

Received May 18, 2021, accepted June 13, 2021, date of publication June 22, 2021, date of current version July 16, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3091647

# Cross-Domain Person Re-Identification Based on Feature Fusion

XIANJUN LUO<sup>1,2</sup>, ZHI OUYANG<sup>1</sup>, NISUO DU<sup>1,2</sup>, JINGKUAN SONG<sup>3</sup>, (Member, IEEE), AND QIN WEI<sup>1</sup>

<sup>1</sup>Guizhou Provincial Key Laboratory of Public Big Data, Guiyang 550025, China

<sup>2</sup>College of Computer Science and Technology, Guizhou University, Guiyang 550025, China

<sup>3</sup>School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China

Corresponding author: Nisuo Du (nsdu@gzu.edu.cn)

This work was supported in part by the Major Scientific and Technological Special Project of Guizhou Province China under Grant 20183002, and in part by the Cultivation Project of Guizhou University [Qian-kehe Platform Talents (2017)] under Grant 5788.

**ABSTRACT** Person re-identification (ReID) is one of the commonly used criminal investigation methods in reconnaissance. Although the current ReID has achieved robust results on single domains, the focus of researches has shifted to cross-domain in recent years, which is caused by domain bias between different datasets. Generative Adversarial Networks (GAN) is used to realize the image style transfer of different datasets to alleviate the effect of cross-domain. However, the existing GAN-based models ignore complete expressions and occlusion of pedestrian characteristics, resulting in low accuracy in feature extraction. To address these issues, we introduce a cross domain model based on feature fusion (FFGAN) to fuse global, local and semantic features to extract more delicate pedestrian features. Before extracting pedestrian features, we preprocess feature maps with a feature erasure block to solve an occlusion issue. Finally, FFGAN enables a more complete visual description of pedestrian characteristics, thereby improving the accuracy of FFGAN in identifying pedestrians. Experimental results show that the effect of FFGAN is significantly improved compared with some advanced cross-domain ReID algorithms.

**INDEX TERMS** Person re-identification, global feature, local features, semantic features, feature erasure, cross domain.

## I. INTRODUCTION

Person re-identification (ReID) is defined as a technology of retrieving all pedestrian images which are similar to a random query image according to their clothing, posture, hair style and other features. After combining with face recognition, ReID is applicable to more real scenes, bringing the cognitive level of artificial intelligence to a new stage. Particularly, as ReID is used in criminal investigation and path tracking, it has received more attention in learning network modeling and accuracy improvement.

Although ReID algorithms based on deep learning have achieved remarkable results, it is still a challenge whether models can be efficiently applied to multiple datasets. For a model such as HOREID proposed by Wang *et al.* [1] which has achieved the Rank-1 identification rate of 94.2% on Market1501, when it is trained and tested on

two different datasets, both generalization and accuracy of models decreased. This cross domain problem is often caused by domain bias such as different lighting [2] and camera angles [3] among several datasets, greatly limiting the generalizability of models to real-life situations. Therefore, the robustness and learning adaptivity of ReID models are significant for theoretic researches and further applications.

In the cross-domain setting, most of models use CNN and other operations (i.e. HSV histogram, direction gradient histogram and *et al.*) to extract global features composed of physical appearances (i.e. color, shape, textural features) [4], [5]. However, these approaches likely miss some pedestrian detail features due to diverse appearances of pedestrians captured by different cameras. In order to further explore fine-grained characteristics, other scholars propose local feature extraction methods [6]. They firstly extract detailed features in several horizontal divided parts of images, and then fuse these features to obtain final features to train ReID models and achieved better performances.

The associate editor coordinating the review of this manuscript and approving it for publication was Sudhakar Radhakrishnan<sup>1</sup>.

Although the neural network extracts low-level details by the local feature extraction method [7], it easily causes semantic ambiguity [8]. For example, a model would like to mistake a long-hair man for a woman or a woman who dresses in a neutral way for a man. Scholars propose semantic features expression to describe pedestrian appearances and reduce misjudgment. However, both levels of features are extracted and used for describing pedestrian features separately, which makes it impossible to effectively obtain the complete visual features and more accurately match the same pedestrian.

Furthermore, when we extracting features, pedestrian images are frequently occluded by foreign objects, such as bicycles, stones, handrails, etc [9]. As the automatic detection does not contain complete appearances of pedestrians, it might cause distinct pedestrians with the same occlusion pattern regarded as the same one or semantic analysis errors. On the other hand, when we extract local features, segmented areas may contain an occluded part which belongs to invalid features, increasing the time of model training. As a result, existing models have difficulty receiving satisfied performance especially when pedestrian images of existing datasets are occluded.

Consequently, we attempt to improve the ReID accuracy by integrating global, local and semantic characteristics in the cross-domain setting. Before extracting these features, we append an image preprocessing module, introducing an attention mechanism to focus on critical features and a feature erasure block to solve the occlusion. Finally, rather than simply converting styles from one domain to another, we also learn the style mapping function in a target domain to consider the interchange of styles between various cameras. In summary, the contributions of the paper are:

- 1) We propose a global, local and semantic feature fusion network, which comprehensively considers these three features. The network effectively extracts detailed features and enriches the visual expression of pedestrian characteristics.
- 2) We introduce a feature erasure block, effectively alleviating the pedestrian occlusion when extracting pedestrian features. We introduce a combined attention mechanism to learn hard-level and soft pixel-level attention at the same time, to maximize the correlation and complementarity between attention and features. It effectively improves the accuracy of feature extraction.
- 3) Our method has great advantages over the most advanced UDA method on two large-scale datasets: Market1501 and DukeMTMC-ReID.

The rest of this paper is organized as follows. Section II briefly introduces related work. Section III first introduces the proposed model and several key modules including style transfer review, image preprocessing and feature fusion. In Section IV, extensive experiments are performed to evaluate the model. Finally, some conclusions are put forward in Section V.

## II. RELATED WORK

To solve the cross-domain problem, many methods rely on GAN [10]–[12] to achieve style transfer [13], [14] from a source domain to a target domain. For instance, Wei *et al.* [15] propose Person Transfer GAN (PTGAN), using a foreground image segmentation network to segment pedestrian foreground in the source domain and transferring the segmented foreground to the background of the target domain. It makes foreground of a generated image consistent but background style as the target domain. Deng *et al.* [16] propose Similarity Preserving GAN (SPGAN), which combines CycleGAN [17] and Siamese Network to achieve style transfer. These methods adopt an image generation technology for data augmentation, and also make the image styles of different datasets consistent to reduce the inaccuracy that is likely to occur when ReID models relearn new image features.

There also exists style gaps in the target domain itself, namely intra-domain variations, which attracts scholars' attention [3], [18]. For example, Ren *et al.* [19] propose the CSGLP model simultaneously to learn the transfer mapping among different camera views and multiple domains, exploring StarGAN [20] as a camera style adaptation network. Zhong *et al.* [21] propose the HHL model which uses StarGAN to transfer images from each camera to the others in target domain. In order to alleviate the distribution difference between cameras, Xuan and Zhang [22] propose a new method IICS for calculating the similarity of images in the target domain. IICS divides the calculation of sample similarity into two branches: intra-camera and inter-camera. The calculation in the camera directly uses the features of CNN to calculate the similarity. The calculation between cameras uses the classification scores of each sample on different cameras as a new feature vector, resulting in more reliable pseudo-labels. IICS uses the pseudo-labels inside and between the cameras to train the model in two stages. Lin *et al.* [23] propose a multi-dataset feature generalization network (MMFA-AAE). MMFA-AAE is based on an adversarial auto-encoder to learn the invariant feature representation of multiple labeled datasets and generalize it to the invisible camera system. The Maximum Mean Difference (MMD) metric is used to align the data distribution of multiple domains. These studies show that cross domain models use the GAN framework to transform source images into target domain styles in a holistic manner for further domain gap reduction.

In the context of cross-domain, the method of feature extraction restricts the accuracy of ReID models. Some scholars use CNN to extract pedestrian global features such as color, shape, texture and etc [1], [24]. For instance, both HHL [21] and ECN [25] model use GAN to transfer images from each camera to the others in target domain and utilize CNN to extract pedestrian global features to identify most of distinctive features, but they cannot distinguish between details and thus miss some features.

In order to extract further fine-grained pedestrian characteristics, some scholars propose the local feature extraction

method. For example, inspired by Part-based Convolutional Baseline (PCB) propose by Sun *et al.* [6], Huang *et al.* [26] propose EANet model consisting of Part Aligned Pooling (PAP) and Part Segmentation (PS) based on multi-task training. Through the alignment of local features and segmentation model training, the performance of models in target domain recognition under cross-domain setting is improved. Zhang *et al.* [27] propose PAST model which also uses PCB method and extracts the local characteristics of pedestrians by feature blocks after extracting global features. These models cleverly use segmentation to extract local features, which increase the pedestrian details and enhancing visual perception of contextual changes in the source and target area. Although local features contain rich detailed features, they have little semantic image understanding and a weak ability to classify images.

To receive the missing semantic information, some scholars introduce semantic feature extraction into traditional methods. For instance, Li *et al.* [28] propose SCGAN consisting of two types of semantic constraints, namely attentional constraints and identity constraints, to enhance the recognition area to ensure the stability of the person's identity and increase the diversity of training samples. The network takes advantage of the complementarity of global features and semantic features, and combines them to describe more complete pedestrian feature expression. However, this approach ignores the significance of local features and thus does not provide a complete pedestrian representation.

Furthermore, when we extract pedestrian features, we often encounter partly occluded pedestrians [29]. Some scholars have designed appropriate feature selection and filtering strategies to reduce the noise caused by occlusion. The usual approach is to reduce the weight of features in occluded areas during feature comparison. For example, Su *et al.* [30] propose the Pose-driven Deep Convolutional (PDC) model, which uses the fully connected layer to automatically learn the weight of each component feature to reduce the occluded area. Some scholars use the method of detecting key points of a pedestrian body to locate non-occluded body regions, and then extract features of the non-occluded body regions to match. For instance, Zhao *et al.* [9] propose SpindleNet, which uses posture estimation to detect key points of a body and divides seven body part regions, comparing feature values according to the pixel layer and retaining features of maximum value to achieve noise suppression. Other scholars use erasure methods to erase part of images. For instance, Ghiasi *et al.* [24] propose DropBlock model to randomly erase an adjacent region of each tensor. All of the methods dealing with occlusion above enhance the recognition ability of deep neural networks for occluded images. However, these methods are rarely applied in the cross-domain setting and some methods demand additional body annotation and reliable pose estimation. Therefore, instead of using these methods, we introduce a feature erasure block. The difference between feature erasure and DropBlock is that the feature erasure drops the occluded area of global feature maps in

preprocessing, while DropBlock randomly drops a different area. By removing the occluded areas, we reduce the impact of occlusion on the performance during the ReID test and avoid the mistake of removing key features due to the randomness of DropBlock.

Therefore, we propose a cross-domain person ReID based on feature fusion network (FFGAN) to solve these issues mentioned above. With the complementarity of global, local and semantic features, we attempt to give full consideration to pedestrians' detailed characteristics and enhance rich contextual variation in character images to facilitate learning of FFGAN. We erase pedestrian occluded areas of images to solve the occlusion when extracting pedestrian features and using the attention mechanism to enhance the ability of network to distinguish between key target areas (pedestrians subject) and background areas. FFGAN fully extracts pedestrian characteristics and effectively enhances the robustness and generalization in the cross-domain setting.

### III. PROPOSED METHOD

#### A. THE MAIN NETWORK

In this section, we present the proposed FFGAN. The structure is shown in Fig. 1. We use ResNet101 as a backbone network and ImageNet as a pre-trained model. Generally, ReID models based on deep learning usually include two steps: feature extraction [31], [32] and metric learning [33], [34]. Before the two steps, we add an image preprocessing step and introduce a feature erasure block to solve the occlusion in this step. It is noteworthy that we do not erase the input image directly, but erase the feature tensor obtained at the preprocessing, simplifying the complexity of the subsequent feature extraction of FFGAN. The erased feature maps are used to extract global, local, and semantic features, respectively. In terms of feature extraction, firstly, we use convolution and global average pooling operations to extract global features. Secondly, the erased feature maps evenly are divided into six regions, and with each region, we add classifier training to each region to extract local features of the six regions respectively. Thirdly, we introduce a multi-level factor decomposition block and a feature selection block to extract semantic features of erased feature maps. After extracting global, local and semantic characteristics of pedestrians respectively, concatenation is used to connect these three feature maps to form the final pedestrian expression feature map. In the metric learning, we use cosine distance to calculate the distance between query images and all gallery images, and sort all the gallery images by distance. Eventually, we calculate Mean Average Precision (mAP) of our network to measure the model effect.

#### B. STYLE TRANSFER REVIEW

In order to achieve camera style transfer in the target domain, we use CycleGAN method. As shown in Fig. 2, our work realizes image styles interchange between two domains. Two sample domains are given:  $S$  (source domain) and  $T$  (target domain). We convert image styles of the domain  $S$

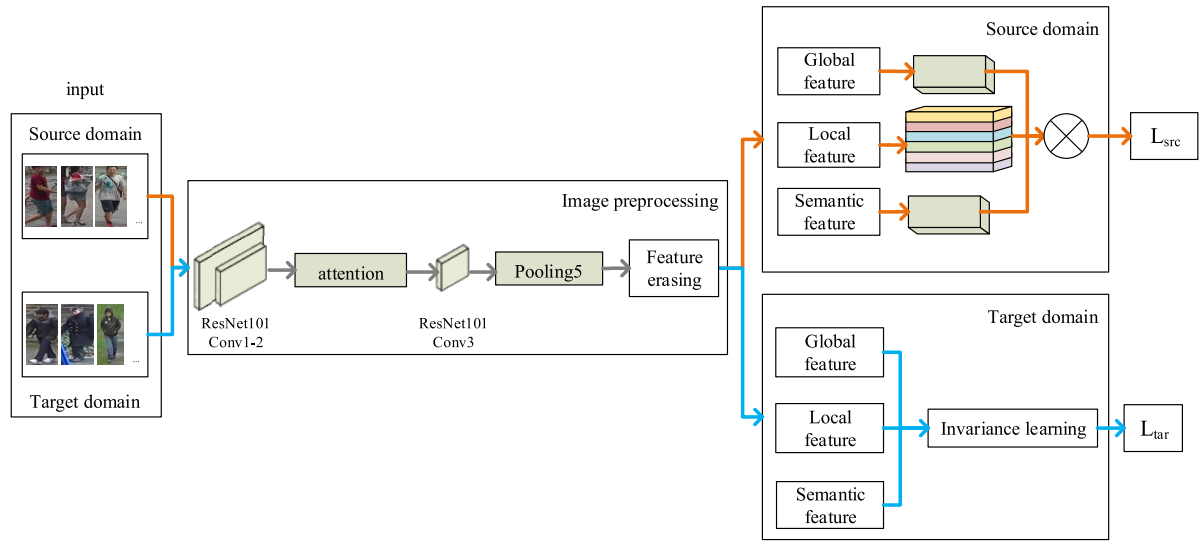


**FIGURE 1.** The framework of the proposed FFGAN. It consists of two parts: image preprocessing module and feature fusion module. In the preprocessing process, we erase the occluded areas. In the following feature fusion process, we extract global, local and semantic features respectively. Next, we fuse these three features to obtain the complete expression of pedestrians. Finally, we transfer the source domain style to the target domain.



**FIGURE 2.** The framework of style transfer. We implement camera style transfer between Market1501 and, such as transferring from camera 1 to 2, 3, 4, 5, 6 and transferring from camera 2 to 1, 3, 4, 5, 6.

into samples of the domain  $T$ . The network learns mapping function  $F$  from  $S$  to  $T$ .  $F$  converts an image  $s$  in  $S$  to an image  $F(s)$  in  $T$ . The discriminator  $D_T$  distinguishes whether the style of generated images belongs to the style of  $T$ . According to the generator and discriminator, we construct a GAN loss. The expression is:

$$L_{GAN} = E_{s \sim P_{data(s)}}[\log(1 - D_T(F(s)))] + E_{t \sim P_{data(t)}}[\log D_T(t)] \quad (1)$$

In addition to transfer image style of source domains to target domains, our model also carries out bidirectional transfer of target domains to source domains. We assume a mapping function  $G$ , which transforms an image  $t$  in  $T$  space into an image  $G(t)$  in  $S$ . Our method learns two mappings of  $F$  and  $G$  simultaneously, and requires  $F(G(t)) \approx t$  and  $G(F(s)) \approx s$ . To achieve this requirement, the loss of cycle consistency is defined as:

$$L_{cyc}(F, D, S, T) = E_{t \sim P_{data(t)}}[\|G(F(s))\|_1] + E_{t \sim P_{data(t)}}[\|F(G(t)) - t\|_1] \quad (2)$$

At the same time, we introduce a discriminator  $D_S$  and define a loss  $L_{GAN}(F, D_S, S, T)$ . The final loss consists of above loss functions. The expression is:

$$L = L_{GAN}(F, D_T, S, T) + L_{cyc} + L_{GAN}(F, D_S, S, T) \quad (3)$$

Finally, we use the transferred dataset to enhance the target dataset.

### C. IMAGE PREPROCESSING

In the process of image preprocessing, as shown in Fig. 3, we introduce a feature erasure block to solve occlusion. We need to present feature tensors computed from a single batch of input images by the backbone network previously. In order for these tensors to contain more pedestrian characteristics, as shown in Fig. 4, we add hard attention, soft attention, spatial attention, and channel attention to the second layer of the backbone network. The soft attention aims at selecting fine-grained significant pixels, while the hard attention is devoted to search for weakly supervised discriminative regions. The spatial attention focuses on different locations of feature maps, ignoring channel domains and treating pedestrian features in each channel equally yet. The channel attention varies for obvious image channels, ignoring local features within each channel. Therefore, we combine the four attention mechanisms to design a model of the attention mechanism for mixed domains. They are essentially complementary in function and highly compatible with each other. The combination of the four attention mechanisms makes the feature extraction more discriminative and robust.

After receiving feature tensors, we erase these occluded areas. We adopt a mask that rubbed off an area in a bottleneck feature map. To find the location of this area, we use human

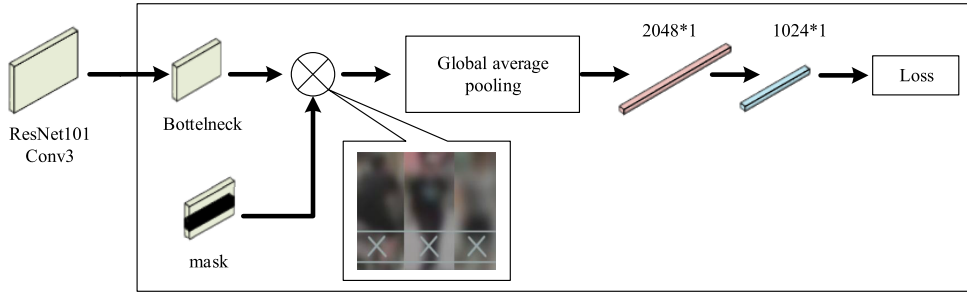


FIGURE 3. The illustration of the feature erasure block.

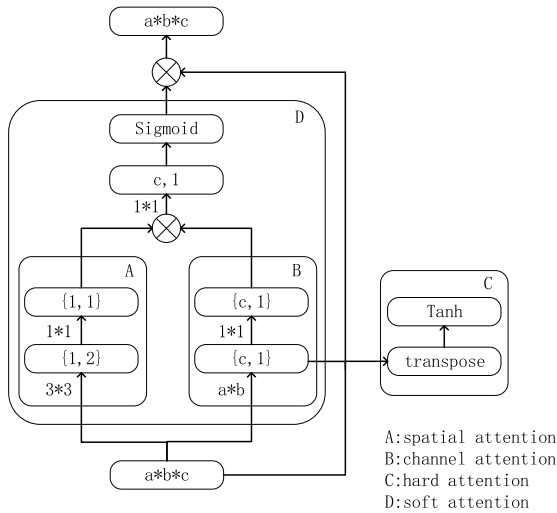


FIGURE 4. The schematic diagram of attention mechanism.

pose estimator to detect pedestrian landmarks and calculate the confidence score  $\delta$  that is set to a threshold value for each landmark. When  $\delta < 0.2$ , it indicates that the block is an occluded area. The erased feature maps go through a global maximum pooling layer to obtain  $2048 \times 1$  features. Subsequently, a convolutional layer is used to reduce the dimension to  $1024 \times 1$  feature vectors. Feature erasure erases the occluded areas that are guided by the mask detected by pose estimation rather than erasing random areas as DropBlock [24].

#### D. FEATURES FUSION

In the process of feature fusion, we need to extract characteristics of pedestrians firstly. As Fig. 1 shows, the network of features fusion model includes three parts: global feature extraction, local feature extraction, and semantic feature extraction.

Firstly, we extract global characteristics of pedestrians using convolution and pooling.

Secondly, in order to extract detail features more effectively, as shown in the Fig. 5, we use block method to extract local features. We obtain a feature tensor  $T$  after erasing features, and then divide it horizontally into six equal parts, obtaining a six column vector  $g$  after average pooling. Next, the model uses a  $1 \times 1$  convolution kernel to reduce dimension to receive a six dimensionality-reduced feature  $h$ .

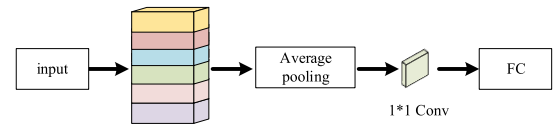


FIGURE 5. The framework of local features extraction block.

Finally, when training  $h$ , we use the sum of cross entropy of six classifiers as a loss function. In testing, we splice six vectors  $g$  or six vectors  $h$  to get a local feature map.

Thirdly, as shown in the Fig. 6, we use a multilayer factorization network to extract semantic features. The semantic feature extraction network consists of multiple blocks ( $B_1, B_2, \dots, B_n$ ) on several semantic levels. Each block contains two substantial components: a set of feature extraction modules ( $FE_1, FE_2, \dots, FE_n$ ) and a feature selection module (FS). Both of components are composed of multiple convolutional layers. In each FE modules, the bottom one represents low-level semantic features such as clothing color, while the top module represents high-level semantic features such as carrying items and gender. The output vectors of these blocks overall provide us with corresponding semantic features. These output vectors are associated with a factor feature which is fused with the last layer feature to obtain pedestrian description features.

After the extraction of global feature, local feature and semantic feature respectively and the up-sampling operation, we conduct scale normalization processing on the three feature maps respectively. Concat-connection of global feature and local feature is carried out as the low-level feature  $f_l$ , and semantic feature is taken as the middle-level feature  $f_m$ . Because the dimensions and channels of the  $f_l$  and  $f_m$  features are inconsistent with each other, and the dimensions of  $f_m$  are relatively high, the up-sampling is used first to reduce the dimension to the same size, and then the  $1 \times 1$  convolutional layer is used to make the number of channels of  $f_l$  and  $f_m$  consistent. In addition to the differences in dimensions and channels between the low-level and middle-level features, the focus points will also be slightly different. The position in an image with a larger feature vector value indicates that the region corresponds to the key feature, such as the low-level feature more emphasis on color, texture, and shape, while the middle-level features emphasize the semantic attributes of pedestrians (whether backpacks, carry items, etc.).

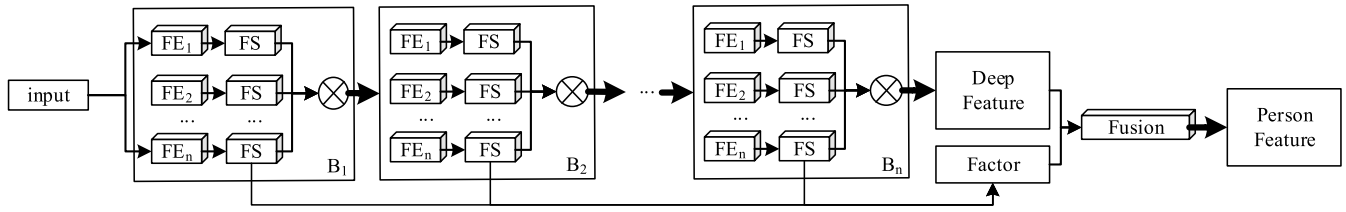


FIGURE 6. The illustration of semantic features extraction.

If we simply use add or concat operation [35], [36] to splice directly, it will inevitably bring about the consequences of reduced feature effectiveness.

Therefore, we use the significant difference and complementarity of these two types of features, and learn from the Hadamard Product method to solve this problem with significant differences. As shown in Fig. 7, the solution is to multiply the corresponding positions of the feature matrix vectors of the low-level features  $f_l = [b_{ij}]_{H \times W}$  and the middle-level features  $f_m = [a_{ij}]_{H \times W}$  from a geometric perspective, increase the feature value of the region with the larger feature value in the original two image vectors, and reduce the feature value of the region with larger difference. And then FFGAN achieves the purpose of extracting common features and reducing significant differences.

$$f_{res} = f_l \otimes f_m = [b_{ij}]_{H \times W}, \quad c_{ij} = a_{ij} * b_{ij} \quad (4)$$

Finally, the calculated results  $f_{res}$  and  $f_m$  are added together to obtain the final expression of pedestrian  $f_{fusion}$ , which contains rich details and semantic features, so as to improve the deficiency of image representation at the bottom of the network. Calculated as follows:

$$f_{fusion} = f_{res} \oplus f_m \quad (5)$$

### E. LOSS FUNCTION

The loss function is mainly composed of three parts: the loss function in source and target domain respectively and the Center Loss [37].

The loss function in the source domain is optimized by the cross-entropy function, which can be expressed as:

$$L_{src} = -\frac{1}{n_s} \sum_{i=1}^{n_s} \log p(a_{s,i} | b_{s,i}) \quad (6)$$

where,  $n_s$  is the number of images in the source domain,  $p(a_{s,i} | b_{s,i})$  is the probability of image  $b_{s,i}$  in the source domain, and  $i$  belongs to the identity of  $a_{s,i}$ .

The loss function of the target domain is composed of three invariant loss functions, namely sample invariance, camera invariance and neighbor invariance.

#### 1) SAMPLE INVARIANCE

In order to increase the distance between pedestrian images with different IDs, the sample invariant loss function is used to constrain ReID model training. Given a target image  $x_{t,i}$ , at the outset, the similarity between the feature of  $x_{t,i}$  and

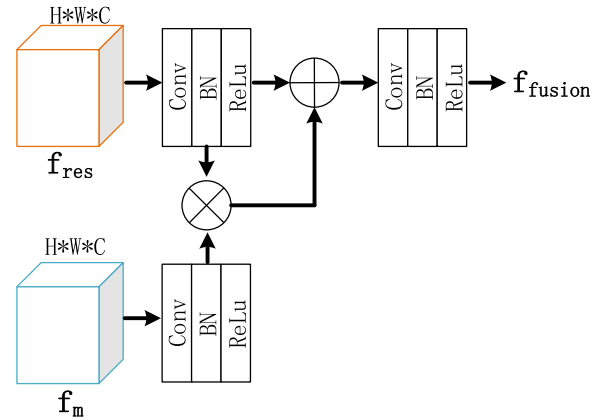


FIGURE 7. The illustration of semantic features extraction.

the feature with a network memory is calculated by cosine distance. Softmax function is used to predict the probability  $p(i | x_{t,i})$  that belongs to the  $i$ -th image class as:

$$p(i | x_{t,i}) = \frac{e^{\frac{W[i]^T h(x_{t,i})}{\omega}}}{\sum_{j=1}^N e^{\frac{W[j]^T h(x_{t,i})}{\omega}}} \quad (7)$$

where,  $\omega$  is a hyperparameter satisfying,  $\omega \in (0,1]$ .

The sample invariance loss function would be expressed as:

$$L_{si} = -\log p(i | x_{t,i}) \quad (8)$$

#### 2) CAMERA INVARIANCE

In the process of camera style-transfer, the pedestrian image with the same ID remains the same after transferring. Camera invariability is used to shorten the distance between original images with the same ID and the transferred images. Each camera is regarded as a domain, and pedestrian images would be converted into the same ID images of other camera styles by using CycleGAN, which would be expressed as:

$$L_{ci} = -\log p(i | x'_{t,i}) \quad (9)$$

#### 3) NEIGHBOR INVARIANCE

Adjacent samples are likely to be similar, so at the beginning, cosine distance is used to calculate the similarity between the image feature  $h(x_{t,i})$  and its stored feature. The pedestrian image index  $Q(x_{t,i}, k)$  which is adjacent to the pedestrian image  $x_{t,i}$ ,  $K$  is found in the stored features.

We define the probability weight of images of the target domain:

$$\omega_{r,j} = \begin{cases} \frac{1}{k} & i \neq j \\ 1 & i = j \end{cases} \quad (10)$$

The neighbor invariance loss function can be expressed as:

$$L_{ni} = - \sum_{j \neq i} \omega_{i,j} \log p(j | x_{t,i}) \quad (11)$$

The target domain loss function can be expressed as:

$$L_{tar} = - \frac{1}{n_t} \sum_{i=1}^{n_t} \sum_j \omega_{i,j} \log p(j | x'_{t,i}) \quad (12)$$

The distance between classes is not obvious and compact enough when different categories in the source domain are mapped to the feature space. Center Loss further restricts generated features to provide with a feature center for each category and minimizes the distance between each sample in Mini-batch and the corresponding center. Besides, it helps to reduce the intra-class distance and significantly increase discriminability of samples in the feature space. Hence, we add Center Loss into the loss function in the source domain.

Center Loss is expressed as:

$$L_c = \frac{1}{2} \sum_{i=1}^m \|x_i - C_{y_i}\|_2^2 \quad (13)$$

Eventually, the total loss function is explained as:

$$L = (1 - \rho)(L_{src} + \sigma L_c) + \rho * L_{tar} \quad (14)$$

## IV. EXPERIMENTAL ANALYSIS

### A. DATASETS AND ASSESSMENT INDICATORS

#### 1) PERSON REID DATASETS

In order to verify the validity of algorithms in this paper, we test on two datasets, Market1501 and DukeMTMC-ReID.

The dataset Market1501 contains 1501 pedestrians captured by 6 cameras. Bounding boxes detect 32,668 pedestrian images. The training set contains 12,936 images of 751 pedestrians and the testing set contains 19,732 images of 750 pedestrians. The query set contains 3,368 pedestrian images.

The dataset DukeMTMC-ReID includes 1404 pedestrians captured by 8 cameras. Bounding boxes detect 36,411 pedestrian images. The training set contains 16,522 images of 702 pedestrians. The testing set contains 17,661 images of 702 pedestrians. The query set contains 2,228 pedestrian images.

#### 2) EVALUATION CRITERIA FOR REID

Two criteria are commonly used for ReID assessment, namely, the cumulative matching characteristic (CMC) and Mean Average Precision (mAP).

### B. EXPERIMENTAL PARAMETER SETTING

We use Pytorch on the NVIDIA TITAN Xp GPU. The programming language is Python 3.7 on Ubuntu 18.04.3 platform.

#### 1) CAMERA STYLE-TRANSFERRED

Market1501 is shot by 6 cameras (6 subdomains). As each camera style of transfer learning needs to be transferred to other cameras, FFGAN training consists of  $C_6^2(15)$  sub-models. DukeMTMC-ReID is filmed by 8 cameras (8 subdomains). FFGAN training required  $C_8^2(28)$  sub-models. During model training, the pixel size of input images is set to 256\*256. Random flipping and random erasing are used for data enhancement. We use Adam optimizer to optimize the model. The learning rates of discriminator and generator are set as 0.0001 and 0.0002 respectively. Learning rates linearly attenuate to zero in the remaining 20 epoches.

#### 2) MAIN NETWORK

To improve network performance, we select ResNet101 as backbone and initialize network weights using pre-trained parameters of ImageNet. In order to extract more pedestrian feature, the image size is adjusted to 256\*128. The datasets are preprocessed by random erasing, random cropping, random flipping and image normalization. 128 samples are taken from a training set for each training. Adam optimizer optimizes the model with a weight attenuation coefficient of 0.0005 and initializes a learning rate of 0.1. Warmup optimizes the learning rate. We set the loss weight  $\rho=0.1$  and  $\sigma=0.05$ .

### C. PERFORMANCE EVALUATION

#### 1) THE EFFECTIVENESS OF SEMANTIC EXTRACTION BLOCK

We evaluate the effect of semantic feature extraction. As reported in Table 1, we introduce “Ours w/AP” to represent our basic model. We add the semantic feature extraction block to the basic model. “Ours w/AP+M” obtain 79.5% in Rank-1 accuracy and 46.7% in mAP accuracy when we transfer from DukeMTMC-ReID to Market1501. This outperforms the baseline model by 4.4% in Rank-1 accuracy and 3.7% in mAP accuracy, respectively. A similar improvement is achieved when the model proposed transfers to DukeMTMC-ReID. Instead of extracting only global or local features, it shows that semantic feature extraction can further extract more powerful features based on extracting global and local features. When fuse these three features, we get a more comprehensive visual representation of pedestrians and guide the training and testing of our model.

#### 2) THE EFFECTIVENESS OF CENTER LOSS BLOCK

We evaluate the effect of Center Loss block. As reported in Table 1, Center Loss block significantly improves performance of the baseline and our basic model. When we append Center Loss block to the basic model and transfer from DukeMTMC-ReID to Market1501, the performance of “ours w/AP+M+C” is 79.9% accuracy in Rank-1 and 47.2%

**TABLE 1. A composite comparison of different modules. A: attention mechanism block, P: local feature extraction block, M: semantic feature extraction block, C: Center Loss Block, B: feature erasure block.**

| Methods         | DukeMTMC-ReID → Market1501 |        |         |      | Market1501 → DukeMTMC-ReID |        |         |      |
|-----------------|----------------------------|--------|---------|------|----------------------------|--------|---------|------|
|                 | Rank-1                     | Rank-5 | Rank-10 | mAP  | Rank-1                     | Rank-5 | Rank-10 | mAP  |
| baseline        | 75.1                       | 87.6   | 91.6    | 43.0 | 63.3                       | 75.8   | 80.4    | 40.4 |
| Ours w/AP+M     | 79.5                       | 89.8   | 92.6    | 46.7 | 66.2                       | 77.3   | 81.1    | 42.3 |
| Ours w/AP+M+C   | 79.9                       | 89.6   | 92.5    | 47.2 | 66.8                       | 77.2   | 80.2    | 43.6 |
| Ours w/AP+M+C+B | 82.1                       | 88.6   | 91.2    | 61.7 | 68.8                       | 77.4   | 80.9    | 55.2 |

**TABLE 2. Compare with the most advanced unsupervised method on the Market1501 and DukeMTMC-ReID datasets. “-” means no results are provided. We report the accuracy of Rank-1 and mAP. The effects of the experiments in this paper are shown in bold.**

| Methods        | DukeMTMC-ReID → Market1501 |             |             |             | Market1501 → DukeMTMC-ReID |             |             |             |
|----------------|----------------------------|-------------|-------------|-------------|----------------------------|-------------|-------------|-------------|
|                | Rank-1                     | Rank-5      | Rank-10     | mAP         | Rank-1                     | Rank-5      | Rank-10     | mAP         |
| LOMO [38]      | 27.2                       | 41.6        | 49.1        | 8.0         | 12.3                       | 21.3        | 26.6        | 4.8         |
| BoW [39]       | 35.8                       | 52.4        | 60.3        | 14.8        | 17.1                       | 28.8        | 34.9        | 3.8         |
| UMDL [41]      | 34.5                       | 52.6        | 59.6        | 12.4        | 18.5                       | 31.4        | 37.6        | 7.3         |
| PUL [42]       | 45.5                       | 60.7        | 66.7        | 20.5        | 30.0                       | 43.3        | 48.5        | 16.4        |
| CAMEL [40]     | 54.5                       | -           | -           | 26.3        | -                          | -           | -           | -           |
| PTGAN [15]     | 38.6                       | -           | 66.1        | -           | 27.4                       | -           | 50.7        | -           |
| SPGAN+LMP [16] | 57.7                       | 75.8        | 82.4        | 26.7        | 46.4                       | 62.3        | 68.0        | 26.2        |
| TJ-AIDL [32]   | 58.2                       | 74.8        | 81.1        | 26.5        | 44.3                       | 59.6        | 65.0        | 23.0        |
| Camstyle [43]  | 58.8                       | 78.2        | 84.3        | 27.4        | 48.4                       | 62.5        | 68.9        | 25.1        |
| HHL [21]       | 62.2                       | 78.8        | 84.0        | 31.4        | 46.9                       | 61.0        | 66.7        | 27.2        |
| ECN [25]       | 75.1                       | 87.6        | 91.6        | 43.0        | 63.3                       | 75.8        | 80.4        | 40.4        |
| UDAP [44]      | 75.8                       | 89.5        | 93.2        | 53.7        | 68.4                       | 80.1        | 83.5        | 49.0        |
| HCT [45]       | 80.0                       | 91.6        | 95.2        | 56.4        | 69.6                       | 83.4        | 87.4        | 50.7        |
| ACT [46]       | 80.5                       | -           | -           | 60.6        | 72.4                       | -           | -           | 54.5        |
| <b>FFGAN</b>   | <b>82.1</b>                | <b>88.6</b> | <b>91.2</b> | <b>61.7</b> | <b>68.8</b>                | <b>77.4</b> | <b>80.9</b> | <b>55.2</b> |

accuracy in mAP, increasing the results of the baseline model by 4.8% in Rank-1 accuracy and 4.2% in mAP accuracy and being higher than “Ours w/AP+M” by 0.4% in Rank-1 accuracy and 0.5% in mAP accuracy, respectively. It indicates that Center Loss makes the distance between pedestrians with the same ID close more similar. It tries to move closer to the sample center, increasing the inter-class distance and reducing the inter-class distance.

### 3) THE EFFECTIVENESS OF FEATURE ERASURE BLOCK

After feature erasure is added in this paper, “Ours w/AP+M+C+B”(FFGAN) gains 82.1% in Rank-1 accuracy and 61.7% in mAP accuracy when we transfer from DukeMTMC-ReID to Market1501, outperforming the results of the baseline model by 7% in Rank-1 accuracy and 18.7% in mAP, respectively. Furthermore, it enhances the results of the “Ours w/AP+M+C” by 2.2% in Rank-1 accuracy and 14.5% in mAP accuracy, respectively. Compared with the previous method of erasing any area of images, FFGAN directly erases the occluded area with the help of the pose estimator. It shows that feature erasure effectively alleviates the influence of occlusion and improve the performance of the model.

### D. COMPARISON WITH STATE-OF-THE-ART METHODS

FFGAN shows Rank-1 accuracy and mAP accuracy for transferring Market1501 to DukeMTMC-ReID and vice versa in Table 2. We compare with two non-transfer learning methods (LOMO [38] and Bow [39]), three domain adaptive methods (CAMEL [40], UMDL [41] and PUL [42]), and six unsupervised domain adaptation methods (PTGAN [15], SPGAN [16], TJ-AIDL [32], CamStyle [43], HHL [21], ECN [25], UDAP [44], ACT [45], HCT [46]).

We achieve 82.1% in Rank-1 accuracy and 61.7% in mAP when transferring from DukeMTMC-ReID to Market1501, 68.8% in Rank-1 accuracy and 55.2% in mAP when transferring from Market1501 to DukeMTMC-ReID.

First, we compare methods based on manual features that use feature to recognize one person with our method, but with negative results. When we transfer from DukeMTMC-ReID to Market1501, LOMO’s Rank-1 accuracy is 27.2% and mAP is 8.0%. BOW’s Rank-1 accuracy is 35.8% and mAP is 14.8%. Compared with these methods, our method boosts the performance over LOMO with margins of 54.9% in Rank-1 accuracy and 53.7% in mAP accuracy. In addition, our method outdoes Bow by 46.9% in Rank-1 accuracy and 46.3% in mAP accuracy.

Next, we compare three methods which use labeled source data to initialize the model and unlabeled target data to train. When we transfer from DukeMTMC-ReID to Market1501, UMDL, PUL and CAMEL obtains a Rank-1 accuracy rate of 34.5%, 45.5% and 54.5%, respectively. Compared with these methods, our method surpasses the best of these methods CAMEL by 27.6% in Rank-1 accuracy and 35.4% in mAP accuracy.

Compared with the most advanced domain adaptation methods, our method is significantly preferable. When we transfer from DukeMTMC-ReID to Market1501, PTGAN, SPGAN+LMP, TJ-AIDL, Camstyle, HHL and ECN obtain a Rank-1 accuracy rate of 38.6%, 57.7%, 58.2%, 58.8%, 62.2%, 75.1%, respectively. We observe that the accuracy of mAP and Rank-1 of our model achieves a significant gain of 18.7% and 14.8% in terms of mAP comparing with ECN [49], respectively. Compared with cross-domain person re-identification models in 2020, when transferred from Market1501 to DukeMTMC-ReID, the Rank-1 and mAP of FFGAN were 6.3% and 8% higher than UDAP, 2.1% and 5.3% higher than HCT, and 1.6% and 0.9% higher than ACT, respectively. These results demonstrate the benefits of merging GAN-based image-level learning [47], [48] and CNN-based feature-level learning [49] into unsupervised cross-domain ReID, and more importantly show that the significant advantages of using feature fusion in learning across domain ReID models.

## V. CONCLUSION

In this paper, we design a cross-domain person re-identification model based on feature fusion (FFGAN). The visual features of the image are enriched by a well-designed scheme of global, local and semantic feature fusion, and then the effects of occlusion are attenuated by a feature erasure block. We perform experimental analyses to verify the robustness of our model. FFGAN achieves competitive advantages over existing GAN-based ReID models for several ReID benchmarks in cross-domain settings, which demonstrates the great advantage of FFGAN for more detailed, discriminative and robust features in cross-domain settings.

At present, many scholars still adopt the method of style transfer for cross-domain problems. The foreground of the migration from the source domain to the target domain does not have the style of the target domain camera. In addition, the transferred image background quality is lower than the original image quality. Pedestrian posture is also a major factor affecting ReID. In the follow-up work, pedestrian posture will be considered for further study on how to improve the quality of images generated by GAN.

## REFERENCES

- [1] G. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, E. Zhou, and J. Sun, "High-order information matters: Learning relation and topology for occluded person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6449–6458.
- [2] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, S. Maneesh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 36–52.
- [3] L. Qi, L. Wang, J. Huo, L. Zhou, Y. Shi, and Y. Gao, "A novel unsupervised camera-aware domain adaptation framework for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Oct. 2019, pp. 8080–8089.
- [4] 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. (CVPR)*, Jun. 2016, pp. 1335–1344.
- [5] B. Chen, Y.-F. Zha, and Y.-Q. Li, "Person re-identification based on convolutional neural network discriminative feature style learning," *Acta Opt. Sinica*, vol. 38, no. 7, 2018, Art. no. 0720001.
- [6] Y.-F. Sun, L. Zheng, and Y. Yang, "Beyond part models: Person retrieval with refined part pooling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Sep. 2019, pp. 501–518.
- [7] Y. Sun, Q. Xu, Y. Li, C. Zhang, Y. Li, S. Wang, and J. Sun, "Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 393–402.
- [8] W. Yu, K. Yang, H. Yao, X. Sun, and P. Xu, "Exploiting the complementary strengths of multi-layer CNN features for image retrieval," *Neurocomputing*, vol. 237, pp. 235–241, May 2017.
- [9] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, "SpindleNet: Person re-identification with human body region guided feature decomposition and fusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1077–1085.
- [10] Z. Yi, H. Zhang, P. Tan, and M. Gong, "DualGAN: Unsupervised dual learning for image-to-image translation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2849–2857.
- [11] H. Judy, T. Eric, P. Taesung, J.-Y. Zhu, I. Phillip, S. Kate, E. Alexei, and D. Trevor, "CyCADA: Cycle-consistent adversarial domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1994–2003.
- [12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 2672–2680.
- [13] M.-Y. Liu, B. Thomas, and K. Jan, "Unsupervised image-to-image translation networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 700–708.
- [14] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 95–104.
- [15] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 79–88.
- [16] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 994–1003.
- [17] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.
- [18] X. Zhu, X. Zhu, M. Li, V. Murino, and S. Gong, "Intra-camera supervised person re-identification: A new benchmark," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1079–1087.
- [19] C.-X. Ren, B. Liang, P. Ge, Y. Zhai, and Z. Lei, "Domain adaptive person re-identification via camera style generation and label propagation," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1290–1302, 2020.
- [20] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.
- [21] Z. Zhong, L. Zheng, S.-Z. Li, and Y. Yang, "Generalizing a person retrieval model hetero and homogeneously," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 176–192.
- [22] S. Xuan and S. Zhang, "Intra-inter camera similarity for unsupervised person re-identification," 2021, *arXiv:2103.11658*. [Online]. Available: <http://arxiv.org/abs/2103.11658>
- [23] S. Lin, C.-T. Li, and A. C. Kot, "Multi-domain adversarial feature generalization for person re-identification," *IEEE Trans. Image Process.*, vol. 30, pp. 1596–1607, 2021.
- [24] G. Hiasi, T.-Y. Lin, and Q.-V. Le, "DropBlock: A regularization method for convolutional networks," in *Proc. Neural Inf. Process. Syst.*, 2018, pp. 10727–10737.

- [25] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 598–607.
- [26] H. Huang, W. Yang, X. Chen, X. Zhao, K. Huang, J. Lin, G. Huang, and D. Du, "EANet: Enhancing alignment for cross-domain person re-identification," 2018, *arXiv:1812.11369*. [Online]. Available: <http://arxiv.org/abs/1812.11369>
- [27] X.-Y. Zhang, J.-W. Cao, C.-H. Shen, and M.-Y. You, "Self training with progressive augmentation for unsupervised cross-domain person re-identification," in *Proc. IEEE Int. Conf. Comput.*, Oct. 2019, pp. 8222–8231.
- [28] X. Li, L. Chen, L. Wang, P. Wu, and W. Tong, "SCGAN: Disentangled representation learning by adding similarity constraint on generative adversarial nets," *IEEE Access*, vol. 7, pp. 147928–147938, 2019.
- [29] 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.
- [30] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3960–3969.
- [31] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 475–491.
- [32] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2275–2284.
- [33] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, and S.-Z. Li, "Embedding deep metric for person re-identification: A study against large variations," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 732–748.
- [34] W. Yu, K. Yang, H. Yao, X. Sun, and P. Xu, "Exploiting the complementary strengths of multi-layer CNN features for image retrieval," *Neurocomputing*, vol. 237, pp. 235–241, May 2017.
- [35] 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.
- [36] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "GLAD: Global-local alignment descriptor for pedestrian retrieval," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 420–428.
- [37] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 499–515.
- [38] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2197–2206.
- [39] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.
- [40] H.-X. Yu, A. Wu, and W.-S. Zheng, "Cross-view asymmetric metric learning for unsupervised person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 994–1002.
- [41] P. Peng, T. Xiang, Y. Wang, M. Pontil, and Y. Tian, "Unsupervised cross-dataset transfer learning for person re-identification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 700–708.
- [42] H.-H. Fan, L. Zheng, C.-G. Yan, and Y. Yang, "Unsupervised person re-identification: Clustering and fine-tuning," in *Proc. ACM Int. Conf. Multimedia*, 2018, pp. 1–18.
- [43] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 5157–5166.
- [44] L. Song, C. Wang, L. Zhang, B. Du, Q. Zhang, C. Huang, and X. Wang, "Unsupervised domain adaptive re-identification: Theory and practice," 2018, *arXiv:1807.11334*. [Online]. Available: <http://arxiv.org/abs/1807.11334>
- [45] F. Yang, K. Li, and Z. Zhong, "Asymmetric co-teaching for unsupervised cross-domain person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12597–12604.
- [46] K. Zeng, M. Ning, Y. Wang, and Y. Guo, "Hierarchical clustering with hard-batch triplet loss for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13657–13665.
- [47] A. Anoosheh, E. Agustsson, R. Timofte, and L. Van Gool, "ComboGAN: Unrestrained scalability for image domain translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 896–8967.

- [48] Y. Huang, Q. Wu, J. Xu, and Y. Zhong, "SBSGAN: Suppression of inter-domain background shift for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9527–9536.
- [49] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3908–3916.



**XIANJUN LUO** was born in Anshun, Guizhou, China, in 1998. She received the bachelor's degree from the Nanjing University of Information Science and Technology, Nanjing, China, in 2018. She is currently pursuing the master's degree with Guizhou University, Guiyang, China. From 2019 to 2020, she was a Visiting Researcher with the University of Electronic Science and Technology of China, Chengdu, China. Her main research interest includes person re-identification.

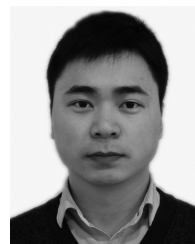
Her focus in the field of person re-identification is to solve the cross-domain issue.



**ZHI OUYANG** was born in Guiyang, Guizhou, China, in 1987. He received the bachelor's degree from Guizhou University, Guiyang, in 2010, and the master's and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 2011 and 2017, respectively. Since 2017, he has been a Teacher with the Guizhou Provincial Key Laboratory of Public Big Data, Guiyang. His main research interests include knowledge management and big data management.



**NISUO DU** was born in Liupanshui, Guizhou, China, in 1986. He received the bachelor's degree from Guizhou University, Guiyang, China, in 2009, and the master's and Ph.D. degrees from the Kansai University of Japan, Osaka, Japan, in 2014 and 2017, respectively. Since 2017, he has been a Teacher with the Guizhou Provincial Key Laboratory of Public Big Data, Guiyang. His research interests include machine learning and analog simulation. His focus is in the field of biological fire simulation.



**JINGKUAN SONG** (Member, IEEE) received the Ph.D. degree from The University of Queensland (UQ), Australia, in 2014, under the supervision of Prof. Heng Tao Shen. From 2016 to 2017, he was a Postdoctoral Research Scientist with Columbia University. From 2014 to 2016, he was as a Research Fellow at the University of Trento. He is currently a Full Professor with the University of Electronic Science and Technology of China (UESTC). His research interests include

large-scale multimedia retrieval, image/video segmentation and image/video understanding using hashing, graph learning, and deep learning techniques. He is a PC member of CVPR'18, MM'18, and IJCAI'18. He is a Guest Editor of IEEE TRANSACTIONS ON MULTIMEDIA (TMM), WWWJ. He was a winner of the Best Paper Award in ICPR, Mexico, in 2016, the Best Student Paper Award in Australian Database Conference, Australia, in 2017, and the Best Paper Honorable Mention Award, Japan, in 2017.



**QIN WEI** was born in Guiyang, Guizhou, China, in 1983. She received the dual bachelor's degree in information management and information systems and economics from Peking University, Beijing, China, in 2005, and the Ph.D. degree in informatics from the University of Illinois at Urbana-Champaign, Illinois, USA, in 2011. From 2014 to 2019, she was the Vice President of the Guizhou Institute of Big Data Industry Development and Application. Her research interest includes multi-source data fusion.