International Journal of Machine Learning and Cybernetics (2023) 14:1951–1965
https://doi.org/10.1007/s13042-022-01739-9

ORIGINAL ARTICLE

Heterogeneous dual network with feature consistency for domain adaptation person re-identification

Hua Zhou[1] · **Jun Kong**[2] · **Min Jiang**[1] · **Tianshan Liu**[3]

Received: 15 February 2022 / Accepted: 1 December 2022 / Published online: 9 December 2022

Abstract
To reduce the noisy pseudo-labels generated by clustering for unsupervised domain adaptation (UDA) person re-identification (re-ID), the method of collaborative training between dual networks has been proposed and proved to be effective. However, most of these methods ignore the coupling problem between dual networks with the same architecture, which makes them inevitably share a high similarity and lack heterogeneity and complementarity. In this paper, we propose a heterogeneous dual network (HDNet) framework with two asymmetric networks, one of which applies convolution with limited receptive fields to obtain local information and the other uses Transformer to capture long-range dependency. Additionally, we propose feature consistency loss (FCL) that does not rely on pseudo-labels. FCL focuses more on the consistency of the sample in the feature space rather than the class prediction space, driving the feature learning of UDA re-ID from the task level to the feature level. Furthermore, we propose an adaptive channel mutual-aware (ACMA) module which contains two branches to focus on the global and local information between channels. We evaluate our proposed method on three popular datasets: DukeMTMC-reID, Market-1501 and MSMT17. Extensive experimental results demonstrate that our method achieves a competitive performance.

Keywords Person re-identification · Unsupervised domain adaptation · Heterogeneous dual network · Feature consistency · Attention

1 Introduction

Person re-identification (re-ID) [1–4] is regarded as a sub-problem of image retrieval, which is widely applied in many fields [5]. Given a monitored pedestrian image, retrieve the pedestrian image under cross-devices. At present, there are proud achievements in supervising person re-identification. However, obtaining a large amount of labeled data is costly and time-consuming. To overcome the insufficiency of labeled training data, many works directly apply the model trained on the large-scale labeled source domain to another unlabeled target domain. Unfortunately, due to domain shift or dataset bias, direct migration across domains usually does not work well. Therefore, unsupervised domain adaptation (UDA) [6–8] is introduced, aiming to adapt the knowledge or patterns learned from the labeled source domain dataset to the unlabeled target domain.

At present, the work of UDA on re-ID is mainly divided into domain shift method [9–11], domain alignment method [12–15] and clustering-based pseudo-label method [16–20]. For example, Fu et al. propose the SSG [15], which reduces the image style difference between two datasets through image segmentation, so as to obtain more robust features. However, the relationship between target samples is often ignored by reducing the difference between domains. In view of this, Zhong et al. propose that ECN-GPP [11] mainly studies the intra-domain variation of the target domain and performs three invariance constraints on the dataset. For the method of clustering-based pseudo-label, Wu et al. propose a group-aware label transfer (GLT) [19] algorithm that can

✉ Jun Kong
kongjun@jiangnan.edu.cn

[1] Jiangsu Provincial Engineering Laboratory of Pattern Recognition and Computational Intelligence, Jiangnan University, Wuxi 214122, China

[2] Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), Jiangnan University, Wuxi 214122, China

[3] Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong 999077, China

correct pseudo-label containing noise online. Subsequently, the multi-label method is derived [21], which mainly turns the re-ID task without label into a multi-classification problem, so as to find the true label of the image. Among them, the clustering-based pseudo-label method has been proved to be more effective and remains the most advanced. Although it achieves state-of-the-art performance in various UDA tasks, its abilities are hindered by noisy pseudo-labels, which are caused by the limited transferability of source-domain features, the unknown number of target-domain identities, and the imperfect results of the clustering algorithm. To address this problem, a dual network framework, Mutual Mean-Teaching (MMT) [22] is proposed, which proposes to use the "synchronized average teaching" framework to optimize pseudo-labels. The core idea is to use more robust "soft" labels to optimize pseudo-labels online. To a certain extent, the problem of pseudo-label noise is solved. However, since the two networks have the identical structure, there will be a coupling problem. That is, they will become more and more similar, which may make them tend to the same kind of noise and lack complementarity. This limits further improvement in performance.

More seriously, as the identities in the testing set are different from the training set, and in the inference process, the re-ID task is to match the pedestrians by extracting the features of the pedestrians. So how to optimize the model in the feature space is crucial. However, the traditional UDA re-ID task ignores the consistency of the samples in the feature space. Specifically, the classification loss is mainly used in the class prediction space, which pays more attention to the similarity relationship between the samples and the labels. It ignores the relationship between the samples in the feature space. In addition, the triplet loss relies on pseudo-labels to select positive and negative samples, so it is sensitive to noisy pseudo-labels, which will lead to confusion in the selection of positive and negative samples, and ultimately lead to the degradation of feature learning.

To overcome the mentioned problems above, we propose a novel framework, namely the heterogeneous dual network (HDNet). The core of the HDNet is to improve the heterogeneity and complementarity of dual networks to solve the problem of network coupling. The overall framework consists of two asymmetric network structures. Both networks use ResNet50 [23] as the basic structure. One of the networks focuses more on global perception by introducing the Transformer Encoder block, while the original network based on convolutional neural network (CNN) focuses on local perception. The heterogeneity of the network allows each other to focus on different characteristics. Moreover, in order to focus on the similarity between samples in the feature space, we propose feature consistency loss (FCL), which replaces task-level similarity learning with feature-level similarity learning. FCL does not depend on any label information and effectively avoids the influence of noise pseudo-labels in the optimization process. Additionally, to enhance the semantic information of the network, we propose an adaptive channel mutual-aware (ACMA) module containing two branches, which are used to focus on the long-range dependency and local perception between channels, respectively. In addition, we introduce the channel shuffle [24] operation to further improve the interaction between cross-channel information.

Our contributions can be summarized in the following four points:

- We present an HDNet framework with two asymmetric network structures by introducing the Transformer Encoder block, which solves the coupling problem of dual networks.
- We propose a feature consistency loss that does not rely on pseudo-labels. It aims to focus on the similar relationship between samples in the feature space.
- An adaptive channel mutual-aware module is proposed, containing two branches to simultaneously focus on the global and local relationships between channels.
- Extensive experiments and ablation studies are conducted to validate the effectiveness of each proposed component and the whole framework.

The rest of this paper is organized as follows. Firstly, we review the related works in section "Related work". Then, a detailed description of the proposed HDNet is given in section "Our approach". Later, we evaluate the proposed method through extensive experiments on the three widely used datasets in section "Experiments". Finally, the conclusions of this work are outlined in section "Conclusion".

2 Related work

Unsupervised Domain Adaptation for Person re-ID. The current mainstream methods for UDA can be divided into three categories. The first is the domain shift method. They [25–28] use generative adversarial networks (GAN) to narrow the gap between the source domain and the target domain. For example, PTGAN [29] is proposed to handle the domain gap problem by transferring the knowledge from source dataset to target dataset. However, due to the poor migration effect and slower convergence speed, researchers begin to turn to new solutions. The second category is the domain alignment method. The objective of domain alignment is to learn domain invariant features. Several attempts leverage semantic attributes to align the feature distribution in the latent space, such as transferable joint attribute-identity deep learning [12], and invariant feature learning [7]. However, these approaches strongly rely on extra attribute

annotations, which require extra labor. The last category is the cluster-based pseudo-label method. They [16–18, 30] first use a pre-trained model on the source domain to generate pseudo-labels for the target domain, and then use the generated pseudo-labels to optimize the network. Finally, the above two steps are trained iteratively until the network converges. Unfortunately, all these methods suffer from large amounts of noise in the pseudo-labels generated by clustering, which hinders the further improvement of network performance.

To solve the above problem, a dual-network framework MMT [22] was proposed to generate more reliable soft pseudo-labels by using dual networks with the same structure. However, due to the coupling problem between symmetric network structures, it is inevitable that the two networks will be inclined to make the same mistake. Therefore, how to improve the difference and reduce homogeneity between dual networks is the starting point for us to propose heterogeneous dual networks.

Transformer for Vision Tasks. Recently, Transformer has shown its advantages over traditional methods in many visual tasks [31–33]. It mainly includes two model architectures. One is a hybrid structure combining CNN and Transformer, such as DETR [34], which opened the curtain for the application of Transformer in vision. The other is a pure Transformer structure. For example, ViT [35] adopted a complete self-attention Transformer structure without CNN, which still achieved the effect of competing with CNN. However, ViT relies on large datasets for pre-training. To overcome this shortcoming, DeiT [36] was proposed which used a teacher-student strategy specific for Transformers to speed up ViT training. Unfortunately, these methods all need to divide the images into multiple non-overlapping patches, which makes the spatial correspondence weaker between the input and intermediate features. However, in re-ID, the spatial alignment is critical for feature learning [37, 38]. To overcome these problems, TransReID [39] was proposed, which was the first Transformer architecture used for re-ID tasks.

Immediately afterwards, a lot of works [33, 40–43] also carried out more in-depth research on Vision Transformer, which reduced the dependence on the dataset and improved the convergence speed of the model. In this paper, we combine Transformer and CNN to achieve the two-way fusion of the global modeling capabilities of Transformer and the local capture capabilities of CNN. As far as we know, this is the first time to explore the global expression capabilities of Transformer in the UDA re-ID.

Attention Mechanism. The attention mechanism [44–46] has received widespread concern due to its ability to select the focus area and produces a more discriminative feature representation. For example, Hu et al. proposed SENet [47], which captured the global relationship between channels by squeeze-and-excitation operations. Subsequently, some studies improved the SE block by capturing more complex channel dependence [48–50] or combining additional spatial attention [51–53]. GSoP [49] introduced a second-order pooling for more effective feature aggregation, while CBAM [53] and scSE [51] improved the SE block by combining additional spatial attention. In order to reduce the computational burden, ECA-Net [54] was proposed to learn effective channels with lower model complexity by considering each channel and its k neighbors to capture local cross-channel interactions. However, all these methods have a common problem that they rarely consider the global and local information between channels simultaneously. To tackle the above-mentioned issue, we propose an adaptive channel mutual-aware (ACMA) module, which pays attention to the long-range dependency and local perception between channels. In addition, in order to strengthen the interaction between cross-channel information, we also introduce the channel shuffle [24] operation, which allows the data of different channels to establish connection without increasing the amount of calculation.

3 Our approach

Overview. In the re-ID task, UDA aims to adapt a source pretrained model to the target dataset with unlabeled target data. The labeled source domain dataset is denoted as $S = \{X_s, Y_s\}$, where X_s and Y_s denote the sample images and the person identities, and the unlabeled target domain dataset is denoted as $T = \{X_t\}$. The source dataset contains N_s sample images with M_s different identities. The N_t sample images in the target-domain T have no identity label available.

The HDNet is designed to solve the coupling problem in the dual-network training process. At the same time, in order to be more suitable for re-ID tasks, we explore the consistency of samples in the feature space. Finally, to further enhance the semantic information of the network, we have proposed an ACMA module. Next, we will describe each of the proposed modules in detail.

3.1 Heterogeneous dual network (HDNet)

Our proposed HDNet framework adopts a two-stage training scheme including supervised learning in source domains and unsupervised adaptation to target domains.

3.1.1 Supervised learning in source domains

In this stage, double deep neural network with different network architectures are first pre-trained in the fully supervised way on the source domain. Giving the source data,

which contains each sample x_i^s and its ground truth identity y_i', these two networks (with weight ω) transform x_i^s into average features $f_a(x_i^s|\omega^a)$ and max features $f_m(x_i^s|\omega^m)$, and outputs predicted probability $p_a(x_i^s|\omega^a)$ and $p_m(x_i^s|\omega^m)$. The network parameters w are then optimized with respect to an identity classification loss and a triplet loss. The identity classification loss is defined as:

$$\mathcal{L}_{id}^z = \frac{1}{N_s}\sum_{i=1}^{N_s} L_{ce}(p_z(x_i^s|\omega^z), y_i'), \tag{1}$$

where z is represented as a or m, and formula (2) (3) are the same. The triplet loss [55] is also defined as:

$$\mathcal{L}_{tri}^z = \frac{1}{N_s}\sum_{i=1}^{N_s} L_{tri}(f_z(x_i^s|\omega^z), y_i'). \tag{2}$$

The whole network is trained with a combination of both losses:

$$\mathcal{L}_{scr} = \mathcal{L}_{id}^z + \mathcal{L}_{tri}^z. \tag{3}$$

With a dual network architecture, the supervised learning thus produces two pre-trained re-ID models.

3.1.2 Unsupervised adaptation in target domain

The adaptation procedure contains two components: Clustering-based hard label fine-tuning and Heterogeneous-network soft label training. Specifically, we study the collaborative training of two heterogeneous networks.

Clustering-based hard label fine-tuning: In each learning iteration, *DBSCAN* and *K-means* clustering are employed in the target domain for pseudo-label prediction. The clustering procedure includes three steps: (1) For each sample in the target domain, the average of the features extracted by the two heterogeneous models is used as the clustering feature. (2) A mini-batch clustering is performed to assign samples into different groups. (3) The produced cluster IDs are used as pseudo-labels $\widetilde{Y}_t$ for the training samples X_t. Then the generated pseudo-labels are used to fine-tune the network through identity classification loss and triplet loss.

However, the pseudo-labels generated by clustering contain a lot of noises, hindering the optimization of the model. To mitigate the pseudo label noise, we propose a Heterogeneous-network framework to generate more reliable pseudo-labels to train the network.

Heterogeneous-network soft label training: The dual network structure is proposed to solve the noise pseudo-label problem, but there is a coupling problem between dual networks with the same architecture. Some early methods [56, 57]try to solve this problem by selecting different training samples, using different initializations or different data augmentation methods. However, because the network structure remains unchanged, as the training progresses, there will inevitably be more and more similar between the dual networks. Later, some methods [58, 59] try to improve the problem by changing the network structure, but they are all at the convolutional level. Limited by the size of the receptive field, the ability of convolution to capture global perception is still insufficient.

To address this problem, we propose a heterogeneous network framework to improve the difference and complementarity between networks. The overall framework is illustrated in Fig. 1. The framework consists of two heterogeneous networks. Both networks use ResNet50 as the backbone network. Among them, network 1 (Net1) directly utilizes the backbone network to obtain local perception, while network 2 (Net2) obtains global perception by introducing a Transformer Encoder block. For the Transformer Encoder, as shown in the lower right corner of Fig. 1, the feature $f \in \mathbb{R}^{B\times C\times H\times W}$ extracted by the backbone network is first flattened into N patches according to the spatial dimension, and the embedded position encoding for each patch is sent to the Transformer Encoder. Each encoder layer consists of an effective multi-head self-attention (EMSA) module [33] and a feed-forward network (FFN), which finally obtain enhanced features with global information. In order to further enhance the semantic information of the network, the local features extracted by Net1 and the global features extracted by Net2 are sent to the ACMA module (datails are mentioned in section "Adaptive Channel Mutual-Aware (ACMA) Module"). Finally, global max pooling (GMP) and global average pooling (GAP) are performed respectively to obtain the final feature map, where the GAP perceives the whole features, while the GMP focuses on the salient features, both of which can be fused to obtain a better description vector. At the same time, in order to alleviate the large amount of noise caused by the pseudo-labels generated by clustering, the Mean Teacher Model [60] is adopted in our framework, where the teacher (with teacher weight θ') is composed of the exponential moving average (EMA) weights of the student (with student weight θ), and the student is supervised by the teacher. Specifically, the parameters of the teacher models of the two networks at current iteration T are denoted as $\theta_1'(T)$ and $\theta_2'(T)$, which are updated as:

$$\begin{aligned}\theta_1'(T) &= \alpha\theta_1'(T-1) + (1-\alpha)\theta_1(T),\\ \theta_2'(T) &= \alpha\theta_2'(T-1) + (1-\alpha)\theta_2(T),\end{aligned} \tag{4}$$

where α is a smoothing coefficient that controls the self-ensembling speed of the Mean Teacher.

Both hard pseudo-labels (100% confidence labels generated by clustering) and soft pseudo-labels (confidence <100% predicted by the teacher network) are used to optimize the model in the whole training process. To

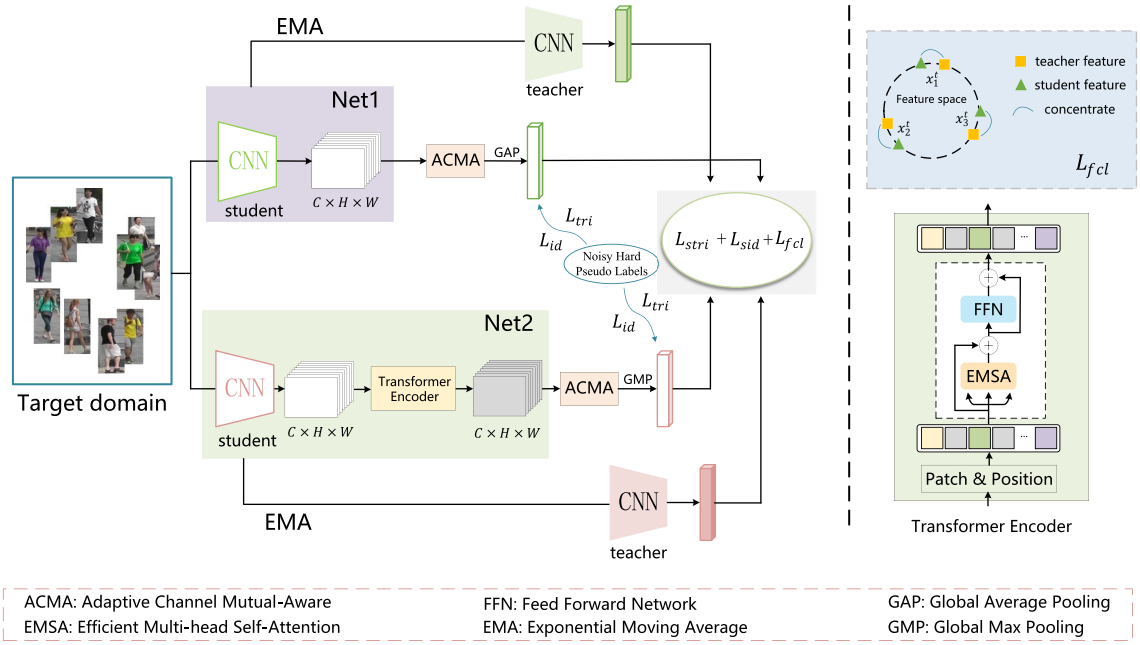

Fig. 1 The framework of our method. First, the framework use two heterogeneous network models pre-trained on the labeled source domain to extract features for the target domain, and then use the hard pseudo-labels generated by clustering and the soft pseudo-labels generated by the opposite teacher model to fine-tune the model with the classification loss ($\mathcal{L}_{id}, \mathcal{L}_{sid}$) and triplet loss ($\mathcal{L}_{tri}, \mathcal{L}_{stri}$), together with a feature consistency loss ($\mathcal{L}_{fcl}$). The above two steps are iteratively performed until the network converges

simultaneously train the dual networks, we provide the same batch of images to two different networks. Each target-domain image can be denoted by x_i^t, and the pseudo-label confidences can be predicted as $p_1(x_i^t|\theta_1)$ and $p_2(x_i^t|\theta_2)$, and the feature transformation function can be expressed as $f_1(x_i^t|\theta_1)$ and $f_2(x_i^t|\theta_2)$.

Therefore, to transfer knowledge from the teacher model to the student model, the class predictions of the teacher model can serve as soft pseudo-labels for training the student model. The probability for each identity i is predicted as $p_1(x_i^t|\theta_1')$ and $p_2(x_i^t|\theta_2')$. So the soft classification loss for optimizing θ_1 and θ_2 with the soft pseudo-labels generated from the teacher network can therefore be formulated as:

$$\begin{aligned}\mathcal{L}_{sid}^t(\theta_1|\theta_2) &= -\frac{1}{N_t}\sum_{i=1}^{N_t}(p_2(x_i^t|\theta_2')logp_1(x_i^t|\theta_1)),\\ \mathcal{L}_{sid}^t(\theta_2|\theta_1) &= -\frac{1}{N_t}\sum_{i=1}^{N_t}(p_1(x_i^t|\theta_1')logp_2(x_i^t|\theta_2)).\end{aligned} \tag{5}$$

To further enhance the discriminating ability of the teacher-student network, we use the features extracted by the teacher model to supervise the features of the student model with a soft triplet loss:

$$\begin{aligned}\mathcal{L}_{stri}^t(\theta_1|\theta_2) &= -\frac{1}{N_t}\sum_{i=1}^{N_t}[Q_i(\theta_1)logQ_i(\theta_2')],\\ \mathcal{L}_{stri}^t(\theta_2|\theta_1) &= -\frac{1}{N_t}\sum_{i=1}^{N_t}[Q_i(\theta_2)logQ_i(\theta_1')],\end{aligned} \tag{6}$$

where $Q_i(\theta)$ denotes the softmax triplet distance of the samples x_i^t. The mini-batch softmax triplet distance from the student is encouraged to be as close as possible to the distance from the teacher by minimizing soft triplet loss. The formula of $Q_i(\theta)$ is as follows:

$$Q_i(\theta) = \frac{\exp(\|f(x_i^t|\theta) - f(x_{i,n}^t|\theta)\|)}{\exp(\|f(x_i^t|\theta) - f(x_{i,p}^t|\theta)\|) + \exp(\|f(x_i^t|\theta) - f(x_{i,n}^t|\theta)\|)}, \tag{7}$$

where $x_{i,p}^t$ denotes the hardest positive sample, and $x_{i,n}^t$ denotes the hardest negative sample of the anchor x_i^t according to the pseudo-labels, $\|\cdot\|$ denotes L_2 distance. At the same time, in order to strengthen the constraint of the teacher model on the student model in the feature space, we propose the feature consistency loss denoted as $\mathcal{L}_{fcl}^t$ and datails are mentioned in section "Feature consistency loss (FCL)".

Therefore, the overall loss function $\mathcal{L}(\theta 1, \theta 2)$ simultaneously optimizes the dual networks, which combines Eq. 1, Eq. 5, Eq. 6, Eq. 9 and is formulated as:

$$\begin{aligned}\mathcal{L}(\theta 1, \theta 2) = &(1-\lambda_{id}^{t})\mathcal{L}_{id}^{t} + \lambda_{id}^{t}\mathcal{L}_{sid}^{t} + (1-\lambda_{tri}^{t})\mathcal{L}_{tri}^{t} \\ &+ \lambda_{tri}^{t}\mathcal{L}_{stri}^{t} + \lambda_{fcl}^{t}\mathcal{L}_{fcl}^{t},\end{aligned} \tag{8}$$

where λ_{id}^{t}, λ_{tri}^{t}, λ_{fcl}^{t} are the weighting parameters.

3.2 Feature consistency loss (FCL)

For the person re-ID task, the identities in testing set are usually different with training set and re-ID is performed as retrieval by matching the extracted features of pedestrian. Therefore, it is meaningful to explore the optimization of feature space. However, the traditional UDA re-ID task ignores the consistency of the sample in the feature space. Specifically, the classification loss only optimizes the network in the class prediction space, and focuses on the relationship between samples and pseudo-labels. In addition, the triplet loss relies on pseudo-labels to select positive and negative samples, so the noise in pseudo-labels will mislead the selection and optimization of samples.

Therefore, to make up for the shortcomings above and further strengthen the constraints, we propose the feature consistency loss (FCL), which is performed in feature space and focuses on similarity relationship among samples. Especially, FCL does not require any label information, which effectively avoids the influence of noise pseudo-labels. FCL is computed to encourage teacher feature to supervise student feature which is formulated as:

$$\begin{aligned}\mathcal{L}_{fcl}^{t}(\theta_1|\theta_2) &= -\frac{1}{N_t}\sum_{i=1}^{N_t}(f_2(x_i^t|\theta_2')log f_1(x_i^t|\theta_1)), \\ \mathcal{L}_{fcl}^{t}(\theta_2|\theta_1) &= -\frac{1}{N_t}\sum_{i=1}^{N_t}(f_1(x_i^t|\theta_1')log f_2(x_i^t|\theta_2)),\end{aligned} \tag{9}$$

where $f(x_i^t|\theta')$ represents the features extracted by the teacher model, and $f(x_i^t|\theta)$ represents the features extracted by the student model. Student network is encouraged by FCL to output features that have the similar distribution as the teacher network. In short, FCL further supplements the classification loss and overcomes the limitation that the triplet loss is affected by noise pseudo-labels.

The feature consistency loss aims to narrow the distance of the same sample features extracted by the student model and the teacher model in the feature space, as shown in the upper right corner of Fig. 1. The triangles represent the features extracted by the teacher model, and the squares represent the features extracted by the student model, and x_i^t represents the i-th sample in the target domain.

3.3 Adaptive channel mutual-aware (ACMA) module

For the person re-ID task, paying more attention to semantic information is very important to improve the performance of the model. In recent years, channel attention has made remarkable achievements in the field of computer vision. However, some existing methods rarely consider the global and local information between channels simultaneously.

To overcome the above problems, we propose the ACMA module, which can maintain the diversity between channels, the structure of which is shown in Fig. 2. This module consists of two branches, which are used to simultaneously focus on the local and global information of the channel. Finally, the two are combined to realize the mutual awareness of different information between channels.

Firstly, to reduce the computational complexity, we divide the feature map $I \in R^{C\times H\times W}$ into two groups according to the channel dimension, where C, H and W represent its number of channels, height and width respectively. And then send the two feature maps $A \in R^{C'\times H\times W}$ and $B \in R^{C'\times H\times W}$ to two branches, where we set $C' = C/2$. In the global branch, A go through GAP for global context modeling, and then enters 1×1 convolution layer to obtain the global relationship weight between channels. Subsequently, multiplying the obtained weight matrix w_g with the original feature map A to obtain the weighted feature map $M \in R^{C'\times H\times W}$. In the local branch, the aggregated features obtained by GAP of B are sent to a 1D convolution of kernel size k to compute the relationship between the local channels, where kernel size k represents the coverage of local cross-channel interaction. Similarly, multiplying the obtained weight matrix w_l with the feature map B, which is represented by the matrix $N \in R^{C'\times H\times W}$. Next, concatenating M and N together, and then in order to further improve the cross-channel information interaction, a channel shuffle operation is performed. Finally, the output matrix $O \in R^{C\times H\times W}$ is obtained. Since the input and output dimensions of the module remain the same, that is a plug-and-play module, it can be applied to different structures. The formula is as follows:

$$O = SC(Concat(F(w_g, A), F(w_l, B))), \tag{10}$$

where $F(\cdot,\cdot)$ represents element-wise multiplication, $Concat(\cdot,\cdot)$ represents stacking feature maps along the channel dimension, and $SC(\cdot,\cdot)$ represents channel shuffle operation.

3.4 Explore the coupling between dual networks

To address the problem of noise pseudo-labels, a dual network mutual learning architecture is proposed. Taking

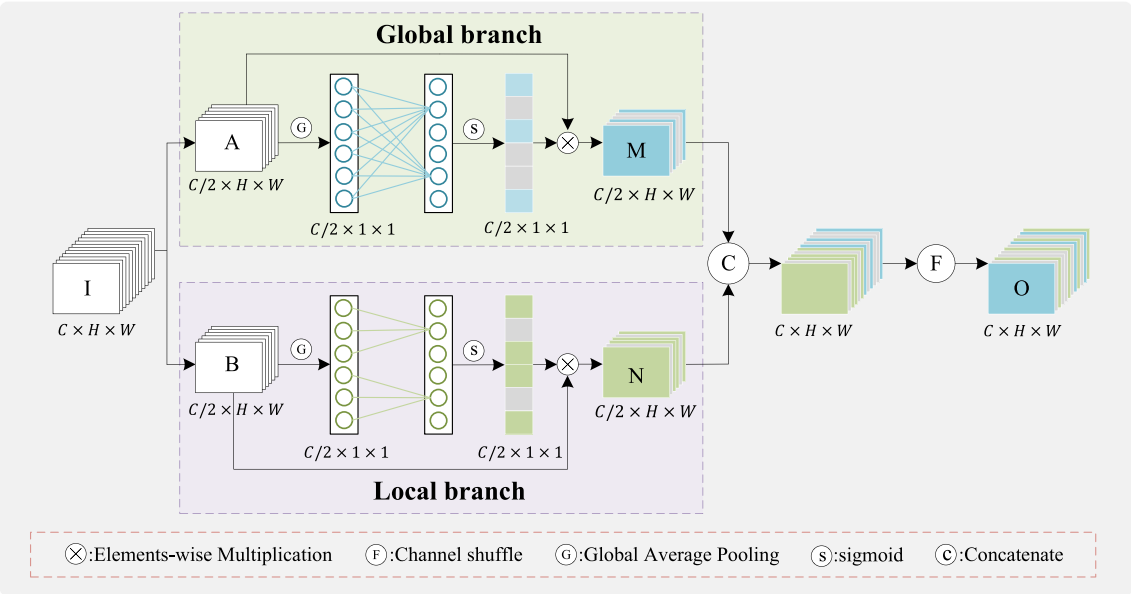

Fig. 2 Overview of proposed Adaptive Channel Mutual-Aware (ACMA) module, which consists of two branches, local branch and global branch

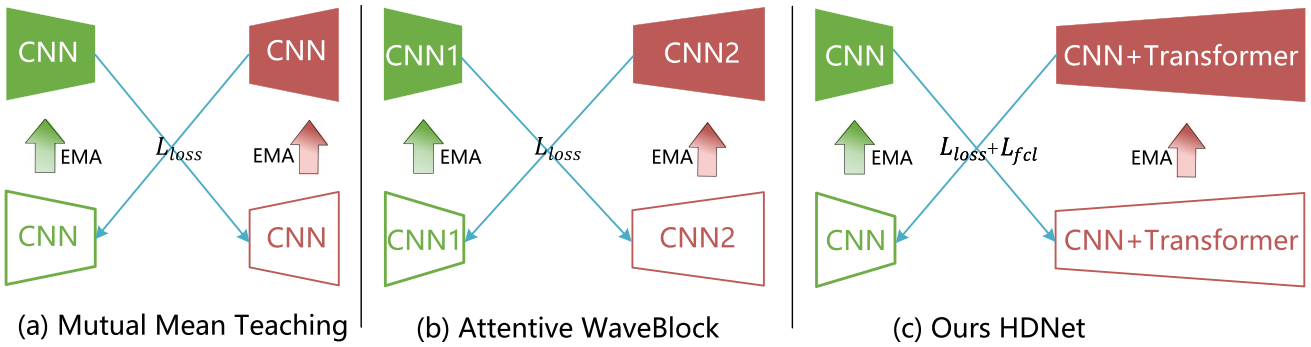

Fig. 3 Comparison of different dual network architectures. Our method follows the training strategy of MMT. The difference is that HDNet uses heterogeneous dual networks for collaborative training and introduces a Transformer module to capture long-range dependency. $\mathcal{L}_{loss}$ represents the classification loss and the triplet loss, $\mathcal{L}_{fcl}$ represents the feature consistency loss

MMT as an example, as shown in Fig. 3a, it uses the temporally average model of two networks to generate more reliable pseudo-labels for collaborative training between two networks. It is a pity that it only uses different seeds to pretrain two networks with exactly the same architecture, it will inevitably make the two networks converge to each other along with training. Thus, to solve this problem, some recent works [58, 59] have improved the dual network architecture, such as AWB [58]. It creates differences (similar ideas exist in other fields [61, 62]) between the two networks by introducing the WaveBlock, and uses the attention mechanism to amplify the differences and find more complementary features, as shown in Fig. 3b. Although these methods alleviate the coupling problem to some extent, they are limited to the convolution level. However, limited by the limited receptive field, CNN's ability to capture global information is still insufficient. Therefore, we introduce the Transformer module to obtain long-range dependency. One network only uses the basic CNN, while the other network combines CNN with Transformer, as shown in Fig. 3c. HDNet not only improves the heterogeneity between dual networks, but also makes up for the defect that CNN can only capture local information. Comparing with convolutions, EMSA in Transformers can capture long-range dependencies and attend diverse information from a global view. Meanwhile, Transformers can preserve the semantic information in interaction among different scale features. Therefore, combining CNN and Transformer to achieve the heterogeneity of the network allows each other to focus on different characteristics.

In addition, the existing dual-network architecture methods only use the classification loss and the triplet loss to optimize the network. However, the classification loss focuses more on the relationship between samples and pseudo-labels, ignoring the relationship between samples. Besides, the triplet loss needs to rely on pseudo-labels to select positive and negative samples. Therefore, in order to further supplement the classification loss and optimize the triplet loss, we propose the feature consistency loss, which pays more attention to the consistency of samples in the feature space and does not need any label information, making it more suitable for UDA re-ID tasks.

4 Experiments

4.1 Datasets and evaluation protocol

We evaluate our method on three datasets that are widely used in person re-ID tasks, including DukeMTMC-reID [63, 64], Market-1501 [65] and MSMT17 [29] . DukeMTMC-reID dataset contains 36411 images of 1812 identities captured by 8 cameras, where the training set has 702 identities and contains 16522 images and the test set has another 702 identities. Market-1501 dataset contains 1501 pedestrians captured by 6 cameras, where the training set has 12936 images of 751 identities and the test set has 19732 images of 750 identities. MSMT17 dataset contains 4101 pedestrians with 126441 bounding boxes captured by 15 cameras. The training set contains 1041 pedestrians with a total of 32621 bounding boxes, while the test set includes 3060 pedestrians with a total of 93820 bounding boxes. For the test set, 11659 bounding boxes were randomly selected as the query, and the other 82161 bounding boxes were used as the gallery. Cumulative Matching Characteristics (CMC) [66] and mean Average Precision (mAP) [65] are used to evaluate the performance. Rank1, Rank5 and Rank10 accuracies in CMC are reported.

4.2 Implementation detail

HDNet is trained by two stages: pre-training in source domains and the adaptation in target domains. For these two stages, all images are resized to 256×128, and traditional image augmentation is performed via random erasing. In addition, we use the Adam [67] optimizer with a weight decay of 0.0005 to optimize the parameters.

Stage 1: Pre-training in source domains: We adopt ResNet50 [23] as the backbone network, and initialize two heterogeneous networks by using parameters pre-trained on the ImageNet [68]. Each mini-batch contains 64 source-domain images of 16 ground-truth identities (4 for each identity). The initial learning rate is set to 0.00035 and is decreased to 1/10 of its previous value on the 40th and 70th epoch in the total 80 epochs.

Stage 2: Adaptation in target domains. The smoothing coefficient α in Eq. 4 is set to 0.999, which is used to initialize and update the Mean Teacher network. The two heterogeneous networks are collaboratively updated by optimizing Eq. 8 with the loss weights $\lambda_{id}^t = 0.5$, $\lambda_{tri}^t = 0.8$ and $\lambda_{fcl}^t = 0.8$. The k in the ACMA module is set to 5. The target domain adaptive training iterates for 40 epochs and the learning rate is fixed to 0.00035. Each mini-batch contains 64 target-domain images of 4 pseudo identities. We use *DBSCAN* and *K-means* clustering methods to assign hard label for our model. The number of pseudo-label classes for four domain adaptive tasks when using *K-means* is 500, 700, 1000 and 1500, respectively. Each epoch consists of 800 training iterations. During testing, we only use one of the best teacher networks on the target domain dataset for feature representations.

4.3 Parameter analysis

In this subsection, we analyze the influence of the hyper-parameter λ_{fcl}^t in Eq. 8 and the k in the ACMA module on the performance of the model.

The λ_{fcl}^t of the $\mathcal{L}_{fcl}^t$. In our experiments, we change one parameter while keeping the others fixed. For Eq. 8, keep $\lambda_{id}^t = 0.5$ and $\lambda_{tri}^t = 0.8$ unchanged, and only change the size of λ_{fcl}^t. To this end, we conduct experiments on the Duke → Market task and set λ_{fcl}^t from 0 to 1. The results are shown in Fig. 4. Coefficient λ_{fcl}^t is used to control the proportion of L_{fcl}^t in the total loss. We can know from section "Feature consistency loss (FCL)" that the FCL aims to focus on the consistency of samples in feature space. When λ_{fcl}^t is too small, the loss function will tend to pay more attention to the relationship between samples and labels, ignoring the relationship between samples, which will affect the performance of the model. Specially, when $\lambda_{fcl}^t = 0$, that is, when FCL is not used, the model has the worst effect. When λ_{fcl}^t is too large, the loss function will pay too much attention to the relationship between samples, which will lead to the decline of model performance. Therefore, it is necessary to find a suitable λ_{fcl}^t. As can be seen from the Fig. 4, our method is insensitive to parameter λ_{fcl}^t. Finally, We take the best result and set $\lambda_{fcl}^t = 0.8$.

The k in the ACMA module. In this section, we evaluate the impact of k value on the ACMA module. According to section "Adaptive Channel Mutual-Aware (ACMA) Module", k represents the coverage of local cross-channel interaction. In the experiment, we set k values from 1 to 9. The results are shown in Fig. 5. It can be seen from the figure that when k value is set to 5, the experimental results reach

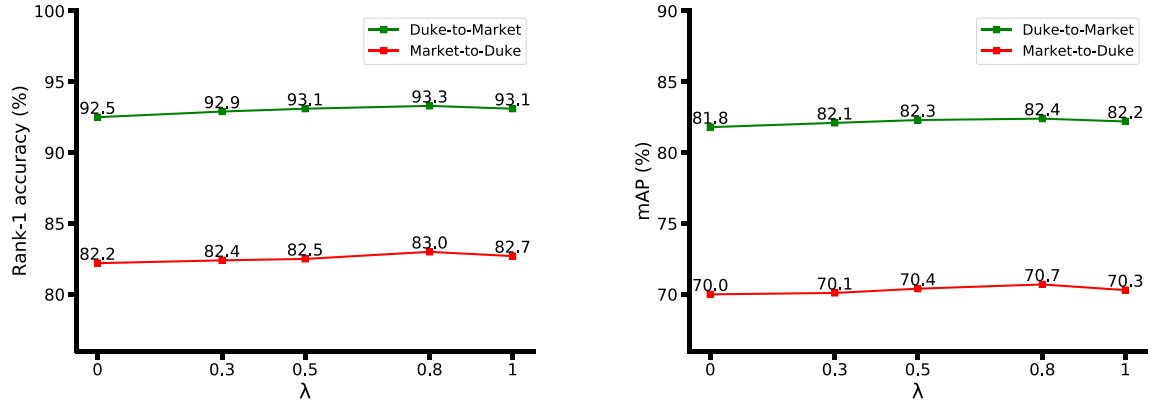

Fig. 4 Evaluation with diferent values of λ_{fcl}^{t} in Eq. 8

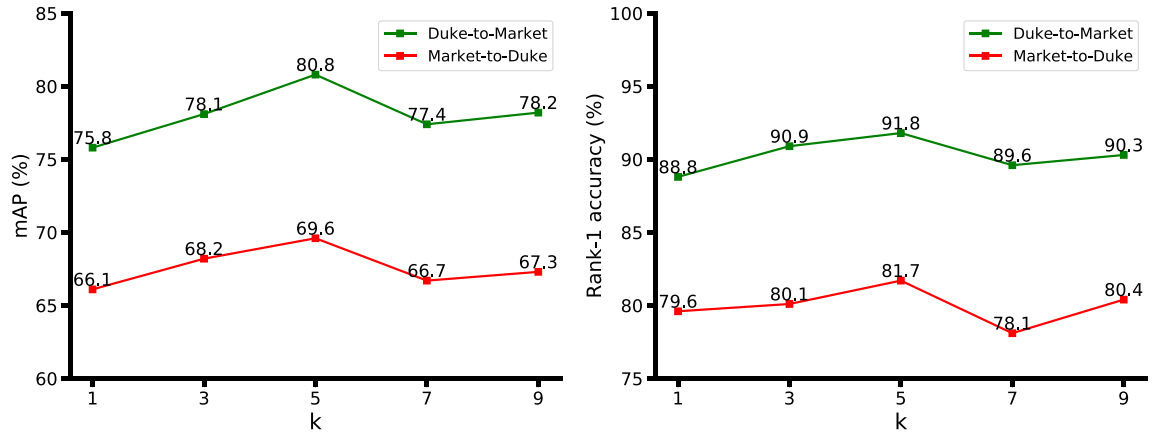

Fig. 5 Evaluation with diferent values of k in the ACMA module

the best, and the results gradually decrease with the continuous increase of k value. When the value of k is less than 5, the results also show a downward trend, which we blame it to the insufficient interaction between channel information to provide more effective information when k is too small. The above results show that the k value has a significant impact on the performance of ACMA module, so it is very important to select an appropriate k value.

4.4 Ablation studies

In this section, we evaluate the effectiveness of the different components of our method and take MMT as the baseline. Results are shown in Table 1. In order to keep consistent with the Baseline, IBN-ResNet50 [69] is also used as the backbone network.

Effectiveness of the Heterogeneous-network framework. We first evaluate the Heterogeneous-network (HN) framework as described in section "Unsupervised adaptation in target domain". For this experiment, we designed a network Baseline + HN that just incorporates the HN into the Baseline. Results are shown in Table 1. We can observe that our proposed HN bring the most significant performance improvement during the adaptation. For Duke → Market, the Baseline + HN achieves a rank-1 accuracy of 92.9% and an mAP of 82.0% which are higher than the Baseline by 2.0% and 5.5%, respectively. The improvement of model performance is explainable,

Table 1 Ablation studies of our proposed HDNet on Duke-to-Market and Market-to-Duke tasks

Methods	Duke→Market(%)		Market→Duke(%)	
	mAP	R1	mAP	R1
Baseline	76.5	90.9	68.7	81.8
Baseline+HN	82.0	92.9	69.9	82.3
Baseline+$\mathcal{L}_{fcl}^{t}$	81.2	92.5	69.8	82.8
Baseline+ACMA	80.8	91.8	69.6	81.7
Our HDNet(full)	82.4	93.3	70.8	83.4

because the asymmetric network structure can improve the complementarity between models, thereby preventing them from biasing towards the same kind of noise and reducing the influence of noise labels on the model, so as to achieve the effect of improving the performance of the model.

Effectiveness of the Feature Consistency Loss. $\mathcal{L}^t_{fcl}$ aims to optimize the model in feature space and pay more attention to the similar relationship between samples. Results are shown in Table 2. First of all, it can be seen from the first row of the Table 2 that the model achieves the worst effect without adding the $\mathcal{L}^t_{fcl}$. When the model is constrained by three losses at the same time, the experimental effect is the best. This also verifies the effectiveness of $\mathcal{L}^t_{fcl}$. To verify the complementary relationship of the $\mathcal{L}^t_{fcl}$ and the $\mathcal{L}^t_{sid}$, compare the first row, the third row and the fourth row in the Table 2, We can see that whether the $\mathcal{L}^t_{fcl}$ or the $\mathcal{L}^t_{sid}$ is used alone, the effect is not as good as the combination of the two, which effectively validate the necessity of the combination of the two. It also shows that the previous methods pay more attention to the similarity of the class prediction space and ignore the similarity relationship between the samples in the feature space, leading to the low performance of the model. Furthermore, in order to illustrate the improvement of $\mathcal{L}^t_{stri}$ by our proposed loss from the experimental results, we add ablation experiments, as shown in the first and second rows of the Table 2. The experimental results reflect the restriction of pseudo-labels on the performance improvement of the model to a certain extent.

Effectiveness of the Adaptive Channel Mutual-Aware module. Our ACMA is a plug-and-play module that can be plugged into any network architecture. We evaluate the the ACMA as described in Sect. 3.3. For this experiment, we design a new network Baseline + ACMA which incorporates ACMA into the network. As shown in Table 1, the incorporation of ACMA improves the person re-ID performance by enhancing the semantic information of the network. For Duke → Market, the Baseline + ACMA achieves a rank-1 accuracy of 91.8% and an mAP of 80.8% which are higher than the Baseline by 0.9% and 4.3%, respectively. The experimental results show that the attention module is effective for enhancing the model generalization and adaptation.

4.5 Comparison with the state-of-the-art methods

We compare our proposed framework with the state-of-the-art methods on the four domain adaptation tasks, Market-to-Duke, Duke-to-Market, Market-to-MSMT17 and Duke-to-MSMT17. In order to better verify the effectiveness of our method, we compare with various methods in unsupervised domain adaptation. The results are shown in Table 3.

First of all, the methods of fine-tuning the network using the pseudo-labels generated by clustering, such as PCB-PAST [14], SSG [15], AD-Cluster [18], Dual-Refinement [76] and GLT [19], our model greatly improves the performance of UDA re-ID. Secondly, compared with the dual-network methods including MMT [22] and NRMT [72], our model also has greater advantages, increasing the mAP by 4.6% and 9.4% respectively in the Duke-to-Market domain adaptive task, 1.1% and 7.6% respectively in the Market-to-Duke domain adaptive task. Furthermore, our HDNet adopts two models to significantly surpass the three-model methods MEB-Net [73] that uses the same backbone, showing a noticeable 5.1% improvement in terms of mAP in the Duke-to-Market domain adaptive task and 3.7% improvements in the Market-to-Duke domain adaptive task. Similarly, we have also surpassed other methods that also use Mean-Teacher, such as UNRN [75] with a 3.0% higher mAP index in the Duke-to-Market domain adaptive task. Moreover, even if the source domain data is also used in the target domain fine-tuning model stage, such as SpCL [74], we still have significant advantages. Finally, compared with the same method of exploring the dual-network coupling problem, in which ACT [56] inputs different training samples to the two networks, and ABMT [59] introduces a bottleneck layer to construct heterogeneous branches, our method still has great advantages. This is explainable because the previous methods are only at the convolution level. Limited by the limited receptive field, CNN cannot capture the global information of pedestrians well, so we use Transformer to capture the long-range dependency to improve the performance of the model.

For the more challenging large-scale dataset MSMT17, our model still performs well. The results are shown in Table 4. It can be seen from the table that the recognition accuracy of our method is significantly higher than that of other methods in recent years. For example, compared with

Table 2 Ablation studies among the three losses on Duke-to-Market and Market-to-Duke tasks

Setting	Methods			Duke→Market(%)		Market→Duke(%)	
	$\mathcal{L}^t_{sid}$	$\mathcal{L}^t_{stri}$	$\mathcal{L}^t_{fcl}$	mAP	R1	mAP	R1
1	✓	✓		76.5	90.9	68.7	81.8
2	✓		✓	78.2	91.0	69.1	81.7
3		✓	✓	80.2	91.8	69.4	81.6
4	✓	✓	✓	81.2	92.5	69.8	82.8

Table 3 Comparison with state-of-the-art methods: For the adaptation on Market-1501 (Market), DukeMTMC-reID (Duke)

Methods	Duke → Market(%)				Market → Duke(%)			
	mAP	R1	R5	R10	mAP	R1	R5	R10
PCB-PAST [14]	54.6	78.4	–	–	54.3	72.4	–	–
SSG [15]	58.3	80.0	90.0	92.4	53.4	73.0	80.6	83.2
MMCL [21]	60.4	84.4	92.8	95.0	51.4	72.4	82.9	85.0
ACT [56]	60.6	80.5	–	–	54.5	72.4	–	–
ECN-GPP [11]	63.8	84.1	92.8	95.4	54.4	74.0	82.9	85.0
JVTC+ [70]	67.2	86.8	95.2	97.1	66.5	80.4	89.9	92.2
AD-Cluster [18]	68.3	86.7	94.4	96.5	54.1	72.6	82.5	85.5
CAIL [71]	71.5	88.1	94.4	96.2	65.2	79.5	88.3	91.4
NRMT [72]	71.7	87.8	94.6	96.5	62.2	77.8	86.9	89.5
MEB-Net [73]	76.0	89.9	96.0	97.5	66.1	79.6	88.3	92.2
MMT [22]	76.5	90.9	96.4	97.9	68.7	81.8	91.2	93.4
SpCL [74]	76.7	90.3	96.2	97.7	68.8	82.9	90.1	92.5
UNRN [75]	78.1	91.9	96.1	97.8	69.1	82.0	90.7	93.5
AWB [58]	81.0	93.5	97.4	98.3	70.9	83.8	92.3	94.0
Dual-Refinement [76]	78.0	90.9	96.4	97.7	67.7	82.1	90.1	92.5
ABMT [59]	78.3	92.5	–	–	69.1	82.0	–	–
GLT [19]	79.5	92.2	96.5	97.8	69.2	82.0	90.2	92.8
Ours ResNet *DBSCAN*	**81.1**	**93.5**	**97.5**	98.0	**69.8**	**82.6**	90.8	93.0
Ours ResNet *K-means*	79.5	92.0	97.2	**98.3**	68.7	81.2	**90.9**	**93.3**
Ours IBN-ResNet *DBSCAN*	**82.4**	**93.3**	**97.7**	**98.6**	70.8	**83.4**	91.7	93.8
Ours IBN-ResNet *K-means*	80.1	92.9	96.7	98.2	**71.1**	82.7	**91.8**	**94.1**

The best performance is shown in bold

Table 4 Comparison with state-of-the-art methods: For the adaptation on Market-1501 (Market), DukeMTMC-reID (Duke) and that on MSMT17

Methods	Market → MSMT17(%)				Duke → MSMT17(%)			
	mAP	R1	R5	R10	mAP	R1	R5	R10
MMCL [21]	15.1	40.8	51.8	56.7	16.2	43.6	54.3	58.9
ECN-GPP [11]	15.2	40.4	53.1	58.7	16.0	42.5	55.9	61.5
NRMT [72]	19.8	43.7	56.5	62.2	20.6	45.2	57.8	63.3
CAIL [71]	20.4	43.7	56.1	61.9	24.3	51.7	64.0	68.9
MMT [22]	22.9	49.2	63.1	68.8	23.3	50.1	63.9	69.8
JVTC+ [70]	25.1	48.6	65.3	68.2	27.5	52.9	70.5	75.9
SpCL [74]	26.8	53.7	65.0	69.8	26.5	53.1	65.8	70.5
UNRN [75]	25.3	52.4	64.7	69.7	26.2	54.9	67.3	70.6
AWB [58]	29.0	57.3	70.7	75.9	28.1	56.8	70.1	75.2
ABMT [59]	23.2	49.2	–	–	26.5	54.3	–	–
Dual-Refinement [76]	25.1	53.3	66.1	71.5	26.9	55.0	68.4	73.2
GLT [19]	26.5	56.6	67.5	72.0	27.7	59.5	70.1	74.2
Ours ResNet *DBSCAN*	25.9	53.4	66.4	**72.1**	26.8	54.6	**70.9**	73.0
Ours ResNet *K-means*	**26.7**	**57.6**	**67.8**	72.0	**28.7**	**56.2**	70.8	**74.3**
Ours IBN-ResNet *DBSCAN*	28.4	58.3	69.7	75.6	29.8	59.4	71.7	74.8
Ours IBN-ResNet *K-means*	**32.6**	**61.5**	**73.4**	**77.8**	**33.7**	**61.5**	**74.0**	**78.6**

The best performance is shown in bold

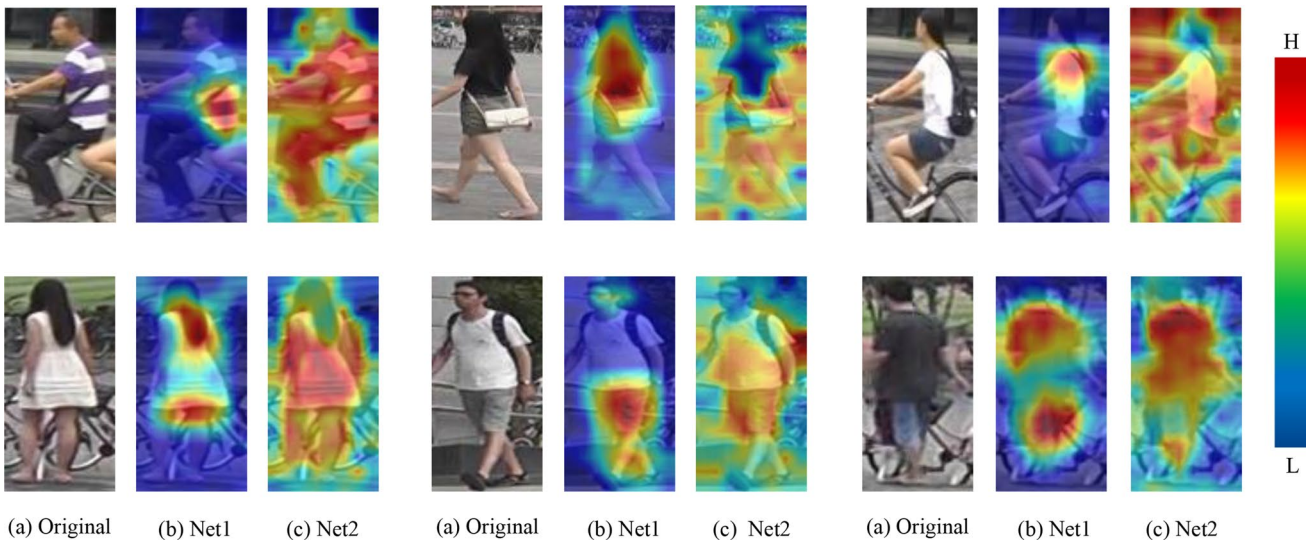

Fig. 6 The gradient-weighted class activation maps of the heterogeneous dual networks. **a** The original image. **b** The original network based on CNN. **c** The introduction of the Transformer encoder block. (H: High values; L: Low values)

the same dual network method ABMT [59], our method achieves 26.7% and 28.7% in mAP accuracy on Market-to-MSMT17 and Duke-to-MSMT17 tasks, which are better than ABMT [59] by 3.5% and 2.2% respectively. All the above experimental results can show the effectiveness of our proposed method. In addition, taking IBN-ResNet50 [69] as the backbone network, the performance of our model is further improved on four domain adaptive tasks.

4.6 Visualization

We evaluate our overall framework from qualitative perspectives. In order to verify the heterogeneity of the dual networks in the HDNet framework, we make gradient-weighted class activation maps. As shown in the Fig. 6, we can see that the two networks pay attention to the different characteristics of pedestrians. Among them, Net1 uses pure CNN to capture the local salient features of pedestrians, and Net2 combines CNN and Transformer to obtain overall pedestrian information. By extracting different feature information, the coupling between the two networks is suppressed. For example, in the first picture in the upper left corner, Net1 pays more attention to the upper body of pedestrian (such as striped jacket), while Net2 pays more attention to the whole pedestrian.

For intuitive understanding, we visualize the top 1–5 retrieval results on the Duke-to-Market task. It can be seen from the Fig. 7 that the Baseline network can not capture the global and local information of pedestrian at the same time. Taking the last pedestrian as an example, the Baseline network only focuses on the local similarity features, such as green clothes and black backpacks, which are wrongly regarded as the target pedestrian, while our HDNet also focuses on the global dissimilarity features, such as blue and gray pants. It effectively avoids similar but mismatched pedestrians, so the retrieval accuracy is greatly improved.

5 Conclusion

In this paper, we propose a heterogeneous dual network framework, which aims to enhance the robustness to pseudo-label noise by improving the heterogeneity of the two networks. In addition, we discuss the issue of feature space consistency, and propose the feature consistency loss, which breaks away from the bondage of pseudo-label. Furthermore, adaptive channel mutual-aware module is proposed to enhance the semantic information of the network. In real scenarios, re-ID tasks can only achieve good performance on the same benchmark. Although UDA re-ID can improve the application of the model in different scenarios to some extent, but it is still a field that has not been fully explored. In the future, how to improve the

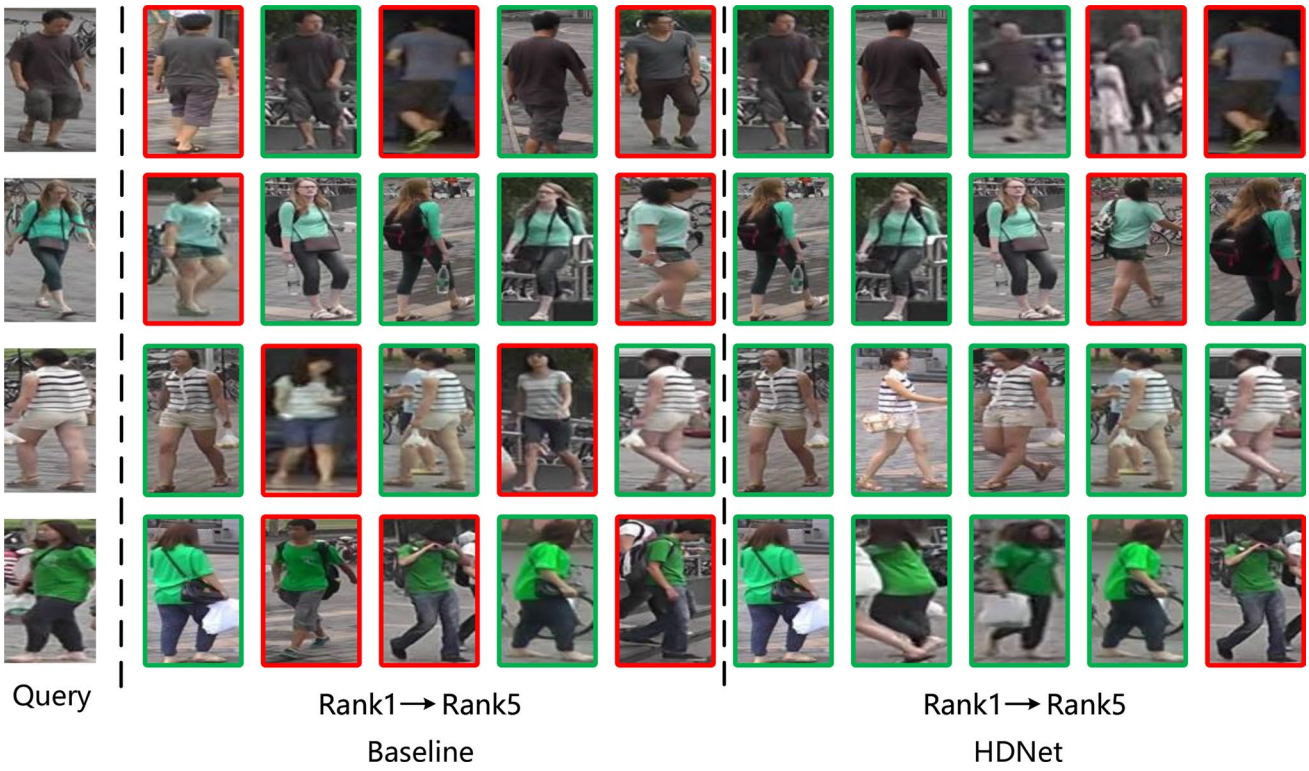

Fig. 7 Visualization of retrieval results. We show the top 1–5 retrieval results of Baseline and HDNet, where the first column represents the query image. They are sorted from left to right according to their similarity to the query image (from high to low). Retrieved images with green and red boxes denote right and wrong results respectively

reliability of pseudo-label is a key exploration direction, which is worthy of further study.

Acknowledgements This work was partially supported by the Fundamental Research Funds for the Central Universities (JUSRP41908), the National Natural Science Foundation of China (61362030, 61201429), China Postdoctoral Science Foundation (2015M581720, 2016M600360), 111 Projects under Grant B12018.

Availability of data and materials The datasets analysed during the current study are available in the Ref [63–65] and [29].

References

1. Zheng L, Yang Y, Hauptmann AG (2016) Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984
2. Han C, Zheng R, Gao C, Sang N (2019) Complementation-reinforced attention network for person re-identification. IEEE Trans Circuits Syst Video Technol 30(10):3433–3445
3. Huang Y, Huang Y, Hu H, Chen D, Su T (2019) Deeply associative two-stage representations learning based on labels interval extension loss and group loss for person re-identification. IEEE Trans Circuits Syst Video Technol 30(12):4526–4539
4. Kong J, He Q, Jiang M, Liu T (2021) Dynamic center aggregation loss with mixed modality for visible-infrared person re-identification. IEEE Signal Process Lett 28:2003–2007
5. Gheisari M, Najafabadi HE, Alzubi JA, Gao J, Wang G, Abbasi AA, Castiglione A (2021) Obpp: an ontology-based framework for privacy-preserving in iot-based smart city. Fut Gen Comput Syst 123:1–13
6. Ding Y, Fan H, Xu M, Yang Y (2020) Adaptive exploration for unsupervised person re-identification. ACM Trans Multimed Comput Commun Appl (TOMM) 16(1):1–19
7. Zhong Z, Zheng L, Luo Z, Li S, Yang Y (2019) Invariance matters: Exemplar memory for domain adaptive person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 598–607
8. Tao X, Kong J, Jiang M, Liu T (2021) Unsupervised domain adaptation by multi-loss gap minimization learning for person re-identification. In: IEEE Transactions on Circuits and Systems for Video Technology
9. Liu J, Zha Z-J, Chen D, Hong R, Wang M (2019) Adaptive transfer network for cross-domain person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 7202–7211
10. Li Y-J, Lin C-S, Lin Y-B, Wang Y-CF (2019) Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 7919–7929
11. Zhong Z, Zheng L, Luo Z, Li S, Yang Y (2020) Learning to adapt invariance in memory for person re-identification. In: IEEE transactions on pattern analysis and machine intelligence
12. Wang J, Zhu X, Gong S, Li W (2018) Transferable joint attribute-identity deep learning for unsupervised person

re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2275–2284
13. Yang F, Yan K, Lu S, Jia H, Xie D, Yu Z, Guo X, Huang F, Gao W (2020) Part-aware progressive unsupervised domain adaptation for person re-identification. IEEE Trans Multimed 23:1681–1695
14. Zhang X, Cao J, Shen C, You M (2019) Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 8222–8231
15. Fu Y, Wei Y, Wang G, Zhou Y, Shi H, Huang TS (2019) Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 6112–6121
16. Song L, Wang C, Zhang L, Du B, Zhang Q, Huang C, Wang X (2020) Unsupervised domain adaptive re-identification: theory and practice. Pattern Recognit 102:107173
17. Kumar D, Siva P, Marchwica P, Wong A (2020) Unsupervised domain adaptation in person re-id via k-reciprocal clustering and large-scale heterogeneous environment synthesis. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp 2645–2654
18. Zhai Y, Lu S, Ye Q, Shan X, Chen J, Ji R, Tian Y (2020) Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 9021–9030
19. Zheng K, Liu W, He L, Mei T, Luo J, Zha Z-J (2021) Group-aware label transfer for domain adaptive person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 5310–5319
20. Babu MV, Alzubi JA, Sekaran R, Patan R, Ramachandran M, Gupta D (2021) An improved idaf-fit clustering based aslpp-rr routing with secure data aggregation in wireless sensor network. Mobile Netw Appl 26(3):1059–1067
21. Wang D, Zhang S (2020) Unsupervised person re-identification via multi-label classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 10981–10990
22. Ge Y, Chen D, Li H (2020) Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In: International Conference on Learning Representations
23. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778
24. Zhang X, Zhou X, Lin M, Sun J (2018) Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 6848–6856
25. Ye M, Shen J, Lin G, Xiang T, Shao L, Hoi SC (2021) Deep learning for person re-identification: a survey and outlook. In: IEEE Transactions on Pattern Analysis and Machine Intelligence
26. Deng W, Zheng L, Ye Q, Kang G, Yang Y, Jiao J (2018) Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 994–1003
27. Tian J, Teng Z, Zhang B, Wang Y, Fan J (2021) Imitating targets from all sides: An unsupervised transfer learning method for person re-identification. Int J Mach Learn Cybern 1–15
28. Xie K, Wu Y, Xiao J, Li J, Xiao G, Cao Y (2021) Unsupervised person re-identification via k-reciprocal encoding and style transfer. Int J Mach Learn Cybern 12(10):2899–2916
29. Wei L, Zhang S, Gao W, Tian Q (2018) Person transfer gan to bridge domain gap for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 79–88
30. Lin Y, Dong X, Zheng L, Yan Y, Yang Y (2019) A bottom-up clustering approach to unsupervised person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp 8738–8745
31. Jiang Y, Chang S, Wang Z (2021) Transgan: Two transformers can make one strong gan. arXiv preprint arXiv:2102.07074
32. Li X, Hou Y, Wang P, Gao Z, Xu M, Li W (2021) Trear: Transformer-based rgb-d egocentric action recognition. In: IEEE Transactions on Cognitive and Developmental Systems
33. Zhang Q, Yang Y (2021) Rest: An efficient transformer for visual recognition. arXiv preprint arXiv:2105.13677
34. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-end object detection with transformers. In: European Conference on Computer Vision. Springer, pp 213–229
35. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2021) An image is worth 16x16 words: Transformers for image recognition at scale. ICLR
36. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H (2021) Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning, pp 10347–10357. PMLR
37. Wang G, Yang S, Liu H, Wang Z, Yang Y, Wang S, Yu G, Zhou E, Sun J (2020) High-order information matters: Learning relation and topology for occluded person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 6449–6458
38. Sun Y, Zheng L, Yang Y, Tian Q, Wang S (2018) Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European Conference on Computer Vision (ECCV), pp 480–496
39. He S, Luo H, Wang P, Wang F, Li H, Jiang W (2021) Transreid: Transformer-based object re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 15013–15022
40. Zhou D, Kang B, Jin X, Yang L, Lian X, Jiang Z, Hou Q, Feng J (2021) Deepvit: Towards deeper vision transformer. arXiv preprint arXiv:2103.11886
41. Lin H, Cheng X, Wu X, Yang F, Shen D, Wang Z, Song Q, Yuan W (2021) Cat: Cross attention in vision transformer. arXiv preprint arXiv:2106.05786
42. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin transformer: Hierarchical vision transformer using shifted windows. International Conference on Computer Vision (ICCV)
43. Chu X, Tian Z, Wang Y, Zhang B, Ren H, Wei X, Xia H, Shen C (2021) Twins: Revisiting the design of spatial attention in vision transformers. In: NeurIPS 2021
44. Chen Y, Zhao H, Hu Z, Peng J (2021) Attention-based context aggregation network for monocular depth estimation. Int J Mach Learn Cybern 12(6):1583–1596
45. Zhang T, Lin H, Tadesse MM, Ren Y, Duan X, Xu B (2021) Chinese medical relation extraction based on multi-hop self-attention mechanism. Int J Mach Learn Cybern 12(2):355–363
46. Movassagh AA, Alzubi JA, Gheisari M, Rahimi M, Mohan S, Abbasi AA, Nabipour N (2021) Artificial neural networks training algorithm integrating invasive weed optimization with differential evolutionary model. J Ambient Intell Hum Comput, 1–9
47. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 7132–7141

48. Cao Y, Xu J, Lin S, Wei F, Hu H (2019) Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pp 0–0
49. Gao Z, Xie J, Wang Q, Li P (2019) Global second-order pooling convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 3024–3033
50. Fu J, Liu J, Tian H, Li Y, Bao Y, Fang Z, Lu H (2019) Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 3146–3154
51. Roy AG, Navab N, Wachinger C (2018) Recalibrating fully convolutional networks with spatial and channel "squeeze and excitation'' blocks. IEEE Trans Med Imaging 38(2):540–549
52. Song X, Jin Z (2021) Domain adaptive attention-based dropout for one-shot person re-identification. Int J Mach Learn Cybern, 1–14
53. Woo S, Park J, Lee J-Y, Kweon IS (2018) Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 3–19
54. Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu, Q (2020) Eca-net: efficient channel attention for deep convolutional neural networks, 2020 ieee. In: CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE
55. Hermans A, Beyer L, Leibe B (2017) In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737
56. Yang F, Li K, Zhong Z, Luo Z, Sun X, Cheng H, Guo X, Huang F, Ji R, Li, S (2020) Asymmetric co-teaching for unsupervised cross-domain person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp 12597–12604
57. Yu X, Han B, Yao J, Niu G, Tsang I, Sugiyama M (2019) How does disagreement help generalization against label corruption? In: International Conference on Machine Learning, pp 7164–7173. PMLR
58. Wang W, Zhao F, Liao S, Shao L (2020) Attentive waveblock: Complementarity-enhanced mutual networks for unsupervised domain adaptation in person re-identification and beyond. arXiv preprint arXiv:2006.06525
59. Chen H, Lagadec B, Bremond F (2021) Enhancing diversity in teacher-student networks via asymmetric branches for unsupervised person re-identification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp 1–10
60. Tarvainen A, Valpola H (2017) Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: NIPS, pp.1195–1204
61. Raveendran AP, Alzubi JA, Sekaran R, Ramachandran M (2022) A high performance scalable fuzzy based modified asymmetric heterogene multiprocessor system on chip (aht-mpsoc) reconfigurable architecture. J Intell Fuzzy Sys 42(2):647–658
62. Hamdoun H, Nazir S, Alzubi JA, Laskot P, Alzubi OA (2021) Performance benefits of network coding for hevc video communications in satellite networks. Iran J Electri Electron Eng (IJEEE) 17(3):1956
63. Ristani E, Solera F, Zou R, Cucchiara R, Tomasi C (2016) Performance measures and a data set for multi-target, multi-camera tracking. In: European Conference on Computer Vision, pp 17–35. Springer
64. Zheng Z, Zheng L, Yang Y (2017) Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: Proceedings of the IEEE International Conference on Computer Vision
65. Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2015) Scalable person re-identification: A benchmark. In: Proceedings of the IEEE International Conference on Computer Vision, pp 1116–1124
66. Gray D, Brennan S, Tao H (2007) Evaluating appearance models for recognition, reacquisition, and tracking. In: Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS), vol. 3, pp 1–7. Citeseer
67. Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980
68. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp 248–255. Ieee
69. Pan X, Luo P, Shi J, Tang X (2018) Two at once: Enhancing learning and generalization capacities via ibn-net. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 464–479
70. Li J, Zhang S (2020) Joint visual and temporal consistency for unsupervised domain adaptive person re-identification. In: European Conference on Computer Vision, pp 483–499. Springer
71. Luo C, Song C, Zhang Z (2020) Generalizing person re-identification by camera-aware invariance learning and cross-domain mixup. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16, pp 224–241. Springer
72. Zhao F, Liao S, Xie G-S, Zhao J, Zhang K, Shao L (2020) Unsupervised domain adaptation with noise resistible mutual-training for person re-identification. In: European Conference on Computer Vision, pp 526–544. Springer
73. Zhai Y, Ye Q, Lu S, Jia M, Ji R, Tian Y (2020) Multiple expert brainstorming for domain adaptive person re-identification. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16, pp 594–611. Springer
74. Ge Y, Zhu F, Chen D, Zhao R, Li H (2020) Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In: Advances in Neural Information Processing Systems
75. Zheng K, Lan C, Zeng W, Zhang Z, Zha Z-J (2020) Exploiting sample uncertainty for domain adaptive person re-identification. arXiv preprint arXiv:2012.08733
76. Dai Y, Liu J, Bai Y, Tong Z, Duan L-Y (2021) Dual-refinement: Joint label and feature refinement for unsupervised domain adaptive person re-identification. IEEE Transactions on Image Processing