Unsupervised Domain Adaptation on Person Reidentification via Dual-Level Asymmetric Mutual Learning

Qiong Wu, Jiahan Li, Pingyang Dai, Qixiang Ye, *Senior Member, IEEE*, Liujuan Cao, *Member, IEEE*, Yongjian Wu, and Rongrong Ji, *Senior Member, IEEE*

***Abstract*— Unsupervised domain adaptation (UDA) person reidentification (Re-ID) aims to identify pedestrian images within an unlabeled target domain with an auxiliary labeled source-domain dataset. Many existing works attempt to recover reliable identity information by considering multiple homogeneous networks. And take these generated labels to train the model in the target domain. However, these homogeneous networks identify people in approximate subspaces and equally exchange their knowledge with others or their mean net to improve their ability, inevitably limiting the scope of available knowledge and putting them into the same mistake. This article proposes a dual-level asymmetric mutual learning (DAML) method to learn discriminative representations from a broader knowledge scope with diverse embedding spaces. Specifically, two heterogeneous networks mutually learn knowledge from asymmetric subspaces through the pseudo label generation in a hard distillation manner. The knowledge transfer between two networks is based on an asymmetric mutual learning (AML) manner. The teacher network learns to identify both the target and source domain while adapting to the target domain distribution based on the knowledge of the student. Meanwhile, the student network is trained on the target dataset and employs the ground-truth label through the knowledge of the teacher. Extensive experiments in Market-1501, CUHK-SYSU, and MSMT17 public datasets verified the superiority of DAML over state-of-the-arts (SOTA).**

***Index Terms*— Person reidentification (Re-ID), retrieval, transfer learning, unsupervised domain adaptation (UDA).**

Manuscript received 28 December 2022; revised 4 August 2023; accepted 15 October 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2022ZD0118202; in part by the National Science Fund for Distinguished Young Scholars under Grant 62025603; in part by the National Natural Science Foundation of China under Grant U21B2037, Grant U22B2051, Grant 62176222, Grant 62176223, Grant 62176226, Grant 62072386, Grant 62072387, Grant 62072389, Grant 62002305, and Grant 62272401; and in part by the Natural Science Foundation of Fujian Province of China under Grant 2021J01002 and Grant 2022J06001. *(Corresponding author: Pingyang Dai.)*

Qiong Wu is with the Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, and the Institute of Artificial Intelligence, Xiamen University, Xiamen 361005, China (e-mail: qiong@stu.xmu.edu.cn).

Jiahan Li is with the Faculty of Computing, Harbin Institute of Technology, Harbin 150009, China (e-mail: jiahan.li@stu.hit.edu.cn).

Pingyang Dai and Liujuan Cao are with the Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, Xiamen 361005, China, and also with the School of Informatics, Xiamen University, Xiamen 361005, China (e-mail: pydai@xmu.edu.cn; caoliujuan@xmu.edu.cn).

Qixiang Ye is with the Peng Cheng Laboratory, Shenzhen 518066, China, and also with the School of Electronics, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: qxye@ucas.ac.cn).

Yongjian Wu is with the Youtu Laboratory, Tencent, Shanghai 200233, China (e-mail: littlekenwu@tencent.com).

Rongrong Ji is with the Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, and the Fujian Engineering Research Center of Trusted Artificial Intelligence Analysis and Application, Institute of Artificial Intelligence, Xiamen University, Xiamen 361005, China, and also with the Peng Cheng Laboratory, Shenzhen 518066, China (e-mail: rrji@xmu.edu.cn).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TNNLS.2023.3326477.

Digital Object Identifier 10.1109/TNNLS.2023.3326477

I. Introduction

PERSON reidentification (Re-ID) [1] aims at matching individual pedestrian images from images captured by different cameras according to identity. This task is challenging because the variations of viewpoints, body poses, illuminations, and backgrounds will influence a person's appearance. Recently, supervised person Re-ID methods [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13] have made impressive progress. However, as the number of images increases and the ensuing scene changes, regular supervised learning approaches are losing their ability to adapt to complex scenarios. The performance of person Re-ID models trained on existing datasets will evidently suffer for person images from a new video surveillance system due to the domain gap. To avoid time-consuming annotations on the new dataset, unsupervised domain adaptation (UDA) is proposed to adapt the model trained on the labeled source-domain dataset to the unlabeled target-domain dataset.

Generating trusted identity information on the target domain is seen as the core of the UDA task. Some UDA Re-ID methods [10], [14], [15], [16], [17] directly apply generative adversarial networks (GANs) [18] to transfer the style of pedestrian images from the source domain to the target while keeping the identities to train the model. However, the complexity of the human form and the limited number of instance in a Re-ID dataset limit the quality of generated images. After abandoning the image generation, some methods [7], [19] introduce the attribute to bridge the domain gap. These methods introduce additional annotation information which defeats the purpose of the UDA Re-ID task. Limited by the missing label on the target domain, others [20], [21], [22], [23] align the distributions of target and source domains while only

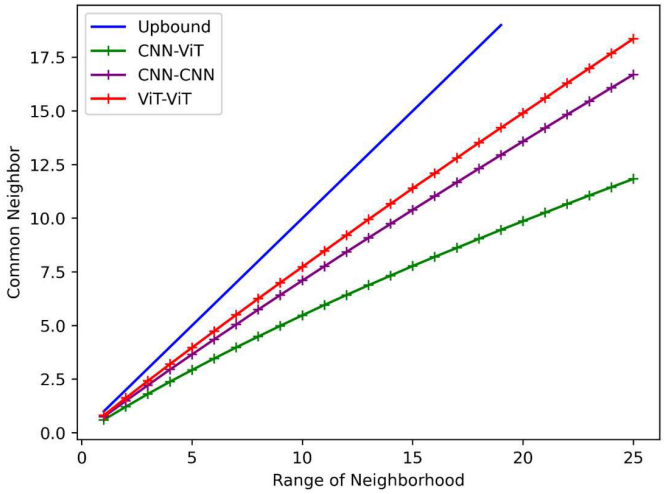


Fig. 1. The statistics of common neighbors between different models. CNN-CNN and ViT-ViT curves denote the average number of common neighbors in the k nearest neighbors of each instance. The features are extracted by two homogeneous networks trained with different initialization. Similarly, CNN-ViT represents the common neighbors between CNN and ViT. Furthermore, upbound refers to the maximum number of neighbors to consider. These models are trained on the CUHK-SYSU dataset in a supervised manner and cluster on the Market1501 dataset. Compared to the CNN-CNN and ViT-ViT, the CNN-ViT contains fewer common neighbors, and the ways they distinguish two individuals are more different than homogeneous networks. It demonstrates that heterogeneous networks address the task in different patterns.

learning classifying on the source. To better adapt the distribution of the target domain and train with the target-domain identity knowledge, various methods [24], [25], [26] apply a clustering algorithm in the target domain to generate the pseudo labels for training in a supervised manner. One of the keys to improving performance is alleviating the influence of noisy labels. In this context, many methods [27], [28], [29] based on clustering algorithms are proposed to rule out the harmony from the noisy labels by introducing more than one framework to predict pseudo labels. They aim to generate knowledge with specific differences in samples and exchange the knowledge among the networks to enhance their ability. Despite encouraging progress, the benefits from the knowledge mined by homogeneous networks are limited.

As shown in Fig. 1, convolutional neural network (CNN)-CNN and vision transformer (ViT)-ViT, these homogeneous networks with similar structures identify pedestrians in a comparable manner, and the relation among the instances are similar. It suggests they use similar patterns to extract pedestrian features, and networks may converge to equal each other. Furthermore, this mode of operation makes it possible for the networks to make the same mistakes and not be able to correct them. Such a design limits the knowledge models can learn from the training set and makes it possible for the networks to repeat mistakes without being able to remedy them. As a result, mining the information from different subspaces is required to broaden the scope of knowledge and generate reliable pseudo labels.

Fig. 1 CNN-ViT illustrates significant variations in the prediction outcomes of heterogeneous networks. This observation motivates us to leverage these differences to extract valuable knowledge from the target domain. We propose dual-level asymmetric mutual learning (DAML), a novel UDA method for person Re-ID that broadens the scope of knowledge for the network by exploiting information from two different subspaces and selectively transferring information between heterogeneous networks. Rather than treating and transferring knowledge equally among all homogeneous networks, DAML employs a selective approach to information transfer between the heterogeneous networks in order to adapt to the target domain.

Specifically, the proposed DAML consists of a CNN as a teacher network and a ViT as a student network. With their heterogeneous constructions, we can mine the knowledge from the target domain in a mutual learning manner. Due to the goals that differ between teacher and student networks, the knowledge transferring between them can be asymmetric. In particular, the CNN that works as a teacher will train on both source and target datasets under the supervision of ground-truth source-domain labels and pseudotarget-domain labels. The former can provide reliable identity information for extracting discriminative feature representation, while the latter will assist the network in adapting the distribution of the target domain. In the pseudo label generation stage, the relationship between two samples is weighted according to their teacher and student features similarity. Moreover, this process wholly exchanges the knowledge learned from two different subspaces. After predicting the identities of input images, the asymmetric constraints between two heterogeneous networks selectively exchange knowledge. The student learns the identity knowledge from the teacher network under the constraints of the target-domain samples. Besides, for the student to learn more from the teacher and better use the ground-truth labels, we transfer the source-domain identity knowledge learned by the teacher to the target domain. In summary, the DAML employs diverse subspaces to generate reliable pseudo label in the target domain and help student adopt ground-truth knowledge in the source domain.

Our main contributions are summarized below.

1) We leverage the differences in prediction patterns to improve the performance of UDA for person Re-ID through mutual learning. Our key insight is that heterogeneous networks inherently exhibit more substantial discrepancies compared to homogeneous networks.
2) We propose a novel DAML method for UDA person Re-ID. The DAML introduces heterogeneous networks to mine knowledge in a broader scope and selectively transfer knowledge for better adaptation.
3) To learn from diverse subspaces, the proposed DAML introduces two heterogeneous networks to mine valuable information from different subspaces and selectively exchange the information between them.
4) To better utilize the knowledge mined by heterogeneous networks and ensure the networks orient to the task, the proposed DAML smoothly update the classifiers in a hard distillation manner and exchange knowledge during training in a soft distillation manner.

II. Related Work

A. Unsupervised Domain Adaptation Person Re-ID

UDA Person Re-ID has attracted increasing attention in recent years due to its effectiveness in reducing manual annotation costs. There are two main categories of methods are proposed to address this issue. First, GAN-based methods aim to transfer samples from the source domain to the target domain without altering their identities. SPGAN [14] and PDA-Net [15] transfer images directly from the source domain to the target domain while maintaining the original identity knowledge. The generated images have a similar style to the target-domain images and are used to train the model under the supervision of their original labels in the source domain. To produce generated images that are more realistic and have more detail, DG-Net [30] and DG-Net++ [17] introduce disentanglement for the generation stage. But, the generation is expensive and the style of generated images may not well fit the target domain.

Rather than transfer images from the source domain to the target domain, HHL [10] transfers target-domain images among the cameras to generate images that have the same identity but contain the difference at the same time. Instead of generation in image level, AEO [31] proposes a novel domain adaptation scheme to learn domain-invariant features in the feature level by an adversarial training manner. Similarly, GSL [32] learns an invariant, discriminative, and domain-agnostic subspace for UDA by a two-stage progressive training strategy. To mine the knowledge in the target domain, the clustering-based methods do not require expensive GAN networks for generation and have achieved state-of-the-art (SOTA) performance to date. PCDA [33] proposes a pseudo labeling curriculum based on a density-based clustering algorithm to mine the knowledge in the target domain. CAT [34] effectively incorporates the discriminative clustering structures in both source and target domains for better adaptation. To further reduce the impact of noisy labels, MMT [27] proposed a mutual learning method providing soft labels. To enhance the complementarity and further depress noise in the pseudo labels, AWS [35] proposed a novel lightweight module that can be integrated into the dual networks of mutual learning. Besides the feature extraction, MMT also can be applied in other field such as translation [36]. For more reliable pseudo labels, SSG [37] clusters samples in three scales and validate each other. MEB-Net [29], respectively, introduces multiple groups of prototypes or homogeneous networks to generate the pseudo labels. Furthermore, MFAG [38] adaptively fuse multiple features for person Re-ID by a graph-based method. UNRN [39] and GLT [28] design a memory bank to save anchors for aligning the distribution and learning identities in a contrastive learning manner. For fully exploring intradomain information and obtaining informative negative samples in the memory bank, MCRN [40] adaptively builds a multicentroid memory. In addition, instead of gaining anchors from samples, TPN [41] trains prototypes and matches them with samples in two domains. Limited by the constraints in the feature level these methods rely on, the models that collaborate to generate pseudo labels are homogeneous. These characteristics determine that the model can only learn similar knowledge from others. Nevertheless, these approaches alleviate the domain gap only considering the single embedding space inevitably makes some mistakes.

B. Transfer Learning

Knowledge distillation makes a student network learns from a strong teacher network to improve the student's ability. The common approaches can be summarized as hard distillation and soft distillation. Soft distillation [42], [43] minimizes the distribution difference between the prediction generated by teacher and student. The soft label generated by the teacher model can alleviate overfitting just like labels smoothing [44]. Unlike the soft, hard-label distillation regards the prediction result of the teacher as a valid label. And positive pairs predicted by the teacher are used to transfer identity knowledge from the teacher to a student network in semisupervised and unsupervised learning tasks. Temporal ensembling [45] put the former networks as the teacher and use memory-saving average predictions for each sample as supervision for the unlabeled samples. To avoid storing predictions for saving memory, Mean Teacher [46] averaged student model weights as the parameter of the teacher. In order to deal with more diverse environments and improve the scalability of the model, some zero-shot transfer learning methods are also proposed for a variety of tasks. TN-ZSTAD [47] utilizes transfer learning to gain knowledge from contrastive language-image pretraining (CLIP) [48] for zero-shot temporal activity detection. To adaptively design the most suitable construction for zero-shot learning, ZeroNAS [49] searches the architectures of the generator and discriminator in a min-max player game via adversarial training. During the training, the predictions made by the teacher are seen as supervision for unlabeled samples. The models consider similar information in these methods because the teacher and the student have the same structure and similar initialization. It makes the networks focus within a certain range and limits the knowledge student can learn. The proposed DAML exchange knowledge utilizes both soft and hard distillation in the different training stages. Thanks to the heterogeneous networks, the proposed DAML gives the student model a broader perspective and can generate pseudo labels from different views.

C. CNN and ViT

Since AlexNet [50] achieve great success on ImageNet [51], a variety of CNNs [52], [53], [54], [55] is proposed to solve different tasks. As transformers [56] were proposed for machine translation and were seen with significant results in many natural language processing (NLP) tasks, the application of self-attention to images is widely concerned. A new model without any convolution, ViTs [57], has been proposed for computer vision tasks and shows its potential. During the calculation process, the CNN keeps the spatial information and can only focus on the surrounding area in one layer due to the nature of convolution. In contrast, ViT emphasizes the correlation between two patches, and its receptive field involves the whole feature map. These differences make the CNN and ViT learn different knowledge from the training set

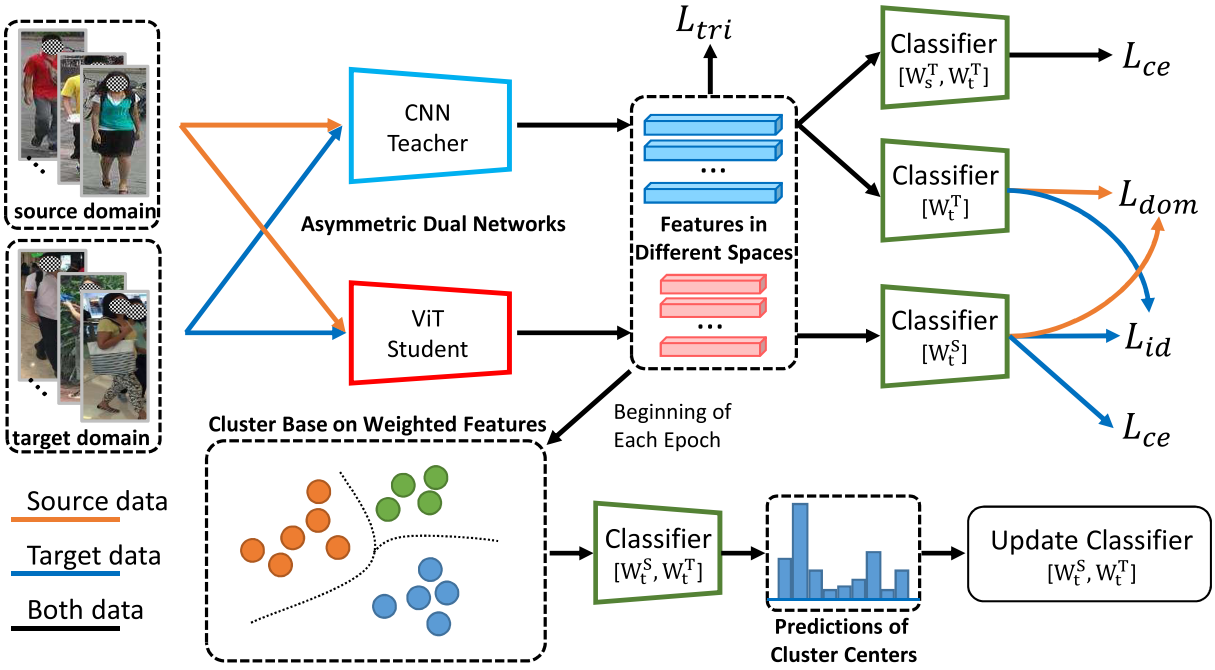


Fig. 2. Overview of our DAML. The teacher network is trained under the supervision of pseudo labels and ground-truth labels for target-domain and source-domain samples. And the student only directly learns knowledge from target-domain samples with pseudo labels. At the beginning of epochs, we first generate the pseudo labels for target dataset, and update the classifiers based on the predictions of cluster centers. To distill the different subspace knowledge from the teacher to the student, $\mathcal{L}_{\text{id}}$ makes the student predictions of target-domain samples close to the teacher. Meanwhile, for student can better adopt the identity knowledge learned by the teacher, we minimize the distribution differences of the same source-domain samples with $\mathcal{L}_{\text{dom}}$.

for the same task. And in this article, we take advantage of this difference to achieve asymmetric distillation, making ViT a better performer with our DAML. The ViT works as a student because the receptive field of a patch in the ViT covers the area that one convolution kernel can consider, not vice-versa.

III. Methodology

A. Overview

The ultimate goal of the UDA person Re-ID is to gain a model work on a target-domain dataset based on a labeled source-domain dataset and an unlabeled target-domain dataset. Let $\mathcal{S} = \{(\mathbf{x}_s^i, \mathbf{y}_s^i)\}_{i=1}^{N_s}$ and $\mathcal{T} = \{\mathbf{x}_t^i\}_{i=1}^{N_t}$, respectively, denote the source-domain images with ground-truth labels and the unlabeled target-domain images, where N_s and N_t are the numbers of samples from these two domains.

As shown in Fig. 2, the DAML method trains the student to extract discriminative representations from two different subspaces to perform the UDA person Re-ID task. First, DAML adopts two heterogeneous networks: teacher CNN $\mathrm{E}_T(\cdot)$ and student ViT $\mathrm{E}_S(\cdot)$ which are pretrained on the source-domain dataset in a supervised manner to extract features in different subspaces. At each epoch, we first group target-domain samples into K classes by the clustering algorithm. The distance between two target-domain samples will be calculated according to the features $\mathrm{E}_T(\mathbf{x}_t^i) = \mathbf{t}_i^T \in \mathbb{R}^{c_T}$ and $\mathrm{E}_S(\mathbf{x}_t^i) = \mathbf{t}_i^S \in \mathbb{R}^{c_S}$ extracted by the teacher and student models with corresponding weights. The $\{\hat{\mathbf{y}}_i\}_{i=1}^{N_t}$ are the pseudo labels for the target-domain samples. Then, for each class center $\mathbf{c}_y$, we generate its prediction with the classifiers $\mathrm{C}(\cdot|\mathbf{W}_t^S)$ and $\mathrm{C}(\cdot|\mathbf{W}_t^T)$ for updating the parameter $\mathbf{W}_t^S$ and $\mathbf{W}_t^T$ in a smooth method.

After that, we train the teacher and the student models with the pseudo labels in a supervised manner. For the teacher model, classifier $\mathrm{C}(\cdot|[\mathbf{W}_s^T, \mathbf{W}_t^T])$ will learn knowledge from both the source domain and the target domain. The classifier $\mathrm{C}(\cdot|\mathbf{W}_t^S)$ for the student model only directly learns the target-domain knowledge. The constraints between two networks transfer the identity knowledge learned by the teacher to the target and help the student learn from diverse subspaces. Finally, we only adopt the features $\mathbf{t}_i^S = \mathrm{E}_S(\mathbf{x}_t^i)$ extracted by the student model for testing.

B. Preliminary

We first revisit the principle of soft knowledge distillation for UDA person Re-ID. This can be expressed as minimizing the discrepancy between the predicted distributions of the student model ($P(\mathbf{x}_t|\theta_S)$) and the teacher model ($P(\mathbf{x}_t|\theta_T)$)

$$\underset{\theta_S}{\operatorname{argmin}} \ \mathcal{D}(P(\mathbf{x}_t|\theta_S)|P(\mathbf{x}_t|\theta_T)). \tag{1}$$

Here, P represents the predicted distributions based on the parameters of the student (θ_S) and teacher (θ_T) models. The distance between the two distributions is measured by the function $\mathcal{D}$. The samples from the target domain are denoted as $\mathbf{x}_t$. By minimizing this discrepancy, the student model can learn the knowledge acquired by the teacher, including both identity and domain-specific knowledge.

The discrepancy between the identification patterns of the student and teacher models, despite having the same pretrained task, indicates the amount of latent knowledge that can be transferred. Therefore, as shown in Fig. 1, leveraging heterogeneous networks in the UDA person Re-ID task provides the student model with a broader range of knowledge.

However, the latent knowledge encompasses both identity and domain information, with the latter negatively impacting performance in the target domain. To alleviate the influence of domain, we transfer knowledge learned by the teacher model to the target domain, and the objective of the proposed *asymmetric mutual learning* (AML) can be defined by

$$\begin{aligned}\underset{\theta_S,\theta_T}{\operatorname{argmin}}\ & \mathcal{D}(P_{\mathrm{Id}}(\mathbf{x}_t|\theta_S)|P_{\mathrm{Id}}(\mathbf{x}_t|\theta_T)) \\ & +\mathcal{D}(P_{\mathrm{Dom}}(\mathbf{x}_s|\theta_T)|P_{\mathrm{Dom}}(\mathbf{x}_s|\theta_S)) \end{aligned} \tag{2}$$

where P_{Id} and P_{Dom} refer to the identity and domain ingredient in the distribution of prediction. And $\mathbf{x}_s$ represents the samples from the source domain. Via learning domain knowledge from the student model, the teacher model can provide identity knowledge that is more beneficial to the student model.

C. Smooth Classifier Update

At the beginning of epochs, we extract the target-domain features $\mathbf{t}_i^T = E_T(\mathbf{x}_t^i)$ and $\mathbf{t}_i^S = E_S(\mathbf{x}_t^i)$ with two heterogeneous networks. To better utilize the knowledge from the two models, we first define the neighborhood of an instance according to its relations in two different subspaces

$$N_i = \left\{ x_j \middle| 1 - \frac{\langle t_i^M, t_j^M \rangle}{\|t_i^M\|_2 \|t_j^M\|_2} < \alpha, M \in \{T, S\} \right\} \tag{3}$$

α here is a hyperparameter. With the limitation of neighbor selection considering both teacher features and student features simultaneously, the neighbors of an instance should be close to it in both subspaces. The above constraint ensures that instances with apparent differences will not be clustered as the same identity because the patterns of the two models applied to recognize an instance are different.

To exploit the information from two different subspaces and make the pseudo labels more reliable, we combine features from heterogeneous networks and define the distance between two samples as

$$d_{i,j} = \begin{cases} 1 - \frac{\langle [\mathbf{t}_i^T, \mathbf{t}_i^S], [\mathbf{t}_j^T, \mathbf{t}_j^S] \rangle}{\|[\mathbf{t}_i^T, \mathbf{t}_i^S]\|_2 \|[\mathbf{t}_j^T, \mathbf{t}_j^S]\|_2}, & x_i \in N_j \text{ and } x_j \in N_i \\ \text{Inf}, & \text{Others} \end{cases} \tag{4}$$

where $[\cdot, \cdot]$ represents the concatenation of two features, and $\langle \cdot, \cdot \rangle$ denotes the dot product between two features. In short, we define the similarity between two samples as the cos similarity between the features constructed by concatenating their teacher feature and student feature. Then, the pseudo labels can be generated based on the relationship among instances with the clustering algorithm.

With the pseudo labels, some methods [27], [39] directly update the classifier by replacing the classifier parameters with the new class centers to adapt the count of classes change. These methods will make the knowledge lost because the class centers may not represent the corresponding class well. To protect the knowledge involved in the classifiers, we update the classifiers more smoothly as follows:

$$\mathbf{W}_t^i = \sum_{k=1}^{\hat{K}} \hat{\mathbf{W}}_t^k \cdot \frac{e^{\mathbf{p}_i^k}}{\sum_{j=1}^{\hat{K}} e^{\mathbf{p}_i^j}} \tag{5}$$

where $\mathbf{W}_t^i$ is the parameters for the ith target-domain identity in the next epoch and $\mathbf{p}_i = \mathrm{C}(\mathbf{c}_i|\hat{\mathbf{W}}_t)$ is the prediction of class center $\mathbf{c}_i$ with the parameters $\hat{\mathbf{W}}_t$ from the last epoch which includes $\hat{K}$ classes. And $\mathbf{W}_t$ denotes classifier parameters in both teacher and student models. Note that the momentum for stochastic gradient descent (SGD) is updated following the parameters in the process.

D. Identity Learning

The core of person Re-ID is identifying the persons. For two heterogeneous networks learning to extract discriminative representation, there are two level objective functions are applied. First, at the feature level, the triplet loss

$$\mathcal{L}_{\mathrm{tri}}(\mathbf{f}) = \frac{1}{n} \sum_{i=1}^{n} \max\{\rho + d_p - d_n, 0\} \tag{6}$$

is applied to guarantee the features can well represent their corresponding samples. Where $\mathbf{f}$ represents a batch of the features, $n = |\mathbf{f}|$ is the size of the batch, ρ is the tiniest margin between the distance to the furthest positive instance d_p and the distance to the nearest negative instance d_n. The relationship between two instances from the source domain depends on the ground-truth labels and the pseudo labels for target-domain samples. Due to the different dimensions of the features extracted by heterogeneous networks, the triplet loss can only be applied in a certain subspace.

Then, in the logits level, we apply the cross-entropy loss with classifiers

$$\mathcal{L}_{\mathrm{Ttid}} = -\frac{1}{n} \sum_{i=1}^{n} \log P\left(\hat{\mathbf{y}}_i \middle| \mathrm{C}\left(\mathbf{t}_i^T \middle| [\mathbf{W}_s^T, \mathbf{W}_t^T]\right)\right) \tag{7}$$

$$\mathcal{L}_{\mathrm{Stid}} = -\frac{1}{n} \sum_{i=1}^{n} \log P\left(\hat{\mathbf{y}}_i \middle| \mathrm{C}\left(\mathbf{t}_i^S \middle| \mathbf{W}_t^S\right)\right) \tag{8}$$

where $\hat{\mathbf{y}}_i$ is the pseudo label for target-domain example $\mathbf{x}_t^i$. The trainable parameters $\mathbf{W}_s^T$, $\mathbf{W}_t^T$, and $\mathbf{W}_t^S$, respectively, denote the classifier parameters for the teacher classifying source-domain samples, the teacher classifying target-domain samples, and the student classifying target-domain samples. Meanwhile, to take advantage of the ground-truth label, the teacher also learns the source-domain knowledge by

$$\mathcal{L}_{\mathrm{Tsid}} = -\frac{1}{n} \sum_{i=1}^{n} \log P\left(\mathbf{y}_i \middle| \mathrm{C}\left(\mathbf{s}_i^T \middle| [\mathbf{W}_s^T, \mathbf{W}_t^T]\right)\right) \tag{9}$$

here, $\mathbf{y}_i$ is the ground-truth label for source-domain sample $\mathbf{x}_s^i$. Note that, with (7) and (9), the classifier $\mathrm{C}(\mathbf{t}_i^T|[\mathbf{W}_s^T, \mathbf{W}_t^T])$ in the teacher has learned both two domain knowledge while classifier $\mathrm{C}(\mathbf{t}_i^T|\mathbf{W}_t^S)$ for the student learns the target-domain knowledge only.

E. Asymmetric Mutual Learning

Compare the structure of the CNN and ViT, the most evident difference is that the ViT can capture long-range information with its cascaded self-attention modules. However, CNN only focuses on the local limited by the size of the convolution kernel. In addition, the CNN inductive bias, which includes assumptions of the data, can involve information

that ViT may not consider and the convolution kernel with a deterministic shape guarantees spatial information. More intuitively, the features extracted by the two networks have different dimensions. It ensures the subspaces learned by the heterogeneous networks are different but makes feature-level constraints unusable. The asymmetric distillation benefit from the difference in the patterns that two heterogeneous networks predict the identity. And focus on twofold: to allow students access to knowledge from the different subspaces and transfer the knowledge from the source to the target.

To make the student learn from different subspaces and take advantage of the reliable source-domain labels, the proposed DAML transfers the identity knowledge from the teacher by reducing the Kullback–Leibler divergence between the predictions of the target-domain features as

$$\mathcal{L}_{\text{id}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{C}\left(\mathbf{t}_i^T \middle| \mathbf{W}_t^T\right) \log \frac{\mathrm{C}\left(\mathbf{t}_i^S \middle| \mathbf{W}_t^S\right)}{\mathrm{C}\left(\mathbf{t}_i^T \middle| \mathbf{W}_t^T\right)} \tag{10}$$

with the above objective function, the student can learn the knowledge from the teacher which adopts knowledge from both source and target domain with (7) and (9). However, domain knowledge is also transferred to students and may harm the performance in the target domain. The ideal way to alleviate the distribution effect is to make the teacher predict identities under the target domain.

Limited by the domain gap, the knowledge learned from the source domain can not be directly applied to the target domain. And the goal of the proposed AML is to gain a student network that adapts to the target domain while benefiting from the source-domain identity knowledge. Making source-domain predictions from the teacher similar to the student will transfer the classifying knowledge learned from the source domain to the target domain

$$\mathcal{L}_{\text{dom}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{C}\left(\mathbf{s}_i^S \middle| \mathbf{W}_t^S\right) \log \frac{\mathrm{C}\left(\mathbf{s}_i^T \middle| \mathbf{W}_t^T\right)}{\mathrm{C}\left(\mathbf{s}_i^S \middle| \mathbf{W}_t^S\right)} \tag{11}$$

here, $\mathbf{W}_t^T$ learns source-domain knowledge with (9) while $\mathbf{W}_t^S$ only learns the knowledge from target domain. Equation (11) focuses on making the teacher predict source-domain samples in the same way as the student. In this way, the student can better adopt identity knowledge from the source domain without a domain gap as much as possible. Compared with (10), the above equation distills the knowledge in a different direction and together make the student model can distinguish pedestrian in the target domain.

F. Optimization

The total loss $\mathcal{L}$ of DAML can be summarized as

$$\begin{aligned}\mathcal{L} = &\left(\mathcal{L}_{\text{Ttid}} + \mathcal{L}_{\text{tri}}(\mathbf{t}^T)\right) + \left(\mathcal{L}_{\text{Stid}} + \mathcal{L}_{\text{tri}}(\mathbf{t}^S)\right) \\ &\times\lambda_1\left(\mathcal{L}_{\text{Tsid}} + \mathcal{L}_{\text{tri}}(\mathbf{s}^T)\right) + \lambda_2\mathcal{L}_{\text{id}} + \lambda_3\mathcal{L}_{\text{dom}}\end{aligned} \tag{12}$$

where λ_1, λ_2, and λ_3 are hyperparameters to balance the contributions of individual loss terms.

IV. Experiments

A. Datasets

Datasets: We evaluated our method on three public datasets **Market-1501** [58], **CUHK-SYSU** [59], and **MSMT17** [16].

1) **Market-1501** contains 32 668 labeled images captured from 1501 identities by six cameras. The training set has 12 936 images of 751 identities. In addition, 3368 query images and 19 732 gallery images from the other 750 identities are used as the testing set.
2) **CUHK-SYSU** includes 33 901 labeled images of 8432 identities taken in diverse scenes. The training set is constructed with 5532 identities having 15 088 images, and the rest is used for testing. There are 2900 images for the query and 6978 images for the gallery in the testing set.
3) **MSMT17** is a large-scale dataset consisting of 126 441 bounding boxes of 4101 identities caught on 12 outdoor and three indoor cameras. Among them, 32 621 images of 1041 identities are used for training and 93 820 of 3060 identities are used for testing.

B. Experiment Setting

1) Performance Metric: As a UDA task, we select one dataset as the source-domain dataset and another as the target-domain dataset. The model is trained with the labeled source-domain training set and adapts the target domain through the unlabeled target-domain training set. Then, the performance is evaluated according to the student network which work on the target-domain testing set. In our experiments, following the standard metrics, we employ the cumulative matching characteristic (CMC) curve and the mean average precision (*m*AP) score. Our experiments report rank-1, rank-5, and rank-10 accuracy and *m*AP scores.

2) Implementation Details: In the most common setting, we select IBN-ResNet-50 [55] as the teacher network and ViT-Base [57] as the student network. The batch size is set to 64 for both the source and target domains. In one batch, the sampler will select 16 identities and four images for each identity according to the ground-truth label or pseudo label for two domains. The input image has a fixed size of 256 × 128.

In the pretraining stage, we first train models 120 epochs on the source-domain dataset. The teacher CNN model is optimized by SGD with an initial learning rate of 1×10^{-2} and weight decay of 5×10^{-4} with a learning rate decays at 40th and 70th epoch. The SGD optimizer is with a momentum of 0.9 and the weight decay of 1×10^{-4} for student ViT. The learning rate is set to 8×10^{-3}, and the cosine schedule is applied. The input images are augmented with random flip and randomly erase with 50% probability.

In the fine-tuning stage, we adopted half the learning rate of the previous stage. Specifically, the learning rate is set to 5×10^{-3} for teacher CNN and 4×10^{-3} for student ViT. And the total number of training epochs is set to 60. The input images for two heterogeneous networks are randomly flipped and erased with 50% probability. We generate the pseudo labels by DBSCAN [60]. For DBSCAN, we select 0.6 as the maximum distance α between neighbors and set the minimal

TABLE I

COMPARISON OF CMC (%) AND *m*AP (%) PERFORMANCES WITH THE SOTA METHODS ON **MARKET-1501**, **CUHK-SYSU**, AND **MSMT17**

Method	Market-1501 → CUHK-SYSU				CUHK-SYSU → Market-1501			
	*m*AP	R1	R5	R10	*m*AP	R1	R5	R10
Directly Transfer (IBN-ResNet-50)	74.1	77.2	85.7	88.7	38.8	63.7	79.4	85.2
Directly Transfer (ViT-Base)	86.0	87.2	94.1	95.0	36.2	60.3	76.5	82.9
UNRN [39](AAAI'21)	62.3	64.1	76.9	82.0	70.9	86.7	92.8	94.5
MMT(IBN-ResNet-50) [27](ICLR'20)	78.4	81.0	89.7	92.2	76.0	88.8	95.2	97.0
MEB-Net [29](ECCV'20)	81.1	83.2	90.9	93.1	69.3	84.0	92.9	95.2
DAML (Ours)	**84.3**	**86.2**	**92.6**	**94.6**	**84.1**	**93.1**	**97.7**	**98.2**
Supervised (IBN-ResNet-50)	90.8	95.2	96.6	89.0	83.0	94.1	97.4	98.4
Supervised (ViT-Base)	93.1	97.2	97.8	92.1	82.3	93.2	97.9	98.8
Method	Market-1501 → MSMT17				CUHK-SYSU → MSMT17			
	*m*AP	R1	R5	R10	*m*AP	R1	R5	R10
Directly Transfer (IBN-ResNet-50)	8.4	23.8	34.5	39.6	10.3	26.3	38.3	44.3
Directly Transfer (ViT-Base)	13.0	33.3	45.3	51.1	12.5	28.2	41.1	47.5
MEB-Net [29](ECCV'20)	20.6	44.1	58.3	64.3	21.3	45.6	59.5	65.6
UNRN [39](AAAI'21)	25.3	52.4	64.7	69.7	12.6	31.1	43.8	49.7
MMT(IBN-ResNet-50) [27](ICLR'20)	26.6	54.4	67.6	72.9	24.0	49.0	63.0	68.6
DAML (Ours)	**41.4**	**65.4**	**76.0**	**80.2**	**44.0**	**67.0**	**78.0**	**81.9**
Supervised (ViT-Base)	54.1	76.6	87.5	90.8	54.1	76.6	87.5	90.8
Supervised (IBN-ResNet-50)	49.9	79.2	88.2	91.3	49.9	79.2	88.2	91.3

TABLE II

ABLATION STUDY IN TERMS OF *m*AP (%) AND CMC (%) ON **CUHK-SYSU (CS)** → **MARKET-1501 (M)**

Method	CS → M	
	*m*AP	R1
IBN-ResNet-50(Directly)	38.8	63.7
ViT-Base(Directly)	36.2	60.3
DAML w/o $\mathcal{L}_{Tsid} + \mathcal{L}_{tri}(\mathbf{s}^T)$	83.6	92.8
DAML w/o $\mathcal{L}_{id}$	83.3	92.3
DAML w/o $\mathcal{L}_{dom}$	83.6	92.5
DAML w/o SCU	81.0	91.0
DAML	84.1	93.1
IBN-ResNet-50(Supervised)	83.0	94.1
ViT-Base(Supervised)	82.3	93.2

number of neighbors to 2 for **CUHK-SYSU** and 4 for others. The hyperparameters λ_1, λ_2, and λ_3 are set to 0.1, 0.7, and 1.2, respectively. At the feature level, the margin ρ for triplet loss is set to 1.2.

C. Comparison With State-of-the-Art Methods

Since Duke University terminated the **DukeMTMC** [61] dataset, which has been widely used for evaluation of UDA person Re-ID task, the comparison becomes difficult. To meet the moral and ethical requirements and provide a new baseline for comparison, we evaluate the performance of some representative works which have official open-source codes based on the **CUHK-SYSU** dataset. And the results in **Market-1501** → **MSMT17** setting is from the authors' reports. We compare our DAML with SOTA UDA person Re-ID approaches. MMT [27] applies two networks that have the same structure for learning from each other with both feature-level and logit-level constraints. MEB-Net [29] introduces three homogeneous networks, and the output of each network is considered comprehensively in the pseudo label generation. Moreover, UNRN [39] designs a memory bank storing class centers from both source and target domains to mitigate the influence of noise labels. As shown in Table I, we evaluate the performance in four different manners, i.e., **Market-1501** → **CUHK-SYSU**, **CUHK-SYSU** → **Market-1501**, **Market-1501** → **MSMT17**, and **CUHK-SYSU** → **MSMT17**.

1) Comparisons on Large-Scale Datasets: The comparison results on **Market-1501** → **MSMT17** and **CUHK-SYSU** → **MSMT17** are shown in the bottom of Table I. The proposed DAML outperforms existing SOTAs by large margins. Specifically, DAML achieves the Rank-1 accuracy of 65.4% and *m*AP of 41.4% in the **Market-1501** → **MSMT17** setting, significantly improving the Rank-1 accuracy by 11.0% and *m*AP by 14.8% over the SOTA MMT. When compared to the SOTAs in **CUHK-SYSU** → **MSMT17** setting, the performance margin between our DAML and MMT is also significantly, e.g., the Rank-1 boost is 18.0%, and the *m*AP boost is 20.0%.

2) Comparisons on Small-Scale Dataset: We also evaluate DAML on two small-scale target-domain datasets settings, **Market-1501** → **CUHK-SYSU** and **CUHK-SYSU** → **Market-1501**, as shown in the top of Table I. Similar to the results on large-scale datasets, DAML consistently outperforms current SOTAs. Specifically, we achieve Rank-1 accuracy of 86.2% and *m*AP of 84.3% in **Market-1501** → **CUHK-SYSU** setting. Compared with the SOTA MEB-Net, the Rank-1 and *m*AP, respectively, improved by 3.0% and 3.2%. Meanwhile Rank-1 accuracy of 93.1% and *m*AP of 84.1% are gained in **CUHK-SYSU** → **Market-1501** setting. It improves the Rank-1 accuracy and *m*AP by 4.3% and 8.1% compared with the SOTA MMT. Note that the performance on **Market-1501** → **CUHK-SYSU** setting is even worse than direct transfer. Because there are only two samples per class in **CUHK-SYSU** on average and it harms the pseudo label generation. We will discuss this problem in Section IV-E4.

The above results demonstrate the outstanding performance of DAML thanks to its ability to learn knowledge from different subspaces and selectively transfer knowledge between two heterogeneous networks for UDA person Re-ID.

TABLE III
INFLUENCE OF DIFFERENT BACKBONES IN TERMS OF *m*AP (%) AND RANK-1 (%) ON **CUHK-SYSU (CS) → MARKET-1501 (M)**

Method	Backbone	Training Parameter	Testing Parameter	CS → M			
				*m*AP	R1	R5	R10
MMT	IBN-ResNet-50 + IBN-ResNet-50	99.8M	24.9M	76.0	88.8	95.2	97.0
MMT	ViT-Base + ViT-Base	345.2M	86.3M	75.2	86.7	94.2	96.4
UNRN	ResNet-50-NL	77.1M	38.5M	70.9	86.7	92.8	94.5
UNRN	ViT-Base	185.4M	86.3M	73.2	86.8	93.2	95.2
DAML	IBN-ResNet-50 + ViT-Base	111.2M	86.3M	84.1	93.1	97.7	98.2

D. Ablation Study

In this section, we conduct ablation experiments on **CUHK-SYSU → Market-1501** setting to assess the contribution of each component by separately removing them from DAML for training and evaluation.

As shown in Table II, when removing $\mathcal{L}_{\text{Tsid}} + \mathcal{L}_{\text{tri}}(\mathbf{s}^T)$, the Rank-1 accuracy drops by 0.3% and *m*AP drops by 0.5%, since the reliable identity information is underutilized. It illustrates that the ability to use the information in the source domain effectively is an essential factor in determining the model's performance. It illustrates the essential to effectively use the information in the source domain. When removing $\mathcal{L}_{\text{id}}$, which helps student network to learn the identity knowledge from the teacher, the performance drops of Rank-1 and *m*AP are 0.8% and 0.8%, respectively, compared with the full DAML. The performance drops due to the ignorance of the knowledge from the different subspaces in the logit level. And the knowledge can still transfer to each other through the pseudo label generation. Similarly, to validate the effectiveness of $\mathcal{L}_{\text{dom}}$, we remove it from DAML. The result also shows the margin of the Rank-1 accuracy by 0.6% and *m*AP by 0.5% to the complete DAML, which demonstrates that $\mathcal{L}_{\text{dom}}$ effectively helps to make the teacher network predict samples in the target domain. When the smooth classifier update (SCU) is not utilized, the parameters for the new classifiers are set as the new class centers. This leads to a decrease in Rank-1 accuracy by 2.1% and a decrease in the *m*AP by 3.1%. In summary, the inclusion of SCU significantly improves the performance. This is mainly attributed to SCU's ability to preserve optimization momentum, thereby enhancing the optimization process. Additionally, SCU aggregates the parameters of the classifiers, ensuring the retention of knowledge obtained from the objective functions. It is worth noting that the efficiency of optimization plays a crucial role in determining the effectiveness of the objective function. Consequently, SCU demonstrates the most substantial progress by addressing this limitation. The results prove that learning the knowledge from different subspaces and taking advantage of correct identity information from the source domain are the two keys to solving UDA person Re-ID.

E. Discussion

1) Influence of Backbone: To meet the requirement of heterogeneous networks in the proposed DAML, we introduce the ViT-Base, which contains more trainable parameters as the backbone. To clarify the source of performance growth, we repeat the experiments of MMT [27] and UNRN [39] while replacing the backbone with ViT-Base. As shown in Table III, "Backbone" represents the construction to extract features in the testing stage. When the ViT-base replaces the backbone, the optimization process follows the setting in TransReID [62]. As shown in Table III, after replacing the backbone with ViT-Base, there is no significant change in results. Limited by the symmetrical design in mutual learning manner and the high similarity in the classifying ways of ViTs, the Rank-1 and *m*AP of MMT dropped 2.1% and 0.8%, respectively. For the UNRN method, which focuses on making pseudo labels reliable through the memory mechanism, the ViT brings 0.1% and 2.3% in Rank-1 and *m*AP with much more parameters. Based on the above experiments, we can conclude that the heterogeneous networks and the asymmetric learning strategy play a major role in the growth of performance.

TABLE IV
ASYMMETRIC DISTILLATION ANALYSIS IN TERMS OF *m*AP (%) AND R1 (%) ON **CUHK-SYSU (CS) → MARKET-1501 (M)**

Method		CS → M	
Teacher	Student	*m*AP	R1
IBN-ResNet-50	ViT-Base	84.1	93.1
ViT-Base	ViT-Base	82.0	91.7
ViT-Base	IBN-ResNet-50	80.1	91.3
IBN-ResNet-50	IBN-ResNet-50	79.7	91.0

2) Heterogeneous Networks Analysis: One of the keys to improving UDA person Re-ID is learning knowledge from the different subspaces. To illustrate the effect of heterogeneous networks, we train the proposed DAML with different combinations of teacher and student. From Table IV, we can figure out that the student with a heterogeneous teacher will achieve better performance. Specifically, the performance of student ViT improved by 0.7% and 1.1% in Rank-1 and *m*AP with the asymmetric teacher. The results in the student CNN is similar, Rank-1 and *m*AP are enhanced by 2.5% and 5.3%. These experimental results strongly prove the necessity of using two heterogeneous networks to work as the teacher and student. And the knowledge from different subspaces has the capacity to help the student to learn broader knowledge.

When ViT is seen as the student, the benefit from the heterogeneous teacher is more evident than the improvement that CNN works as the student. The heterogeneous teacher for ViT brings in 1.4% and 2.1% on Rank-1 and *m*AP. It only improves Rank-1 and *m*AP in 0.3% and 0.4% for the student CNN. This phenomenon can be ascribed to the difference in the range of receptive field of these two networks that the former can consider the relationship between any two areas, but the size of convolution kernels limits the latter. It gives ViT has the ability to learn the pattern that CNN applied to classify identities but not vice versa.

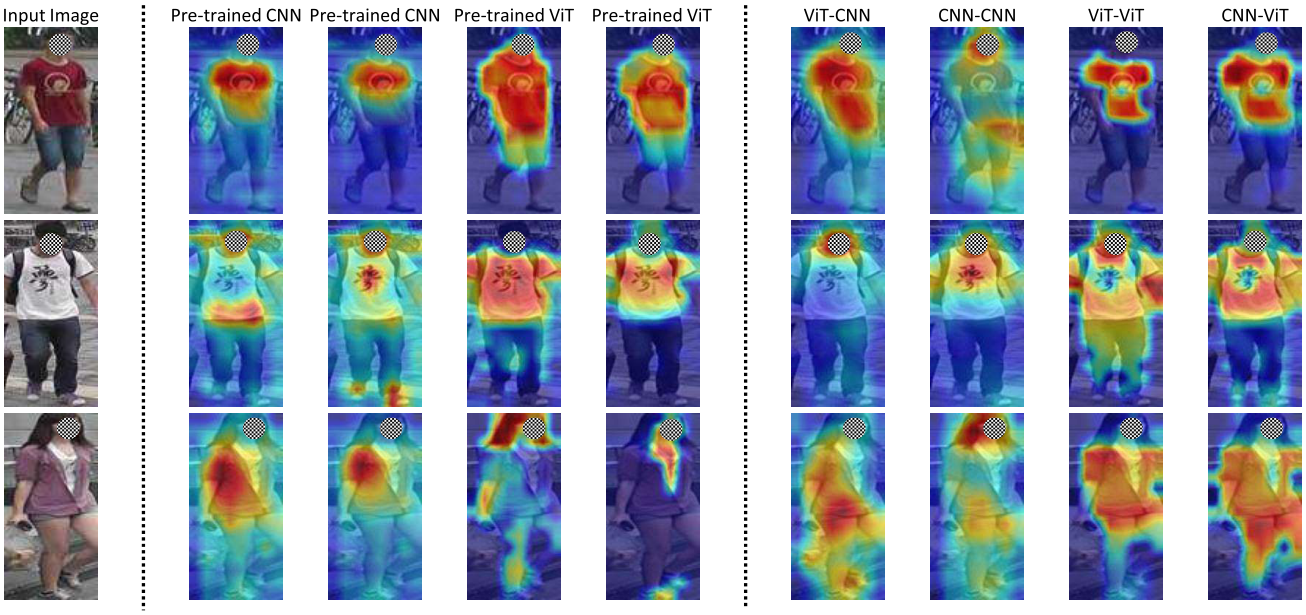


Fig. 3. Visualization results of the models on **Market-1501**. For each line, we show an input image, the area considered by the pretrained ViT, the pretrained CNN, and the different combinations of teacher and student in turn.

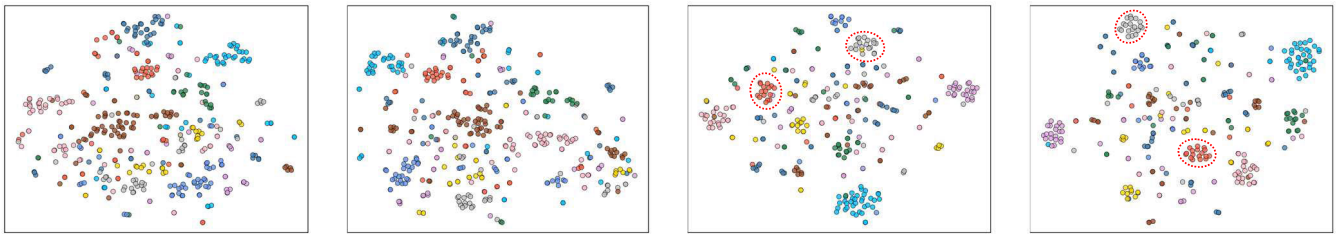


Fig. 4. Visualization results of the feature distribution on the target domain of **CUHK-SYSU** → **Market-1501** setting by t-SNE. The pretrained CNN and ViT represent directly evaluate the models trained on the source domain on the target. And DAML CNN and ViT denote evaluating the teacher CNN and the student ViT that being transferred by the proposed DAML. Each color represents an identity. (a) Pretrained CNN. (b) Pretrained ViT. (c) DAML CNN. (d) DAML ViT.

3) Visualization: The proposed DAML can make the student learn the knowledge from different subspaces. To further illustrate the effectiveness of DAML, which can selectively transfer the knowledge between two networks, we apply Score-CAM [63] to visualize the pixelwise attention areas on **CUHK-SYSU** → **Market-1501** setting. Fig. 3 visualizes individual attention patterns for the three people from the target domain, where each column represents the attention area of the pretrained CNN, ViT, and the different combinations of teacher-student. From the first two columns, we can observe that the classifying patterns of CNN and ViT are different, which states the difference between their embedding spaces. With these discrepancies, the heterogeneous networks can be improved by learning knowledge from heterogeneous networks. In the last four columns, we can find that the networks mutual learning with heterogeneous networks can better consider the individual by the whole pedestrian while also taking into account many details that identify the persons more efficiently. On the contrary, the recognition patterns of the networks that mutual learning with the same network have no significant change. The visualization demonstrates the function of DAML in learning the knowledge from different subspaces improving the performance of the student.

We visualize the distribution of features in the target domain using the t-distributed stochastic neighbor embedding (t-SNE). As depicted in Fig. 4(a) and (b), the features extracted by the pretrained models exhibit a lack of distinct boundaries among different identities. However, upon applying the DAML method to transfer the models to the target domain, the features exhibit clear clustering patterns, as evident in Fig. 4(c) and (d). The pseudo label generation facilitated by knowledge exchange enables the teacher CNN to effectively classify pedestrians in the target domain. Furthermore, in the context of AML, the student ViT does not directly acquire ground-truth knowledge from the source domain. Consequently, the student ViT demonstrates improved performance in the target domain, as evidenced by the well-defined clusters within the red circle in Fig. 4(c) and (d). In conclusion, the DAML method facilitates efficient model transfer to the target domain.

4) Samples Augment: Due to the small number of samples for each class in **CUHK-SYSU**, the performance of clustering algorithm is severely limited The simplest and most direct way to address this problem is by augmenting the samples with random erase [64] and random crop, which can generate new samples while keeping the original identity when extracting features for the clustering algorithm. As shown in Table V, the performance with augmented data is better than the original. Specifically, the Rank-1 and *m*AP are enhanced by 4.4% and 5.0% when applying the random crop method. Similarly, the random erase improves Rank-1 and *m*AP by 4.4% and

TABLE V

INFLUENCE OF SAMPLE AUGMENT FOR CLUSTERING ALGORITHM IN TERMS OF *m*AP (%) AND CMC (%) ON **MARKET-1501 (M) → CUHK-SYSU (CS)**

Method	Repeat Times	M → CS			
		*m*AP	R1	R5	R10
Directly Transfer	-	86.0	87.2	94.1	95.0
DAML	0	84.3	86.2	92.6	94.6
+ Random Crop	1	89.3	90.6	95.5	96.9
	2	89.1	90.4	95.4	96.6
+ Random Erase	1	88.3	90.0	95.0	96.3
	2	89.2	90.6	95.5	96.6
+ Random Crop + Random Erase	1	88.5	89.8	95.4	96.8
	2	87.6	89.0	95.2	96.6
Supervised	-	90.8	95.2	96.6	89.0

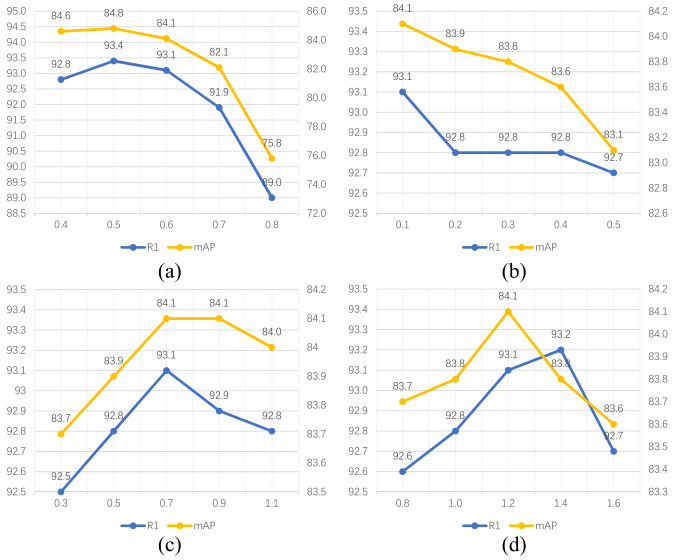


Fig. 5. The impact of hyperparameters (a) α, (b) λ_1, (c) λ_2, and (d) λ_3. The performances are evaluated on **CUHK-SYSU → Market-1501** setting.

4.9%. The above experiment results state the necessity of enough samples for each class in the clustering algorithm. Nevertheless, applying both random crop and random erase is not as effective as applying only one. The Rank-1 and *m*AP are only increased by 3.6% and 4.2%. It suggests that the excessive augment method may harm identity knowledge and reduce the benefits from the augmented samples.

Compared to datasets collected for research, the number of identities and the number of samples in each identity are unknown in a real-world system. And this information cannot be counted on the raw data unless annotated on them. However, one of the advantages of UDA Re-ID is avoiding the annotation on the target domain, and it means the class-dependent super parameters are not available. Because clustering algorithms elapse a long time to run on large datasets and take up most of the total training time, the super parameter selection experiments may be unacceptable for real-world systems. Thus, a method without any clustering algorithm that still can mine identity knowledge from the target domain may make more sense for applying to the real world. On the other hand, an efficient data augment method can restore the UDA methods based on clustering algorithm.

5) Impact of Hyperparameters: We further conduct a hyperparameter-sensitive analysis for DAML. Fig. 5 illustrates the impact of various hyperparameters, including α, λ_1, λ_2, and λ_3, on the Rank-1 accuracy and *m*AP. The experimental results demonstrate that DAML is reasonably robust to the values used to control knowledge transfer, namely λ_1, λ_2, and λ_3. Specifically, as λ_1 increases, which controls the learning of identity information from the source domain by the teacher model, the performance gradually decreases. This indicates that the domain gap between the source domain and the target domain can negatively affect the performance. For λ_2 and λ_3, which control the transfer process in AML, the curves exhibit clear unimodal characteristics. Importantly, the performance fluctuation caused by all the hyperparameters is within 1.2% compared to the overall performance. Regarding the maximum distance between neighbors, denoted as α, it is closely related to the quality of clustering. The performance experiences a significant decrease when deviating from the parameter settings used in previous methods [27], [29], [39]. Furthermore, an inappropriate setting for the minimum number of neighbors in DBSCAN can lead to incorrect pseudo labels, and models gain a disastrous performance. It is worth noting that the cluster algorithm is not the focus of this article.

V. CONCLUSION

In this article, we proposed the DAML to learn knowledge from a broader scope via AML with heterogeneous networks for UDA person Re-ID. Our method aims to learn the knowledge from various subspaces and transfer the identity knowledge from the source to the target domain. The former can improve feature expressiveness while also rectifying potential faults during training. The latter takes full advantage of the identity knowledge from the source domain to improve the performance in the target domain. Specifically, DAML first generates the pseudo labels according to the features extracted by heterogeneous networks, which are more reliable due to the consideration of various subspaces. And the knowledge from two subspaces can be exchanged in a hard distillation manner in this process. With the SCU, the classifiers can maintain the knowledge from the last epoch. Then, the teacher will train on both source-domain and target-domain datasets to utilize the ground-truth label and transfer the knowledge to the target domain with the domain knowledge from the student. To better adapt to the target domain, the student only trained on the target-domain dataset and benefited from the guidance from the teacher, which had learned the source-domain knowledge. Experiments on four different experiment settings prove essential to learn the knowledge from various subspaces and demonstrate the effectiveness of the proposed DAML for UDA person Re-ID.

REFERENCES

[1] S. Gong, M. Cristani, C. C. Loy, and T. M. Hospedales, "The re-identification challenge," in *Person Re-Identification*. Springer, 2014, pp. 1–20.

[2] W. Li, X. Zhu, and S. Gong, "Scalable person re-identification by harmonious attention," *Int. J. Comput. Vis.*, vol. 128, no. 6, pp. 1635–1653, Jun. 2020.

[3] Y. Sun, L. Zheng, Y. Li, Y. Yang, Q. Tian, and S. Wang, "Learning part-based convolutional features for person re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 902–917, Mar. 2021.

[4] J. Yin, A. Wu, and W.-S. Zheng, "Fine-grained person re-identification," *Int. J. Comput. Vis.*, vol. 128, no. 6, pp. 1654–1672, Jun. 2020.

[5] 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. ECCV*, 2018, pp. 501–518.

[6] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, "Relation-aware global attention for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3183–3192.

[7] 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.

[8] A. Wu, W.-S. Zheng, and J.-H. Lai, "Unsupervised person re-identification by camera-aware similarity consistency learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6921–6930.

[9] H.-X. Yu and W.-S. Zheng, "Weakly supervised discriminative feature learning with state information for person identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5527–5537.

[10] Z. Zhong, L. Zheng, S. Li, and Y. Yang, "Generalizing a person retrieval model Hetero- and homogeneously," in *Proc. ECCV*, 2018, pp. 176–192.

[11] C. Han et al., "Re-ID driven localization refinement for person search," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 9813–822.

[12] Q. Zhou, B. Zhong, X. Liu, and R. Ji, "Attention-based neural architecture search for person re-identification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 11, pp. 6627–6639, Nov. 2022.

[13] K. Li, Z. Ding, K. Li, Y. Zhang, and Y. Fu, "Vehicle and person re-identification with support neighbor loss," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 826–838, Feb. 2022.

[14] 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. CVPR*, 2018, pp. 994–1003.

[15] Y.-J. Li, C.-S. Lin, Y.-B. Lin, and Y. 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. 7918–7928.

[16] 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.

[17] Y. Zou, X. Yang, Z. Yu, B. V. K. V. Kumar, and J. Kautz, "Joint disentangling and adaptation for cross-domain person re-identification," in *Proc. ECCV*, vol. 12347, 2020, pp. 87–104.

[18] I. J. Goodfellow et al., "Generative adversarial networks," 2014, *arXiv:1406.2661*.

[19] X. Chang, Y. Yang, T. Xiang, and T. M. Hospedales, "Disjoint label space transfer learning with common factorised space," in *Proc. AAAI*, 2019, pp. 3288–3295.

[20] 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. 8079–8088.

[21] 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.

[22] F. Yang et al., "Joint noise-tolerant learning and meta camera shift adaptation for unsupervised person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4853–4862.

[23] P. Dai et al., "Disentangling task-oriented representations for unsupervised domain adaptation," *IEEE Trans. Image Process.*, vol. 31, pp. 1012–1026, 2022.

[24] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang, "A bottom-up clustering approach to unsupervised person re-identification," in *Proc. AAAI*, 2019, pp. 8738–8745.

[25] 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. 8221–8230.

[26] Z. Bai, Z. Wang, J. Wang, D. Hu, and E. Ding, "Unsupervised multi-source domain adaptation for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12909–12918.

[27] Y. Ge, D. Chen, and H. Li, "Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification," in *Proc. ICLR*, 2020, pp. 1–15.

[28] K. Zheng, W. Liu, L. He, T. Mei, J. Luo, and Z.-J. Zha, "Group-aware label transfer for domain adaptive person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5306–5315.

[29] Y. Zhai, Q. Ye, S. Lu, M. Jia, R. Ji, and Y. Tian, "Multiple expert brainstorming for domain adaptive person re-identification," in *Proc. ECCV*, 2020, pp. 594–611.

[30] 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/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2133–2142.

[31] A. Ma, J. Li, K. Lu, L. Zhu, and H. T. Shen, "Adversarial entropy optimization for unsupervised domain adaptation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 11, pp. 6263–6274, Nov. 2022.

[32] L. Zhang, J. Fu, S. Wang, D. Zhang, Z. Dong, and C. L. P. Chen, "Guide subspace learning for unsupervised domain adaptation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 9, pp. 3374–3388, Sep. 2020.

[33] J. Choi, M. Jeong, T. Kim, and C. Kim, "Pseudo-labeling curriculum for unsupervised domain adaptation," in *Proc. BMVC*, 2019, p. 67.

[34] Z. Deng, Y. Luo, and J. Zhu, "Cluster alignment with a teacher for unsupervised domain adaptation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9943–9952.

[35] W. Wang, F. Zhao, S. Liao, and L. Shao, "Attentive WaveBlock: Complementarity-enhanced mutual networks for unsupervised domain adaptation in person re-identification and beyond," *IEEE Trans. Image Process.*, vol. 31, pp. 1532–1544, 2022.

[36] M. Li, P.-Y. Huang, X. Chang, J. Hu, Y. Yang, and A. Hauptmann, "Video pivoting unsupervised multi-modal machine translation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3918–3932, Mar. 2023.

[37] Y. Fu et al., "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. 6111–6120.

[38] R. Zhou, X. Chang, L. Shi, Y.-D. Shen, Y. Yang, and F. Nie, "Person reidentification via multi-feature fusion with adaptive graph learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1592–1601, May 2020.

[39] K. Zheng, C. Lan, W. Zeng, Z. Zhang, and Z. Zha, "Exploiting sample uncertainty for domain adaptive person re-identification," in *Proc. AAAI*, 2021, pp. 3538–3546.

[40] Y. Wu et al., "Multi-centroid representation network for domain adaptive person Re-ID," in *Proc. AAAI*, 2022, pp. 2750–2758.

[41] Y. Pan, T. Yao, Y. Li, Y. Wang, C.-W. Ngo, and T. Mei, "Transferrable prototypical networks for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2234–2242.

[42] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[43] L. Wei, A. Xiao, L. Xie, X. Zhang, X. Chen, and Q. Tian, "Circumventing outliers of autoaugment with knowledge distillation," in *Proc. ECCV*, 2020, pp. 608–625.

[44] L. Yuan, F. E. H. Tay, G. Li, T. Wang, and J. Feng, "Revisiting knowledge distillation via label smoothing regularization," in *Proc. CVPR*, 2020, pp. 3902–3910.

[45] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *Proc. 5th ICLR*, Toulon, France, 2017, pp. 1–13.

[46] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. ICLR*, 2017, pp. 1–10.

[47] L. Zhang et al., "TN-ZSTAD: Transferable network for zero-shot temporal activity detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3848–3861, Mar. 2023.

[48] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.

[49] C. Yan et al., "ZeroNAS: Differentiable generative adversarial networks search for zero-shot learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9733–9740, Dec. 2022.

[50] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1106–1114.

[51] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[52] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
[53] 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.
[54] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.
[55] X. Pan, P. Luo, J. Shi, and X. Tang, "Two at once: Enhancing learning and generalization capacities via IBN-Net," in *Proc. ECCV*, 2018, pp. 484–500.
[56] A. Vaswani et al., "Attention is all you need," in *Proc. NIPS*, 2017, pp. 5998–6008.
[57] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021, pp. 1–22.
[58] 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.
[59] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "End-to-end deep learning for person search," 2016, *arXiv:1604.01850*.
[60] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. KDD*, 1996, pp. 226–231.
[61] E. Ristani, F. Solera, R. S. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. ECCV*, 2016, pp. 17–35.
[62] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "TransReID: Transformer-based object re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14993–15002.
[63] H. Wang et al., "Score-CAM: Score-weighted visual explanations for convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 111–119.
[64] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI*, 2020, pp. 13001–13008.

Qiong Wu received the B.E. degree in computer science and technology from Xiamen University, Xiamen, China, in 2020, where he is currently pursuing the Ph.D. degree with the Institute of Artificial Intelligence and the Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China.

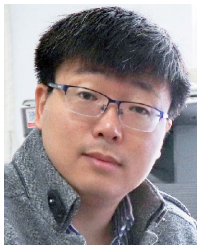
His research interests include pattern recognition, machine learning, and computer vision.

Jiahan Li is currently pursuing the Ph.D. degree with the Harbin Institute of Technology, Harbin, China.

She has authored articles in international journals, including IEEE SIGNAL PROCESSING LETTERS (SPL). Her current research interests include image translation and image restoration.

Pingyang Dai received the M.S. degree in computer science and the Ph.D. degree in automation from Xiamen University, Xiamen, China, in 2003 and 2013, respectively.

He is currently a Senior Engineer with the Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, and the School of Informatics, Xiamen University. His research interests include computer vision and machine learning.



Qixiang Ye (Senior Member, IEEE) received the B.S. and M.S. degrees in mechanical and electrical engineering from the Harbin Institute of Technology, Harbin, China, in 1999 and 2001, respectively, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2006.

He was a Visiting Assistant Professor with the Institute of Advanced Computer Studies (UMIACS), University of Maryland, College Park, MD, USA, in 2013, and a Visiting Scholar with IIID, Duke University, Durham, NC, USA, in 2016. Since 2016, he has been a Professor with the University of Chinese Academy of Sciences, Beijing. He has authored more than 100 papers in refereed conferences and journals, including the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE International Conference on Computer Vision (ICCV), European Conference on Computer Vision (ECCV), the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, the IEEE TRANSACTIONS ON IMAGE PROCESSING, and the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. His research interests include image processing, visual object detection, and machine learning.

Dr. Ye received the Sony Outstanding Paper Award.

Liujuan Cao (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the School of Computer Science and Technology, Harbin Engineering University, Harbin, China, in 2005, 2008, and 2013, respectively.

She is currently an Associate Professor with Xiamen University, Xiamen, China. She has authored over 40 papers in the top and major tired journals and conferences, including Conference on Computer Vision and Pattern Recognition (CVPR) and IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP). Her research interests include computer vision and pattern recognition.

Dr. Cao is also the Financial Chair for the IEEE International Workshop on Multimedia Signal Processing (MMSP) 2015, the Workshop Chair for the ACM International Conference on Internet Multimedia Computing and Service (ICIMCS) 2016, and the Local Chair for the Visual and Learning Seminar 2017.

Yongjian Wu received the master's degree in computer science from Wuhan University, Wuhan, China, in 2008.

He is currently an Expert Researcher and the Director of the Youtu Laboratory, Tencent Company Ltd., Shanghai, China. His research interests include face recognition, image understanding, and large-scale data processing.

Rongrong Ji (Senior Member, IEEE) is currently a Nanqiang Distinguished Professor at Xiamen University, Xiamen, China; the Deputy Director of the Office of Science and Technology, Xiamen University; and the Director of the Media Analytics and Computing Laboratory, Xiamen University. He has authored 50+ articles in ACM/IEEE TRANSACTIONS, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI) and *International Journal of Computer Vision* (IJCV), and 100+ full papers on top-tier conferences, such as Conference on Computer Vision and Pattern Recognition (CVPR) and Advances in Neural Information Processing Systems (NeurIPS). His research interests include computer vision, multimedia analysis, and machine learning.

Dr. Ji is also an Advisory Member for Artificial Intelligence Construction in the Electronic Information Education Committee of the National Ministry of Education. He was a recipient of the Best Paper Award of ACM Multimedia 2011. He was recognized as the National Science Foundation for Excellent Young Scholar in 2014, the National Ten Thousand Plan for Young Top Talent in 2017, and the National Science Foundation for Distinguished Young Scholar in 2020. He has served as the Area Chair for top-tier conferences, such as CVPR and ACM Multimedia. His publications have got over 20K citations in Google Scholar.