IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 16, 2021, accepted February 24, 2021, date of publication February 26, 2021, date of current version March 23, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3062281

Person Re-Identification by Effective Features and Self-Optimized Pseudo-Label

MING-XIANG HE[1,2], JIN-FANG GAO[1], GUAN LI[1,2], AND YOU-ZHI XIN[1]
[1]College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China
[2]National Virtual Simulation Experiment Center, Shandong University of Science and Technology, Qingdao 266590, China

Corresponding author: Guan Li (liguan0105@163.com)

This work was supported in part by the Industry-University Cooperation Education Program of the Ministry of Education under Grant 201901035016, and in part by the 2020 Qingdao Social Science Planning Research Project under Grant QDSKL2001143.

ABSTRACT With the development of deep learning, person re-identification (ReID) has been widely concerned and studied. At present, in practical application, there are three main problems in person ReID: first, it is difficult to locate the target person because the person is frequently partially occluded in crowed scenes; second, it is difficult to match the target person due to the similarity of the target person and other pedestrian features; third, the problem of model performance degradation caused by the large style discrepancies across domain/datasets. These three problems greatly limit the application of person ReID in real scenes. To solve these problems, we proposed a person ReID method based on effective features and self-optimized pseudo-label. Firstly, we designed a feature aggregation module which combines mask channel and pose channel to accurately extract the global saliency features, so as to solve the occlusion problem; secondly, we designed a head-shoulder feature auxiliary module to enhance the feature representation of the head-shoulder, so as to solve the problem of similarity between the target person and other pedestrian features; finally, we designed a self-optimized pseudo-label training module to improves the generalization ability of the model, so as to solve the problem of different styles in the cross-domain environment. Extensive contrast experiments with the state-of-the-art methods on multiple person re-ID datasets show that our method leads to significant improvement, which prove the effectiveness of our method.

INDEX TERMS Person re-identification, deep learning, saliency feature, head-shoulder feature, pseudo-label.

I. INTRODUCTION

The task of person re-identification (ReID) is to identify the target person from a large set collected by multiple non-overlapping cameras [1]. In recent years, with the development of deep learning, person ReID has attracted extensive attention and research [2]–[5], and has made significant progress in both academia and industry [6], and was applied to the real scenes, such as stations and supermarkets [7]. However, there are some problems in the practical application, mainly including three aspects. The first aspect is the problem that the target person to be identified is occluded, which is referred to as the occlusion problem. It includes two situations in which the target person is occluded by objects and occluded by other pedestrians, as shown in Figure 1. In (a), part of the target person to be identified is occluded by the car; in (b), part of the target person is occluded by another pedestrian. This is common, especially in crowded scenes. The occlusion problem greatly affects the performance of person ReID methods and has been widely studied. The second aspect is the problem that the target person and other pedestrians have similar features, which is referred to as the similarity problem. For example, people like to wear white or light-colored clothes in summer, while people like to wear black or dark-colored clothes in winter, as shown in Figure 1 (c). In the strong light, weak light, and low resolution of the captured image, there is often a serious lack of clothing feature attributes, which makes it difficult to distinguish between pedestrians with similar clothing features, and brings great challenge to person ReID. The third aspect is the problem of domain style difference caused by cross-domain, which is referred to as the cross-domain problem. There is generally a great difference between the styles of the current application domain and the source domain, resulting in a decrease in model performance.

Because the model is trained and tested on the specified dataset, in the actual deployment, there are often significant differences between the domain and the source dataset due to

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

FIGURE 1. (a) Part of the target person's body is occluded by a vehicle, (b) part of the target person's body is occluded by other pedestrians, (c) the characteristics of different people are similar.

environmental differences such as lighting, background, and so on.

At present, most of the methods cannot solve the above problems well and cannot be deployed to the real scene. For example, attention-based methods [8]–[11] can eliminate the influence of background, but cannot cope with the situation that the target is occluded by other pedestrians. The part-based method [12]–[15] achieves strong performance by matching parts to parts, but they require strict alignment in advance. The pose-based method [16], [17] uses the existing posture estimators to enhance the corresponding local features, while the semantic information method [18]–[20] realizes the matching by defining the persons' attributes. However, the two methods mainly rely on the external decoration and the attributes of clothing, and they perform poorly in the case of similar features and severe attribute loss. The two-stream network method based on appearance feature extraction stream and part heatmap extraction stream [21]–[23] take too long inference time and cannot extract accurate heatmap in crowded conditions. The domain adaptation method based on unsupervised learning [24]–[28] transforms the labeled source image into the image with the target style, while the pseudo-label method based on clustering [29]–[34] optimizes the network by fine-tuning the pseudo label through clustering. These methods have a certain improvement in performance. However, most of these methods have the problem of label noise or image noise, which leads to the serious inseparability of positive and negative sample pairs, resulting in poor retrieval and sorting results.

To solve the above problems, this article proposes a person ReID method based on effective features and self-optimized pseudo-label. The method consists of three parts: the feature aggregation module combining mask channel and pose channel, the head-shoulder feature auxiliary module, and the self-optimized pseudo-label training module. Inspired by the foreground perception Pyramid [35] and the Mask-guided method [36], we proposed a feature aggregation module to solve the occlusion problem. It generates the corresponding mask-guided heatmap through the mask channel, predicts the keypoints heatmap of the target person through the pose channel, and then carries out the weighted fusion of the two kinds of heatmaps to obtain the saliency heatmap; finally, two heatmaps are aggregated to generate global features. Inspired by the feature segmentation method [13] and the head-shoulder representation method [37], we proposed the head-shoulder feature auxiliary module to solve the similarity problem. It first uses the head-shoulder positioning layer to cut the head-shoulder part; then the head-shoulder features are input into the attention layer to obtain the head-shoulder feature representation; finally, we enhance the feature representation of the head-shoulder area by fusing global features and head-shoulder features. Inspired by clustering-based pseudo-label methods [29]–[32], we proposed a self-optimized pseudo-label training method to solve the cross-domain problem. Firstly, the distance distribution of positive and negative samples is modeled by the normal distribution which can update the mean and variance; secondly, a momentum updating mechanism is used to maintain the variables in the global distribution; finally, we adopted a distribution-based hard-mining loss function to update the relevant losses in time to optimize the network in each batch.

The main contributions of this article are as follows:

1) A feature aggregation method combining pose channel and mask channel is proposed to solve the occlusion problem.

2) A head-shoulder feature auxiliary method is proposed to enhance the feature representation of the head-shoulder to solve the similarity problem.

3) A self-optimized pseudo-label training method is designed to solve the cross-domain problem.

The rest of this article is structured as follows: in chapter two, we introduced the work related to person ReID; in chapter three, we introduced the method proposed in this article in detail; in chapter four, we introduced the data sets, evaluation indicators, tricks used in the experiment, and conducted the ablation study; in chapter five, we carried out a lot of experiments to prove the superiority of this method; in chapter six, we summarized and prospected the research work of this article.

II. RELATED WORK

In recent years, person ReID based on deep learning has made great progress. However, there are some problems in the application of real scenes, which greatly limits the deployment of person ReID in reality. These problems have attracted extensive attention, and a lot of research work has been carried out around them.

A. OCCLUSION PROBLEM

The occlusion problem greatly affects the performance of person ReID methods and has been widely studied. Attention-based methods [8]–[10] enhance the output features and make them focus on the foreground human body, which has a good performance in eliminating the background. However, this method cannot eliminate the body part of other pedestrians, and cannot deal with the situation that the target person is occluded by other pedestrians. Part-based methods use the local-to-local matching strategy to deal with the problem of incomplete target persons. For example, He and Wang [35] proposed a local patch-level matching model based on dictionary learning and

explicit patch ambiguity modeling, and introduced a global part-based matching model to provide complementary spatial layout information. But this method requires repeatedly calculating the feature, so the calculation cost is extremely high. Wang *et al.* [13] proposed a Part-based Convolutional Baseline (PCB) network, which outputs convolutional features composed of several segmental features, and pays attention to the consistency of each part to solve the occlusion problem. However, none of these methods can skip the alignment step, greatly increasing the computational cost, and the alignment is a challenging task. There is another method based on mask-guided, it uses person masks containing body shape information to help eliminate the background clutter at the pixel level for person ReID. For example, Kalayeh *et al.* [21] proposed a model combining human semantic information and person ReID, which combines the source image and person masks as input to reduce the influence of appearance changes such as illumination, posture, and occlusion. There is also a two-stream network method. For example, Suh *et al.* [23] proposed a network composed of an appearance map extraction stream and a body part map extraction stream, using a bilinear mapping of the corresponding local appearance and body part descriptors to get a part-aligned map. Although this method can solve the occlusion problem, it largely depends on accurate pedestrian segmentation and takes a lot of time to infer external cues.

B. SIMILARITY PROBLEM

The features of the target person are similar to those of other pedestrians, and due to the loss of feature attributes during camera capture, it brings great difficulty to re-identify pedestrians. At present, the popular solution is based on semantic information, which relies on external attribute information for identification. For example, Layne *et al.* [18] defined 15 semantic features, including hairstyle, shirt type, shoes, and so on, and used support vector machine to obtain various attribute features and integrated them with the underlying features to achieve better differentiation. Liu *et al.* [19] described various attributes of clothing in detail. Shi *et al.* [20] suggested learning attributes from existing fashion photography datasets, including colors, textures, and category labels. However, this method relies heavily on external attributes, and some external attributes are accidental. Besides, due to the low illumination, angle of view, and resolution of the captured image, the attributes will be seriously lost, resulting in extremely unstable performance. The part-based method [12], [13] can also be used to solve the similarity problem. It extracts local details by dividing features into several horizontal grids, and achieves strong performance by matching parts with parts, but they need strict alignment in advance. Li *et al.* [38] proposed a multi-position learning method with head-shoulder part as input, which aims to solve the problem of person ReID in crowded environment. However, this method mainly focuses on the change of pose, and there are few kinds of research on the head-shoulder part, feature extraction, and feature fusion based on global representation. It also performs poorly when the target is physically intact because it requires only the head and shoulders as the input.

C. CROSS-DOMAIN PROBLEM

Due to the differences in style between the practical application domain and the model training dataset, the performance of the model will decline significantly in the application. It is the most effective solution to annotate each new environment and retrain the model. However, the cost is too high to meet the actual situation. At present, the popular solution is the pseudo-label estimation method based on clustering. The basic idea of this method is to use the similarity of unlabeled samples to predict pseudo labels by feature clustering, and then use the pseudo labels for fine-tuning, to achieve the purpose of optimizing the network. This kind of clustering and training process usually alternates until the model is stable. For example, Ni *et al.* [39] proposed a relative metric learning method based on clustering and projection vector learning; Wu *et al.* [40] proposed a hierarchical clustering algorithm for inter-camera and cross-camera shooting, and Fan *et al.* [29] added a selection operation between clustering and fine-tuning to improve the optimization effect of the model. However, these models are usually interfered by noise pseudo-label. There is the serious inseparability of the distance distribution of the positive and negative samples, and there are unsatisfactory distances of small inter-classes and large inter-classes, which seriously affects the accuracy of identification results.

III. PROPOSED METHOD

We proposed a person ReID method based on effective features and self-optimized pseudo-label in this article, which provides a solution to the problems encountered in the actual deployment of person ReID. In this chapter, we will begin with an overview of our method, then go into detail about each module in the method.

A. OVERVIEW OF THE PROPOSED METHOD

The overall network structure of the method is shown in Figure 2, which is divided into three modules: the feature aggregation module combining mask channel and pose channel, the head-shoulder feature auxiliary module, and the self-optimized pseudo-label training module.

Given a person to be queried, the purpose of person ReID is to recognize all related persons from the massive pedestrian gallery by way of feature matching. Therefore, it is particularly important to accurately extract target features. In our method, two modules focus on feature extraction, namely the feature aggregation module and the head-shoulder feature auxiliary module. The feature aggregation module uses a mask channel to reduce the influence of background on pedestrian features through pixel-level instance segmentation. However, the mask channel only focuses on the influence of background, and can't deal with the situation that the target person is obscured by other pedestrian's body parts.

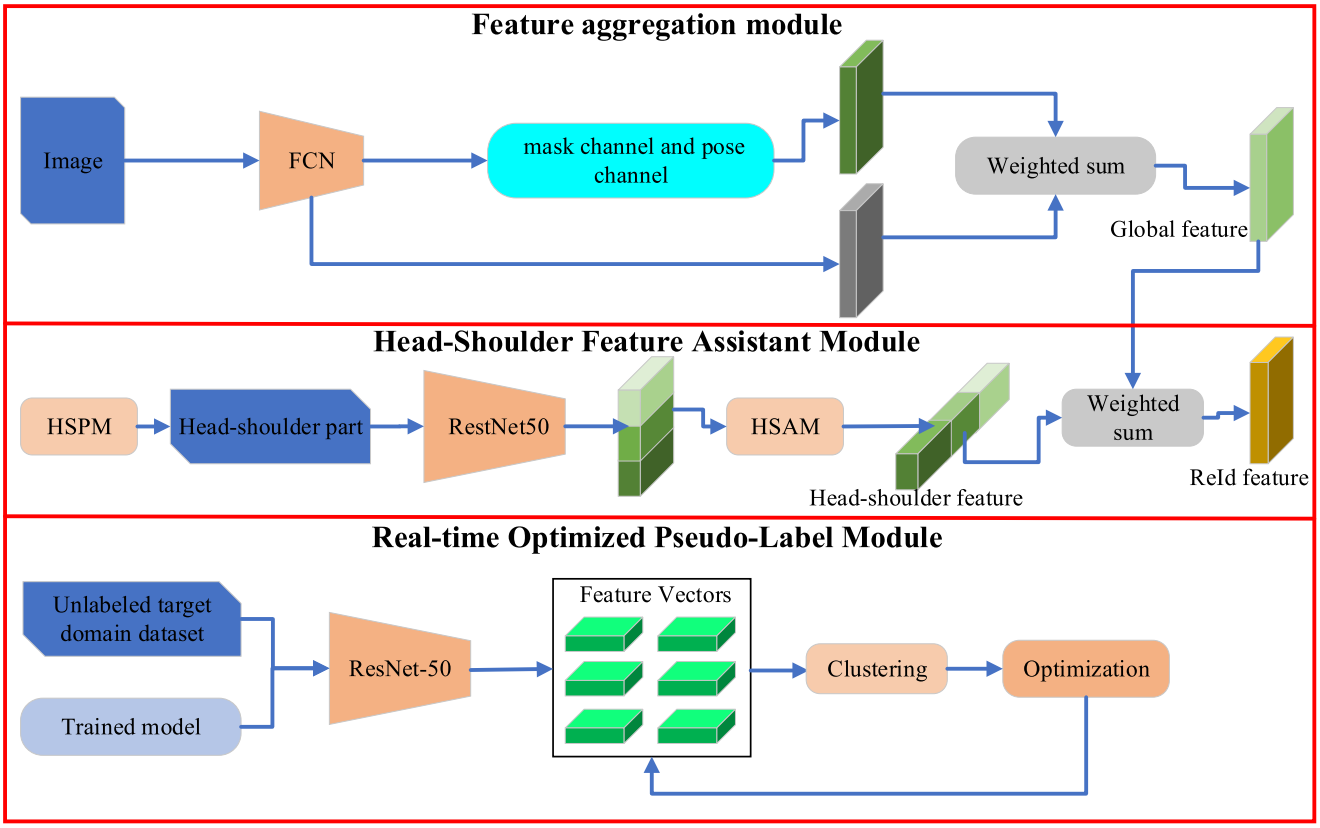


FIGURE 2. The overall network structure of the method proposed in this article. The global feature is the global feature output by the feature aggregation module, the head-shoulder feature is the head shoulder feature output by the head shoulder module, and the reid feature is the final extracted target character feature. HSPM is the head shoulder positioning submodule in the head shoulder feature enhancement module, and HSAM is the head shoulder attention submodule.

Therefore, we added a pose channel to reduce the effect of occlusion by other pedestrians through the prediction of key points of human posture. The results of the mask channel and pose channel are fused to produce accurate global features. The head-shoulder feature auxiliary module mainly enhances the feature representation of the head and shoulders to solve the similarity problem. Because the head and shoulders contain rich and recognizable features, and the head-shoulder part is easy to segmentation. The module can assign different weights according to the monotonicity of features, and fuse the head-shoulder features with the output of the global feature by the feature aggregation module to output the final features. Through these two modules, we can get accurate and recognizable target features. Figure 3 shows the comparison of the ReId feature map extracted by our feature extraction method and other feature extraction methods. It can be seen from the image that our method can extract richer features.

Although we designed an excellent network to extract person features, there is a large style difference between the actual application domain and the source domain, which challenges the generalization ability of the model. Therefore, we designed a self-optimized pseudo-label training module based on clustering, so as to adapt to different domain styles. Different from the common pseudo-label method based on clustering, our method can realize real-time active network optimization network by promoting the separation of positive and negative samples.

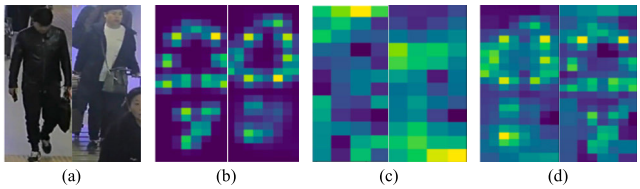


FIGURE 3. Contrast images of feature maps extracted by different methods, where (a) Is the input image; (b) Is the pose-based method; (c) Is the mask-based method, and (d) Is our method.

Generally, our method mainly focuses on the extraction of character features and improvement of the generalization ability of the model, including three aspects of innovation, which will be described in detail below.

B. FEATURE AGGREGATION MODULE

It is a common situation that the target person is occluded. However, most methods only focus on eliminating the influence of the background and ignore the situation that the target person may be partially occluded by other pedestrians.

Therefore, we designed a global feature aggregation method combining mask channel and pose channel to solve the occlusion problem.

The method consists of three submodules: spatial feature extraction module, guided heatmap prediction module, and global feature aggregation module. Firstly, the full convolutional network (FCN) was used to extract spatial features, and then mask channel and pose channel were used to predict mask-guided heatmap and pose-guided heatmap. Finally, the saliency heatmap was generated and aggregated to generate global features. The structure of the module is shown in Figure 4.

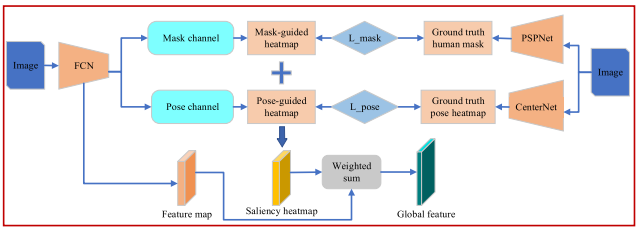


FIGURE 4. Network structure of feature aggregation module, L_{mask} and L_{pose} is the regression loss function, salience heatmap is the fusion of mask guided heatmap and pose guided heatmap, and global feature is the final output global feature of this module.

Traditional neural networks use full connection layer or average pooling layer to aggregate global features, which are seriously affected by background and occlusion. For example, in the crowded scene, the bounding box generated by the detector for the target person is usually rough and contains a lot of exterior scene information and occluded information, which seriously interferes with the subsequent feature extraction. To obtain a more accurate feature representation of the target person, we used a full convolutional network (FCN) based on ResNet-50 [44] as the backbone of feature extraction, abandoning all the full connection layers and retaining only the convolution and pooling layers. At this time, the FCN can still retain the spatial coordinate information, and can achieve the non-interference extraction of background, occluded features, and target person features.

To reduce the influence of background on person features, we used a mask channel to obtain the foreground probability of spatial features. The output spatial features generate the mask-guided heatmap of corresponding size through the mask channel. We regard the spatial feature classification of the background and foreground person as a problem of probability prediction. The PSPNet [48] was used to guild the mask channel to generate the mask-guided heatmap, and we used $M_p \in \mathbb{R}^{W\times H}$ to denote it, where W and H are the width and height of the output feature map, respectively. We used PSPNet to generate ground truth human mask, and we used M_{gt} to denote it. Our purpose is to use M_{gt} for the regression of M_p, with the regression loss function being L_{msak}. The calculation is as shown in Equation (1).

$$L_{msak} = \left\| M_p - M_{gt} \right\|_F \tag{1}$$

However, the mask-guided heatmap generated by the mask channel cannot distinguish the target person from the occluded person, so we used a pose channel to generate the heatmap of the target person, which only focuses on the target person. Similarly, the output spatial features generate the pose-guided heatmap of corresponding size through the pose channel. We used CenterNet [47] to guide the pose channel to generate the pose-guided heatmap. The key points in human pose estimation include 13 human joints. We used $P_p \in \mathbb{R}^{W\times H\times 13}$ to denote a pose-guided heatmap, $P_{gt} \in \mathbb{R}^{W\times H\times 13}$ to denote the generated ground truth pose heatmap generated by sampling prediction in CenterNet [47]. Our purpose is to use P_p to regress P_{gt}, with the regression loss function being L_{pose}. The calculation is shown in Equation (2).

$$L_{pose} = \left\| P_p - P_{gt} \right\|_F \tag{2}$$

The mask-guided heatmap only focuses on human parts, and it cannot solve the problem that the target person is occluded by other pedestrians. The pose-guided heatmap focuses on the identified person. Thus, we fuse the mask-guided heatmap and the pose-guided heatmap to generate more accurate saliency heatmap, which can be used to aggregate the output spatial feature map. Through the weighted sum operation, the network finally generates a global saliency feature representation. In training, we used a softmax cross-entropy loss L_{id} and a triple loss L_{tri}. The calculation of the final loss function is shown in Equation (3).

$$L = \lambda_{mask} L_{mask} + \lambda_{pose} L_{pose} + L_{id} + L_{tri} \tag{3}$$

In the experiment, let λ_{mask} and λ_{pose} be 0.05 and 0.1, respectively.

C. HEAD-SHOULDER FEATURE AUXILIARY MODULE

The head-shoulder feature auxiliary module can enhance the head-shoulder feature representation, solve the problem that the target person is similar to other pedestrian features, and improve the accuracy of person ReID results. The structure of the module is shown in Figure 5.

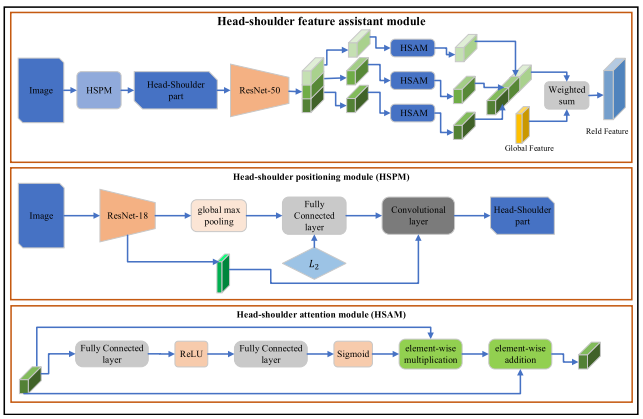


FIGURE 5. The network structure of the head-shoulder feature auxiliary module, L_2 is L_2 loss function, HSPM is the head-shoulder positioning submodule, HSAM is the head-shoulder attention submodule, the global feature is the output of the feature aggregation module, and Reid feature is the final output character feature.

We first designed a head-shoulder positioning module to locate and cut the head-shoulder part of the input image.

Because different feature mapping channels represent different meanings and different feature space positions represent different semantics, we designed a head-shoulder attention module to enhance the head-shoulder representation in the channel and space. It includes a generalized mean pooling, a fully connected layer for dimensionality reduction, a ReLU activation function, another fully connected layer for dimensionality raising, and a sigmoid activation function. It is applied to each horizontal slice and finally connected to generate a head-shoulder area feature map with the size of $C \times 1 \times 1$.

Most of the existing person ReID methods directly connect global features and local features, and ignore the relationship between feature weight and input conditions. To solve this problem, an adaptive feature weight fusion module is designed, and different weights are assigned to global and local features according to the input conditions to carry out feature fusion. Specifically, if a person's clothing features are monotonous, more weight is assigned to the person's head-shoulder part. The fusion method of the global feature and the head-shoulder feature is shown in Equation (4).

$$f = (w_1 * f_g) \circledast (w_2 * f_h) \tag{4}$$

where, w_1 and w_2 are weights, f_g represents global features, f_h represents the feature of the head-shoulder area and $*$ represents the element-wise multiplication.

D. SELF-OPTIMIZED PSEUDO-LABEL MODULE

In the actual deployment, the local domain style is usually quite different from the source domain style, resulting in a serious decline in the performance of the model. To solve this problem, the popular method is pseudo-label estimation based on clustering, but this method is usually interfered by noise pseudo labels, and the distance between positive and negative samples is seriously inseparable. Our purpose is to improve the effect of model optimization in pseudo-label training by solving the indivisibility of the global distance between positive and negative samples. Accordingly, we designed a clustering-based pseudo-label training method that can be optimized in real-time. In the training process, a global distance distribution separation method is used to optimize the network in real-time to enhance the performance of the model. The structure of the module is shown in Figure 6.

The clustering-based method usually includes three steps: the first step is model pre-training, which aims to use the data from the source dataset to pre-train the person ReID network for feature learning; the second step is to assign pseudo label to the target domain data by clustering them; the third step is to fine-tune the network with pseudo-label to improve the generalization ability of the model. The second and third phases are performed alternately in multiple iterations. In the third stage, we used triplet-based losses for optimization.

Triple-based loss is calculated for the similarity of samples by optimizing the distance between the anchor and positive samples to be less than the distance between the anchor and negative samples. However, it cannot guarantee that the relative order of the distance distribution of a large number of sample pairs is correct, so we need to optimize it. We found that the distance distributions in the person ReID method were similar to the normal distribution, so we used the normal distribution to model the distance distribution. In particular, we assume that the mean and variance at the beginning are both 2. Meanwhile, we designed an algorithm to promote the separation of positive and negative sample pairs.

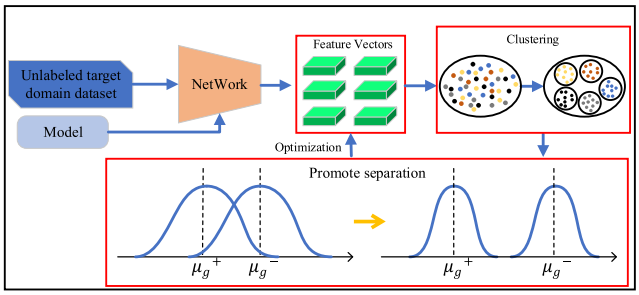


FIGURE 6. The structure of the self-optimized pseudo-label training module.

In the training process, instead of searching all samples at the same time, the convolutional neural network back propagates the calculated losses in mini-batch to optimize the network. Based on this, in each batch, we update the distance distribution of positive and negative sample pairs and the corresponding loss by changing the distribution parameters. We use $N(\mu^+, \sigma^{+2})$ and $N(\mu^-, \sigma^{-2})$ to represent the distance distribution of positive and negative sample pairs respectively. The momentum update formula is shown in Equation (5) - (8).

$$\begin{cases} \lambda\mu^+ + (1-\lambda)\,\mu_l^+ \Rightarrow \mu^+ \\ \lambda\sigma^{+2} + (1-\lambda)\sigma_l^{+2} \Rightarrow \sigma^{+2} \end{cases} \tag{5}$$

$$\begin{cases} \lambda\mu^- + (1-\lambda)\,\mu_l^- \Rightarrow \mu^- \\ \lambda\sigma^{-2} + (1-\lambda)\sigma^{-2} \Rightarrow \sigma^{-2} \end{cases} \tag{6}$$

$$\mu_l^+ = \frac{1}{N^+}\sum_{i=1}^{N^+} d_i^+, \quad \sigma_l^{+2} = \frac{1}{N^+}\sum_{i=1}^{N^+} (d_i^+ + \mu^+)^2 \tag{7}$$

$$\mu_l^- = \frac{1}{N^-}\sum_{i=1}^{N^-} d_i^-, \quad \sigma_l^{-2} = \frac{1}{N^-}\sum_{i=1}^{N^-} (d_i^- + \mu^-)^2 \tag{8}$$

where $\lambda \in [0, 1)$ is the momentum coefficient. In this experiment, we set it as 0.9, N^+ and N^- represent the number of positive and negative sample pairs respectively, d_i represents the Euclidean distance of the ith sample pair.

To promote the separation of positive and negative sample pairs, and to better facilitate the separation of the two distributions, we hoped that the hard samples located in the potential overlapping areas could also be separated. Therefore, we used a distribution-based separation loss function of hard mining to optimize the network. The calculation is shown

in Equation (9).

$$L = Softplus\left(\mu^{+} - \mu^{-}\right) + \lambda_g\left(\sigma^{+2} + \sigma^{-2}\right) + L_H \quad (9)$$

$$L_H = \lambda_h\, Softplus(\left(\mu^{+} + k\sigma^{+}\right) - \left(\mu^{-} - k\sigma^{-}\right)) \quad (10)$$

where λ_g and λ_h are both hyper-parameters used to balance the importance of mean value and variance. In the experiment, we set them as 0.3 and 0.7, respectively, and k is the hyper-parameter controlling the hard mining power. The larger k is, the wider the coverage area is. In this experiment, we set the value of k as 3.

IV. EXPERIMENTATION

In this chapter, we described the experimental details, datasets, evaluation indicators, and tricks used in our experiments, and conducted the ablation study to prove the effectiveness of our above improvement.

A. EXPERIMENTAL DETAILS

We use ResNet50 [44] as the backbone of our baseline, use triplet loss and cross entropy loss to train the network, and initialize ResNet50 with pre-trained parameters on ImageNet [50]. Set the batch size to 192. In each minibatch, we randomly sample 16 identities and 4 images. Adjust the pixels of each image to 256 × 128. The initial learning rate is set to 3.5e-4, and is reduced by 0.1 at the 40th and 70th epoch. We use the Adam algorithm [51] for back propagation to optimize the network.

In our experiment, based on NVIDIA 2080 GPU, we build a PyTorch deep learning framework with Python 3.7 development environment under Linux operating system. All models were trained and tested in this environment. The models were trained with 200 epochs.

B. DATASETS

We used Market1501 [42], DukeMTMC-reID [43], MSMT17 [41], occluded dataset and similarity dataset as the evaluation datasets of this article.

1) MARKET1501

This dataset contains 1,501 pedestrians captured by six cameras and 32,668 detected pedestrian bounding boxes. Each pedestrian is captured by at least 2 cameras, and there may be multiple images in one camera. The training set has 751 people, including 12,936 images; the test set has 750 people with 19,732 images.

2) DUKEMTMC-REID

This dataset is collected by 8 synchronous cameras, including training set of 16,522 images from 702 people, a test set of 2,228 images from 702 people, and a search gallery of 17,661 images.

3) MSMT17

This dataset is collected by 15 cameras and contains 126,441 bounding boxes for 4,101 pedestrians. The training set contains 1,041 pedestrians with 32,621 bounding boxes, and the test set contains 3,060 pedestrians with 93,820 bounding boxes. For the test set, 11,659 bounding boxes are randomly selected as query, while the other 82,161 bounding boxes are selected as the gallery.

4) OCCLUSION DATASET

Partial-REID dataset [47] contains 900 images with 60 identities, and each id consists of 5 full-body images, 5 partial images, and 5 occlusion images.

Occluded-REID dataset [45] consists of 2000 images of 200 occluded persons. Each identity has 5 full-body person images and 5 occlusion person images with different types of sever occlusions.

Occluded-Duke MTMC dataset [46] is the largest occluded person re-id dataset to date. The testing set contains 1,110 identities, including 17,661 gallery images and 2,210 query images. It is more difficult and practical since both probe and gallery images have occlusions. And it has rich variations, including different viewpoints and various obstacles, including cars, bicycles, trees, and other peoples.

5) SIMILARITY DATASET

This dataset is the test dataset of similarity experiments with the images of people with similar features selected from other datasets. The test set contains 800 images of target persons with similar features in 80 scenes. Each target person has 5 full-body images and 5 images with different perspectives and backgrounds.

C. EVALUATION INDICATORS

In this article, we adopted the standard indicator of most person ReID methods, namely the cumulative matching curve (CMC) and the mean Average Precision (mAP).

D. DATA AUGMENTATION

The dataset of person ReID is always limited, and it is impossible to capture every real-scene image. Therefore, it is necessary to extend the existing training data and improve the generalization ability of the model. In the experiment, we used the following methods for data augmentation.

1) RANDOM ERASING AUGMENTATION (REA)

It is a data augmentation method of random erasure. In simple terms, randomly select an area in the image and mark it with noise mask. The mask can be black, gray, or random noise, as shown in Figure 7. Random erasure is a way of data augmentation, which can reduce the degree of model overfitting and thus improve the performance of the model.

2) BRIGHTNESS CONVERSION

It is a method to fully simulate different lighting environments by adjusting the contrast, saturation, and hue of the image.

E. SELECTION OF TRICKS

This article used several tricks to improve the performance of our model, as shown below:

FIGURE 7. (a) Original image, (b) Random erasing augmentation, (c) Brightness conversion.

1) WARMUP LEARNING

Standard Baseline uses the common step-down learning rate, with an initial learning rate of 3.5e-4, a total of 120 epochs are trained, and the learning rate is reduced in the 40th and 70th epochs. However, initializing the network with a large learning rate may cause the network to vibrate to a suboptimal space, because the gradient of the initial stage of the network is large. Warmup's strategy is to use a gradually increasing learning rate to initialize the network at the initial stage, and gradually initialize to better search space. In this article, we used the simplest linear strategy, that is, the first 10 epochs gradually increase from 0 to the initial learning rate [48].

2) LAST STRIDE

The last layer convolution of each block in the ResNet50 backbone has a down sampling process, that is, the stride of the last layer convolution is 2. If a 256×128 image is input normally, the network will output an 8×4 feature map. In general, increasing size can improve performance. Here, we change the stride of the last convolution to 1, which does not need to increase the number of parameters or change the parameter structure of the model, but will expand the feature map size to 16×8. A larger feature map can extract more fine-grained features, thus improving the performance of the model [49].

3) LARGE BATCH SIZE

Using a large batch size can improve the stability of training and get a better result. Here we change the training batch size from 64 to 192.

F. ABLATION STUDY

To analyze the effectiveness of individual modules in our method, we test the performance of each module incrementally. The reason is that each improved method is not completely independent. Some improved methods are effective when applied individually, but they may be ineffective or even negative when being combined. It is difficult to make a comprehensive analysis because there are too many combinations of improvement methods. Therefore, we carried out ablation studies to prove the rationality of our method combination. We designed five different network structures: FCN + occluded module, FCN + similarity module, FCN + pseudo-label module, FCN + occluded module + similarity module, and FCN + occluded module + similarity module + pseudo-label module. We trained and tested them on three datasets, three times for each test, and took the average values. The experimental results are shown in Table 1.

TABLE 1. The results of the ablation study.

	Method	Market1501		DukeMTMC		MSMT17	
		mAP	R-1	mAP	R-1	mAP	R-1
A	Baseline	81.3	91.2	73.5	87.1	49.2	77.2
B	A + Occlusion module	82.7	92.6	73.9	87.8	52.3	79.1
C	A + Similar module	82.1	92.8	74.1	87.9	52.5	78.5
D	A + Pseudo-Label model	82.6	92.4	74.3	88.1	52.1	79.6
E	A + Occlusion + Similar	83.4	93.7	74.7	88.4	52.7	79.2
F	E + Pseudo-Label model	**83.7**	**95.6**	**75.2**	**89.2**	**53.4**	**81.3**

A→B We added the occluded module based on the baseline. From Table 1, we can see that the performance of the three datasets has been improved. This is because the module can not only reduce the influence of the background, but also deal with the situation that part of the body of the target person is occluded, and can accurately extract the characteristics of the target person.

A→C We added the similarity module based on the baseline. As can be seen from Table 1, we obtained higher scores on the three datasets. It shows that our similarity module can make good use of the features of the head and shoulders to enhance the person's feature representation, and better deal with the situation that the target person is similar to other pedestrians.

A→D based on the baseline, we added the pseudo-label training module for optimization. The pseudo-label module optimizes the network in real-time by promoting the separation of positive and negative samples in the process of pseudo-label training. The scores on the three datasets are all improved, which proves that our pseudo-label training module can improve the generalization ability of the model.

A→E We added the occluded and similarity modules based on the baseline. Both of the two modules focus on feature extraction, and the similarity module further enhances the representation of head-shoulder based on the output of the global feature by the occluded module to deal with the similarity problem. It can be seen from the results that the score of the combination of the two modules is improved to a certain extent compared with that of the single module, which shows the effectiveness of the combination.

E→F Based on occluded and similarity modules, we also added the pseudo-label module. Based on extracting more effective features, the model generalization ability is improved through the self-optimized pseudo-label module and makes the model more robust. It can be seen from the table that the performance of the model is at its peak.

TABLE 2. The experimental results of occlusion problem.

Occlusion test dataset	Method	Market1501 train		DukeMTMC train		MSMT17 train	
		mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Partial-REID dataset	PCB	46.1	48.4	49.4	51.7	57.8	61.3
	MaskReID	20.7	24.3	24.4	28.8	35.4	40.7
	PABR	50.3	63.6	54.8	65.3	66.6	74.5
	OSNet	54.8	64.4	57.3	66.8	68.4	76.3
	PTGAN	10.9	13.4	14.4	17.3	27.4	31.3
	SPGAN	20.3	21.2	23.6	25.9	34.5	39.3
	Baseline	46.7	49.6	48.7	51.3	57.8	64.4
	Our method	**71.3**	**78.2**	**72.3**	**79.4**	**78.6**	**84.2**
Occluded-REID dataset	PCB	31.7	35.3	35.6	39.8	44.3	47.6
	MaskReID	19.3	21.6	24.2	26.6	35.5	38.2
	PABR	42.6	50.8	47.7	54.9	56.8	65.4
	OSNet	47.6	54.3	51.2	57.7	62.3	66.9
	PTGAN	6.3	9.7	11.6	15.3	24.6	26.2
	SPGAN	15.6	17.7	21.4	26.7	30.4	35.7
	Baseline	42.7	46.4	45.5	48.8	52.7	55.3
	Our method	**64.2**	**74.8**	**68.8**	**76.7**	**74.6**	**81.8**
Occluded-Duke MTMC dataset	PCB	27.6	34.3	32.2	36.8	44.3	46.9
	MaskReID	16.2	17.6	20.4	31.9	31.6	35.4
	PABR	27.7	32.2	31.8	37.3	45.3	54.8
	OSNet	34.5	39.8	36.6	42.8	49.7	57.3
	PTGAN	9.3	13.6	10.2	15.6	24.3	30.6
	SPGAN	10.2	14.4	14.6	20.1	28.5	33.7
	Baseline	20.6	25.9	24.3	31.1	36.6	40.3
	Our method	**42.3**	**49.2**	**47.7**	**54.3**	**60.2**	**64.4**

The experiment proves that our three modules are effective whether used alone or in combination, and the performance is at its peak when they are combined.

V. ANALYSIS OF EXPERIMENTAL RESULTS

In this chapter, we compared our method with several popular pedestrian recognition methods, including PTGAN [26], SPGAN [24], SPReID [21], MaskReID [22], PCB [12] and PABR [23]. All the methods were trained and tested in the same environment with the same training strategy. We verified the ability of our method to solve each problem from three aspects: occluded problem, similarity problem, and cross-domain problem.

A. THE OCCLUDSION PROBLEM EXPERIMENT

To verify the effectiveness of our method in the occlusion problem, we trained each method with three large datasets [41]–[43] and tested each method on three occlusion datasets. Each test was carried out three times, and the average values were taken. The experimental results are shown in Table 2. Figure 8 is a histogram of Rank-1 scores tested on Occluded-Duke MTMC dataset, which clearly shows the performance comparison of different methods on the occlusion problem.

It can be seen from Table 2 that PABR and OSNet also show good performance on the Partial-REID dataset, because the dataset only includes 600 images from 60 people, and the occlusion situation is relatively simple. However, as the size of the test set increases and the occlusion situation becomes more abundant, their performance drops severely, which shows that they cannot handle the occlusion situation in the actual environment well. The method proposed in this article can obtain excellent results whether it uses different data sets for training or testing on different occlusion test sets. It can be seen that the method can adapt to complex and changeable occlusion conditions and can effectively solve the occlusion problem.

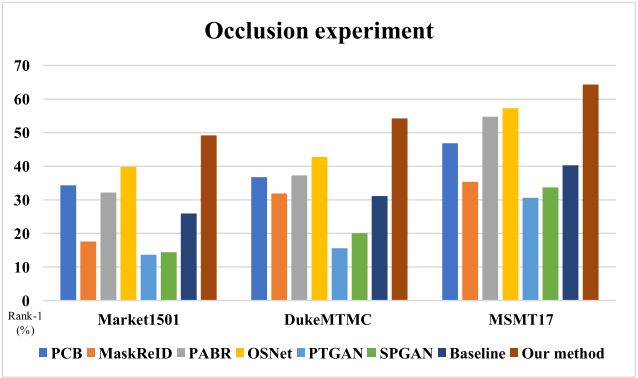


FIGURE 8. After training with different data sets [41]–[43], the rank-1 score histogram of each method on Occluded-Duke MTMC dataset.

B. THE SIMILARITY PROBLEM EXPERIMENT

To verify the effectiveness of our method on the similarity problems, we trained each method on three large data sets [41]–[43], conducted three tests on our similarity data

TABLE 3. The experimental results of similarity problem.

Method	Market1501 train		DukeMTMC train		MSMT17 train	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
PCB	69.4	78.3	71.5	80.3	73.4	82.3
MaskReID	62.7	71.4	65.3	73.6	66.5	75.1
PABR	78.3	88.2	77.4	87.6	78.6	89.4
OSNet	81.7	89.6	81.4	89.3	82.7	90.2
PTGAN	47.3	56.8	49.5	64.3	51.4	66.2
SPGAN	58.5	67.4	62.3	71.2	62.7	72.5
Baseline	79.4	87.2	75.7	83.2	79.4	87.2
Our method	**84.3**	**92.1**	**82.6**	**89.7**	**84.3**	**92.1**

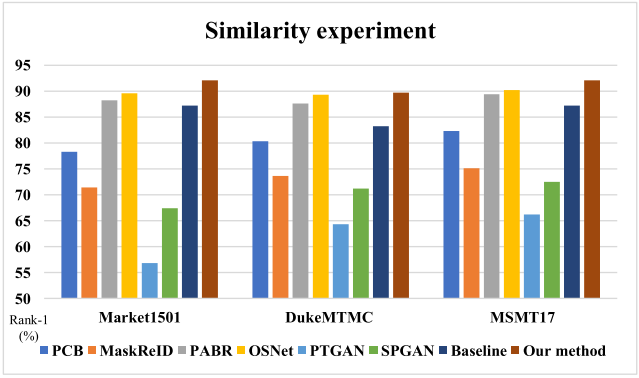


FIGURE 9. After training with different data sets [41]–[43], the rank-1 score histogram of each method on similar data sets.

sets, and took the average values. The experimental results are shown in Table 3. Figure 9 is the Rank-1 score histogram of the similarity contrast experiments, which clearly shows the performance comparison of different methods on the similarity problem.

Table 3 shows that our method achieved the best result by using different training sets, which proves that our method proposed in the paper can solve the similarity problem well. Although PABR and OSNet methods achieved good results, they still appear inferior compared with our scores, not to mention that they need longer reasoning time than ours.

C. THE CROSS-DOMAIN PROBLEM EXPERIMENT

To verify the effectiveness of our method when faced with different domain styles, we used three large data sets [41]–[43] for experiments. First, we trained on one dataset, and then tested on the other two datasets. Similarly, we conducted three experiments for each test and took the average values accordingly. The results are shown in Table 4. Figure 10 is the Rank-1 score histogram of the cross-domain experiments, which clearly shows the performance comparison of different methods on the cross-domain problem.

As shown in Table 4, our method achieved the highest score when tested on the other two data sets, regardless of

TABLE 4. The experimental results of cross-domain problem.

Train set	Mtehod	Test set			
		DukeMTMC		MSMT17	
		mAP	Rank-1	mAP	Rank-1
Market1501	PCB	31.5	43.4	21.8	32.7
	MaskReID	23.7	34.6	12.9	25.4
	PABR	47.3	54.7	33.6	42.9
	OSNet	53.2	61.4	38.4	52.7
	PTGAN	22.6	31.5	13.7	18.6
	SPGAN	31.3	47.3	19.5	32.4
	Baseline	52.5	63.2	41.5	53.6
	OurMethod	**57.4**	**66.3**	**43.7**	**58.4**
DukeMTMC	PCB	28.6	41.7	22.7	36.2
	MaskReID	21.8	34.7	17.4	26.3
	PABR	43.5	54.2	36.3	48.7
	OSNet	45.3	59.5	41.5	56.3
	PTGAN	32.4	39.1	28.4	42.3
	SPGAN	38.5	47.9	36.2	45.8
	Baseline	44.6	69.3	43.2	55.7
	OurMethod	**51.4**	**74.6**	**46.4**	**61.3**
MSMT17	PCB	35.4	47.2	28.7	43.2
	MaskReID	24.8	35.2	21.6	33.9
	PABR	46.5	62.3	41.2	55.8
	OSNet	53.4	68.3	42.3	61.5
	PTGAN	41.6	52.7	31.4	43.7
	SPGAN	46.3	57.2	39.6	51.7
	Baseline	54.6	71.3	45.3	71.6
	OurMethod	**58.2**	**79.2**	**48.6**	**78.8**

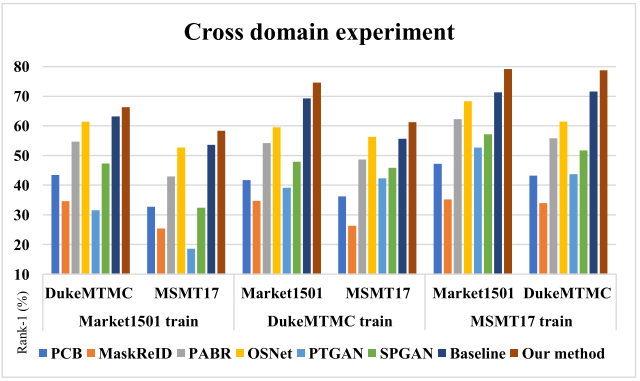


FIGURE 10. Rank-1 score histogram of each method on the other two data sets after training with one data set.

which dataset it was trained on. It is proved that our proposed method is robust to the cross-domain problem.

Finally, Figure 11 shows a query result of our method in each group of contrast experiments. From the top 5 of the

FIGURE 11. Our method uses one query result in each group of experiments, and each query corresponds to the rank list of top5.

rank list corresponding to each query, it can be seen that our method has strong robustness and high accuracy.

In general, our method can solve the occlusion problem, the similarity problem of features between the target person and other pedestrians, as well as cross-domain problems, and it has strong robustness.

VI. CONCLUSION

In this article, we proposed a person ReID method based on effective features and self-optimized pseudo-label. This method extracts accurate and recognizable features through the feature aggregation module and head-shoulder feature auxiliary module to solve the problems of occlusion and similarity. We designed a self-optimized pseudo-label training method to improve the generalization ability of the model and solve the cross-domain problem. Many contrast experiments on three public datasets and self-designed occlusion and similarity datasets show that the proposed method is superior to other advanced methods in most cases, which proves the superiority of the proposed method. This method provides a robust scheme for the deployment of person ReID in reality. In person ReID tasks, the efficiency problem cannot be ignored. In the future, we will research the efficiency issue, and try to improve the detection efficiency of the model on the premise of ensuring accuracy.

REFERENCES

[1] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps, and R. J. Radke, "A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 523–536, Mar. 2019.

[2] Y.-C. Chen, X. Zhu, W.-S. Zheng, and J.-H. Lai, "Person re-identification by camera correlation aware feature augmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 392–408, Feb. 2018.

[3] S. Zhang, Q. Zhang, X. Wei, Y. Zhang, and Y. Xia, "Person re-identification with triplet focal loss," *IEEE Access*, vol. 6, pp. 78092–78099, 2018.

[4] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1239–1248.

[5] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3492–3506, Jul. 2017.

[6] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1335–1344.

[7] H. Wang, H. Du, Y. Zhao, and J. Yan, "A comprehensive overview of person re-identification approaches," *IEEE Access*, vol. 8, pp. 45556–45583, 2020.

[8] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2285–2294.

[9] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. C. Kot, and G. Wang, "Dual attention matching network for context-aware feature sequence based person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5363–5372.

[10] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2119–2128.

[11] S. Jiao, J. Wang, G. Hu, Z. Pan, L. Du, and J. Zhang, "Joint attention mechanism for person re-identification," *IEEE Access*, vol. 7, pp. 90497–90506, 2019.

[12] 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)," 2017, *arXiv:1711.09349*. [Online]. Available: http://arxiv.org/abs/1711.09349

[13] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 274–282.

[14] Y. Zhang, X. Gu, J. Tang, K. Cheng, and S. Tan, "Part-based attribute-aware network for person re-identification," *IEEE Access*, vol. 7, pp. 53585–53595, 2019.

[15] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong, "Partial person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 4678–4686.

[16] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, "Spindle Net: Person re-identification with human body region guided feature decomposition and fusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 907–915.

[17] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose invariant embedding for deep person re-identification," 2017, *arXiv:1701.07732*. [Online]. Available: http://arxiv.org/abs/1701.07732

[18] R. Layne, T. Hospedales, and S. Gong, "Person re-identification by attributes," in *Proc. Brit. Mach. Vis. Conf.*, 2012, pp. 24.1–24.11.

[19] C. Liu, S. Gong, C. C. Loy, and X. Lin, "Person re-identification: What features are important?" in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 391–401.

[20] Z. Shi, T. M. Hospedales, and T. Xiang, "Transferring a semantic representation for person re-identification and search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4184–4193.

[21] M. M. Kalayeh, E. Basaran, M. Gokmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1062–1071.

[22] L. Qi, J. Huo, L. Wang, Y. Shi, and Y. Gao, "MaskReID: A mask based deep ranking neural network for person re-identification," 2018, *arXiv:1804.03864*. [Online]. Available: http://arxiv.org/abs/1804.03864

[23] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee, "Part-aligned bilinear representations for person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 1–20.

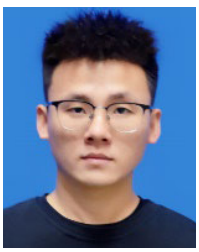
[24] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 994–1003.

[25] J. Liu, Z.-J. Zha, D. Chen, R. Hong, and M. Wang, "Adaptive transfer network for cross-domain person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7202–7211.
[26] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 79–88.
[27] Y. Ding, H. Fan, M. Xu, and Y. Yang, "Adaptive exploration for unsupervised person re-identification," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 1, pp. 1–19, Apr. 2020.
[28] S. Lin, H. Li, C.-T. Li, and A. C. Kot, "Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification," 2018, *arXiv:1807.01440*. [Online]. Available: http://arxiv.org/abs/1807.01440
[29] H. Fan, L. Zheng, C. Yan, and Y. Yang, "Unsupervised person re-identification: Clustering and fine-tuning," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 14, no. 4, p. 83, Oct. 2018.
[30] Y. Fu, Y. Wei, G. Wang, Y. Zhou, H. Shi, U. Uiuc, and T. Huang, "Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6112–6121.
[31] L. Song, C. Wang, L. Zhang, B. Du, Q. Zhang, C. Huang, and X. Wang, "Unsupervised domain adaptive re-identification: Theory and practice," *Pattern Recognit.*, vol. 102, Jun. 2020, Art. no. 107173.
[32] F. Yang, K. Li, Z. Zhong, Z. Luo, X. Sun, H. Cheng, X. Guo, F. Huang, R. Ji, and S. Li, "Asymmetric co-teaching for unsupervised cross-domain person re-identification," in *Proc. AAAI*, 2020, pp. 12597–12604.
[33] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang, "A bottom-up clustering approach to unsupervised person re-identification," in *Proc. AAAI*, vol. 33, Aug. 2019, pp. 8738–8745.
[34] J. Lv and X. Wang, "Cross-dataset person re-identification using similarity preserved generative adversarial networks," in *Proc. Int. Conf. Knowl. Sci., Eng. Manage.* Cham, Switzerland: Springer, 2018, pp. 171–183.
[35] H. Lingxiao, Y. Wang, W. Liu, H. Zhao, Z. Sun, and J. Feng, "Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification," in *Proc. ICCV*, Seoul, South Korea, Oct. 2019, pp. 8450–8459.
[36] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1179–1188.
[37] J. Garcia, N. Martinel, C. Micheloni, and A. Gardel, "Person re-identification ranking optimisation by discriminant context information analysis," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1305–1313.
[38] J. Li, Y. Zhai, Y. Wang, Y. Shi, and Y. Tian, "Multi-pose learning based head-shoulder re-identification," in *Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, Miami, FL, USA, Apr. 2018, pp. 238–243.
[39] T. Ni, Z. Ding, F. Chen, and H. Wang, "Relative distance metric leaning based on clustering centralization and projection vectors learning for person re-identification," *IEEE Access*, vol. 6, pp. 11405–11411, 2018.
[40] A. Wu, W.-S. Zheng, and J.-H. Lai, "Distilled camera-aware self training for semi-supervised person re-identification," *IEEE Access*, vol. 7, pp. 156752–156763, 2019.
[41] J. Lv, W. Chen, Q. Li, and C. Yang, "Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7948–7956.
[42] 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.*, Dec. 2015, pp. 1116–1124.
[43] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline *in vitro*," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3754–3762.
[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[45] J. Zhuo, Z. Chen, J. Lai, and G. Wang, "Occluded person re-identification," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, San Diego, CA, USA, Jul. 2018, pp. 1–6.
[46] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, "Pose-guided feature alignment for occluded person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 542–551.
[47] X. Zhou, D. Wang, and P. Krähenbähl, "Objects as points," 2019, *arXiv:1904.07850*. [Online]. Available: http://arxiv.org/abs/1904.07850
[48] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.
[49] Y. Liu and Y. Ding, "An improved baseline for person re-identification," in *Proc. Int. Conf. Artif. Intell. Pattern Recognit.*, 2019, pp. 46–49.
[50] 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. 46–49.
[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2014, pp. 1–15.

MING-XIANG HE is currently a Professor with the College of Computer Science and Engineering, Shandong University of Science and Technology, China. He is also a member of the National Virtual Simulation Experiment Center, Shandong University of Science and Technology. His current research interests include image processing, artificial intelligence, and database systems.

JIN-FANG GAO is currently pursuing the master's degree in software engineering with the Shandong University of Science and Technology, China. Her current research interests include deep learning and object detection and text recognition.

GUAN LI is currently an Associate Professor with the College of Computer Science and Engineering, Shandong University of Science and Technology, China. She is also a member of the National Virtual Simulation Experiment Center, Shandong University of Science and Technology. Her current research interests include intelligent information processing, artificial intelligence, and information safety.



YOU-ZHI XIN is currently pursuing the master's degree in software engineering with the Shandong University of Science and Technology, China. His current research interests include deep learning and object detection and text recognition.