_1 and M_2, where M_1' and M_2' are the respective models obtained by using different random data augmentation.

(2) Mutual mean-teaching of multiple network model within different structures. The asymmetric mutual mean-teaching of the models in the first stage is that of M_1, M_2, and M_1', M_2', to obtain excellent models 1 and 2, which perform well in the target domain. The second stage is in-depth asymmetric mutual mean-teaching of these two excellent models to finally get the best performing best model, then use of this best model to extract 2048-dimensional feature maps from person images to facilitate subsequent feature matches.

(3) Multiple clustering and mutual teaching. As shown in Fig. 3, we extract features from the pre-trained network on the target domain and use k-means [30] and DBSCAN [31] for mutual teaching of these features using two clustering methods to obtain reliable clustering pseudo-labels, which are used as supervision signals to supervise and train the network until convergence.

B. PRE-TRAINING LEARNING ON THE SOURCE DOMAIN

The pre-trained model on the source domain uses label smoothing classification loss and soft triplet loss to update network parameters. The model parameter of network M_k is marked as θ_k, the image $X_{S,j}$ is input to M_k to obtain the output feature as $F\left(X_{S,j} \mid \theta_k\right)$, S is the source domain, T is the target domain, the probability that the output corresponds to the predicted value j is $P_j\left(X_{S,j} \mid \theta_k\right)$, and the classification loss of label smoothing is defined as

$$\mathcal{L}_{S,id}^{k} = \frac{1}{N_s} \Sigma_{i=1}^{N_s} \Sigma_{j=1}^{M_s} q_j \log p_j\left(X_{s,i} \mid \theta_k\right), \tag{1}$$

$$q_j = \begin{cases} 1 - \delta + \dfrac{\delta}{M_s}, & j = Y_{S,j} \\ \dfrac{\delta}{M_s}, & j \neq Y_{S,j} \end{cases}. \tag{2}$$

where N_s is the number of pictures in the source domain, M_s is the number of person IDs in the source domain, and δ is a small constant, which is set to 0.1 to prevent the model from trusting the training set too much. The soft triplet loss

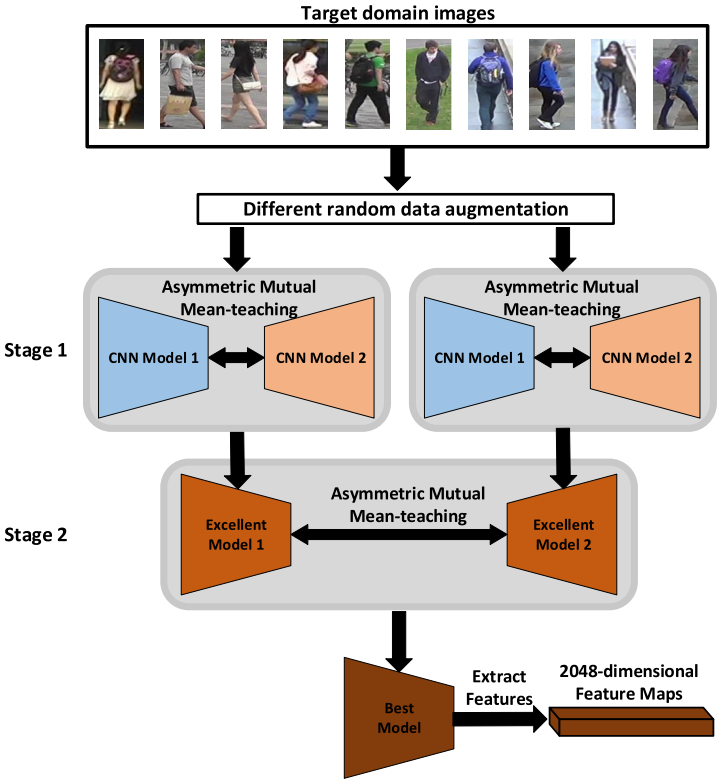

FIGURE 2. Framework of proposed AMMT method. The two CNN Model 1 and CNN Model 2 in the Stage 1 are input through different data augmentation, CNN Model 1 and CNN Model 2 represent ResNet50-IBN-a and ResNeSt-50, respectively. Then the two well-performing models obtained in the Stage 1 are used for asymmetric mutual mean-teaching again, this part is Stage 2. Finally, Best Model is obtained for feature extraction and testing.

is defined as

$$\mathcal{L}^k_{S,stri} = -\frac{1}{N_s} * \Sigma_{i=1}^{N_s} log \times \frac{e^{\left\|F(X_{S,i}|\theta_k)-F\left(X_{S,i^-}|\theta_k\right)\right\|}}{e^{\left\|F(X_{S,i}|\theta_k)-F\left(X_{S,i^+}|\theta_k\right)\right\|} + e^{\left\|F(X_{S,i}|\theta_k)-F\left(X_{S,i^-}|\theta_k\right)\right\|}}. \tag{3}$$

where X_{S,i^+} is the most difficult positive sample of $X_{S,i}$ in the source domain dataset, X_{S,i^-} is the most difficult negative sample, and $\|.\|$ is the distance metric for calculated features. We use Euclidean distance. Hence, the overall loss function in the pre-training model is defined as

$$\mathcal{L}^k_S = \mathcal{L}^k_{S,id} + \mathcal{L}^k_{S,tri}. \tag{4}$$

where $\mathcal{L}^k_S$ is the loss function corresponding to network model M_k in the source domain. Network models M_1, M_2 and $M_1^{'}, M_2^{'}$ are trained on the source domain dataset for mutual mean-teaching in asymmetric model.

C. ASYMMETRIC MUTUAL MEAN-TEACHING OF MULTIPLE NETWORK MODEL WITHIN DIFFERENT STRUCTURES

Fig. 2 shows the asymmetric model mutual mean-teaching. It using the network models CNN Model 1 and CNN Model 2 are denoted as M_1 and M_2, respectively. The initialization parameters are trained on the source domain. If the features output by M_1 and M_2 are directly used for mutual supervision, the predictions of these two networks may converge to be equal to each other, they will lose their output independence, and the complementarity in the process of mutual mean-teaching will be greatly reduced, resulting in classification errors and noise of pseudo-labels may be amplified during training.

To avoid error amplification, we use the average models M_1^* and M_2^* corresponding to M_1 and M_2 to generate reliable soft pseudo-labels, and use the output of M_1^* and M_2^* to supervise M_2 and M_1 respectively. The parameters corresponding to M_1 and M_2 are denoted as θ_1 and θ_2. The corresponding parameters of M_1^* and M_2^* are denoted as θ_1^* and θ_2^*. The parameters of the average model are updated as follows

$$\theta_1^{*(T)} = \alpha\theta_1^{*(T)} + (1-\alpha)\,\theta_1^{(T-1)}, \tag{5}$$

$$\theta_2^{*(T)} = \alpha\theta_2^{*(T)} + (1-\alpha)\,\theta_2^{(T-1)}. \tag{6}$$

where $\theta_k^{*(T)}$ is the parameter of M_k^* in the T-th iteration, $\theta_1^{(T-1)}$ is the parameter of M_k^* in the previous iteration (T-1), where α is a scale factor to be within the range [0, 1), the initial average parameters $\theta_k^{*(0)}$ is equal to $\theta_k^{(0)}$.

In person re-ID tasks, classification loss and triplet loss are usually used to jointly learn and update the parameters of the network, making it more robust. The feature maps output by M_1^* and M_2^* are denoted as $F\left(X_{T,i} \mid \theta_1^*\right)$ and $F\left(X_{T,i} \mid \theta_2^*\right)$ respectively, and are used as each other's supervision signals, by calculating mutual ID loss and mutual triplet loss back-propagation updates the parameters of M_2 and M_1. Mutual ID loss is defined as

$$\mathcal{L}_{mid}^{t}\left(\theta_1^* \mid \theta_2\right) = -\frac{1}{N_t}\Sigma_{i=1}^{N_t}[C\left(F\left(X'_{T,i} \mid \theta_1^*\right)\right)] \cdot \log\left[C\left(F\left(X_{T,i} \mid \theta_2\right)\right)\right], \tag{7}$$

$$\mathcal{L}_{mid}^{t}\left(\theta_2^* \mid \theta_1\right) = -\frac{1}{N_t}\Sigma_{i=1}^{N_t}[C\left(F\left(X_{T,i} \mid \theta_2^*\right)\right)] \cdot \log\left[C\left(F\left(X'_{T,i} \mid \theta_1\right)\right)\right]. \tag{8}$$

where $\boldsymbol{C}(\cdot)$ represents the softmax processing of the feature to obtain the soft pseudo-labels, so that the pseudo-labels predicted by the two networks are supervised by each other and can learn more diversified features. Networks will not stall each other. Mutual triplet loss is determined as

$$\mathcal{L}_{mtri}^{t}\left(\theta_1^* \mid \theta_2\right) = -\frac{1}{N_t}\Sigma_{i=1}^{N_t}[D_i\left(\theta_2\right)\log D_i\left(\theta_2\right) + \left(1 - D_i\left(\theta_1^*\right)\right)\log\left(1 - D_i\left(\theta_1^*\right)\right)], \tag{9}$$

$$\mathcal{L}_{mtri}^{t}\left(\theta_2^* \mid \theta_1\right) = -\frac{1}{N_t}\Sigma_{i=1}^{N_t}[D_i\left(\theta_1\right) logD_i\left(\theta_1\right) + \left(1 - D_i\left(\theta_2^*\right)\right)\log\left(1 - D_i\left(\theta_2^*\right)\right)]. \tag{10}$$

where

$$D_i\left(\theta_k\right) = \frac{e^{\left\|F\left(X_{T,i} \mid \theta_k\right) - F\left(X_{T,i^-} \mid \theta_k\right)\right\|}}{e^{\left\|F\left(X_{T,i} \mid \theta_k\right) - F\left(X_{T,i^+} \mid \theta_k\right)\right\|} + e^{\left\|F\left(X_{T,i} \mid \theta_k\right) - F\left(X_{T,i^-} \mid \theta_k\right)\right\|}}. \tag{11}$$

$D_i\left(\theta_k\right)$ denotes the softmax of the feature distance between negative sample pairs. The difference between the mutual triplet loss and general triplet loss is that the mutual triplet loss function is designed to learn to approximate the ratio of the distance between the positive sample pair and the negative sample pair distance, which is more suitable for unsupervised learning features.

D. ASYMMETRIC CLUSTERING MUTUAL TEACHING AND AMMT ALGORITHM DESCRIPTION

Fig. 4 shows the asymmetric mutual teaching of the multiple methods in Fig. 3. The k-means is particularly susceptible to outliers. When the algorithm traverses the centroid, the outliers have a significant impact on the movement of the centroid before reaching stability and convergence, and the number of categories needs to be set. DBSCAN does not require specifying the number of clusters, avoids outliers, and works very well in clusters of any shape and size. Furthermore, using these two different types of clustering methods can enhance the diversity of pseudo-labels. It is conducive to the mutual screening of clustering pseudo-labels to obtain labels that are closer to the true value.

Therefore, we use k-means and DBSCAN for mutual clustering teaching. Each clustering algorithm has a certain amount of noise in the cluster categories, causing bias in cluster centroids, and large errors in subsequent pseudo-label supervision. The Mutual Intersection in Fig. 4 is used to mutually select the pseudo-labels obtained by the two clustering methods. The process is Eq.12 and Eq.13, the cluster categories L_K^i obtained by k-means and L_D^j obtained by DBSCAN should be learned separately, i.e.,

$$L_K^{\prime i} = \Sigma_{i=1}^{N_K}\Sigma_{j=1}^{N_D}\max(L_K^i \cap L_D^j), \tag{12}$$

$$L_K^{*i} = \begin{cases} L_K^{\prime i}, & \frac{L_K^{\prime i}}{L_K^i} \geq \xi \\ L_K^i, & \frac{L_K^{\prime i}}{L_K^i} < \xi \end{cases}. \tag{13}$$

where L_K^i is the i-th class obtained using k-means clustering, and L_D^j is the j-th class obtained using DBSCAN clustering, ξ is set to 0.5, L_K^{*i} is the clustering result with the highest confidence.

After reducing the noise of each cluster category through mutual teaching, reliable cluster centroids and pseudo-labels are obtained. These reliable cluster centroids obtained from each cluster are used to dynamically update *Features*1 and *Features*2. The obtained pseudo-labels are used to dynamically update the parameters of M_1 and M_2 through ID loss, triplet loss, and centroid triplet loss. The ID loss is defined as

$$\mathcal{L}_{id}^{T}(\theta_k) = \frac{1}{N_T}\Sigma_{i=1}^{N_T}\Sigma_{j=1}^{M_T} q_j \log p_j\left(X_{T,i} \mid \theta_k\right), \tag{14}$$

$$q_j = \begin{cases} 1 - \delta + \frac{\delta}{M_T}, & j = \tilde{y}_j^t \\ \frac{\delta}{M_T}, & j \neq \tilde{y}_j^t \end{cases}. \tag{15}$$

where N_T is the number of pictures in the target domain, M_T is the number of person IDs in the source domain, δ is a small constant, and $\tilde{y}_j^t$ is the cluster with high-confidence pseudo-labels obtained through multiple clustering in mutual teaching, i.e., the pseudo-label corresponding to the identity $\boldsymbol{j}$ in the target domain. The soft triplet loss is defined as

$$\mathcal{L}_{tri}^{T}(\theta_k) = -\frac{1}{N_T} * \Sigma_{i=1}^{N_T}\log \frac{e^{\left\|F\left(X_{T,i} \mid \theta_k\right) - F\left(X_{T,i^-} \mid \theta_k\right)\right\|}}{e^{\left\|F\left(X_{T,i} \mid \theta_k\right) - F\left(X_{T,i^+} \mid \theta_k\right)\right\|} + e^{\left\|F\left(X_{T,i} \mid \theta_k\right) - F\left(X_{T,i^-} \mid \theta_k\right)\right\|}}. \tag{16}$$

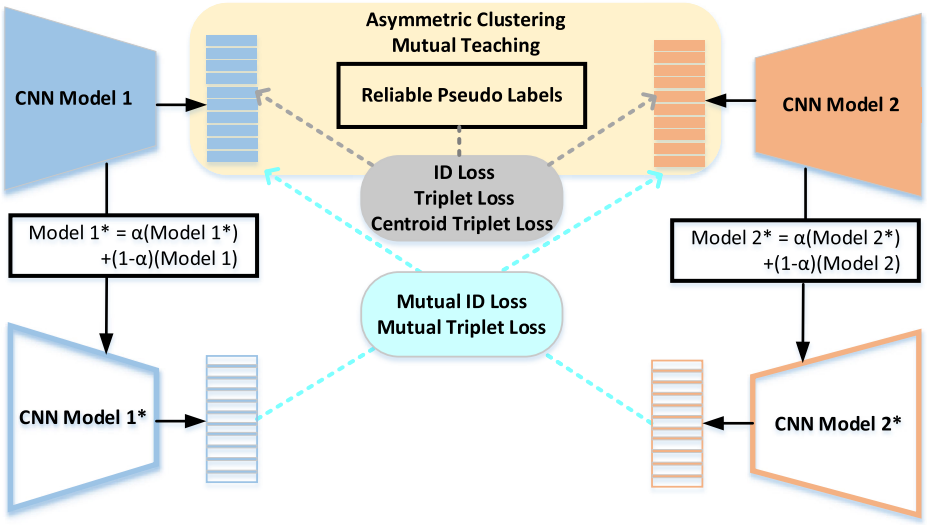

FIGURE 3. Details of asymmetric mutual mean-teaching. First, the features extracted by CNN Model 1 and CNN Model 2 are used to obtain reliable pseudo-labels through Asymmetric Clustering Mutual Teaching, and then these reliable pseudo-labels are used as signals to supervise themselves. CNN Model 1* and CNN Model 2* are the corresponding average models of CNN Model 1 and CNN Model 2, and their output features or pseudo-labels are used as mutual supervision signals.

where X_{T,i^+} is the most difficult positive sample of $X_{T,i}$ in the target domain dataset, X_{T,i^-} is the most difficult negative sample, and $\|.\|$ is the distance measure used to calculate the distance between features.

The traditional triplet loss uses image features to narrow or push the distance between positive and negative samples to distinguish the network model. We replace the image features with the cluster centroid features with higher confidence, which can reduce the noise of the unsupervised network output. The centroid triplet loss is defined as

$$\mathcal{L}_{CTL}^{T}(\theta_k) = \frac{1}{N_T}\Sigma_{i=1}^{N_T}\max \times (0, \left\|Centroid\left(X_{T,i}\right) - F\left(X_{T,i^+} \mid \theta_k\right)\right\| + m - \left\|Centroid\left(X_{T,i}\right) - F\left(X_{T,i^-} \mid \theta_k\right)\right\|). \quad (17)$$

where $\theta_k = \theta_1, \theta_2, \theta_1^*, \theta_2^*$.

The overall loss is defined as

$$\mathcal{L}_{ALL} = \mathcal{L}_{mid}^{t} + \mathcal{L}_{mtri}^{t} + \mathcal{L}_{id}^{T} + \mathcal{L}_{tri}^{T} + \mathcal{L}_{CTL}^{T}. \quad (18)$$

The algorithm of the AMMT is summarized in Algorithm 1.

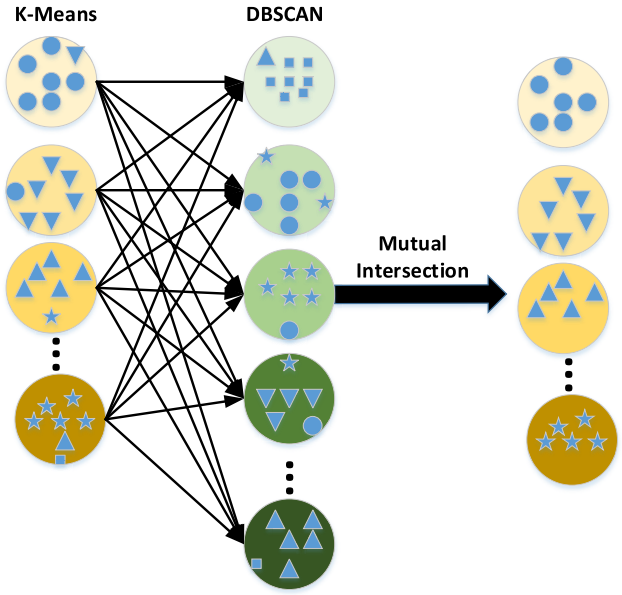

FIGURE 4. Details of asymmetric mutual teaching in multiple clustering. The use of K-means and DBSCAN two clustering algorithms can enhance the quality of pseudo-labels. Mutual Intersection is the process of mutual selection of two clustering methods.

IV. EXPERIMENTS

A. DATASET

To demonstrate the superiority of the proposed method, we conducted experiments on three public person re-ID datasets, Market1501, DukeMTMC-reID (DukeMTMC), and MSMT17, which are described in Table 1.

The Market1501 used 5 high-resolution cameras and 1 low-resolution camera to capture 1501 different person. The training set contains 751 person IDs, the test set contains 750 pedestrian IDs. The DukeMTMC is a high-resolution video data set collected by 8 synchronized cameras. The number of person IDs is 1404, and the training set contains 702 person IDs. The test set contains 702 person IDs. The MSMT17 data set has the advantages of a large number of person IDs and a large number of cameras. The dataset is taken on campus with 15 cameras in different weather conditions and different seasons. There are 4101 person IDs in total, the training set contains 1041 person IDs, and the test set contains 3060 person IDs.

Algorithm 1 Asymmetric Mutual Mean-Teaching Network (AMMT)

Input: Source domain dataset $D_S = \{X_S, Y_S\}$, target domain $D_T = \{X_T\}$.
Input: Asymmetric network model M_1 and M_2, M_1' and M_2'.
Output: Best model parameters θ_k.

```
Initialize pre-trained the weights θ1 and θ2 of M1 and M2, the weights θ1' and θ2' of M1' and M2'.
Stage 1:
for each epoch do
    Extract feature on D_T: F(X_T | θ_k)
    Generate reliable pseudo-labels ỹ_j^t of X_T from asymmetric clustering mutual teaching by Eq.12 and 13.
    for each iteration do
        Calculate soft-labels from each temporally average model by Eq.9 and 10.
        Joint update parameters θ_k by the gradient descent of the objective function Eq.18.
        Update temporally average model weights θ_k following Eq.5 and 6.
    end for
end for
Get Excellent Model 1 and Excellent Model 2.
Stage 2:
Input Excellent Model 1 and Excellent Model 2.
Do Stage 1.
Return Best model parameters θ_k.
```

TABLE 1. Public person re-identification dataset information.

Dataset	Cam	Person ID	Training	Testing	Query
Market1501	6	1501	12936	19732	3368
DukeMTMC-reID	8	1404	16522	17661	2228
MSMT17	15	4101	32621	82161	11659

B. IMPLEMENTATION DETAILS

1) EXPERIMENTAL HARDWARE ENVIRONMENT

The experimental hardware and software environment included an Ubuntu 18.04 operating system, PyTorch 1.2.0 deep learning framework, Python 3.7 programming language, TITAN V GPU, and 11 GB memory.

2) NETWORK MODEL

The two network models used by the proposed method were ResNet50-IBN-a and ResNeSt-50, using layers 1-4 of the network model. Adaptive average pooling and batch normalization for the output features made the features smoother and ensured that the next feature comparison was more predictive. The final 2048-dimensional features were input to the prediction layer to obtain the person identity results of the predicted classification. At the same time, 2048-dimensional features were used for feature similarity measurement and predicted person identity features for classification loss function backpropagation to update network parameters. We used ImageNet pre-training parameters to initialize the network weights, first pre-training on the source domain and then updating the network weight parameters again.

3) TRAINING

To speed up the training and inference speed, we resized all input person images to 256×128. The same images were input to two networks, but the data augmentation methods, such as random flip and random erase, were different. Each batch size was set to 64 and contained 16 randomly selected person identities, each corresponding to four instance images. The gradient optimizer selected the adaptive gradient optimizer (Adam), the momentum was set to 0.9, and the weight attenuation coefficient was 0.0005. In the first stage, the learning rate of the mutual mean-teaching of two different network models pre-trained on the source domain was 0.0035, and the number of epochs was 40. In the second stage, the two networks in the first stage after the first round of mutual mean-teaching were used for mutual mean-teaching again. The learning rate was 0.002, with 10 epochs.

4) CLUSTERING

We used k-means and DBSCAN for clustering, where the k-means clustering category was set to 500, and the ε-neighborhood distance threshold in DBSCNA was determined in the first epoch according to the mean value after sorting the feature distance matrix. min samples was set to 4, and the distance calculation method metric was precomputed. The mutual teaching threshold ξ of the two clustering methods was set to 0.5.

5) TESTING

We only used the best one model after asymmetric mutual mean-teaching to test and inference, and does not need multiple complex models.

C. COMPARISON WITH STATE-OF-THE-ART METHODS

To verify the effectiveness of AMMT, we selected the method based on image-based domain transfer [1], [3], [2], [36]–[38], pseudo-labels based on image and feature similarity [4], [6], [7], [32]–[35], and cluster-based pseudo-labels [8]–[10], [12], [16], [17], [39]–[41]. From the comparison of the experimental results in Tables 2 and 3, it can be found that the method in this article has an advantage over these methods.

Comparing the methods based on image-based domain transfer and pseudo-labels based on image and feature similarity, AMMT performs much better than these methods.

We mainly compare methods based on clustering pseudo-labels, and the experiments for the comparison all have 500 clustering categories, which is more comparable. More

TABLE 2. Comparison with state-of-the-art methods on Market-1501 and DukeMTMC datasets. mAP (%), Rank-1 (%), Rank-5 (%) and Rank-10 (%) are reported. The missing value is denoted as '–', which is not reported in original paper. Bold text represent the best result in the same evaluation standard.

Methods		DukeMTMC to Market1501				Market1501 to DukeMTMC			
		mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
Pseudo labels method based on image and feature similarity	MMFA [32]	27.4	56.7	75.0	81.8	24.7	45.3	59.8	66.3
	TJ-AIDL [33]	26.5	58.2	74.8	81.1	23.0	44.3	59.6	65.0
	UCDA-CCE [34]	30.9	60.4	-	-	31.0	47.7	-	-
	ECN [6]	43.0	75.1	87.6	91.6	40.4	63.3	75.8	80.4
	MAR [7]	40.0	67.7	81.9	-	48.0	67 . 1	79.8	-
	ECN++ [35]	63.8	84.1	92.8	95.4	54.4	74.0	83.7	87.4
	MMCL [4]	60.4	84.4	92.8	95.0	51.4	72.4	82.9	85.0
Image-based domain transfer method	SPGAN [1]	22.3	41.1	56.6	63.0	22.8	51.5	70.1	76.8
	HHL [36]	31.4	62.2	78.8	84.0	27.2	46.9	61.0	66.7
	ATNet [37]	25.6	55.7	73.2	79.4	24.9	45.1	59.5	64.2
	CamStyle [38]	27.4	58.8	78.2	84.3	25.1	48.4	62.5	68.9
	PAD-Net [2]	47.6	75.2	86.3	90.2	45.1	63.2	77.0	82.5
	SNR [3]	61.7	82.8	-	-	58.1	76.3	-	-
Cluster-based pseudo labels method	PUL [16]	20.5	45.5	60.7	66.7	16.4	30.0	43.4	48.5
	BUC [8]	38.3	66.2	79.6	84.5	27.5	47.4	62.6	68.4
	UDAP [39]	53.7	75.8	89.5	93.2	49.0	68.4	80.1	83.5
	PCB-PAST [17]	54.6	78.4	-	-	54.3	72.4	-	-
	SSG [9]	58.3	80.0	90.0	92.4	53.4	73.0	80.6	83.2
	AD-Cluster [10]	68.3	86.7	94.4	96.5	54.1	72.6	82.5	85.5
	Co-teaching-500 [40]	71.7	87.8	95.0	96.5	61.7	77.6	88.0	90.7
	AE [41]	58.0	81.6	91.9	94.6	46.7	67.9	79.2	83.6
	MMT-500 [12] (ResNet50-IBN-a)	76.5	90.9	96.4	97.9	65.7	79.3	89.1	92.4
	Proposed AMMT-500 (ResNet50-IBN-a)	82.5	93.1	97.4	98.4	70.4	82.3	**91.1**	93.0
	Proposed AMMT-500 (ResNeSt-50)	**83.3**	**93.2**	**97.7**	**98.6**	**70.9**	**82.5**	90.9	**93.8**

TABLE 3. Comparison with state-of-the-art methods on MSMT17 datasets. mAP (%), Rank-1 (%), Rank-5 (%) and Rank-10 (%) are reported. The missing value is denoted as '–', which is not reported in original paper. Bold text represents the best result in the same evaluation standard.

Methods	Market1501 to MSMT17				DukeMTMC to MSMT17			
	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
PTGAN [15]	2.9	10.2	-	24.4	3.3	11.8	-	27.4
ENC [6]	8.5	25.3	36.3	42.1	10.2	30.2	41.5	46.8
MMCL [4]	15.1	40.8	51.8	56.7	16.2	43.6	54.3	58.9
ECN++ [35]	15.2	40.4	53.1	58.7	16.0	42.5	55.9	61.5
SSG [9]	13.2	31.6	-	49.6	13.3	32.2	-	51.2
AE [41]	9.2	25.5	37.3	42.6	11.7	32.3	44.4	50.1
MMT-500 [12] (ResNet50-IBN-a)	19.6	43.3	**56.1**	**61.6**	23.3	**50.0**	**62.8**	**68.4**
Proposed AMMT-500 (ResNet50-IBN-a)	23.7	43.4	55.7	61.5	28.1	49.0	61.6	66.9
Proposed AMMT-500 (ResNeSt-50)	**23.9**	**43.6**	**56.1**	61.4	**28.4**	49.4	62.1	67.4

importantly, we used no additional manual labeling data on the target domain, and did not use post-processing techniques such as re-ranking.

1) PERFORMANCE ON Market1501

The mAP and Rank-1 of the proposed AMMT method reached 83.3% and 93.2% when DukeMTMC transfers

TABLE 4. Comparison of ablation experimental results of asymmetric network model mutual mean-teaching on Market1501. ResNet50-IBN-a and ResNeSt-50 represent the backbone network used, Stage 1 means to conduct the first stage of asymmetric mutual mean teaching, Stage 2 means to conduct the second stage of asymmetric mutual mean teaching. The best results are bolded.

Methods		Model	DukeMTMC to Market1501			
			mAP	Rank-1	Rank-5	Rank-10
Symmetric Model	Direct Transfer	ResNet50-IBN-a	30.2	59.5	74.5	80.5
		ResNeSt-50	33.4	62.6	78.1	83.5
	Stage 1	ResNet50-IBN-a	74.6	89.4	95.9	97.6
		ResNeSt-50	77.2	91.3	96.0	97.2
	Stage 2	Symmetric Model Stage 1 Based	80.2	92.4	97.5	98.5
Asymmetric Model	Stage 1	ResNet50-IBN-a and ResNeSt-50	80.0	91.7	96.8	98.0
		ResNet50-IBN-a and ResNeSt-50	79.4	92.0	96.5	97.6
	Stage 2	Asymmetric Model Stage 1 Based	**81.9**	**92.9**	**97.8**	**98.6**

TABLE 5. Comparison of ablation experiment results of asymmetric network model mutual mean-teaching on DukeMTMC. ResNet50-IBN-a and ResNeSt-50 represent the backbone network used, Stage 1 means to conduct the first stage of asymmetric mutual mean teaching, Stage 2 means to conduct the second stage of asymmetric mutual mean teaching. The best results are bolded.

Methods		Model	Market1501 to DukeMTMC			
			mAP	Rank-1	Rank-5	Rank-10
Symmetric Model	Direct Transfer	ResNet50-IBN-a	31.0	49.2	65.5	70.9
		ResNeSt-50	32.8	52.3	66.5	71.2
	Stage 1	ResNet50-IBN-a	64.9	79.2	87.2	89.7
		ResNeSt-50	66.0	80.3	88.2	90.4
	Stage 2	Symmetric Model Stage 1 Based	68.8	81.7	90.3	93.0
Asymmetric Model	Stage 1	ResNet50-IBN-a and ResNeSt-50	66.8	79.6	88.9	91.2
		ResNet50-IBN-a and ResNeSt-50	67.9	81.0	88.9	91.2
	Stage 2	Asymmetric Model Stage 1 Based	**69.9**	**82.3**	**91.2**	**93.4**

to Market1501, which outperforms the state-of-the-art clustering-based MMT-500 by 6.8% and 2.3%, respectively. Compared with co-teaching-500, AMMT outperforms by 11.6% and 5.4%, respectively.

2) PERFORMANCE ON DukeMTMC

The mAP and Rank-1 of the proposed AMMT method reached 70.9% and 82.5% when Market1501 transfers to DukeMTMC, which outperforms MMT-500 by 5.2% and 3.2%, respectively. Compared with co-teaching-500, AMMT outperforms by 9.2% and 4.9%, respectively.

3) PERFORMANCE ON MSMT17

Table 3 shows the experimental results of Market1501 and DukeMTMC transfer to MSMT17, respectively. The mAP and Rank-1 of AMMT reached 23.9% and 43.6% when Market1501 transfers to MSMT17, which surpass MMT-500 by 4.3% and 0.3%. And for DukeMTMC transfers to MSMT17, AMMT reached 28.4% and 49.4%, which surpass MMT-500 by 5.1% in mAP, but performs poorly on Rank-1. The reason for this is that the MSMT17 dataset is large and the person ID number is 4101, when using k-means clustering, the number of clustering categories set is 500, while the number of clustering categories for DBSCAN is about 1000 each time, resulting in a large difference in the mutual teaching.

D. ABLATION STUDIES

To verify the effectiveness of the proposed asymmetric mutual mean-teaching of network model and the effectiveness of asymmetric mutual teaching of clustering, 11 sets of comparative experiments were performed on the DukeMTMC and Market1501 datasets.

1) EFFECTIVENESS OF ASYMMETRIC MUTUAL MEAN-TEACHING BETWEEN NETWORK MODEL

Tables 4 and 5 show the results of performance comparison experiments using the symmetric network model mutual mean-teaching and asymmetric network model mutual mean-teaching. "Symmetric Model" means using the same network models for deep mutual learning, and "Asymmetric Model" means using two different network models. From the tables, we can see the following.

a: ASYMMETRIC MUTUAL MEAN-TEACHING IN STAGE 1

On the Market1501 dataset, using ResNet50-IBN-a and ResNeSt-50 for mutual mean-teaching improved mAP by 5.4% and 2.2%, respectively, compared to the symmetric network model, and Rank-1 increased by 2.3% and 0.7% respectively. On the DukeMTMC dataset, mAP accuracy increased by 1.9% and 1.9%, respectively, and Rank-1 increased by 0.4% and 0.7%.

TABLE 6. The influence of different Ks on Market1501 and DukeMTMC datasets. K represents the number of clustering categories in the k-means method. ResNet50-IBN-a and ResNeSt-50 represent the backbone network used. The best results are bolded.

K	Methods	DukeMTMC to Market1501				Market1501 to DukeMTMC			
		mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
K=500	MMT-500 [12] (ResNet50-IBN-a)	76.5	90.9	96.4	97.9	65.7	79.3	89.1	92.4
	Proposed AMMT-500 (ResNet50-IBN-a)	82.5	93.1	97.4	98.4	70.4	82.3	91.1	93.0
	Proposed AMMT-500 (ResNeSt-50)	**83.3**	93.2	97.7	**98.6**	**70.9**	**82.5**	90.9	**93.8**
K=700	MMT-700 [12] (ResNet50-IBN-a)	74.5	91.1	96.5	98.2	68.7	81.8	**91.2**	93.4
	Proposed AMMT-700 (ResNet50-IBN-a)	82.5	**93.5**	**97.8**	**98.6**	70.8	81.8	91.0	93.1
	Proposed AMMT-700 (ResNeSt-50)	81.6	92.9	97.5	98.5	70.3	81.9	90.4	93.0
K=900	MMT-900 [12] (ResNet50-IBN-a)	72.7	91.2	96.3	98.0	67.3	80.8	90.3	93.0
	Proposed AMMT-900 (ResNet50-IBN-a)	80.6	**93.5**	97.6	98.5	69.2	81.6	90.2	92.8
	Proposed AMMT-900 (ResNeSt-50)	79.9	93.0	**97.8**	98.5	69.0	81.9	90.1	92.9

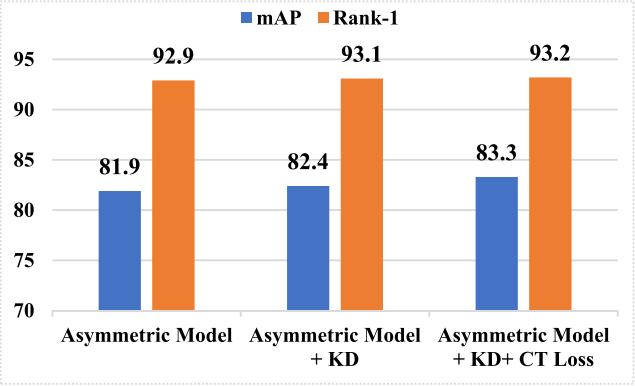

FIGURE 5. Comparison of ablation experiment results of asymmetric clustering method mutual teaching on Market1501.

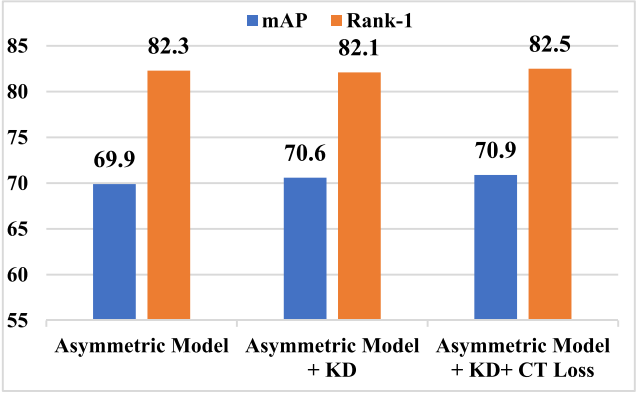

FIGURE 6. Comparison of ablation experiment results of asymmetric clustering method mutual teaching on DukeMTMC.

b: ASYMMETRIC MUTUAL MEAN-TEACHING IN STAGE 2

On the Market1501 dataset, using ResNet50-IBN-a and ResNeSt-50, mutual mean-teaching improved by 1.7% on mAP and 0.5% on Rank-1 compared to the symmetric network model. On the DukeMTMC dataset, mAP accuracy increased by 1.1%, and Rank-1 increased by 0.6%.

2) EFFECTIVENESS OF ASYMMETRIC MUTUAL MEAN-TEACHING IN STAGE 2

From Tables 4 and 5, we can see that when the symmetrical two network models were used, the second stage of mutual mean-teaching was used, and mAP and Rank-1 increased by 3.0%-5.6% and 1.1%-3.0%, respectively, on Market1501. On DukeMTMC, mAP and Rank-1 increased by 2.8%-3.9% and 1.4%-2.5%, respectively.

Using asymmetric two network models, the second stage of mutual mean-teaching was used, and mAP and Rank-1 increased by 1.5%-1.9% and 0.9%-1.2%, respectively, on Market1501. On DukeMTMC, mAP and Rank-1 increased by 2.0%-3.1% and 1.3%-2.7%, respectively.

3) EFFECTIVENESS OF ASYMMETRIC MUTUAL TEACHING IN CLUSTERING

To verify the effectiveness of using asymmetric clustering algorithms to learn from each other, a cross-domain ablation experiment was conducted on Market1501 and DukeMTMC. The experimental results are shown in Fig. 5 and Fig. 6, where Asymmetric Model represents a model that uses multiple network models for mutual mean-teaching. KD means k-means and DBSCAN clustering algorithms are used for

FIGURE 7. Visualization experiment results of proposed AMMT: (a) Market1501; (b) DukeMTMC.

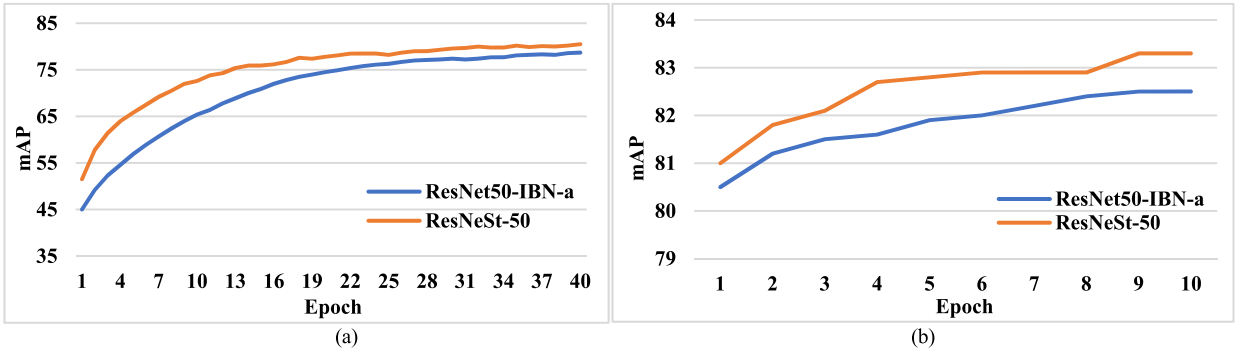

FIGURE 8. Evaluation with different epochs on Market1501: (a) asymmetric mutual mean-teaching in stage 1; (b) asymmetric mutual mean-teaching in stage 2.

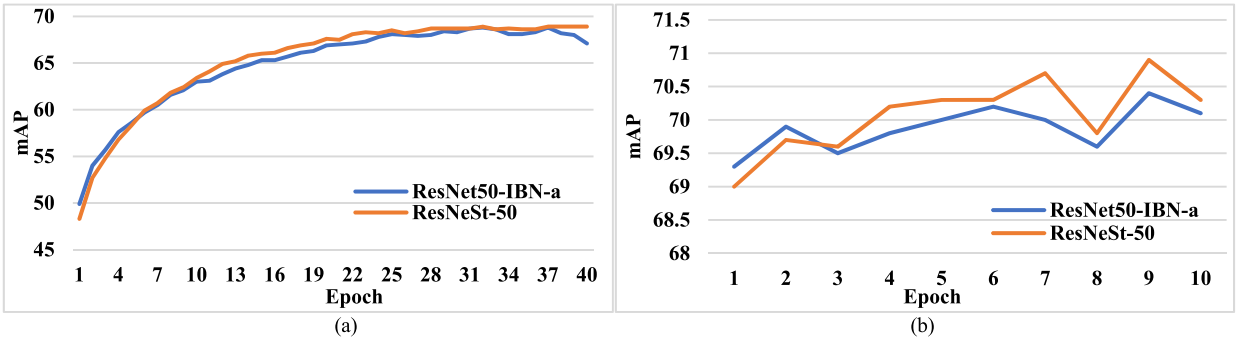

FIGURE 9. Evaluation with different epochs on DukeMTMC: (a) asymmetric mutual mean-teaching in stage 1; (b) asymmetric mutual mean-teaching in stage 2.

mutual teaching. CT Loss is our proposed centroid triplet loss function. As can be seen, using asymmetric clustering method for mutual teaching, mAP and Rank-1 increased by 0.5% and 0.2%, respectively, on Market1501, compared to a single clustering algorithm, and mAP increased by 0.7% on DukeMTMC. Using CT Loss increased mAP and Rank-1 by 0.9% and 0.1%, respectively, on Market1501, and increased mAP and Rank-1 by 0.3% and 0.4%, respectively, on DukeMTMC. These six sets of comparative experiments demonstrated that the use of multi-clustering algorithms in mutual teaching can effectively improve the confidence of clustering pseudo-labels.

4) THE INFLUENCE OF THE NUMBER OF K-MEANS CLUSTERING CATEGORIES (K)

Since the AMMT is a cluster-based UDA method, the number of k-means clustering categories (K) is of importance to the accuracy. The experimental results are shown in Table 6, compared with MMT, our proposed AMMT performs well at different k values.

Both the ResNet50-IBN-a model and the ResNeSt-50 model used by AMMT in the inference stage have higher accuracy rates than the MMT with ResNet50-IBN-a as the backbone model. And the accuracy of the ResNet50-IBN-a model and the ResNeSt-50 model is not much different,

indicating that AMMT can promote the common progress of the two models and improve each other. When K takes 500 and 700, the accuracy is higher than K takes 900 on Market1501 and DukeMTMC datasets.

E. VISUALIZATION OF RESULTS AND DISCUSSION

Fig. 7 shows the visualization results of our proposed AMMT on Market1501 and DukeMTMC. It can be seen from the figure that the results of Rank-1 to Rank-5 can be accurately identified. Some of the reasons for incorrect identification may be similar clothing, and failing to focus on distinguishing local features. Fig. 8 and Fig. 9, respectively, show the changes of mAP during the training process of AMMT on Marke1501 and DukeMTMC. From this we can see that in the first stage of the asymmetric mutual mean-teaching, the network has basically converged at the 40th epoch, and has even produced overfitting, but after the second stage of asymmetric mutual mean-teaching, the best performance of the network model is reached around the 10th epoch.

V. CONCLUSION

We proposed an asymmetric mutual mean-teaching method to solve the unsupervised domain adaptive person re-identification task. In terms of asymmetric network models, two asymmetric network models with different structures were used for mutual mean-teaching to enhance the generalization ability of the network's diversity. In terms of asymmetric clustering, two different clustering algorithms were used for mutual teaching to enhance the confidence of clustering pseudo-labels and reduce the noise of pseudo-labels. At the same time, the triple loss was improved and changed to centroid triple loss to adapt to the high-confidence pseudo-labels, only one model with good performance was used in the reasoning test, and complex model integrated reasoning was not required. A large number of mutual unsupervised cross-domain ablation experiments were conducted with AMMT on three datasets, which demonstrated the effectiveness of AMMT. Compared to the most advanced methods, it can obtain a higher mAP accuracy and Rank-1 hit rate. In future research, we will use the AMMT idea to further explore the field of occluded re-ID.

REFERENCES

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

[2] Y.-J. Li, C.-S. Lin, Y.-B. Lin, and Y.-C.-F. Wang, "Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7919–7929.

[3] X. Jin, C. Lan, W. Zeng, Z. Chen, and L. Zhang, "Style normalization and restitution for generalizable person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3143–3152.

[4] D. Wang and S. Zhang, "Unsupervised person re-identification via multi-label classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10981–10990.

[5] F. Yang, K. Yan, S. Lu, H. Jia, D. Xie, Z. Yu, X. Guo, F. Huang, and W. Gao, "Part-aware progressive unsupervised domain adaptation for person re-identification," *IEEE Trans. Multimedia*, early access, Jun. 12, 2020, doi: 10.1109/TMM.2020.3001522.

[6] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 598–607.

[7] 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/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2148–2157.

[8] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang, "A bottom-up clustering approach to unsupervised person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 8738–8745.

[9] Y. Fu, Y. Wei, G. Wang, Y. Zhou, H. Shi, U. Uiuc, and T. Huang, "Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6112–6121.

[10] Y. Zhai, S. Lu, Q. Ye, X. Shan, J. Chen, R. Ji, and Y. Tian, "AD-cluster: Augmented discriminative clustering for domain adaptive person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9021–9030.

[11] X. Jin, C. Lan, W. Zeng, and Z. Chen, "Global distance-distributions separation for unsupervised person re-identification," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 735–751.

[12] Y. Ge, D. Chen, and H. Li, "Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification," 2020, *arXiv:2001.01526*. [Online]. Available: http://arxiv.org/abs/2001.01526

[13] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.

[14] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 17–35.

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

[16] 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, pp. 1–18, Nov. 2018.

[17] X. Zhang, J. Cao, C. Shen, and M. You, "Self-training with progressive augmentation for unsupervised cross-domain person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8222–8231.

[18] F. Yang, K. Li, Z. Zhong, Z. Luo, X. Sun, H. Cheng, X. Guo, F. Huang, R. Ji, and S. Li, "Asymmetric co-teaching for unsupervised cross-domain person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 12597–12604.

[19] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*. [Online]. Available: http://arxiv.org/abs/1503.02531

[20] T. Chen, I. Goodfellow, and J. Shlens, "Net2Net: Accelerating learning via knowledge transfer," 2015, *arXiv:1511.05641*. [Online]. Available: http://arxiv.org/abs/1511.05641

[21] Y. Li, J. Yang, Y. Song, L. Cao, J. Luo, and L.-J. Li, "Learning from noisy labels with distillation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1910–1918.

[22] 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.

[23] H. Bagherinezhad, M. Horton, M. Rastegari, and A. Farhadi, "Label refinery: Improving ImageNet classification through label progression," 2018, *arXiv:1805.02641*. [Online]. Available: http://arxiv.org/abs/1805.02641

[24] R. Anil, G. Pereyra, A. Passos, R. Ormandi, G. E. Dahl, and G. E. Hinton, "Large scale distributed neural network training through online distillation," 2018, *arXiv:1804.03235*. [Online]. Available: http://arxiv.org/abs/1804.03235

[25] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," 2017, *arXiv:1703.01780*. [Online]. Available: http://arxiv.org/abs/1703.01780

[26] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4320–4328.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[28] X. Pan, P. Luo, J. Shi, and X. Tang, "Two at once: Enhancing learning and generalization capacities via IBN-Net," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 464–479.
[29] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, M. Li, and A. Smola, "ResNeSt: Split-attention networks," 2020, *arXiv:2004.08955*. [Online]. Available: http://arxiv.org/abs/2004.08955
[30] K. Krishna and M. N. Murty, "Genetic K-means algorithm," *IEEE Trans. Syst., Man, Cybern. B. Cybern.*, vol. 29, no. 3, pp. 433–439, Jun. 1999.
[31] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. KDD*, 1996, vol. 96, no. 34, pp. 226–231.
[32] S. Lin, H. Li, C.-T. Li, and A. C. Kot, "Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification," 2018, *arXiv:1807.01440*. [Online]. Available: http://arxiv.org/abs/1807.01440
[33] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2275–2284.
[34] L. Qi, L. Wang, J. Huo, L. Zhou, Y. Shi, and Y. Gao, "A novel unsupervised camera-aware domain adaptation framework for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8080–8089.
[35] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Learning to adapt invariance in memory for person re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Feb. 28, 2020, doi: 10.1109/TPAMI.2020.2976933.
[36] Z. Zhong, L. Zheng, S. Li, and Y. Yang, "Generalizing a person retrieval model hetero- and homogeneously," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 172–188.
[37] J. Liu, Z.-J. Zha, D. Chen, R. Hong, and M. Wang, "Adaptive transfer network for cross-domain person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7202–7211.
[38] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "CamStyle: A novel data augmentation method for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1176–1190, Mar. 2019.
[39] L. Song, C. Wang, L. Zhang, B. Du, Q. Zhang, C. Huang, and X. Wang, "Unsupervised domain adaptive re-identification: Theory and practice," *Pattern Recognit.*, vol. 102, Jun. 2020, Art. no. 107173.
[40] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," 2018, *arXiv:1804.06872*. [Online]. Available: http://arxiv.org/abs/1804.06872
[41] Y. Ding, H. Fan, M. Xu, and Y. Yang, "Adaptive exploration for unsupervised person re-identification," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 1, pp. 1–19, Apr. 2020.
[42] B. Zoph, G. Ghiasi, T.-Y. Lin, Y. Cui, H. Liu, E. D. Cubuk, and Q. V. Le, "Rethinking pre-training and self-training," 2020, *arXiv:2006.06882*. [Online]. Available: http://arxiv.org/abs/2006.06882
[43] X. Dong, L. Zheng, F. Ma, Y. Yang, and D. Meng, "Few-example object detection with model communication," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1641–1654, Jul. 2019.
[44] X. Dong and Y. Yang, "Teacher supervises students how to learn from partially labeled images for facial landmark detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 783–792.
[45] L. Liu, T. Zhou, G. Long, J. Jiang, X. Dong, and C. Zhang, "Isometric propagation network for generalized zero-shot learning," 2021, *arXiv:2102.02038*. [Online]. Available: http://arxiv.org/abs/2102.02038
[46] M. Ye, J. Li, A. J. Ma, L. Zheng, and P. C. Yuen, "Dynamic graph co-matching for unsupervised video-based person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2976–2990, Jun. 2019.
[47] M. Ye, J. Shen, X. Zhang, P. C. Yuen, and S.-F. Chang, "Augmentation invariant and instance spreading feature for softmax embedding," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Aug. 3, 2020, doi: 10.1109/TPAMI.2020.3013379.
[48] M. Ye, X. Lan, and P. C. Yuen, "Robust anchor embedding for unsupervised video person re-identification in the wild," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 170–186.
[49] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jan. 26, 2021, doi: 10.1109/TPAMI.2021.3054775.
[50] X. Wang, S. Wang, H. Shi, J. Wang, and T. Mei, "Co-mining: Deep face recognition with noisy labels," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9358–9367.
[51] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Exploit the unknown gradually: One-shot video-based person re-identification by step-wise learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5177–5186.
[52] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Bian, and Y. Yang, "Progressive learning for person re-identification with one example," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2872–2881, Jun. 2019.
[53] Y. Lin, Y. Wu, C. Yan, M. Xu, and Y. Yang, "Unsupervised person re-identification via cross-camera similarity exploration," *IEEE Trans. Image Process.*, vol. 29, pp. 5481–5490, 2020.

YACHAO DONG received the B.E. degree from the Henan University of Technology, Henan, China, in 2018. He is currently pursuing the master's degree with the Beijing Key Laboratory of Information Service Engineering, Beijing Union University, Beijing. His research interests include artificial intelligence and deep learning.

HONGZHE LIU received the B.E. degree from Chinese Marine University, China, in 1995, the M.A.Sc. degree from California State University at Long Beach, Long Beach, CA, USA, in 2000, and the Ph.D. degree from Beijing Jiaotong University, Beijing, China. She is currently the Leader of the Beijing Key Laboratory of Information Service Engineering and a Professor with Beijing Union University. Her research interests include artificial intelligence, visual intelligence, cognitive computing, and visual computing. She is a member of the Chinese Computer Society and the Vice Secretary General of the Network Application Branch of China Computer User Association. She serves as many domestics and foreign periodicals reviewer. She is also a Reviewer of the National Natural Fund.

CHENG XU (Member, IEEE) received the B.E. and M.A.Sc. degrees from the Beijing Key Laboratory of Information Service Engineering, Beijing Union University, China, in 2012 and 2015, respectively, and the Ph.D. degree from the Beijing University of Posts and Telecommunications (BUPT), China. He is currently a Lecturer with Beijing Union University. His research interests include the Internet of Vehicles, intelligent driving, data intelligent, and data security.