IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 28, 2021, accepted October 10, 2021, date of publication October 13, 2021, date of current version October 25, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3119576

Multiscale Reference-Aided Attentive Feature Aggregation for Person Re-Identification

LI XU AND XIANG FU

[1]Key Laboratory of Jiangxi Province for Image Processing and Pattern Recognition, Nanchang Hangkong University, Nanchang 330063, China
[2]School of Software, Nanchang Hangkong University, Nanchang 330063, China

Corresponding author: Xiang Fu (fxfb163@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61763033, Grant 61866028, and Grant 61762067; and in part by the Jiangxi Provincial Department of Science and Technology under Grant 20203BBGL73222.

ABSTRACT In person re-identification (Re-ID), increasing the diversity of pedestrian features can improve recognition accuracy. In standard convolutional neural networks (CNNs), the receptive fields of neurons in each layer are designed to have the same size. Therefore, in complex pedestrian re-identification tasks, the standard CNNs extract local features but are unable to obtain satisfactory results for global features extracted from the images. Local feature learning methods are helpful for obtaining more abundant features, which focus on the most significant local features and ignore the correlations between features of various parts of the human body. To solve the above problems, a new multiscale reference-aided attentive feature aggregation (MS-RAFA) mechanism is proposed, consisting of three main modules. First, to extract the most significant local features and strengthen the correlations between the features of various parts of the human body, an autoselect module (ASM) is designed, an attentional mechanism that can stack the structural information and spatial relations to form new features. Then, to realize multiscale feature fusion of the multiple output branches of the backbone network and increase feature diversity, we propose a multilayer feature fusion module (MFFM), which enables the model to mine the features hidden by salient features and to learn features better. Finally, to supervise the MFFM and make the network obtain better recognition features, we propose a multiple supervision mechanism. Finally, experimental results demonstrate that our proposed method outperforms the state-of-the-art methods on three large-scale datasets.

INDEX TERMS Feature correlation, multiscale reference-aided, multilayer feature fusion, person re-identification.

I. INTRODUCTION

Person re-identification (Re-ID), which forms the core of video surveillance technology, implements image processing, computer vision, pattern recognition, machine learning and other related technologies to solve cross-camera and cross-scene pedestrian retrieval problems. Re-ID utilizes the spatiotemporal continuity of images to continuously track pedestrians across cameras. Visual feature-based recognition methods are more reliable than those based on biological information, such as carrying items or clothing, and can be used more reliably in Re-ID [1]–[4]. With the popularity of video capture systems, video-based Re-ID also achieves more robust performance. Many scholars have developed improved pedestrian re-recognition methods and achieved very good results. However, in cases involving different visual points, low image resolution, illumination changes, unconstrained attitude change and occlusion, the recognition effect is not ideal [5]–[9].

In recent years, deep learning, represented by convolutional neural networks (CNNs), has been successfully applied to the field of pedestrian re-recognition. CNNs are constructed by researchers with certain prior knowledge, and a large number of stacked convolutional cores are used to extract regional features. At present, there are many efficient feature extraction CNNs, such as GoogLeNet [10]–[13], ResNet [14], and VGGNet [15]. Feature extraction methods that implement feature learning algorithms can obtain better pedestrian representations. In the field of Re-ID, most algorithms use the feature of the last layer of the network to realize pedestrian recognition, which achieves good results but also has some defects [26]. Each CNN layer focuses on a different piece of information, low-level features such as texture and shape features focus on shallow information of

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

the object of interest, while deep layer features focus more on the semantic information. Therefore, only using the last layer features weakens the recognition effect. To better verify this observation, Figure 1 depicts a visualization of the feature map of different layers of ResNet50. Here, layer 1, layer 2, layer 3 and layer 4 represent the final output feature of the corresponding layers of ResNet50. Each layer focuses on different significant features, but the detailed information of some local features, such as clothing color and shoes, is not sufficiently extracted. These local details can improve recognition accuracy, but deep neural networks have difficulty in selectively focusing on these details.

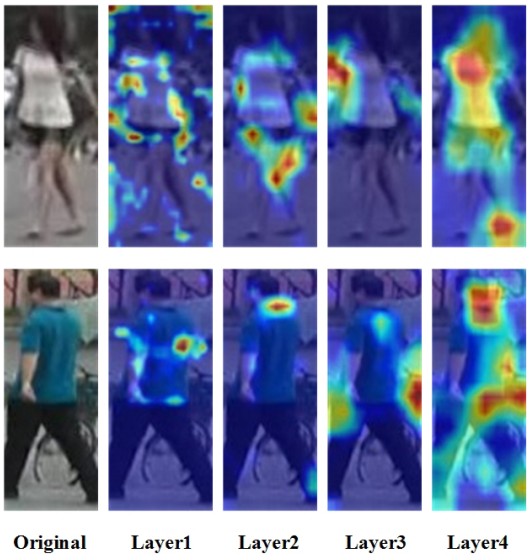

FIGURE 1. CNN network visualization.

Based on the above considerations, some works have obtained valuable detailed information by focusing the network on local areas to extract global features. These methods can be summarized as follows. (1) With the attention mechanism method, partial alignment is achieved by enhancing the distinguishing areas and suppressing the background to reduce background interference. Many works learn attention using convolutional operations with small receptive fields on feature maps [2], [22]–[24]. However, to intuitively determine whether a feature node is important, one should know the features of global scope, which facilitates the comparisons needed for decision-making. In addition, if the various features are linked arbitrarily, some significant discriminant features that do not show obvious strength will be masked by other significant features. (2) In fringe segmentation-based methods, the human image is segmented into fixed horizontal stripes, and the finer-grained local salient features of each stripe are studied [16]–[19]. Although the above methods are effective, they have high requirements for image alignment. Moreover, the positioning ability of the model is poor, and the redundancy between features obtained from different regions is relatively high. (3) Methods based on automatic positioning attempt to locate body parts by learning a grid [4], [20], [21]. The above methods are detection networks that need extra training, and when there is considerable background noise in the location, the complexity of the whole model is increased.

To address the above deficiencies, in this paper we present a new multiscale reference-aided attentive feature aggregation (MS-RAFA) mechanism that enables the network to adaptively extract all potential salient pedestrian features. More specifically, we propose an autoselect module (ASM) to mine local and global information in different stages of the backbone network, which solves the problem of insufficient salient feature extraction. Then, we present a multilayer feature fusion module (MFFM), which is used to better aggregate the low- and high-level features of the backbone. The MFFM uses an adaptive selection mechanism to select effective features in different stages from multiscale features, which solves the problem of feature redundancy, and models the object from a global scope. Additionally, multiple monitoring mechanisms are used to supervise and teach the MFFM so that the network can obtain better identifying features. It is worth noting that the end-to-end training method is used, and no additional training network is required, reducing the complexity of the model. The specific content will be introduced in Section 3.

To summarize, our proposed work makes the following contributions:

- We introduce a novel multiscale reference-aided attentive feature aggregation mechanism (MS-RAFA) that can mine all potential salient features stage-by-stage and integrate these discriminative salience features with the global feature, forming the final diverse pedestrian feature representation.
- We devise an autoselect module (ASM), an attention mechanism placed on a backbone network that can optimize the backbone network features. This module extracts global and local appearance information compactly, stacks structural information and spatial relationships to form new features, and uses that information as input to the next stage.
- We incorporate a multilayer feature fusion module (MFFM) to extract and fuse low- and high-level features. The MFFM is a nonlinear dynamic selection mechanism that allows each neuron to adjust the size of its receptive field adaptively according to multiple input information scales.
- We propose a multiple supervision mechanism to verify the necessity of joining the MFFM and the final effect value of the network.
- Extensive experiments on CUHK03 [46], Market1501 [47] and DukeMTMC-ReID [48] demonstrate that our method significantly outperforms existing state-of-the-art methods.

II. RELATED WORK

Pedestrian re-recognition technology can be divided into five steps [25]: data collection, bounding box generation, training data annotation, model training and pedestrian retrieval.

Due to the continuing improvements in computing power, many deep learning-based methods have been developed in recent years to solve pedestrian re-recognition tasks. This section will introduce the most representative works related to ours.

A. LOCAL FEATURE LEARNING

Varior *et al.* [27] used a twin network to divide a pair of input images horizontally into several blocks. Then, several segmented image blocks were sent to a long short-term memory (LSTM) network in sequence [28], and the local features of all image blocks were fused to obtain the final feature representation. An image structure analysis method [29], [30] was adopted to obtain the corresponding parts of features, such as head, chest, legs and shoes, and the color features of each part were extracted for matching. Zhang *et al.* [18] designed a dynamic alignment network to automatically align image blocks from top to bottom. In terms of the spatial alignment of the human body, Zheng *et al.* [31] used spatial transformer networks (STNs) [32] to directly segment and align the original image. The STN was then also used to transform the shallow features extracted by CNNs to spatially align human body features. Reference [33] proposed a method of horizontal pyramid matching (HPM), which divides a pedestrian picture into 1, 2, 4 and 8 subparts horizontally. Reference [70] proposed a Multi-Granularity Network based on Local Context aware Correlation Feature (MGN_CACF) based on the ResNet50-IBN-a backbone, which is split into four branches.

Because local feature learning is only used to obtain and combine information of different parts of the human body, the trained network has insufficient generalizability. However, the information of each part of the human body has strong semantic correlations, which are helpful for teaching the network to learn better representation. If only part of the feature graph is learned, the correlation information between each part of the feature will be lost.

B. FINE-GRAINED INFORMATION LEARNING

One challenge in pedestrian recognition is distinguishing those with similar appearances. Reference [34] proposed a densely semantically aligned (DSA) model to map human body features to three-dimensional space. However, this method often requires both front and back pictures of the same person, which limits its applicability. Zhang *et al.* [37] proposed a bilinear CNN model using two networks, VGG-D and VGG-M, as the joint benchmark network and achieved a good effect without using bounding box marking information. Reference [23] proposed an activation mapping method that judged the activation area by the loss function of the overlapping activation penalty to continuously expand the spatial perception range of the CNN. Reference [35] proposed an interaction-and-aggregation block (IA-Block), which can not only obtain pixel-level fine-grained information but also introduce channel information to obtain a more comprehensive feature representation. Reference [36] integrated attributes into features and proposed an attribute-driven feature separation and time-aggregated pedestrian re-recognition method. Referenced [71] proposed leveraging the stability of person attributes to guide the learning of discriminative domain-invariant features (DIFs) and align attributes with corresponding local visual features.

Because the classification networks of these methods have strong feature representation ability, they can achieve better results in conventional image classification. However, in the study of Re-ID, the difference in some pedestrians' appearances is actually very subtle, so the effect is not ideal. The common solution is to use the network weights pretrained on ImageNet as the initial weights and then fine-tune them on a fine-grained classification dataset to obtain the final classification network.

C. ATTENTION MECHANISM LEARNING

The essence of the attention mechanism [38] is to imitate the human visual signal processing mechanism to selectively observe part of the area while ignoring other visible information. Li *et al.* [39] proposed a spatiotemporal attention model that uses multiple spatial attention models to ensure that each learns different parts of the body. Reference [24] used an attention diagram to determine whether an unnoticed area contains features that could provide a judgment basis to obtain complete human body features. Reference [40] proposed a pose-guided feature alignment (PGFA) method to obtain the features of the area where the body parts are connected. However, this method only focuses on unshaded parts and fails to identify shaded parts adequately. Reference [9] proposed a spatiotemporal completion network (STCNet) to solve the problem of pedestrian lower body occlusion. The spatial generator generates the frames that need to be completed, and then the temporal attention generator finds the adjacent key frames for completion. Reference [72] proposed a weighted aggregation strategy to impart a strong multiview reason ability to the imaginative reasoning module (IRM) and to classify and aggregate the single-view features of the same pedestrian.

Although the abovementioned attention mechanisms have achieved certain effects in pedestrian re-recognition, their main purpose is to extract the most significant features and suppress less obvious features. In this way, feature diversity is reduced, so the extracted features may be insufficient.

III. PROPOSED METHOD

We propose a new multiscale reference-aided attentive feature aggregation mechanism (MS-RAFA), which includes three main modules: the autoselect module (ASM), multilayer feature fusion module (MFFM) and the multiple supervision module. The framework is shown in Figure 2. ASM is an attentional mechanism and sits on a backbone network, which is used to solve the problem of insufficient feature extraction of the backbone network. MFFM operates the output features of each layer of the backbone network and selects useful information for fusion to obtain more features.

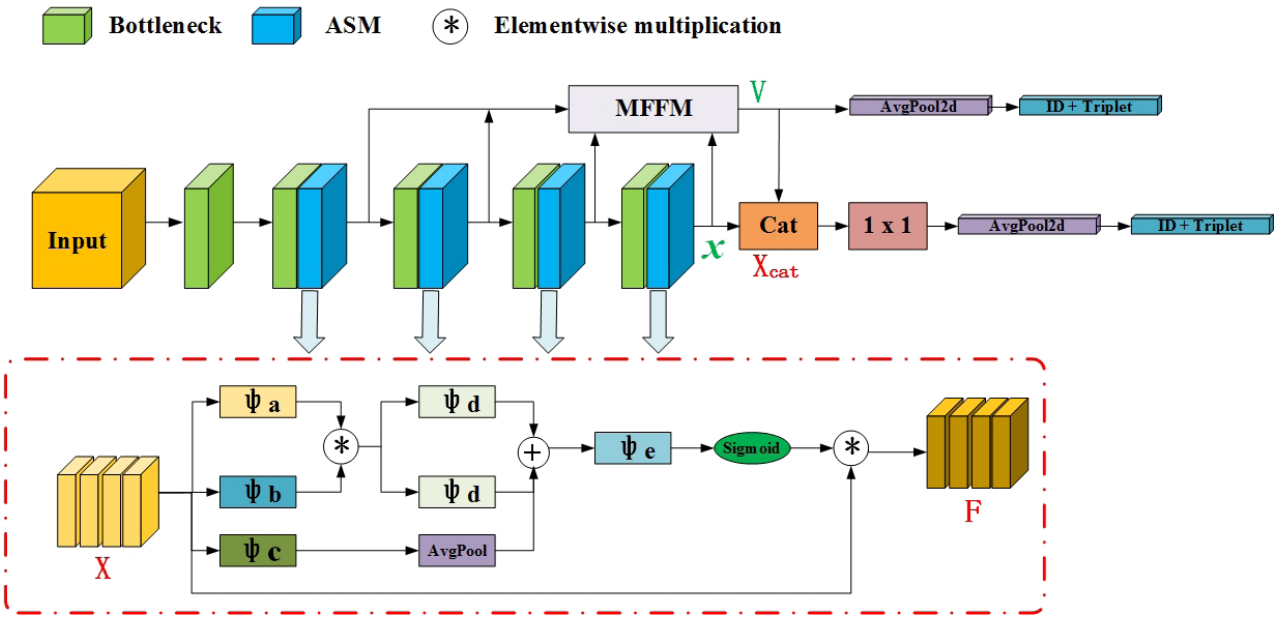

FIGURE 2. Multiscale reference-aided attentive feature aggregation mechanism. After modifying the ResNet50 backbone, we adopt several separate feature autoselection phases. In the training process, ID loss and triplet loss are used to supervise the learning in the two stages, and the final loss is the sum of the two supervision losses. In the test, the features of all the different stages are linked together as the final pedestrian descriptor image.

The multiple supervision module supervises the optimization direction of the MFFM module and guides the MFFM to select effective pedestrian features for fusion.

A. AUTOSELECT MODULE

The goal of Re-ID is to recognize the same pedestrians from multiple cameras, but often the differences among these people are not sufficiently large. Most pedestrians are distinguished by clothing color, height, belongings and other information. The convolutional unit in a CNNs only focuses on the region of the convolutional kernel in the neighborhood each time. Although the receptive field becomes increasingly larger in later periods, it still learns from information in the local region, thus neglecting the contribution of other global regions to the current region. If the use of only local information cannot obtain the differences between objects well, the correlation between the features of various parts of the human body will be ignored, limiting the ability to improve the performance of the model. However, global learning extracts features from the global information of each pedestrian picture. These features have no spatial information and can easily lose details, which is not conducive to pedestrian recognition.

Considering the above, this paper proposes an ASM that uses the channel attention mechanism to obtain local information. Then, global average pooling is used to obtain global information. The local and global information are superimposed to obtain more accurate pedestrian information. Through these operations, effective object features can be better selected. The specific ASM structure is shown in the red dotted box in Figure 2. the ASM is placed into the backbone branch to compensate for its insufficient feature extraction ability.

Given an intermediate feature tensor of width W, height H, and channel number C in CNN layer $X \in R^{C\times H\times W}$, after the ASM has conducted a series of operations, a new feature graph F of size $C \times H \times W$ is obtained. ψ_a, ψ_b, ψ_c, ψ_d and ψ_e are all theoretically 1×1 convolutions that can flexibly change the dimensions of the data. After the size of ψ_a is changed to $C \times W \times H$, the size of ψ_b and ψ_c are changed to $C \times H \times W$, ψ_d changes size to $1 \times C \times 1$, and ψ_e changes size to $3C \times 1 \times 1$. The 1×1 convolution has different effects in different locations. Its main purpose is to multiply or add the features of the same input, increase the similarity and extract features. At the end of the module output, the sigmoid activation function is added to increase the nonlinear expression ability of the module and make the model more consistent with the data.

In Re-ID, it is possible to compute the global attention by using local convolutional operations. The main operation of the ASM is to take the C-dimensional eigenvector of each spatial position as a feature node, each of which uses the similarity relation function $f = R(x, y)$ to obtain the similarity between features of other location nodes and form raster data. Through raster scanning of spatial positions, N feature nodes are expressed as feature set $V = \{X_i \in R^C, i = 1, \cdots, N\}$. The similarity relation $R_{i,j}$ between any two nodes i and j in set V can be defined as the dot product similarity between the nodes:

$$Ri, j = f(xi, xj) = \alpha(xi)^T \cdot \omega \cdot \beta(xj) \quad (1)$$

where α and β are two embedded functions shared between feature nodes, and ω is an operation function for changing the size of the image. We first transform the input tensor $X^{C\times H\times W}$ into $X_1 \in R^{C\times W\times H}$, then implement

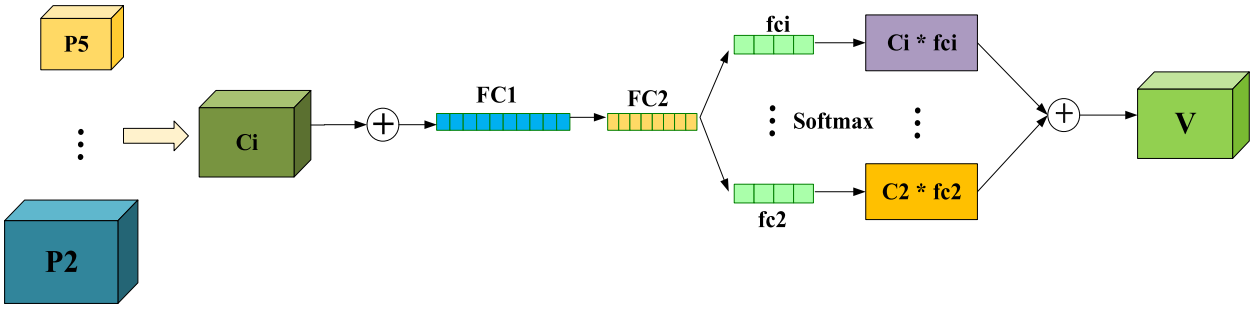

FIGURE 3. Multilayer feature fusion module.

1×1 convolution and BN layer, and finally activate the ReLU function to transform X_1. We obtain and use the affinity matrix $R_{i,j} \in R^{C\times C\times 1}$ to represent the pairable relations of all nodes and then transform these pairable relations to obtain the global features. The local information of the C-dimensional feature vector of each spatial position can be obtained by the ASM module:

$$Loi = \alpha \cdot xi \cdot \varphi i \tag{2}$$

where α is the 1×1 convolution function and φ_i is an adaptive average pooling function. After the operation, the input tensor $X^{C\times H\times W}$ is changed to $X_2^{C\times 1\times 1}$.

Global scopes contain rich structural and semantic information, while local features contain the most significant information. Using the ASM, they can be stacked, and valuable knowledge can be mined to infer spatial attention. Then, the spatial concern value of the i-th feature node Gi, obtained through the modeling function, is defined as:

$$Gi = Sigmoid(\alpha \cdot B \cdot R) \tag{3}$$

where B is the BN function and R represents the ReLU operation. The sigmoid function causes the output to take on a value between 0 and 1. As a result, the weight of each feature in the spatial structure is different, and the probability of the feature that needs to be given more attention is larger, which is conducive to feature extraction and model learning.

To learn the attention of the i-th and j-th feature nodes, in addition to the pair relation term $R_{i,j}$, the feature X_i is added. Using the global scope structure information and local original information related to this feature, the final output $\bar{F}$ can be expressed as:

$$\bar{F}i, j = [Gi, Xi] \tag{4}$$

Experimental results show that the ASM proposed in this paper can effectively extract the global and local hidden features. Moreover, it can be used in different neural network frameworks. Most importantly, it can improve Re-ID accuracy.

B. MULTILAYER FEATURE FUSION MODULE

Fusion and analysis of features at different levels [41] can help in semantic segmentation, classification and detection. Common fusion operations are performed at the pixel level, such as addition or concatenation, but the performance gains are limited and lack semantic information. To aggregate features at different scales from different branches and retain the features in the final representation, inspired by long-term dependence on mechanisms to fuse multilayer features [42], we add an MFFM. As shown in Figure 2, the structural design of the MFFM is derived from SKNet [43], and the specific structure diagram is shown in Figure 3.

The input of the MFFM is the output feature of the ASM in each stage of the backbone network. In a CNNs, the bottom layer contains shallow information such as position and shape information, while the top layer contains deep semantic information. To obtain effective pedestrian information at different stages, we proposed the MFFM method to fuse these different depths of information. The MFFM is a nonlinear module that can dynamically select features. It allows the input of multiple neurons of different sizes, each of which adjusts the size of its acceptance domain according to the information scale, and then outputs features of uniform size.

As shown in Figure 2, the backbone network is divided into five layers, each with different sizes and information. Except for the first layer, the other four layers are followed by an ASM. Features that pass through the ASM are the source of the multilayer features that will enter the MFFM. In Figure 3, P_2-P_5 represents the multilayer feature input, in which the lower-level feature map is larger and contains more feature information, while the higher-level feature map is smaller in size but with a larger object. The study found that as the object grows larger, most neurons gather more information from the larger kernel pathway. This suggests that the nonlinear dynamic selection mechanism in the MFFM has superior performance in object recognition, which enables neurons to adaptively adjust their receptive field size. P_2-P_5 are convolved to change the number of channels and the size of the feature map so that the output feature, denoted as $C_i(i = 3, 4, 5)$, is uniform in size. For the input, a given intermediate feature tensor $P_2 \in R^{C\times H\times W}$ with width W, height H, and number of channels C is:

$$\begin{aligned} C2 &= Relu(W1 \cdot Maxpool2d \cdot W2 \cdot P2), \\ C3 &= Relu(W1 \cdot W2 \cdot P3), \\ C4 &= Relu(W2 \cdot P4), \\ C5 &= Relu(W2 \cdot P5) \end{aligned} \tag{5}$$

where W_1 and W_2 are implemented by convolution followed by BN. The final size of C_i is the same as that of P_4, therefore, the feature map tensor $C_i \in R^{C\times H\times W}$ is obtained by taking the size of P_4 as the standard.

After unifying transformation to obtain the same size, to highlight the significant features, we add C_i elementwise to form a new aggregation feature L, which combines low- and high-level features. Then, *FC1* uses global average pooling to operate on L, and the generated channel statistics are embedded in the global information. Here the input $L \in R^{C\times H\times W}$, and the output $FC1 \in R^{C\times 1\times 1}$. Specifically, the elements in *FC1* are calculated by spatial dimension contraction of L:

$$FC1 = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} L(i,j) \tag{6}$$

Furthermore, a compact feature $FC_2 \in R^{d\times 1}$ is created, which is attenuated by *FC1* to achieve precise and adaptive selection guidance. This is achieved through a simple convolutional layer, which improves efficiency by reducing dimensions:

$$FC2 = \alpha \cdot B(R \cdot FC1^{d\times c}) \tag{7}$$

where α denotes 1×1 convolution, B denotes BN, and R is the ReLU operation. To study the effect of dimension d on model efficiency, we use a reduction ratio r to control its value:

$$d = \max(\mathrm{C}/r, m) \tag{8}$$

where m represents the minimum value of d ($m = 256$ is fixed in our experiment), r is set to 4, and then the feature vector is expanded to the size of the feature before attenuation.

The *FC1* and *FC2* operators combine and aggregate the information from multiple paths to obtain a global and comprehensive representation for the selection weights. Next, the weight *FCi* of each channel is obtained through a softmax operation (in which the importance is expressed by the probability of each channel). Then, *FCi* is multiplied with the original feature *Ci* to obtain the feature weight of each channel. This is the equivalent of a select operation, it aggregates the feature maps of differently sized kernels according to the selection weights. Finally, the weight feature map obtained is added element by element, and the final weight feature V is obtained through the multilayer feature fusion module as:

$$V = Cat(Ci * fci), \quad i = 2, 3, 4, 5 \tag{9}$$

The size of the feature map of V is the same as that of the output feature graph of the last layer of the backbone network. Through this soft attention method, adaptive kernel selection is performed to obtain the potentially significant features among the global features, increasing feature diversity and improve object recognition efficiency and effectiveness.

C. MULTIPLE SUPERVISION MECHANISM

As shown in Figure 2, to increase the information diversity and improve information interactivity, the features output from the MFFM and the last ASM are added elementwise to form a new feature, $X_{cat} \in R^{2C\times H\times W}$, with twice the number of channels. Then, the feature map tensor $V \in R^{C\times H\times W}$ output from the MFFM module and the feature tensor $x \in R^{C\times H\times W}$ output from the backbone network are combined by the add operation. The subsequent 1×1 convolution changes the number of channels from 2C to C to reduce the number of channels in the feature map. Previously, the fully connected layer FC had been used for information classification, but its computational complexity is high. Therefore, we replace it with global average pooling, which calculates a weighted sum of the front layer features, takes the internal average of each feature map, and turns each feature map into a value.

The MFFM aggregates multiscale pedestrian features to obtain more robust information. The feature information involved in this module is quite different, aggregating these features will not achieve good recognition performance. Therefore, a multiple monitoring mechanism that uses real pedestrian information to monitor the MFFM and adaptively select effective information is designed. As shown in Table 8, an ablation experiment demonstrates that supervision is necessary. After obtaining the features supervised by the MFFM, the multiple supervision mechanism is merged with the features of the backbone network for the last layer, and then another supervision is implemented.

We add two loss functions to the final effect value of the monitoring network. The identification loss function obtains the predicted logit value of the image, similar to the classification loss, and is defined as:

$$Lid = \sum_{\mathrm{i=1}}^{\mathrm{N}} -qi\ log(pi) \begin{cases} qi = \varepsilon/N, & y \neq i \\ qi = 1 - \varepsilon\dfrac{N-1}{N}, & y = i \end{cases} \tag{10}$$

where y and P_i represent the real ID tag and predicted logit value of the classification, respectively. N represents the number of classes, qi is the proposed smoothing label and $\varepsilon = 0.1$ [11]. Considering the goal of Re-ID, which is to find the most similar sequences of people from a gallery of images, the idea of metric learning is introduced to enable networks to find useful features for similarity measurement. Therefore, triplet loss is adopted to improve the final ranking performance, which is defined as:

$$Ltp = \sum N\ [dpos - dneg] + \tag{11}$$

where d_{pos} is the feature distance of the same identity and d_{neg} is the distance of different identities. N is the batch size of the triplet samples, and $[\cdot]_+$ represents *Max (·, 0)*. The purpose of triplet loss is to ensure that the distance between the positive sample pairs is less than the distance between the negative sample pairs. here, it ensures that the distance between similar features and the positive sample is close. Note that the distance is measured by the Euclidean distance in the design of this paper.

In this paper, the model undergoes supervised learning twice, first for the features obtained by the MFFM and second for the fusion features of the last layer of the backbone

network. The final total loss of the model is the sum of the two losses and can be written as:

$$Loss = Lidi + Ltpi, \quad (i = 1, 2) \tag{12}$$

where i indicates the i-th supervised learning. Finally, the monitoring mechanism is used to supervise the results and verify the effectiveness of the proposed method.

IV. EXPERIMENTS

A. EXPERIMENTAL DETAILS

We use ResNet50 as our backbone network. The total batch size is set to 64, and two GPU graphics cards are shared. The batch_size on each graphics card is set to the value that automatically divides the total batch_size.

We use common data augmentation strategies: random clipping [5], horizontal flipping and random erasure [44]. The input of all datasets is changed to images with a uniform size of 256 × 128, and the backbone network is pretrained on ImageNet [45]. Using the Adam optimizer, all models are trained with a total of 600 epochs, the recording of parameters starts from epoch = 320, and a new file is recorded every 40 epochs. The learning rate is 8×10^{-4}, and the weight decay is 5×10^{-4}.

The experiment is performed on three public re-identification datasets: CUHK03 [46], Market1501 [47] and DukeMTMC-ReID (a subset of the DukeMTMC [48] dataset). The details of the datasets are shown in Table 1. To compare the performance of this method with that of existing Re-ID methods, we use the rank index in cumulative matching characteristics (CMC) and mean average precision (mAP) as the evaluation index of each queried image.

TABLE 1. Dataset statistics.

Dataset	CUHK03	Market1501	DukeMTMC-ReID
Training-IDs	767	751	702
Querry-IDs	700	750	702
Gallery-IDs	700	751	1110
Camera	2	6	8
Images	28192	32668	36411

B. EXPERIMENTAL RESULTS

In this section, the final experimental results for the proposed method on different datasets are highlighted and compared with the results for other methods.

Table 2 shows the results of different methods on the Market1501 dataset. Note that all methods listed in the table are based on the ResNet50 backbone network. i) The methods are from different journals and published in different years. The oldest papers were published in 2018, and the latest were published in 2021. ii) Our experimental results are superior to those of other methods, in terms of both mAP and Rank-1 accuracy. Our mAP value is 89.1%, and the Rank-1 accuracy is 95.8%. The mAP value was 5.4% higher than the baseline, while the Rank-1 value was 1.6% higher.

TABLE 2. Comparison with the most advanced methods on the Market1501 dataset (%).

Method	mAP	Rank-1
FC[1](19)	86.2	95.2
NL[55](18)	87.4	95.6
CBAM[56](18)	85.6	94.8
SE[57](18)	86.0	95.2
SNL[58](19)	87.3	95.7
MHN-6[22](19)	85.0	95.1
BAT-net[59](19)	84.7	95.1
MGN(w flip)[17](19)	86.9	95.7
JDGL[60](19)	86.0	94.8
DSA-reID[34](19)	87.6	95.7
OSNet[61](19)	84.9	94.8
RGA[49](20)	87.5	96.0
SCSN[50](20)	88.5	95.7
SAN[51](20)	88.0	96.1
ISP[52](20)	88.6	95.3
INTACT[53](20)	-	88.1
M^3+ResNet50[54](20)	82.6	95.4
HOReID[63](20)	84.9	94.2
MGN_CACF[72](21)	88.5	95.7
DDB[73](21)	88.2	95.7
LM1+ LM2[74](21)	87.3	95.1
SFNet[75](21)	87.7	95.3
OSNet[76](21)	86.7	94.8
Baseline	83.7	94.2
MS-RAFA(ours)	**89.1**	**95.8**

CUHK03 is a more challenging dataset than Market1501 and DukeMTMC-ReID. This is because i) CUHK03 has fewer samples and contains serious viewpoint variations and occlusion problems; and ii) the annotation of bounding boxes marked by the object detection algorithm has location offsets. All methods listed in Table 3 are also based on the ResNet50 backbone network. The experimental results of our method, both mAP and Rank-1, are significantly better than those of the other methods. The mAP and Rank-1 values for the labelled category are 79.6% and 83.9%, which are 10.6% and 10.1% higher than the baseline, respectively. The mAP and Rank-1 values of the detected category are 77.2% and 82.2%, respectively, which are 11.7% higher than both the mAP and Rank-1 of the baseline.

Table 4 shows the results of the different methods on the DukeMTMC-ReID dataset. To further verify the effectiveness of our method, we select methods with different backbone networks, for example, ResNet101, DenseNet121 and SEResNet101. Similar to the other two datasets, the proposed MS-RAFA also achieves the best results in terms

TABLE 3. Comparison with the most advanced methods on the CUHK03 dataset (%).

Method	CUHK03			
	labeled		detected	
	mAP	Rank-1	mAP	Rank-1
MHN-6[22](19)	72.4	77.2	65.4	71.7
BAT-net[59](19)	76.1	78.6	73.2	76.2
MGN(w flip)[17](19)	67.4	68.0	66.0	66.8
DSA-reID[34](19)	75.2	78.9	73.1	78.2
OSNet[61](19)	-	-	67.8	72.3
BFE[62](19)	76.7	79.4	73.5	76.4
CBAM[56](18)	73.0	78.0	-	-
FC[1](19)	73.2	78.4	-	-
NL[55](18)	72.6	76.6	-	-
SE[57](18)	71.9	76.3	-	-
SNL[58](19)	72.4	77.4	-	-
RGA[49](20)	75.6	79.3	-	-
SAN[51](20)	76.4	80.1	-	-
ISP[52](20)	74.1	76.5	71.4	75.2
MGN_CACF[72](21)	73.8	78.0	-	-
DDB[73](21)	78.0	80.4	74.8	77.2
LM1+ LM2[74](21)	70.7	72.0	67.7	69.4
SFNet[75](21)	72.6	75.0	70.0	72.1
OSNet[76](21)	69.3	73.3	-	-
Baseline	69.0	73.8	65.5	70.5
MS-RAFA (ours)	**79.6**	**83.9**	**77.2**	**82.2**

of Rank-1 accuracy and mAP. Our results exceed those of MGN_CACF [72] (21) by 0.4%.

Through these comparisons, it can be seen that the method proposed in this paper is the most effective. The experimental results clearly demonstrate that mining potential features and integrating these complementary features leads to great performance advantages.

C. ABLATION EXPERIMENT

To demonstrate the experimental results of our MS-RAFA, we conduct an incremental evaluation of its modules on the Market1501 dataset. The dataset consists of complex scenes and contains much information, so the experimental results when testing our model are more convincing. We again use ResNet50 as the backbone network. In Figure 2, the MFFM module fuses the output of the four layers after the output of that contain the ASM module. After that, the feature V outputted from the MFFM module is fused with the feature outputted through the backbone network (which we simply denote as x). Finally, it is supervised by ID and triplet loss twice. It is worth noting that the final total loss is the sum of the two supervised losses.

To better verify the experimental effect of each module on the overall network, we conduct experiments on each module, and the results are shown in Table 5. For the three modules, the result of removing any one module is lower than the result

TABLE 4. Comparison with the most advanced methods on the DukeMTMC-ReID dataset (%).

Method	Backbone	mAP	Rank-1
MHN(PCB)[22](19)	ResNet50	77.2	89.1
BFE[62](19)	ResNet50	75.9	88.9
CASN(PCB)[24](19)	ResNet50	73.7	87.7
DCDS[64](19)	ResNet101	75.5	87.5
AANet[65] (19)	ResNet152	74.2	87.6
PSE+ECN[7](18)	ResNet50	75.7	84.5
IANet[35](19)	ResNet50	73.4	83.1
VPM[66](19)	ResNet50	72.6	83.6
SPReID[67](18)	ResNet152	73.3	85.9
Tricks[68](19)	SEResNet101	78.0	87.5
SCSN[50](20)	ResNet50	79.0	90.1
SAN [51](20)	ResNet50	75.5	87.9
INTACT[53](20)	ResNet50	-	81.2
M^3+ResNet50[54](20)	HA-CNN	72.2	87.1
HOReID[63](20)	ResNet50	75.6	86.9
MGN_CACF[72](21)	ResNet50	79.5	90.3
DDB[73](21)	ResNet50	78.9	89.8
LM1+ LM2[74](21)	ResNet50	76.8	87.1
SFNet[75](21)	ResNet50	78.1	88.4
OSNet[76](21)	ResNet50	76.6	88.7
Baseline	ResNet50	71.8	85.9
MS-RAFA (ours)	**ResNet50**	**79.9**	**89.5**

TABLE 5. Impact of each module on network performance (%) and MS: multiple supervision module.

ASM	MFFM	MS	mAP	Rank -1	Rank -5	Rank -10
-	-	-	83.7	94.2	-	-
√	-	√	88.2	95.5	98.6	99.1
-	√	√	87.7	95.2	98.5	99.2
√	√	-	88.8	95.8	98.7	99.2
√	√	√	89.1	95.8	98.7	99.1

of combining them all together, which shows that the three modules are complementary. After implementing the ASM, the mAP value improved by 1.4%, which shows that the module leads to the extraction of more effective target features through the identification of local and global features. After implementing the MFFM, the mAP value improved by 0.9%. This is because the MFFM can make reasonable use of both high- and low-level features. The two proposed modules complement each other and extract more hidden features. After using the multiple supervision module, the mAP value improved by 0.3%. This indicates that multiple supervision helps detect and teach the MFFM to obtain better recognition features. Therefore, this reasonably demonstrated the validity of our three proposed modules.

The feature map outputted by the first layer of the ResNet50 network model is very large (7×7 convolution), contains very complex information, and has a large number of parameters and a large number of computations. Therefore, we do not add the ASM after the first layer, as shown in the red dotted box in Figure 2. Instead, ASM modules are added after the last four layers (i.e., layer2, layer3, layer4 and layer5).

TABLE 6. Experimental results from imposing internal changes to the ASM (%) and the effectiveness of the global relation representation (Rel.) and the feature itself (Ori.). w/o: without.

Method	mAP	Rank-1	Rank-5	Rank-10
w/o Rel.	87.5	95.5	98.6	99.1
w/o Ori.	88.1	95.4	98.6	99.2
Both	89.1	95.8	98.7	99.1

TABLE 7. Internal ablation experiment on the MFFM module(%) and w/o: without.

Method	mAP	Rank-1	Rank-5	Rank-10
w/o layer5	87.8	95.2	98.2	98.9
w/o layer54	87.2	95.2	98.6	99.1
w/o layer543	87.0	95.3	98.6	99.1
w/o layer5432	88.2	95.5	98.6	99.1
All	89.1	95.8	98.7	99.1

Table 6 shows the results on ablation experiments on the ASM structure to verify its effect on model performance. For learning attention, without taking the proposed global relation (Rel.) as part of the input, the scheme is inferior to our scheme that includes the ASM by 1.6% in mAP. In addition, without taking the feature itself (Ori.) as part of the input, the scheme is inferior to our ASM scheme by 1.0% in mAP. The combination of the two structures achieves better performance. This suggests that the main improvement in the experimental results comes from the new design for learning attentional relations, in which modeling of global relations provides better performance.

To further verify the fusion effectiveness of the MFFM module, we implement the following experiments with internal detail changes, and its results are shown in Table 7. The MFFM module fuses the output of the four layers that contain the ASM module. Here, we remove some of the fusion branches to verify the fusion effectiveness of the MFFM on the low- and high-level features. The terms w/o layer5, w/o layer54, w/o layer543, and w/o layer5432 indicate removal of the corresponding fused layers in turn. For example, w/o layer54 indicates that the output of the branches of the fifth and fourth layers do not enter the MFFM module for fusion. w/o layer5432 is equivalent to using only the ASM in the model.

In Table 7, as the number of fusion layers decreases, the performance of the model gradually decreases. The decreases by 2.1%, and the Rank-1 value decreases by 0.6%, which indicates that the information inside the MFFM module is complementary. Specifically, the MFFM module acquires features of different levels and integrates them into a new feature. It is effective because the formation of new features has a certain positive effect on the extraction of global features. It also forms information complementarity between different layers to obtain more implicit features.

To verify the fusion of the V and x features, as well as the effectiveness of the supervision of ID and triplet loss twice, we set up a comparison experiment, as shown in Table 8, where w/o VLoss and w/o xloss represent removal of the

TABLE 8. Verification of the effectiveness of the supervision mechanism(%) and w/o: without.

Method	mAP	Rank-1	Rank-5	Rank-10
w/o VLoss	87.3	95.2	98.5	99.2
w/o xloss	87.7	95.4	98.6	99.3
w/o xvCat	88.8	95.8	98.7	99.2
Both	89.1	95.8	98.7	99.1

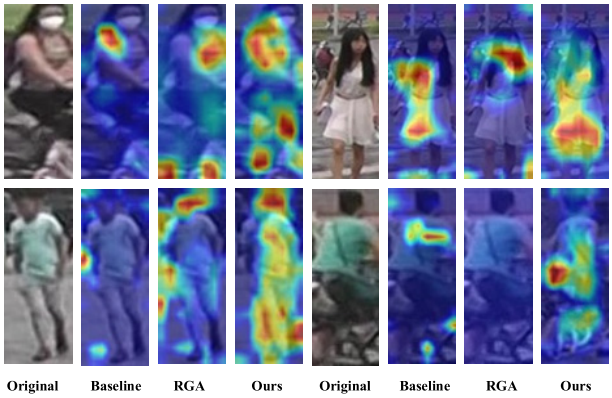

FIGURE 4. Grad-CAM visualization according to gradient responses: baseline vs. RGA vs. ours.

corresponding loss, xvCat represents the fusion operation on the output features of the MFFM and the features of the last layer of the backbone network, and both refers to the complete multiple supervision mechanism. After using VLoss, the mAP value improves by 1.8%, and after using xLoss, the mAP value improves by 1.4%. Therefore, adoption of both branches is most effective for loss supervision and information fusion. The mAP and Rank-1 values for this condition are the best in the comparative experiment.

D. VISUALIZATION OF ATTENTION

Similar to RGA [49], we apply the Grad-CAM (gradient-weighted class activation mapping) [69] tool to the baseline model and to our model for qualitative analysis. The Grad-CAM tool identifies areas that the network deems important, as shown in a comparison between the models in Figure 4 The Grad-CAM masks of our proposed model cover the human area better than the baseline model, allowing the network to focus on larger regions of different parts of the body. Compared with RGA [49], which uses spatial and channel attention mechanisms, our method more clearly reflects the extraction of more salient and implicit features. This is the result of the aggregation and mining of multiscale attention features from global-scale structural information.

V. CONCLUSION

For the re-identification of pedestrians, a new multiscale reference-aided attentive feature aggregation (MS-RAFA) mechanism was proposed to learn more distinct features. First, to extract the most significant local and global features and strengthen the correlation between the features of various

parts of the human body, an attention mechanism called the autoselection module (ASM) was designed. Then, to extract and fuse low- and high-level features, an multilayer feature fusion module (MFFM) was proposed. The MFFM is an independent branch that fuses each layer of output from the combination of backbone networks, which enables the model to extract hidden features concealed by salient features and to learn them better. Finally, the effectiveness of the model was verified by multiple supervision mechanisms to obtain better model performance. Extensive ablation studies demonstrated the high efficiency and state-of-the-art performance of our design.

REFERENCES

[1] Y. Liu, Z. Yuan, W. Zhou, and H. Li, "Spatial and temporal mutual promotion for video-based person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 8786–8793.
[2] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2285–2294.
[3] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Mancs: A multi-task attentional network with curriculum sampling for person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 365–381.
[4] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3219–3228.
[5] Y. Wang, L. Wang, Y. You, X. Zou, V. Chen, S. Li, G. Huang, B. Hariharan, and K. Q. Weinberger, "Resource aware person re-identification across multiple resolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8042–8051.
[6] Y. Huang, Z.-J. Zha, X. Fu, and W. Zhang, "Illumination-invariant person re-identification," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 365–373.
[7] M. S. Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 420–429.
[8] H. Huang, D. Li, Z. Zhang, X. Chen, and K. Huang, "Adversarially occluded samples for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5098–5107.
[9] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "VRSTC: Occlusion-free video person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7183–7192.
[10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
[12] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[13] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-V4, inception-resnet and the impact of residual connections on learning," in *Conf. Artif. Intell.*, 2016, pp. 4278–4284.
[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[16] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 480–496.
[17] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proc. ACM Multimedia Conf. Multimedia Conf.*, Oct. 2018, pp. 274–282.
[18] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun, "AlignedReID: Surpassing human-level performance in person re-identification," 2017, *arXiv:1711.08184*. [Online]. Available: http://arxiv.org/abs/1711.08184
[19] F. Zheng, C. Deng, X. Sun, X. Jiang, X. Guo, Z. Yu, F. Huang, and R. Ji, "Pyramidal person re-identification via multi-loss dynamic training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8514–8522.
[20] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 384–393.
[21] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian, "Deep representation learning with part loss for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2860–2871, Jun. 2019.
[22] B. Chen, W. Deng, and J. Hu, "Mixed high-order attention network for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 371–381.
[23] W. Yang, H. Huang, Z. Zhang, X. Chen, K. Huang, and S. Zhang, "Towards rich feature discovery with class activation maps augmentation for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1389–1398.
[24] M. Zheng, S. Karanam, Z. Wu, and R. J. Radke, "Re-identification with consistent attentive Siamese networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5735–5744.
[25] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jan. 26, 2021, doi: 10.1109/TPAMI.2021.3054775.
[26] F. Wang, C. Zhang, S. Chen, G. Ying, and J. Lv, "Engineering hand-designed and deeply-learned features for person re-identification," *Pattern Recognit. Lett.*, vol. 130, pp. 293–298, Feb. 2020.
[27] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A Siamese long short-term memory architecture for human re-identification," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 135–153.
[28] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[29] D. S. Cheng and M. Cristani, "Person re-identification by articulated appearance matching," in *Person Re-Identification*. Springer, 2014, pp. 139–160.
[30] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *Proc. BMVC*, 2011, vol. 1, no. 2, p. 6.
[31] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 3037–3045, Oct. 2018.
[32] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. NIPS*, vol. 28, 2015, pp. 2017–2025.
[33] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, and T. Huang, "Horizontal pyramid matching for person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8295–8302.
[34] Z. Zhang, C. Lan, W. Zeng, and Z. Chen, "Densely semantically aligned person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 667–676.
[35] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Interaction-and-aggregation network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9317–9326.
[36] Y. Zhao, X. Shen, Z. Jin, H. Lu, and X.-S. Hua, "Attribute-driven feature disentangling and temporal aggregation for video person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4913–4922.
[37] X. Zhang, F. Zhou, Y. Lin, and S. Zhang, "Embedding label structures for fine-grained feature representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1114–1123.
[38] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 842–850.
[39] S. Li, S. Bak, P. Carr, and X. Wang, "Diversity regularized spatiotemporal attention for video-based person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 369–378.
[40] 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.

[41] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Mar. 2006, pp. 2169–2178.
[42] Z. Zhu, M. Xu, S. Bai, T. Huang, and X. Bai, "Asymmetric non-local neural networks for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 593–602.
[43] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 510–519.
[44] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 13001–13008.
[45] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[46] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 152–159.
[47] 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.
[48] 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. (ICCV)*, Oct. 2017, pp. 3754–3762.
[49] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, "Relation-aware global attention for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3186–3195.
[50] X. Chen, C. Fu, Y. Zhao, F. Zheng, J. Song, R. Ji, and Y. Yang, "Salience-guided cascaded suppression network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3300–3310.
[51] X. Jin, C. Lan, W. Zeng, G. Wei, and Z. Chen, "Semantics-aligned representation learning for person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 11173–11180.
[52] K. Zhu, H. Guo, Z. Liu, M. Tang, and J. Wang, "Identity-guided human semantic parsing for person re-identification," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K., Aug. 2020, pp. 346–363.
[53] Z. Cheng, Q. Dong, S. Gong, and X. Zhu, "Inter-task association critic for cross-resolution person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2605–2615.
[54] J. Zhou, B. Su, and Y. Wu, "Online joint multi-metric adaptation from frequent sharing-subset mining for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2909–2918.
[55] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.
[56] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
[57] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[58] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–10.
[59] P. Fang, J. Zhou, S. Roy, L. Petersson, and M. Harandi, "Bilinear attention networks for person retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8030–8039.
[60] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2138–2147.
[61] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3702–3712.
[62] Z. Dai, M. Chen, X. Gu, S. Zhu, and P. Tan, "Batch DropBlock network for person re-identification and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3691–3701.
[63] 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.
[64] L. T. Alemu, M. Shah, and M. Pelillo, "Deep constrained dominant sets for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9855–9864.
[65] C.-P. Tay, S. Roy, and K.-H. Yap, "AANet: Attribute attention network for person re-identifications," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7134–7143.
[66] 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.
[67] M. M. Kalayeh, E. Basaran, M. Gokmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1062–1071.
[68] F. Hong, D. Huang, and G. Chen, "Interaction-aware factorization machines for recommender systems," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 3804–3811.
[69] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
[70] Y. Hou, C. Chen, and Y. Li, "Efficient multi-granularity network based on local context-aware correlation feature for person re-identification," in *Proc. 6th Int. Conf. Intell. Comput. Signal Process. (ICSP)*, Apr. 2021, pp. 553–556, doi: 10.1109/ICSP51882.2021.9408832.
[71] H. Li, Y. Chen, D. Tao, Z. Yu, and G. Qi, "Attribute-aligned domain-invariant feature learning for unsupervised domain adaptation person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 1480–1494, 2021, doi: 10.1109/TIFS.2020.3036800.
[72] H. Li, N. Dong, Z. Yu, D. Tao, and G. Qi, "Triple adversarial learning and multi-view imaginative reasoning for unsupervised domain adaptation person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Jul. 26, 2021, doi: 10.1109/TCSVT.2021.3099943.
[73] X. Wu, B. Xie, Y. Zhang, S. Zhao, and S. Zhang, "Construction of diverse DropBlock branches for person re-identification," *IEEE Trans. Cognit. Develop. Syst.*, early access, Jul. 12, 2021, doi: 10.1109/TCDS.2021.3096546.
[74] X. Fan, J. Zhang, and Y. Lin, "Person re-identification based on mutual learning with embedded noise block," *IEEE Access*, vol. 9, pp. 129229–129239, 2021, doi: 10.1109/ACCESS.2021.3102450.
[75] M. Fu, S. Sun, H. Gao, D. Wang, X. Tong, Q. Liu, and Q. Liang, "Improving person re-identification using a self-focusing network in Internet of Things," "Improving person re-identification using a self-focusing network in Internet of Things," *IEEE Internet Things J.*, early access, May 31, 2021, doi: 10.1109/JIOT.2021.3084978.
[76] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Learning generalisable omni-scale representations for person re-identification," "Learning generalisable omni-scale representations for person re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Mar. 26, 2021, doi: 10.1109/TPAMI.2021.3069237.

LI XU is currently pursuing the master's degree with Nanchang Hangkong University. Her research interests include computer vision, image processing, and person re-identification.

XIANG FU received the Ph.D. degree from Xidian University, in 2008. He is currently an Associate Professor with Nanchang Hangkong University. His research interests include computer vision, image processing, and pattern recognition.