IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 February 2024, accepted 10 April 2024, date of publication 16 April 2024, date of current version 24 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3389698

RESEARCH ARTICLE

Hierarchical Attentive Feature Aggregation for Person Re-Identification

HUSHENG DONG[1,2] AND PING LU[3]

[1]School of Computer Engineering, Suzhou Vocational University, Suzhou 215104, China
[2]Jiangsu Province Support Software Engineering Research and Development Center for Modern Information Technology Application in Enterprise, Suzhou 215104, China
[3]School of Information Technology, Suzhou Institute of Trade and Commerce, Suzhou 215009, China

Corresponding author: Husheng Dong (hsdong2012@gmail.com)

This work was supported in part by the Research Funds of Suzhou Vocational University under Grant SVU2021YY03; and in part by the Science and Technology Program of Suzhou under Grant SS202151, Grant SNG2021037, and Grant SNGD202307.

ABSTRACT Recent efforts on person re-identification have shown promising results by learning discriminative features via the multi-branch network. To further boost feature discrimination, attention mechanism has also been extensively employed. However, the branches on the main level rarely communicate with others in existing branching models, which may compromise the ability of mining diverse features. To mitigate this issue, a novel framework called Hierarchical Attentive Feature Aggregation (Hi-AFA) is proposed. In Hi-AFA, a hierarchical aggregation mechanism is applied to learn attentive features. The current feature map is not only fed into the next stage, but also aggregated into another branch, leading to hierarchical feature flows along depth and parallel branches. We also present a simple Feature Suppression Operation (FSO) and a Lightweight Dual Attention Module (LDAM) to guide feature learning. The FSO can partially erase the salient features already discovered, such that more potential clues can be mined by other branches with the help of LDAM. By this manner, the branches could cooperate to mine richer and more diverse feature representations. The hierarchical aggregation and multi-granularity feature learning are integrated into a unified architecture that builds upon OSNet, resulting a resource-economical and effective person re-identification model. Extensive experiments on four mainstream datasets, including Market-1501, DukeMTMC-reID, MSMT17, and CUHK03, are conducted to validate the effectiveness of the proposed method, and results show that state-of-the-art performance is achieved.

INDEX TERMS Attention, diverse features, feature aggregation, person re-identification.

I. INTRODUCTION

Person re-identification aims to match a specific person captured by non-overlapping cameras, or across time using the same camera. In many surveillance applications, such as cross-camera tracking [1] and multi-person association [2], person re-identification serves as a fundamental technique and it is generally considered as an image retrieval problem. Despite great progress in recent years, person re-identification still remains an open research challenge. Due to large appearance variation arising from viewpoint changes, varying illumination conditions, occlusion, and complex background, it is rather difficult to match cross-view image pairs.

Extracting discriminative features that fully characterize the query person, and distinguish from others at the same time, is of vital importance for any person re-identification systems. Owing to remarkable ability of learning discriminative features, solutions based on Convolutional Neural Networks (CNNs) have become the mainstream for person re-identification [3], [4]. In practice, because global features are prone to ignore the information of small regions, it has been a trend to fuse global features with part-based local features [5], [6]. These local features are generally learned from multi-branch architectures with supervision, they can help re-identification models focus on fine-grained details

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

in each individual local part. Thus higher performance can be achieved when comparing to merely use global features [7], [8].

To further enhance the discrimination of feature representations, the visual attention mechanism has also been introduced into person re-identification [9], [10], [11], [12]. By endowing more distinguishable patterns with higher weights, attention mechanism equips networks with the ability of laying emphasis on more informative regions. In the meantime, the irrelevant background interference would be suppressed. Therefore, representations strengthened by attention mechanism can better represent pedestrian images and provide more distinguishable information.

Despite observed effectiveness of adopting local features and visual attention mechanism, there are two shortcomings of most existing person re-identification approaches. First, the branches on the main level rarely communicate with others in existing branching networks, the ability of finding potential clues remains improvement. Second, the widely used branching architecture usually brings high computational cost at the time of boosting performance. Especially in some works like [6] and [7] that several convolutional blocks are duplicated, or in [10] and [11] that heavy matrix multiplications are executed for attentions, the model complexity may increase greatly.

In this paper, we propose to address above problems by hierarchically aggregating features based on the Omni-Scale Network (OSNet) [13]. Technically, we first introduce a hierarchical feature aggregation strategy to progressively combine multi-scale features. The pre-stage feature map is not only fed into the next stage in current branch, but also aggregated into another parallel branch. In this way, the semantic and detail information at different stages and different branches are aggregated. During aggregation, a Feature Suppression Operation (FSO) is applied to partially erase feature maps with the aim of mining more diversified features. Intuitively, the erased regions generally correspond to the areas where network has strong activations, so other potential clues would stand out in the next branch. As a result, the branches are forced to work together, and all salient features can be extracted in a branch-by-branch manner. Besides, we also design a novel lightweight attention module to guide feature learning. Comparing to other typical attentions, the number of parameters and computation complexity are significantly reduced. To better leverage the multi-branch structure, the final feature maps in each branch are processed via different pooling strategy to obtain global, multi-granularity part-based, and channel-based features.

We name our model Hierarchical Attentive Feature Aggregation (Hi-AFA). Taking the advantage of lightweight OSNet [13] architecture, the number of parameters is kept in a low magnitude by using it as backbone. We note that our Hi-AFA is not restricted to the usage of OSNet, other lightweight architectures can also be employed as the backbone.

The main contributions of our work can be summarized as follows:

(1) We design a novel hierarchical feature aggregation framework (Hi-AFA), which aims to generate more discriminative features by combining the features of different levels and branches. By partially erasing feature maps via Feature Suppression Operation (FSO), the branches can cooperate to mine richer and more diversified features.

(2) We design a Lightweight Dual Attention Module (LDAM), which contains two complementary parts: Spatial Attention Module (SAM) and Channel Attention Module (CAM). Due to the adoption of group convolution, it has much less parameters than existing attentions, and the computational cost is quite low.

We integrate Hi-AFA and LDAM into the OSNet, forming a resource-economical and effective multi-branch network. From the branches, diverse features are computed for person re-identification. We conduct extensive experiments on four public person re-identification datasets. The proposed method achieves better performance or comparable results to a broad range of existing models, while keeping much lower model complexity.

The rest of this work is organized as follows. Section II briefly reviews related works. In Section III, the structure of Hi-AFA and LDAM will be elaborated. Section IV presents the experimental evaluations and some discussions. Finally, the whole work is concluded in Section V.

II. RELATED WORK

As one of the most active research areas in computer vision, a large number of solutions have been reported for person re-identification [14], [15]. In this section, we will briefly review some closely related works, including local feature learning, attention mechanism, and feature aggregation.

A. LOCAL FEATURE LEARNING FOR PERSON RE-IDENTIFICATION

The prevailing success of deep learning has made person reidentification no-exception. The earlier approaches based on deep learning, such as [3], [16], [17], and [18], naively applied CNN backbones to extract global features. Due to the limitation of being prone to ignore local information from small regions [19], more and more works focus on learning local features.

To obtain local features, the works in [20] and [21] firstly partitioned pedestrian images according to some predefined rules, and then computed local features from each sub-image separately. This approach is easy to implement, but the predefined partitions are not often ideally aligned with human body parts. Instead of using rough partition strategy, some methods extracted body part features via external clues like pose estimation and human part parsing. In [22], Zhang et al. constructed densely semantically aligned part images to assist feature learning. Rao et al. [23] learned multi-scale skeleton representations. However, these methods need to detect key

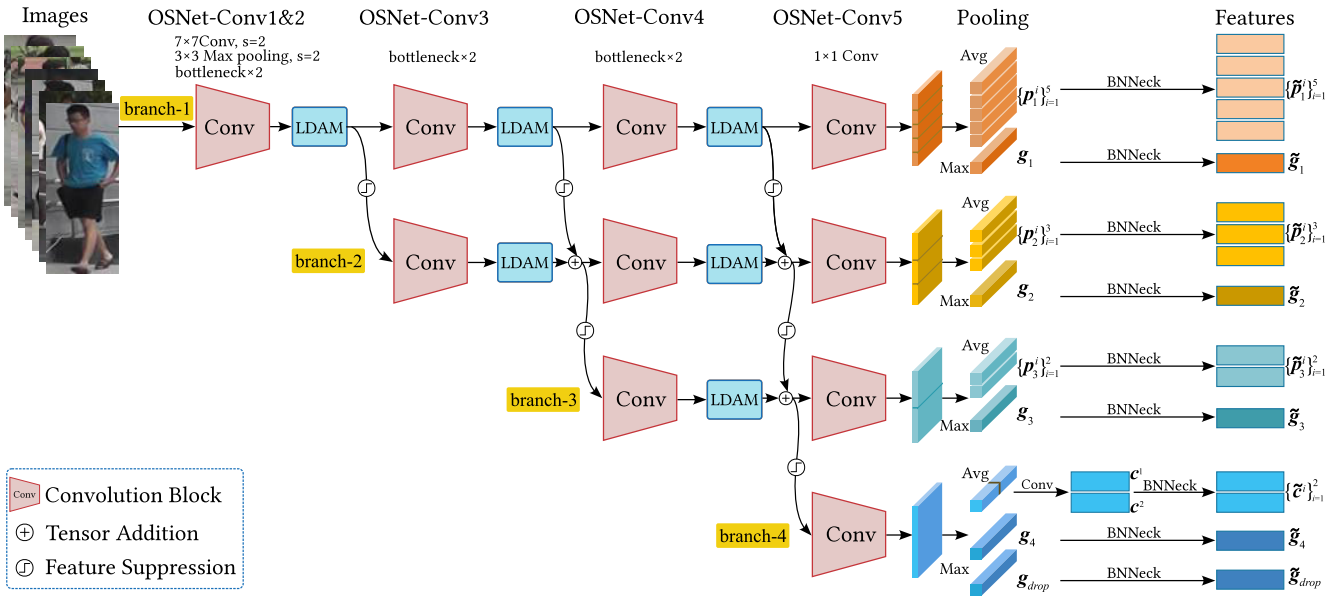


FIGURE 1. The architecture of Hierarchical Attentive Feature Aggregation (Hi-AFA) model. The OSNet is used as the backbone, and its transition stages are omitted for simplicity. There are four parallel branches in Hi-AFA, and their numbers of convolution blocks gradually decrease to 1 from branch-1 to branch-4. The feature maps are not only fed into the next convolution block in current branch, but also aggregated into next branch after suppression. Multi-granularity part-based local features and global features are computed from the first three branches. For branch-4, global and channel-based features are extracted, and DropBlock is applied to obtain another feature tensor. All pooled feature volumes are further forwarded to BNNeck to produce final embeddings.

points or perform semantic parsing with additional models, extra computation cost is inevitable [24].

Recently, splitting feature maps into a bunch of spatial parts has become the mainstream [4], [6], [7], [25], [26]. Generally speaking, the feature maps are obtained by multi-branch deep architectures first, and multi-granularity features are then acquired by pooling with different sizes. Part-based Convolutional Baseline (PCB) [4] is a typical representative of this type, which splits the last feature map into horizontal stripes of the same size. Multiple Granularity Networks (MGN) [7] improved PCB by adding a global branch to utilize the global features. Pyramid [6] learned multi-granularity features by dividing the final feature map into a pyramidal partition set. Although impressive performance is achieved, the branches mainly work separately in these works, the capability of mining diverse features is limited. While in Hi-AFA this is addressed by the aggregation structure assisted with feature suppression operation.

B. ATTENTION MECHANISM IN PERSON RE-IDENTIFICATION

The attention mechanism has also been introduced to person re-identification after success in other computer vision tasks like visual question answering [27] and scene segmentation [28]. As attention can guide model to focus on informative features while suppress irrelevant ones, it well matches the goal of handling challenges in person re-identification.

Directly incorporating a separate stream of spatial attention in deep networks is a common strategy for feature enhancement [29]. Li et al. [9] proposed a multi-granularity attention selection mechanism to better select region of interest. Si et al. [29] captured spatial dependencies among different pedestrian images by incorporating a correlation attention module. Chen et al. [30] learned the attention with counterfactual causality which can measure the attention quality and provide supervisory signal to guide learning process. Xun et al. [31] designed a local attention guided network to extract approximate semantic local features of human body parts. To better model long range dependencies, second order non-local attentions are computed in [8] and [11]. However, one potential limitation is that the computation cost is a bit high.

Channel-wise attention [32] has also been introduced to explore the correlations among different channels, the combination of spatial attention and channel attention can enhance feature representation further [10], [33]. To this end, Zhang et al. [34] captured the global structural information for better attention learning via mining pairwise correlations among feature positions and channels. Chen et al. [10] applied orthogonal regularizations to enforce diversity on attention maps. In [35], an attention-guided mask module was proposed to address occlusion problem. In [36], holistic and partial attentions are jointly learned to increase the feature robustness against pose variations.

C. FEATURE AGGREGATION FOR PERSON RE-IDENTIFICATION

Feature aggregation is a common strategy to make full use of features. In deep architectures like ResNet [37] and DenseNet [38], feature aggregation plays a vital role in relieving the vanishing gradient problem for feasible optimization.

In person re-identification, a number of solutions with feature aggregation have been reported [12], [39], [40], [41].

Chen et al. [12] employed a salience suppression strategy to mine diverse visual clues at different stages. Xu et al. [42] aggregated the predictions of multiple networks to mimic the decision process of multi-experts. Fu et al. [43] designed an iterative impression aggregation module to update features for similarity computation. Hou et al. [44] proposed to enhance feature representations by selectively aggregating correlated spatial and channel features. The typical two-stream network is employed to fuse the features extracted from different spaces in [45] and [46]. Based on the Vision Transformer (ViT) with impressive capability of exploiting structural patterns, Zhang et al. [47] proposed a hierarchical and iterative structure to refine and aggregate multi-level features. Wang et al. [48] proposed a neighbor transformer network to model interactions across all input images. However, one shortcoming of ViT based methods is that they are thirsty for training samples [49].

The proposed Hi-AFA learns local features via a multi-branch architecture and it splits feature maps into horizontal parts. To guide feature learning, both spatial and channel-wise attentions are included to build a lightweight dual attention module. Due to the branching architecture, Hi-AFA might look like PyConv [50], FractalNet [51], CliqueNet [52], and BranchyNet [53] at first glance. However, the branches of FractalNet are trained alternately, which implies the sub-paths still work separately in essence. The parameters in CliqueNet [52] are recurrently updated many times, the computational cost is too high. For BranchyNet and PyConv, there are no aggregations to utilize features of different stages. The Hi-AFA is also related to [41] and [47] that sharing the same idea of aggregating intermediate features. But there are notable differences with Hi-AFA: (1) A Feature Suppression Operation (FSO) is applied to partially erase feature maps, thereby allowing the network to discover diverse visual clues. (2) The attentive features at intermediate stages are aggregated along both depth and parallel branches. (3) Multi-granularity part-based and channel-based features are extracted from the branches for better utilization.

III. METHODOLOGY

Let $\mathcal{T} = \{\boldsymbol{I}_i, y_i\}_{i=1}^{n}$ be a set of training images, where $\boldsymbol{I}_i \in \mathbb{R}^{H\times W\times 3}$ is the ith pedestrian image with corresponding label $y_i \in \{1, 2, \cdots, c\}$ and c is the number of identities. For each image, our goal is to compute its rich and diverse feature representations via a multi-branch architecture. To achieve this goal, the proposed Hi-AFA relies on a given CNN backbone and enriches it with hierarchical aggregation branches. By this manner, more potential clues can be mined for fine-grained cross-view matching. The overall architecture of Hi-AFA is illustrated in Figure 1.

Due to outstanding ability of feature extraction, the off-the-shelf OSNet [13] is utilized as the backbone of Hi-AFA. Similar to PyConv [50], multiple filters are utilized to learn diverse features in each convolutional block of OSNet. There are five convolutional blocks in OSNet, which will be referred to as Conv1 to Conv5 hereafter, and the key component of them is the bottleneck illustrated in Figure 2. The Conv1 block contains a standard 7×7 convolution layer and a 3×3 max pooling layer, both are conducted with stride 2. From Conv2 to Conv4, each contains two bottlenecks. A transition block, which serves as down sampler, is followed after Conv2 and Conv3. The Conv5 block contains a 1×1 convolution only. Benefiting from the design of multiple convolutional feature streams in bottleneck, OSNet [13] outperforms ResNet50 [37] and its variants (e.g., PyramidNet [54]) with much lower model complexity on the re-identification task.

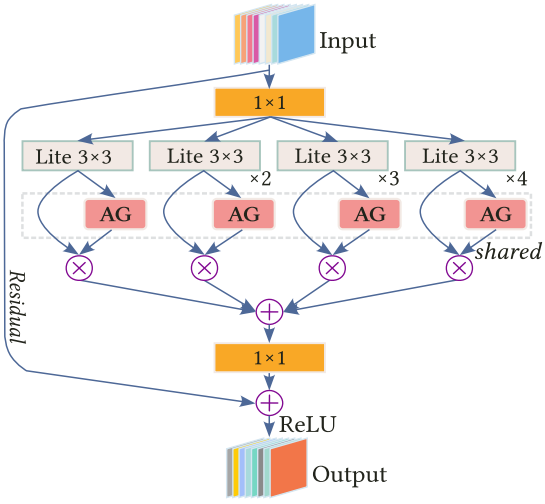


FIGURE 2. The bottleneck of OSNet [13]. The Lite 3 × 3 convolution consists of a 1 × 1 convolution, a depth-wise 3 × 3 convolution, Batch normalization, and ReLU activation. AG means aggregation gate, which is a learnable neural network. ×2, ×3, and ×4 represent the Lite 3 × 3 convolution is repeated 2, 3, and 4 times.

Our Hi-AFA can be roughly divided into three parts: the common OSNet-Conv1&2 blocks, hierarchical attentive feature aggregation, and final feature processing. Images are first passed through OSNet backbone, up until its Conv3 block. After forwarding images through the initial layers, the network forms an upper triangle structure of multiple branches, which comprise the remaining layers of OSNet up to Conv5 block. By this design, the layers up to Conv3 are shared by all the branches. This concept has been employed in a few person re-identification solutions like [7], [25], and [26], which can decrease model size effectively. Finally, the feature volumes in each branch are pooled with different size, such that we can obtain multi-granularity features. The part-based local features are computed via average pooling, and max pooling is utilized to get global features. The key components of Hi-AFA are detailed in the following.

A. HIERARCHICAL FEATURE AGGREGATION

It has been demonstrated multi-scale feature aggregation can help to improve person re-identification performance [42], [44], [47]. However, traditional aggregation operations

generally only consider aggregating high- and low-level features. Few efforts have been devoted to the cooperation of branches for potential clues mining in multi-branch architecture. In this work, the proposed hierarchical feature aggregation aims to combine features from different branches, such that richer and more diverse features can be explored.

As shown in Figure 1, from branch-1 to branch-4 the numbers of convolutional blocks gradually decrease to 1 due to aggregation structure. And extra links are added between adjacent branches in Hi-AFA, which makes it differ from previous multi-branch network with independent branches. By this design, the feature stream also flows along parallel branches for aggregation. As a consequence, the branches are forced to cooperate with each other.

Let $\mathcal{F}_l : \mathbb{R}^{H\times W\times C} \mapsto \mathbb{R}^d$ be the feature extraction function parameterized by a set of trainable parameters $\mathcal{W}_l$, where $l \in \{1, 2, \cdots, 5\}$ is the stage index of OSNet [13] backbone, H, W, C are the height, width, and channels of a tensor. The feature representation of an image $\boldsymbol{I}$ at lth stage ($l \geq 2$) can be denoted as $\boldsymbol{X}_{b,l} = \mathcal{F}_{b,l}\left(\boldsymbol{I}; \mathcal{W}_{b,l}\right)$, where $b \in \{1, \cdots, 4\}$ is the branch index in Hi-AFA. If we denote $\mathcal{F}_{b,l}\left(\boldsymbol{I}; \mathcal{W}_{b,l}\right) = \mathbf{0}$ ($2 \leq b = l \leq 4$), then the feature aggregation can be formulated as

$$\begin{aligned}\boldsymbol{X}_{b,l} =& \mathcal{A}\left(\mathcal{F}_{b,l-1}\left(\boldsymbol{I}; \mathcal{W}_{b,l-1}\right)\right) + \mathcal{S}\left(\mathcal{A}\left(\mathcal{F}_{b-1,l}\left(\boldsymbol{I}; \mathcal{W}_{b-1,l}\right)\right)\right) \\ & 2 \leq b \leq 4, b+1 \leq l \leq 5, \end{aligned} \tag{1}$$

where $\mathcal{A}(\cdot)$ and $\mathcal{S}(\cdot)$ represent the computation of attention and FSO respectively. Here we introduce FSO first, the attention will be detailed in the next section.

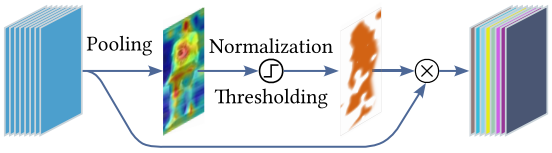


FIGURE 3. A schematic of the proposed feature suppression operation.

Although the branches are enforced to cooperate in the aggregation structure, they may fall into the trivial salient features if no extra guidance is provided. To address this problem, FSO is particularly applied to attentive features before aggregation, which functions a little like the dropout. But unlike dropout that randomly chooses units to deactivate, FSO only filters out high responses, so as to suppress the salient features discovered in previous branch. Despite some information loss due to the thresholding process, the branches are endowed with the ability to mine more potential visual clues for visual matching, and this is critical to the re-identification task.

As illustrated in Figure 3, we first apply channel-wise average pooling to get averaged 2-D feature map $\boldsymbol{Y}_{b,l}$ given $\boldsymbol{X}_{b,l}$, and obtain its normalized version $\bar{\boldsymbol{Y}}_{b,l}$ by min-max normalization. Then, we compute a thresholding mask $\boldsymbol{M}_{b,l}$ as follows:

$$\boldsymbol{M}_{b,l}(x, y) = \begin{cases} 0, & if\, \bar{\boldsymbol{Y}}_{b,l}(x, y) > \tau \\ 1, & otherwise \end{cases} \tag{2}$$

where $\tau \in (0, 1]$ is a thresholding parameter assigned manually, and $\bar{\boldsymbol{Y}}_{b,l}(x, y)$ stands for the intensity value at position (x, y). With obtained $\boldsymbol{M}_{b,l}$, the suppressed features $\widetilde{\boldsymbol{Y}}_{b,l}$ can be computed as $\widetilde{\boldsymbol{Y}}^c_{b,l} = \boldsymbol{Y}^c_{b,l} \otimes \boldsymbol{M}_{b,l}$ (c is the channel index of $\boldsymbol{Y}_{b,l}$, and $\otimes$ represents the elementwise multiplication). Because there are only simple average pooling, normalization, and thresholding operation in FSO, it can be performed efficiently.

Based on the hierarchical aggregation structure, diversified features can be obtained for person re-identification. First, the multi-level attentive features in different branches are recurrently aggregated, thus diversified information can be utilized. Second, potential important features may stand out in the next branch after the previous salient feature being suppressed. The network is thereby enabled to extract all potential useful features branch-by-branch.

B. LIGHTWEIGHT DUAL ATTENTION MODULE

The proposed Lightweight Dual Attention Module (LDAM) can be viewed as a variant of the classical Convolutional Block Attention Module (CBAM) [55], which consists of Channel Attention Module (CAM) and Spatial Attention Module (SAM). The two types of attention modules work in a complementary manner to enhance feature representations. CAM explores the correlation between channel features, and SAM aims to capture and aggregate semantically related spatial features. But LDAM differs from CBAM in attention computation process, especially the group convolution employed in CAM and SAM, which leads to much less parameters than CBAM. As a result, the computational cost is quite low. Besides, the softmax activation is used in LDAM, other than sigmoid in CBAM. The detail of LDAM is as follows.

1) CHANNEL ATTENTION MODULE

It is well known that each channel map of high-level convolutional feature can be viewed as a class-specific response, and the responses are generally semantic-related. In person re-identification task, it will contribute to better fine-grained recognition if some channels sharing similar sematic contexts (e.g., foreground and background) are more correlated. Thus, we group and aggregate those semantically correlated channels by explicitly exploiting the interdependencies between channel maps.

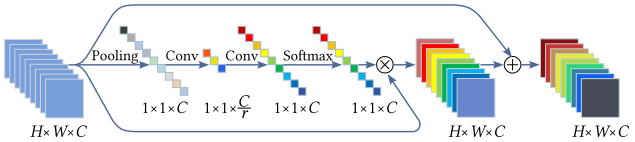


FIGURE 4. Structure of channel attention module.

The structure of CAM is illustrated in Figure 4. Given a local feature tensor $\boldsymbol{X} \in \mathbb{R}^{H\times W\times C}$, we first squeeze the spatial dimension with average pooling and max pooling. It is known that average pooling can well retain structural information, but it is easily distracted by background interference. Max pooling overcomes this problem by focusing on the most salient part, while the cost is some structural information loss. In CAM, we jointly use them to obtain two context descriptors of $\boldsymbol{x}_{avg} \in \mathbb{R}^{1\times 1\times C}$ and $\boldsymbol{x}_{max} \in \mathbb{R}^{1\times 1\times C}$, and aggregate them via summation to obtain $\widetilde{\boldsymbol{x}} = \boldsymbol{x}_{avg} + \boldsymbol{x}_{max}$. Then, we use group convolution to squeeze the channel size of $\widetilde{\boldsymbol{x}}$ to C/r, where r is a shrinkage parameter. After dividing $\widetilde{\boldsymbol{x}}$ to g independent fractions, we apply $1 \times 1 \times C/g$ filters on each of them and concatenate the resulting intermediate descriptors. By such group convolution, we can achieve the typical convolution with much less parameters. Similarly, a second group convolution layer is applied to restore the channel size to C. At last, a softmax activation is applied. The whole procedure of CAM can be formulated as

$$h = \text{softmax}\left(\text{gconv}_2\left(\text{gconv}_1\left(\widetilde{\boldsymbol{x}}\right)\right)\right), \tag{3}$$

where $\text{gconv}_1(\cdot)$ and $\text{gconv}_2(\cdot)$ represent the two group convolutions. Finally, we can obtain the output of CAM by

$$\boldsymbol{A}_{ch} = \gamma \boldsymbol{X} \otimes \boldsymbol{h} + \boldsymbol{X}, \tag{4}$$

where γ is a hyperparameter to adjust the impact of CAM. In equation (4), each position of $\boldsymbol{X}$ is multiplied with $\boldsymbol{h}$ along the channel dimension.

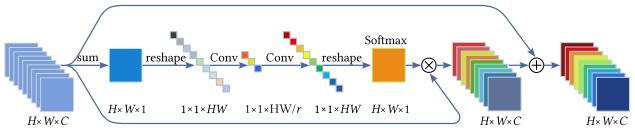


FIGURE 5. Structure of spatial attention module.

2) SPATIAL ATTENTION MODULE

An illustration of SAM is shown in Figure 5. In contrast to CAM, SAM captures and aggregates related features in the spatial domain. Given a local feature map $\boldsymbol{X}$ with size $H \times W \times C$, SAM first obtains a 2-D matrix $\boldsymbol{M} \in \mathbb{R}^{H\times W}$ by summation over the channels for each spatial position, i.e., $\boldsymbol{M}(x, y) = \sum_{c=0}^{C} \boldsymbol{X}^c(x, y)$. Here, $\boldsymbol{X}^c$ represents the submap of $\boldsymbol{X}$ at cth channel. Then $\boldsymbol{M}$ is reshaped to $1 \times 1 \times HW$ for convenience of applying two sets of 1×1 convolutions. Similar to the two group convolution layers in CAM, a context descriptor with the shape of $1 \times 1 \times HW/r$ is obtained after the first convolution, and the second one restores its shape back to $1 \times 1 \times HW$. After that, a softmax function is applied, and 2-D attention map $\boldsymbol{H} \in \mathbb{R}^{H\times W}$ is obtained by restoring the shape back. The value at each position of $\boldsymbol{H}$ indicates the degree of importance for that location. Formally, the spatial attention map $\boldsymbol{H}$ is computed as

$$\boldsymbol{H} = \text{vec}^{-1}\left(\text{softmax}\left(\text{gconv}_2\left(\text{gconv}_1\left(\text{vec}\left(\boldsymbol{M}\right)\right)\right)\right)\right), \tag{5}$$

where $\text{vec}(\cdot)$ and $\text{vec}^{-1}(\cdot)$ represent the vectorization of a 2-D matrix and its inverse operation, respectively. With $\boldsymbol{H}$, the output of SAM can be computed as

$$\boldsymbol{A}_{sp} = \gamma \sum_{c=1}^{C} \boldsymbol{X}^c \otimes \boldsymbol{H} + \boldsymbol{X}. \tag{6}$$

It can be found that there are only simple operations of pooling, 1×1 convolution, and softmax in LDAM, the computational cost is rather low. For CAM and SAM both, two group convolutional layers are applied, which make LDAM differ from CBAM [55] in structure. The usage of group convolutions follows the squeeze and excitation process in SENet [32], which can enable network increase its sensitivity to informative features while greatly reduce the parameters. Consequently, the channel and spatial dependencies will be better modeled. Besides, softmax rather than sigmoid activation function is used in two types of attention modules. This is because softmax can encourage filters to learn diverse features, hence making the model more robust.

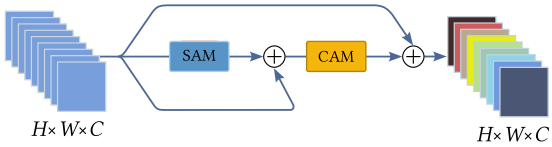


FIGURE 6. Sequential combination of SAM and CAM in LDAM.

It has been proved that sequential combination of SAM and CAM can lead to better performance [34], [55], we follow the same scheme to place SAM in front of CAM for attention learning (see Figure 6 for illustration). Due to the lightweight design of LDAM, it is quite flexible and can be easily plugged into networks multiple times if necessary.

C. FEATURE PROCESSING IN EACH BRANCH

In order to learn multi-granularity features and make a better usage of them, we employ a simple partition strategy to obtain global, part-based, and channel-based features. The final feature maps in each branch are equally partitioned with different size to get multiple granularity local features. Both global and local features are extracted from each branch. In addition, we also extract channel-based features via channel partition.

To extract part-based local features, we simply divide the final feature map into n_b submaps according to the number of convolutional blocks in each branch. That is, n_b equals 5, 3, and 2, from branch-1 to branch-3. The local features $\left\{\boldsymbol{p}_b^i\right\}_{i=1}^{n_b}$ ($\mathcal{B} = \{1, 2, 3\}, b \in \mathcal{B}$) are all acquired by spatial average pooling, and their shapes are of $24 \times 8 \times 512$. Additionally, we use max pooling on the initial feature maps, obtaining global representations $\left\{\boldsymbol{g}_b\right\}$ ($b \in \mathcal{B}$) of 512-dimension. The hybrid usage of average and max pooling here can help to retain structural information and obtain robust global feature simultaneously.

FIGURE 7. Example images randomly chosen from three benchmark datasets. Images in each row are of the same person in each dataset.

For branch-4, we first aggregate the information by global max pooling on the tensor, resulting a vector $\boldsymbol{g}_4 \in \mathbb{R}^{512}$. We also apply the mask computed via DropBlock [56] to the feature map, the resulting tensor is further applied with global max pooling. This leads to another vector $\boldsymbol{g}_{drop} \in \mathbb{R}^{512}$. In addition to $\boldsymbol{g}_4$ and $\boldsymbol{g}_{drop}$, two channel-based feature vectors are also extracted. After reducing the original feature map using average pooling, we split the resulting 512-dimension vector into two sub vectors and each has a length of 256. Then, 1×1 convolution is used to rescale them to 512-dimension, by which two channel-based vectors $\boldsymbol{c}^1 \in \mathbb{R}^{512}$ and $\boldsymbol{c}^2 \in \mathbb{R}^{512}$ are obtained.

During training, the global features in $\mathcal{R} = \{\boldsymbol{g}_{drop}, \boldsymbol{g}_{b'}\}$ ($\mathcal{B}' = \{1, 2, 3, 4\}, b' \in \mathcal{B}'$) will be fed into a ranking loss to learn distance metrics. We also use BNNeck [57] to obtain $\mathcal{I} = \{\widetilde{\boldsymbol{g}}_{drop}, \widetilde{\boldsymbol{g}}_{b'}, \widetilde{\boldsymbol{p}}_b^i, \widetilde{\boldsymbol{c}}^k\}$ ($b' \in \mathcal{B}', b \in \mathcal{B}, 1 \le i \le n_b$, $k \in \{1, 2\}$), by which identity classifiers will be learned. The BNNeck is comprised of a batch normalization and a fully connected layer with number-of-classes units. During inference, the network without BNNeck and classifiers will be used as a feature extractor, which is utilized to extract features for all query and gallery images. Then, Euclidian distance is calculated to perform a standard information retrieval.

D. LOSS FUNCTIONS

The combination of identification loss, ranking loss, and center loss [57] is adopted for the optimization of network parameters.

The cross-entropy with label smoothing [58] is used as identification loss, which treats each identity as a distinct class. In each minibatch, the label smoothed cross-entropy is defined as

$$\mathcal{L}_{xe} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\left((1-\epsilon)\, y_i^k + \frac{\epsilon}{K}\right)\log\left(p_i^k\right), \quad (7)$$

where $\epsilon \in (0, 1)$ is a smoothing parameter, N is the mini-batch size, K is the number of identities, y_i^k and p_i^k are the ground-truth and predicted probability respectively.

For computation of ranking loss, the multi-similarity [59] is utilized. As a pair-based list-wise loss function, multi-similarity loss integrates pair mining and soft weighting scheme into a single-framework. The multi-similarity loss is computed as

$$\mathcal{L}_{ms} = -\frac{1}{N}\sum_{i=1}^{N}\{\frac{1}{\alpha}\log\left[1 + \sum_{k\in\mathcal{P}_i}\exp\left(-\alpha\left(S_{ik} - \lambda\right)\right)\right] + \frac{1}{\beta}\log\left[1 + \sum_{k\in\mathcal{N}_i}\exp\left(\beta\left(S_{ik} - \lambda\right)\right)\right]\}, \quad (8)$$

where $S_{ik} = \langle \boldsymbol{\psi_i}, \boldsymbol{\psi_k} \rangle$ is the dot product of feature vectors $\boldsymbol{\psi}_i$ and $\boldsymbol{\psi}_k$, α, β, and λ are manually set hyper-parameters, $\mathcal{P}_i$ and $\mathcal{N}_i$ are the selected positive and negative pairs for an anchor $\boldsymbol{\psi}_i$.

To enhance the compactness of each identity cluster, the center loss [57] is also included, which is defined as

$$\mathcal{L}_{ce} = \frac{1}{2}\sum_{i=1}^{N}\|\boldsymbol{\psi}_i - \boldsymbol{c}_{y_i}\|_2^2, \quad (9)$$

where $\boldsymbol{c}_{y_i}$ denotes the center of class y_i.

During training, the final loss function is

$$\mathcal{L} = \lambda_{xe}\sum_{\boldsymbol{\psi}\in\mathcal{I}}\mathcal{L}_{xe} + \lambda_{ms}\sum_{\boldsymbol{\psi}\in\mathcal{R}}\mathcal{L}_{ms} + \lambda_{ce}\sum_{\boldsymbol{\psi}\in\mathcal{I}\cup\mathcal{R}}\mathcal{L}_{ce}, \quad (10)$$

where λ_{xe}, λ_{ms}, and λ_{ce} are suitable weights that can be obtained by grid search. The identification loss L_{xe}, ranking loss $\mathcal{L}_{ms}$, and center loss $\mathcal{L}_{ce}$ are computed over $\mathcal{I}$, $\mathcal{R}$, and $\mathcal{I} \cup \mathcal{R}$, separately.

IV. EXPERIMENTS

In this section, we report the experimental results of the proposed Hi-AFA on four mainstream person re-identification datasets, including Market-1501 [60], DukeMTMC-reID [61], MSMT17 [62], and CUHK03 [63]. Figure 7 shows some randomly selected images. We compare Hi-AFA with a line of state-of-the-art solutions, and conduct extensive ablation studies to investigate the effectiveness of each component.

A. DATASETS

We conduct experiments on the following four widely used person re-identification datasets.

Market-1501 [60] is currently the most popular person re-identification dataset, which is captured by six cameras. This dataset contains 1,501 identities with 32,668 bounding boxes obtained by the Deform Part Model (DPM) detector. The

training set contains 751 identities with 12,936 images, and in the testing set there are 750 identities with 3,368 query images and 19,732 gallery images.

DukeMTMC-reID [61] contains 36,441 images of 1,404 pedestrians captured by eight cameras. A total of 16,552 images belonging to 702 identities make up the training set, and the remaining 702 identities along with 408 distractors make up the testing set. In the testing set, there are 2,268 query images and 17,661 gallery images respectively.

MSMT17 [62] is collected by twelve outdoor and three indoor cameras. There are 4,101 identities with a total of 126,441 images. It is divided into a training set of 32,621 images and a testing set of 93,820 images. Due to its massive scale, more complex and dynamic sciences, it is much more challenging to perform person re-identification on MSMT17.

CUHK03 [63] consists of 14,097 pedestrian images of 1,467 identities captured from two disjoint camera views. There are two types of bounding boxes in CUHK03, one is obtained by human annotation, and the other is detected by DPM. We adopt the splitting protocol of 767/700 identities for training and testing on this dataset.

B. EXPERIMENTAL SETTINGS

1) IMPLEMENTATION DETAILS

The OSNet [13] initialized with the weights pretrained on ImageNet is used as our backbone. All images are resized to 384×128 pixels such that more detailed information can be captured. For both training and testing, the input images are normalized to channel-wise zero mean and a standard variation of 1. During training, we adopt a data augmentation strategy of random cropping, horizontal flip, as well as random erasing. The model is trained 200 epochs with a batch size of 64. Each mini-batch consists of 8 identities, with 8 instances per identity. The Adam optimizer with $\epsilon = 1 \times 10^{-8}$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$ is used for training. The learning rate is set to 8×10^{-4} with a weight decay of 5×10^{-4}. In LDAM, the shrinkage parameter r is set to 8, and a group size of $g = 8$ is used for group convolution. The hyper-parameters α, β, and λ in equation (10) are set to 2, 40, and 0.5. The balance parameters λ_{xe}, λ_{ms}, and λ_{ce} in equation (10) are empirically set to 0.5, 0.5, and 5×10^{-4}. We use the same settings for all considered datasets.

2) EVALUATION METRICS

The Cumulative Matching Characteristic (CMC) at top ranks and mean Average Precision (mAP) are reported as evaluation metrics. The value at different ranks of CMC shows the re-identification accuracy by counting the query identities among the top n results. The mAP reflects the overall re-identification accuracy by calculating the area under the precision-recall curve. We note that all experiments are conducted under the single-shot scenario.

C. COMPARISON WITH STATE-OF-THE-ART METHODS

Table 1 shows the performance of our proposed Hi-AFA and other state-of-the-arts on Market-1501, DukeMTMC-reID, MSMT17, and CUHK03. The compared methods can be generally grouped into three categories: discriminative feature learning based (top of the table), attention based (middle of the table), and transformer based (bottom). We report the mAP and CMC values at Rank-1/5 for comparison. We observe that Hi-AFA achieves superior performance on multiple benchmarks or competitive results compared to previous methods.

1) RESULTS ON MARKET-1501

Our Hi-AFA achieves 91.8% mAP and 97.0%/99.0% Rank-1/5 accuracies on this dataset. Comparing to the previous best Rank-1 96.3% reported by LightMBN [25], the improvement is 0.7%. Although the mAP of Hi-AFA is lower than previous best ABD+NFormer [48], it still ranks the second. Note that the stunning mAP of ABD+NFormer mainly comes from NFormer, which improves the mAP of ABD-Net [10] from 88.3% to 93.0%. As NFormer can be viewed as a post-processing module, some higher mAP is natural. We also conduct experiments with Hi-AFA+NFormer. For each image, the features extracted via Hi-AFA are concatenated to a representation vector. NFormer is then applied to all vectors in a mini-batch to obtain their final representations. Following [48], the number of neighbors is also set to 20 in Hi-AFA+NFormer. The obtained mAP and Rank-1 accuracy are as high as 95.4%/97.2% on Market-1501, exceeding other methods significantly.

Compared to the two representative feature learning based methods of Pyramid [6] and MGN [7], the improvements of mAP and Rank-1 accuracy are 3.6%/4.9% and 1.3%/1.3%. Because Hi-AFA shares similar branching structure with them, we believe the improvements should be attributed to the aggregation structure and attention modules. Among the methods based on attention or transformer, IANet [44], SCSN [12], and HAT [47] all embrace the aggregation strategy to make better use of multi-scale features. Whereas our Hi-AFA outperforms all of them, which demonstrates the encouraging ability of learning discriminative features in Hi-AFA.

2) RESULTS ON DUKEMTMC-REID

Hi-AFA achieves competitive results on this dataset. The mAP of Hi-AFA is 82.9%, which ranks the second among all methods. The highest score is 85.7%, reported by ABD+NFormer [48] again. On the most important Rank-1, Hi-AFA achieves the same score with AdaSP [67] and BPB(Res50-IBN) [70], all report 91.7% matching accuracy. When Hi-AFA is welded with NFormer, the mAP and Rank-1 are improved to 91.1% and 94.0%, outperforming all others significantly. Compared with SCSN [12] and HAT [47] that aggregate information via cascaded attentions or transformers, the superiority of Hi-AFA is obvious. The mAP and Rank-1 are improved by 3.9%/1.5% and 0.7%/1.3%. Both of them have to undertake heavy computation burden to mine diverse features, while in Hi-AFA

TABLE 1. Performance comparison of Hi-AFA with the state-of-the-art methods on Market-1501, DukeMTMC-reID, MSMT, and CUHK03 datasets. R1/5 indicates Rank-1/5 accuracy. In each column, the highest score is marked in bold, and the second-best is underlined.

Method	Market-1501			DukeMTMC-reID			MSMT17			CUHK03-(L/D)	
	mAP	R1	R5	mAP	R1	R5	mAP	R1	R5	mAP	R1
AlignedReID [5]	79.1	91.8	-	69.7	82.1	-	43.7	69.8	-	-/59.6	-/61.5
PCB+RPP [4]	80.9	93.3	97.4	68.1	82.9	90.3	-	-	-	-/63.7	-57.5
HOReID [64]	84.9	94.2	-	75.6	86.9	-	-	-	-	-	-
OSNet [13]	84.9	94.8	-	73.5	88.6	-	52.9	78.7	-	-/67.8	-/72.3
CDNet [65]	86.0	95.1	-	76.8	88.6	-	54.7	78.9	-	-	-
FED [66]	86.3	95.0	-	78.0	89.4	-	-	-	-	-	-
ICE [30]	86.6	95.1	98.3	76.5	88.2	94.1	50.4	76.4	86.6	-	-
MGN [7]	86.9	95.7	-	78.4	88.7	-	-	-	-	67.4/66	68.0/66.8
DSA [22]	87.6	95.7	-	74.3	86.2	-	-	-	-	75.2/73.1	78.9/78.2
C2F [8]	87.7	94.8	97.2	74.9	87.4	92.1	-	-	-	79.3/84.1	80.6/81.3
Pyramid [6]	88.2	95.7	98.4	79.0	89.0	-	-	-	-	76.9/74.8	78.9/78.9
AdaSP [67]	89.8	95.5	-	83.0	91.7	-	67.1	85.5	-	82.4/80.1	84.6/82.0
LightMBN [25]	91.2	96.3	-	-	-	-	-	-	-	85.1/82.4	87.2/84.9
HA-CNN [9]	75.7	91.2	-	63.8	80.5	-	-	-	-	41/38.6	44.4/41.7
IANet [44]	83.1	94.4	-	73.4	87.1	-	46.8	75.5	85.5	-	-
MHAN [68]	85.0	95.1	98.1	77.2	89.1	94.6	-	-	-	72.4/65.4	77.2/71.7
DCA [69]	87.5	94.7	-	80.1	89.0	-	64.0	83.1	-	-	-
DAAF [36]	87.9	95.1	-	77.9	87.9	-	-	-	-	67.6/63.1	69/64.9
ABD-Net [10]	88.3	95.6	-	78.6	86.0	-	60.8	82.3	90.6	-	-
RGA [34]	88.4	96.1	-	-	-	-	57.5	80.3	-	77.4/74.5	81.1/79.6
BPB(Res50-IBN) [70]	88.4	95.7	-	81.3	91.7	-	-	-	-	-	-
SCSN [12]	88.5	95.7	-	79.0	91.0	-	58.0	83.0	91.2	84.0/81.0	86.8/84.7
CAL [71]	89.5	95.5	98.5	80.5	90.0	96.1	64.0	84.2	92.0	-	-
APNet [72]	90.5	96.2	98.8	81.5	90.4	95.6	63.5	83.7	91.7	85.3/81.5	87.4/83.0
PAT [73]	88.0	95.4	-	78.2	88.8	-	-	-	-	-	-
TransReID [74]	89.5	95.2	-	82.6	90.7	-	69.4	86.2	-	-	-
HAT [47]	89.5	95.6	-	81.4	90.4	-	-	-	-	80.0/75.5	82.6/79.1
MSINet [75]	89.6	95.3	-	-	-	-	59.6	81.0	-	-	-
PHA [76]	90.2	96.1	-	-	-	-	68.9	86.1	-	83.0/80.3	84.5/83.2
ABD+NFormer [48]	93.0	95.7	-	85.7	90.6	-	62.2	80.8	-	79.1/76.4	80.6/79.0
Hi-AFA	91.8	97.0	99.0	82.9	91.7	96.2	71.9	87.6	94.3	85.4/83.6	87.9/85.5
Hi-AFA+NFormer	**95.4**	**97.2**	**99.3**	**91.1**	**94.0**	**96.4**	**76.7**	**90.2**	**94.6**	**88.7/86.4**	**89.5/88.6**

this is achieved by simple but effective hierarchical feature aggregation and FSO.

3) RESULTS ON MSMT17

Our Hi-AFA achieves the best mAP (71.9%) and Rank-1 (87.6%) over all previous competitors. The previous best is TransReID [74], which reports 69.4% mAP and 86.2% Rank-1 accuracy. Although TransReID benefits from the transformer-based learning structure, Hi-AFA outperforms it with 2.5%/1.4%. On top of that, much higher performance of 76.7% mAP and 90.2% Rank-1 accuracy can be obtained by Hi-AFA+NFormer. From Table 1, we can also observe that Hi-AFA has obvious superiority over other multi-branch feature learning based and attention-based models. Take the feature learning based AdaSP [67] for example, its mAP and Rank-1 are 67.1% and 85.5%, while our Hi-AFA exceeds it by 4.8% and 2.1%. When compared with attention based DCA [69], the improvements are even higher. The results on MSMT17 demonstrates the scalability of Hi-AFA on such a huge person re-identification benchmark.

4) RESULTS ON CUHK03

As shown in Table 1, Hi-AFA achieves the best in terms of both mAP and Rank-1 accuracy, which gives 85.4%/83.6% mAP and 87.9%/85.5% Rank-1 matching accuracy on labeled and detected settings respectively. The previous best was reported by APNet [72], which gives 85.3%/81.5% mAP and 87.4%/83.0% Rank-1 accuracy. The improvements are 0.1%/2.1% for mAP, and 0.5%/2.5% for Rank-1 accuracy. With the support of NFormer, the results can be boosted to 88.7%/86.4% and 89.5%/88.6%. Compared to the backbone OSNet [13], Hi-AFA improves the mAP and Rank-1 accuracy by as large as 15.8% and 13.2% under the detected setting, which justifies the superiority of aggregating attentive features.

D. ABLATION STUDY

In the following, we systematically investigate the effectiveness of each key component of Hi-AFA, namely hierarchical feature aggregation, FSO, LDAM, along with the final feature processing. Experiments are conducted on all four considered datasets. On CUHK03, only the labeled version (CUHK03-L)

TABLE 2. Results of different sub-models and backbones on four considered datasets (%), the highest score in each column is marked in bold.

Branches	Market-1501		DukeMTMC-reID		MSMT17		CUHK03-L	
	mAP	R1	mAP	R1	mAP	R1	mAP	R1
OSNet [13]	84.9	94.8	73.5	88.6	52.9	78.7	69.3	73.5
branch-1	87.6	95.3	76.5	89.4	58.2	82.4	77.6	81.2
branch-{1, 2}	90.1	96.4	79.2	90.5	66.8	85.5	80.5	83.2
branch-{1, 2, 3}	91.5	96.8	82.3	91.5	71.0	86.9	84.2	86.4
branch-{1, 2, 3, 4}	**91.8**	**97.0**	**82.9**	**91.7**	**71.9**	**87.6**	**85.4**	**87.9**
Hi-AFA-*BrIndep*	90.6	96.5	82.0	91.3	70.2	86.4	83.5	86.1
baseline(ResNet-50)	83.7	94.5	71.6	87.3	51.7	76.0	69.0	73.2
Hi-AFA(ResNet-50)	90.2	96.0	81.4	90.5	69.2	85.1	84.3	86.5
baseline(DenseNet-169)	84.4	94.6	72.5	87.9	51.9	76.1	70.6	74.1
Hi-AFA(DenseNet-169)	90.8	96.2	82.2	90.7	69.5	85.2	84.9	87.2

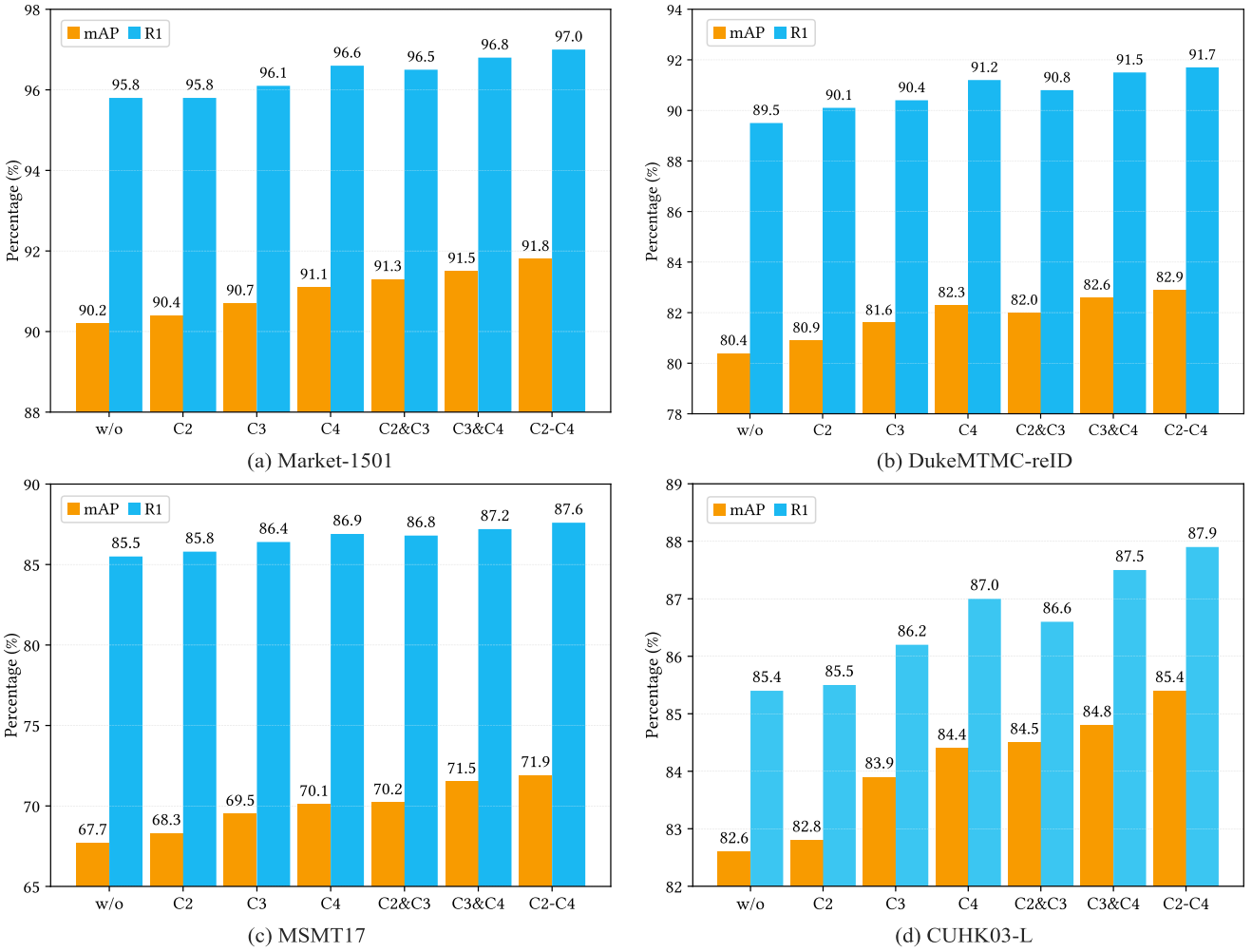


FIGURE 8. Performance comparison of Hi-AFA under different FSO embedding settings. w/o means without FSO, C*i* means FSO is embedded after the *i*th convolution block, and C2-C4 means from convolution block 2 to 4.

is considered, since the two types of bounding boxes are from same source. The results are obtained with only one setting changed and the rest remain the same.

1) EFFECT OF HIERARCHICAL FEATURE AGGREGATION

The hierarchical feature aggregation structure plays an important role in the proposed Hi-AFA model. To investigate its effectiveness, different sub-models of Hi-AFA are evaluated. We use the branch-1 in Hi-AFA as basic model, and then gradually add other branches to it. The Hi-AFA with independent branches (denoted as Hi-AFA-*BrIndep*) and backbone OSNet [13] are also evaluated for comparison.[1] Note that in Hi-AFA-*BrIndep*, only the first links between branches are kept, all later ones are discarded. Thus the branches work independently.

[1]OSNet did not report results on CUHK03-L. We obtain them by ourselves.

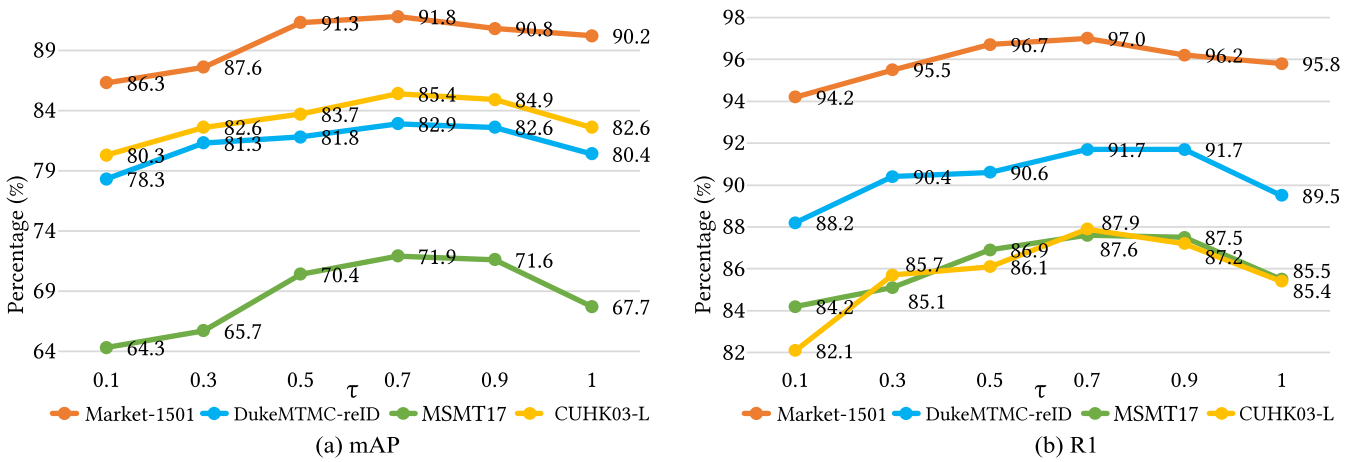


FIGURE 9. Variation of mAP (a), and Rank-1 accuracy (b) with respect to parameter τ on each dataset.

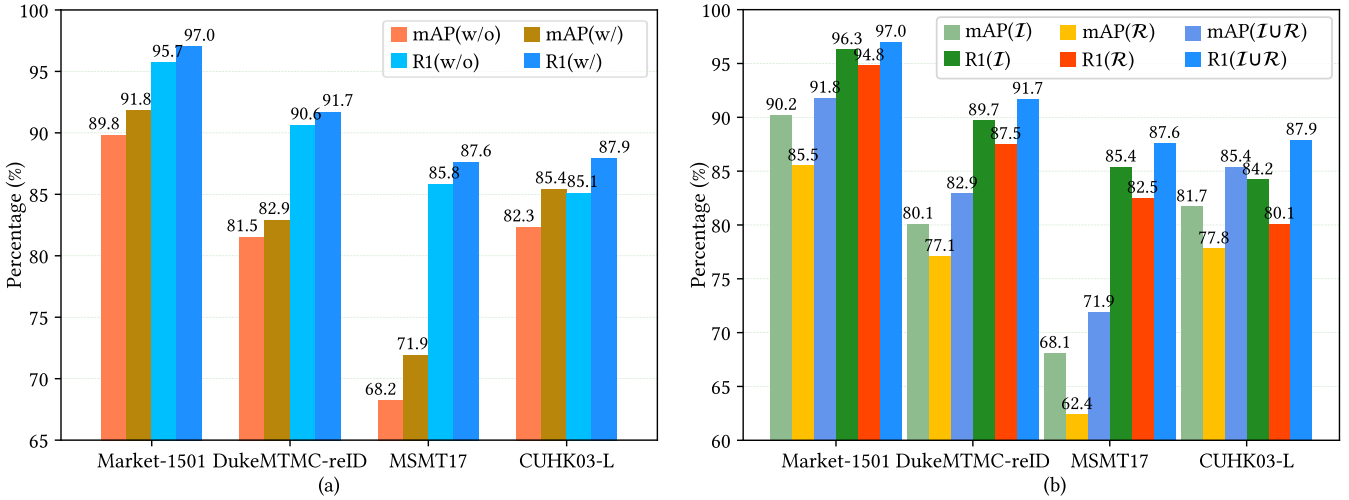


FIGURE 10. Performance comparison of (a) Hi-AFA with (w/) and without (w/o) LDAM, (b) $\mathcal{I}$ and $\mathcal{R}$ feature sets.

Table 2 demonstrates the results of each sub-model. We observe that with merely branch-1 quite encouraging results can be achieved. For example, it gives 87.6% mAP and 95.3% Rank-1 accuracy on Market-1501, which are 2.7% and 0.5% higher that the results of backbone OSNet [13]. By gradually adding other branches, the re-identification performance increases accordingly on all datasets. This proves that the feature aggregation structure in Hi-AFA can lead to significant performance improvements. From Table 2 we can also find that the mAP and Rank-1 of Hi-AFA-*BrIndep* are obviously lower than full-state Hi-AFA with all links (i.e., branch-{1, 2, 3, 4}). This indicates that the lateral links between adjacent branches are vital to the final re-identification performance, because they could enforce the branches cooperate with each other to explore more potential clues. While in Hi-AFA-*BrIndep* the branches work independently with no correspondence, the performance drops in consequence.

In the bottom of Table 2, the results of Hi-AFA with two other widely used backbones of ResNet-50 [37] and DenseNet-169 [38], are also reported. We first evaluate them as backlines, and then apply our Hi-AFA in these backbones. We observe that consistent improvements can be achieved on both of them, which indicates Hi-AFA is effective for different backbones. In general, the DenseNet-169 performs slightly better than ResNet-50, but they are all inferior to OSNet. Therefore, OSNet is our first choice of backbone.

2) EFFECT OF FSO

To demonstrate the effect of feature suppression, we evaluate Hi-AFA with different FSO embedding strategies, including without FSO (w/o), the main architecture equipped with FSO after Conv2 to Conv4 in backbone network (C2, C3, and C4), and different combinations of them at consecutive stages (C2&C3, C3&C4, and C2-C4).

From the evaluation results shown in Figure 8, we can draw the following observations. (1) FSO can boost the re-identification performance effectively. With FSO embedded, both mAP and Rank-1 accuracy can be obviously improved. For instance, even the weakest embedding strategy of C2 can bring 0.2% mAP gain on Market-1501 dataset. (2) The later stage FSO is embedded, the higher performance gain will be acquired. This is a natural result. It is well known that the higher-stage convolutional features are more

TABLE 3. Comparison of different attentions (%).

Attention	Params(M)	FLOPs(G)	Market1501		DukeMTMC-reID		MSMT17		CUHK03-L	
			mAP	R1	mAP	R1	mAP	R1	mAP	R1
LDAM	0.26	0.006	91.8	97.0	82.9	91.7	71.9	87.6	85.4	87.9
CBAM [55]	8.39	0.27	90.8	96.3	82.2	91.3	69.1	86.3	84.8	87.1
RGA [34]	2.66	79.89	92.3	97.2	83.4	92.2	72.6	87.9	85.9	88.5
Nonlocal [77]	5.25	32.23	91.4	96.5	81.8	91.6	70.4	87.6	84.3	86.6

category-related than shallow layers. By embedding FSO into latter stages of CNN backbone, more diverse and discriminant features can be obtained, thus resulting better matching results. (3) The combination of FSOs can further boost the re-identification performance. Similar to the usage of single-stage FSO, the combination of C3&C4 also performs better than C2&C3, demonstrating the superiority of later feature suppression again. By plugging FSO into all stages, C2-C4 gives the highest results on all datasets. Comparing to the model without FSO, the improvements are 1.6%/1.2%, 2.5%/2.2%, 4.2%/2.1%, and 2.8%/2.5% respectively. This comparison justifies the effectiveness of mining diverse features by FSO.

3) FEATURE SUPPRESSION THRESHOLD ANALYSIS

The parameter threshold τ in FSO controls the degrees of feature suppression operation in Hi-AFA, so it is of vital importance to choose a proper threshold. With a low threshold, too much features will be erased, which is harmful to feature learning. On the contrary, a high threshold may limit the removal of enough features, the branches cannot cooperate well to mine new significant ones. To carefully choose the optimal value of threshold τ, we conduct experiments by varying its value from 0.1 to 1 and plot the corresponding mAP and Rank-1 in Figure 9. It can be observed that results on four datasets generally present a similar trend. Both mAP and Rank-1 accuracy increase when threshold τ grows larger at the first stage, and highest scores are obtained roughly at $\tau = 0.7$. But when τ keeps increasing, the performance begins to degrade. Therefore, we set τ to 0.7 for performance consideration.

4) EFFECT OF LDAM

In the proposed Hi-AFA, LDAM plays an important role of guiding feature learning. To investigate its effectiveness, we conduct comparative experiments of Hi-AFA with and without LDAM. Under the setting of Hi-AFA without LDAM, all attention modules are removed for a clean comparison. The result is shown in Figure 10 (a). It can be found that, Hi-AFA consistently outperforms the model without LDAM by a large margin. With the guidance of LDAM, the mAP is improved by 2.0%, 1.4%, 3.7%, 3.1%, and Rank-1 accuracy is also promoted by 1.3%, 1.1%, 1.8%, 2.8% on each dataset. This demonstrates that LDAM can effectively guide Hi-AFA to learn discriminative and robust features for cross-view matching.

In addition to experiments of utilizing LDAM or not, three other attentions including CBAM [55], RGA [34], and Nonlocal [77] are also compared with LDAM. We use the same Hi-AFA architecture and replace LDAM with these attentions to conduct experiments. The performance comparison is shown in Table 3. We can observe that RGA [34] performs consistently better than others due to its consideration of structural relationship between human body parts. It outperforms the second best by 0.5%/0.2%, 0.5%/0.5%, 0.7%/0.3%, and 0.5%/0.6% on each dataset. Although the performance of LDAM is a bit lower than RGA [34], it performs better than CBAM [55] and Nonlocal [77]. Since LDAM and CBAM have similar architectures, we think the performance improvement should be mainly attributed to the group convolution which endows attention with more flexibility.

Given a tensor of shape $H \times W \times C$, the computational complexity of LDAM is $\mathcal{O}((H^2W^2 + C^2)/(gr))$, and it is $\mathcal{O}(HWC + C^3/r)$ for CBAM. RGA and Nonlocal are at the same level of $\mathcal{O}(H^2W^2C + HWC^2)$, which is much higher than the former two. Owing to the group convolution in LDAM, its complexity is the lowest. In Table 3, we also present the Floating-Point Operations (FLOPs) and Parameters (Params) of each attention. The results are obtained by feeding each attention with an input tensor of shape $32 \times 24 \times 8 \times 2048$. It can be found that there are only 0.26M parameters in LDAM, and the FLOPs are merely 0.006G, which is quite lightweight. On the contrary, there are heavy matrix multiplications in Nonlocal [77] and RGA [34], the FLOPs of them amount to as high as 32.23G and 79.89G, respectively. From the view of performance, RGA [34] should be the best choice for guiding feature learning. However, when our perspective shifts to the model size and computational cost, lightweight attentions will be more welcome, and the proposed LDAM is a good compromise.

5) EFFECT OF FINAL FEATURE PROCESSING

In Hi-AFA, two feature sets are obtained finally, namely $\mathcal{I}$ and $\mathcal{R}$. $\mathcal{R}$ consists of all global features which are obtained by max pooling and DropBlock. Features in $\mathcal{I}$ contain two groups, one is obtained by applying BNNeck to Features in $\mathcal{R}$, and the other group contains spatial- and channel-wise partitioned local features. To investigate the effect of such combination of global features, spatial- and channel-wise local features, we conduct experiments with $\mathcal{I}$, $\mathcal{R}$, and $\mathcal{I} \cup \mathcal{R}$, respectively. The results are presented in Figure 10 (b). It can be seen that much higher performances are obtained with $\mathcal{I}$ than with $\mathcal{R}$, which means that the diverse local features are more discriminant than global ones. Besides, we can find that

TABLE 4. Comparison of performance with different feature settings (%).

Feature	Market1501		DukeMTMC-reID		MSMT17		CUHK03-L	
	mAP	R1	mAP	R1	mAP	R1	mAP	R1
baseline	91.5	96.8	82.4	91.5	71.1	87.0	84.3	86.5
+ g_{drop}	91.7	96.8	82.6	91.5	71.5	87.3	85.2	87.7
+ $\{c^i\}_{i=1}^2$	91.7	96.9	82.7	91.7	71.6	87.5	84.9	87.4
all	91.8	97.0	82.9	91.7	71.9	87.6	85.4	87.9

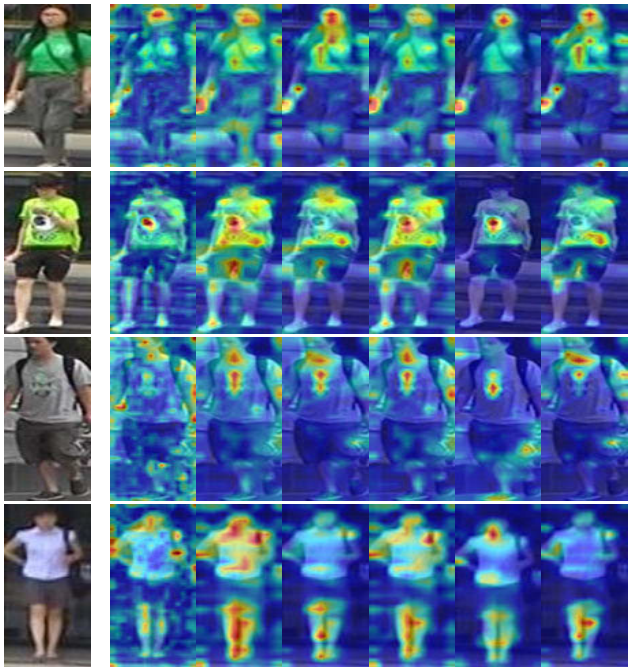


FIGURE 11. *Visualization of attention maps in Hi-AFA. B*i*C*j *indicates the attention map of* j*th convolution block in branch-*i.

$\mathcal{I} \cup \mathcal{R}$ significantly outperforms $\mathcal{I}$ or $\mathcal{R}$ alone, demonstrating the importance of joint usage of global and local features. Note that we apply identification loss to features in $\mathcal{I}$ and ranking loss to $\mathcal{R}$, and both of them are supervised by center loss. In such way, the features can be fully utilized and the advantages of different losses are fully exploited.

To validate the effectiveness of DropBlock and channel-wise features, we first use all features except $\{\boldsymbol{g}_{drop}, \boldsymbol{c}^1, \boldsymbol{c}^2\}$ as baseline, and then add $\boldsymbol{g}_{drop}$, $\{\boldsymbol{c}^1, \boldsymbol{c}^2\}$, and both of them for evaluation. The results are shown in Table 4, it can be found that each of DropBlock and channel features can bring certain performance promotion. When both $\boldsymbol{g}_{drop}$ and channel-wise features are added, the mAP and rank-1 accuracy are improved by 0.3%/0.2%, 0.5%/0.2%, 0.8%/0.6%, and 1.1%/1.4% on each dataset. This indicates that better generalization can be obtained with them.

E. VISUALIZATION OF ATTENTION MAPS

To investigate the attended image regions of each attention module, we use Grad-CAM [78] to visualize the attention maps for qualitative analysis. In all branches, the attention maps after each attention module are generated.

TABLE 5. Comparison of model size and complexity.

Model	Params(M)	Memory(MB)	FLOPs(G)
OSNet [13]	2.19	10.41	1.47
MGN [7]	74.38	179.55	11.94
Pyramid [6]	31.06	122.72	6.12
ABD-Net [10]	49.91	148.12	10.12
MHAN [68]	30.36	186.18	24.55
Hi-AFA	12.76	55.83	2.24

As shown in Figure 11, we can observe that the attentions at convolution block 2 are relatively coarse, multiple parts are of high importance in every attention map. When going deeper, they become more concentrated, forming few blobs on salient parts. For attention maps at the same stage, the attended areas are generally consistent but differs from each other in detail. Take B1C4, B2C4, and B3C4 in last row for example, besides the commonly highlighted legs, they focus on left shoulder, head, and right elbow, respectively. This proves the capability of mining diverse salient features of different branches. Therefore, they can greatly help to distinguish visual similar pedestrians in person re-identification task.

F. MODEL COMPLEXITY

The idea of learning diverse features via multi-branch architecture is quite popular in person re-identification. It enables networks to focus on different person features in individual branches. However, such branching strategy brings higher computational cost at the time of boosting re-identification performance. Although our Hi-AFA also embraces the branching strategy, the reduction of computational complexity is considered in the first place. In either the backbone or attention module, much less parameters are required. In Table 5, the space complexity and model size of Hi-AFA, some other branching models, as well as the backbone OSNet [13] are listed, in terms of FLOPs, Params, and Memory size. We can find that there are only 12.76M parameters in the proposed Hi-AFA, the consumption of memory is 55.83MB, and the FLOPs are about 2.24G. Although it is about 6 times larger than the backbone OSNet [13], Hi-AFA is still quite slim when comparing to other branching models.

V. CONCLUSION

In this paper, we present a novel Hierarchical Attentive Feature Aggregation (Hi-AFA) network to address the challenging person re-identification task. In Hi-AFA, the features are aggregated not only along the depth, but also the parallel branches. In such way, the branches can work together to mine more diverse and richer features for fine-grained recognition. To guide the feature learning, we design a lightweight dual attention module, in which much less parameters are required. With the aim of capturing essential person features, we extract global, channel-based and multi-granularity part-based features from the distinct branches.

Due to the usage of lightweight backbone and attention module, the overall model complexity of Hi-AFA is kept on a lower level than state-of-the-art models, but superior or comparable performance is obtained on four mainstream person re-identification datasets. Ablation analysis is also performed to investigate the insight of the proposed model. The backbone of Hi-AFA is not restricted to OSNet, other lightweight deep convolutional models can also be utilized. In future work, we will continue the research on more effective and lighter person re-identification.

REFERENCES

[1] 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.

[2] M. Vo, E. Yumer, K. Sunkavalli, S. Hadap, Y. Sheikh, and S. G. Narasimhan, "Self-supervised multi-view person association and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2794–2808, Aug. 2021.

[3] Y. Sun, L. Zheng, W. Deng, and S. Wang, "SVDNet for pedestrian retrieval," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3820–3828.

[4] 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.

[5] H. Luo, W. Jiang, X. Zhang, X. Fan, J. Qian, and C. Zhang, "AlignedReID++: Dynamically matching local information for person re-identification," *Pattern Recognit.*, vol. 94, pp. 53–61, Oct. 2019.

[6] 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. 8506–8514.

[7] 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.

[8] A. Zhang, Y. Gao, Y. Niu, W. Liu, and Y. Zhou, "Coarse-to-fine person re-identification with auxiliary-domain classification and second-order information bottleneck," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 598–608.

[9] 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.

[10] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, and Z. Wang, "ABD-Net: Attentive but diverse person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8350–8360.

[11] B. Bryan, Y. Gong, Y. Zhang, and C. Poellabauer, "Second-order non-local attention networks for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3759–3768.

[12] 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. 3297–3307.

[13] 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. 3701–3711.

[14] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 2872–2893, Jun. 2022.

[15] Z. Ming, M. Zhu, X. Wang, J. Zhu, J. Cheng, C. Gao, Y. Yang, and X. Wei, "Deep learning-based person re-identification methods: A survey and outlook of recent works," *Image Vis. Comput.*, vol. 119, Mar. 2022, Art. no. 104394.

[16] X. Bai, M. Yang, T. Huang, Z. Dou, R. Yu, and Y. Xu, "Deep-person: Learning discriminative deep features for person re-identification," *Pattern Recognit.*, vol. 98, Feb. 2020, Art. no. 107036.

[17] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose-invariant embedding for deep person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4500–4509, Sep. 2019.

[18] I. Lasri, A. Riadsolh, and M. Elbelkacemi, "Facial emotion recognition of deaf and hard-of-hearing students for engagement detection using deep learning," *Educ. Inf. Technol.*, vol. 28, no. 4, pp. 4069–4092, Apr. 2023.

[19] C. Ding, K. Wang, P. Wang, and D. Tao, "Multi-task learning with coarse priors for robust part-aware person re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1474–1488, Mar. 2022.

[20] F. Yang, K. Yan, S. Lu, H. Jia, X. Xie, and W. Gao, "Attention driven person re-identification," *Pattern Recognit.*, vol. 86, pp. 143–155, Feb. 2019.

[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] 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.

[23] H. Rao, X. Hu, J. Cheng, and B. Hu, "SM-SGE: A self-supervised multi-scale skeleton graph encoding framework for person re-identification," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1812–1820.

[24] J. Li, S. Zhang, Q. Tian, M. Wang, and W. Gao, "Pose-guided representation learning for person re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 2, pp. 622–635, Feb. 2022.

[25] F. Herzog, X. Ji, T. Teepe, S. Hörmann, J. Gilg, and G. Rigoll, "Lightweight multi-branch network for person re-identification," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 1129–1133.

[26] G. Wang, Y. Yuan, J. Li, S. Ge, and X. Zhou, "Receptive multi-granularity representation for person re-identification," *IEEE Trans. Image Process.*, vol. 29, pp. 6096–6109, 2020.

[27] S. Whitehead, H. Wu, H. Ji, R. Feris, and K. Saenko, "Separating skills and concepts for novel visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5628–5637.

[28] T. Wu, J. Huang, G. Gao, X. Wei, X. Wei, X. Luo, and C. H. Liu, "Embedded discriminative attention mechanism for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16760–16769.

[29] T. Si, F. He, H. Wu, and Y. Duan, "Spatial-driven features based on image dependencies for person re-identification," *Pattern Recognit.*, vol. 124, Apr. 2022, Art. no. 108462.

[30] H. Chen, B. Lagadec, and F. Bremond, "ICE: Inter-instance contrastive encoding for unsupervised person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14940–14949.

[31] X. Gong, Z. Yao, X. Li, Y. Fan, B. Luo, J. Fan, and B. Lao, "LAG-Net: Multi-granularity network for person re-identification via local attention system," *IEEE Trans. Multimedia*, vol. 24, pp. 217–229, 2022.

[32] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[33] 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.*, 2018, pp. 365–381.

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

[35] P. Chen, W. Liu, P. Dai, J. Liu, Q. Ye, M. Xu, Q. Chen, and R. Ji, "Occlude them all: Occlusion-aware attention network for occluded person re-ID," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11813–11822.

[36] Y. Chen, H. Wang, X. Sun, B. Fan, C. Tang, and H. Zeng, "Deep attention aware feature learning for person re-identification," *Pattern Recognit.*, vol. 126, Jun. 2022, Art. no. 108567.

[37] 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.

[38] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[39] H. Wang, L. Jiao, S. Yang, L. Li, and Z. Wang, "Simple and effective: Spatial rescaling for person reidentification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 1, pp. 145–156, Jan. 2022.

[40] S. Zhou, J. Wang, D. Meng, Y. Liang, Y. Gong, and N. Zheng, "Discriminative feature learning with foreground attention for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4671–4684, Sep. 2019.

[41] N. Martinel, G. L. Foresti, and C. Micheloni, "Deep pyramidal pooling with attention for person re-identification," *IEEE Trans. Image Process.*, vol. 29, pp. 7306–7316, 2020.

[42] B. Xu, J. Liang, L. He, and Z. Sun, "Mimic embedding via adaptive aggregation: Learning generalizable person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 372–388.

[43] D. Fu, B. Xin, J. Wang, D. Chen, J. Bao, G. Hua, and H. Li, "Improving person re-identification with iterative impression aggregation," *IEEE Trans. Image Process.*, vol. 29, pp. 9559–9571, 2020.

[44] 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. 9309–9318.

[45] J. Li, S. Zhang, and T. Huang, "Multi-scale temporal cues learning for video person re-identification," *IEEE Trans. Image Process.*, vol. 29, pp. 4461–4473, 2020.

[46] W. Zhang, Z. Li, H. Du, J. Tong, and Z. Liu, "Dual-stream feature fusion network for person re-identification," *Eng. Appl. Artif. Intell.*, vol. 131, May 2024, Art. no. 107888.

[47] G. Zhang, P. Zhang, J. Qi, and H. Lu, "HAT: Hierarchical aggregation transformers for person re-identification," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 516–525.

[48] H. Wang, J. Shen, Y. Liu, Y. Gao, and E. Gavves, "NFormer: Robust person re-identification with neighbor transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7287–7297.

[49] D. Li, S. Chen, Y. Zhong, and L. Ma, "DiP: Learning discriminative implicit parts for person re-identification," 2022, *arXiv:2212.13906*.

[50] I. Cosmin Duta, L. Liu, F. Zhu, and L. Shao, "Pyramidal convolution: Rethinking convolutional neural networks for visual recognition," 2020, *arXiv:2006.11538*.

[51] G. Larsson, M. Maire, and G. Shakhnarovich, "FractalNet: Ultra-deep neural networks without residuals," 2016, *arXiv:1605.07648*.

[52] Y. Yang, Z. Zhong, T. Shen, and Z. Lin, "Convolutional neural networks with alternately updated clique," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2413–2422.

[53] S. Teerapittayanon, B. McDanel, and H. T. Kung, "BranchyNet: Fast inference via early exiting from deep neural networks," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 2464–2469.

[54] D. Han, J. Kim, and J. Kim, "Deep pyramidal residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6307–6315.

[55] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.

[56] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "DropBlock: A regularization method for convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, Dec. 2018, pp. 10750–10760.

[57] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu, "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Trans. Multimedia*, vol. 22, no. 10, pp. 2597–2609, Oct. 2020.

[58] 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.

[59] X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott, "Multi-similarity loss with general pair weighting for deep metric learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5017–5025.

[60] 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.

[61] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 17–35.

[62] 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.

[63] 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.

[64] 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. 6448–6457.

[65] H. Li, G. Wu, and W.-S. Zheng, "Combined depth space based architecture search for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6725–6734.

[66] Z. Wang, F. Zhu, S. Tang, R. Zhao, L. He, and J. Song, "Feature erasing and diffusion network for occluded person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4744–4753.

[67] X. Zhou, Y. Zhong, Z. Cheng, F. Liang, and L. Ma, "Adaptive sparse pairwise loss for object re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19691–19701.

[68] 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.

[69] H. Zhu, W. Ke, D. Li, J. Liu, L. Tian, and Y. Shan, "Dual cross-attention learning for fine-grained visual categorization and object re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4682–4692.

[70] V. Somers, C. D. Vleeschouwer, and A. Alahi, "Body part-based representation learning for occluded person re-identification," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 1613–1623.

[71] Y. Rao, G. Chen, J. Lu, and J. Zhou, "Counterfactual attention learning for fine-grained visual categorization and re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 1005–1014.

[72] G. Chen, T. Gu, J. Lu, J.-A. Bao, and J. Zhou, "Person re-identification via attention pyramid," *IEEE Trans. Image Process.*, vol. 30, pp. 7663–7676, 2021.

[73] Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang, and F. Wu, "Diverse part discovery: Occluded person re-identification with part-aware transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2897–2906.

[74] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "TransReID: Transformer-based object re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14993–15002.

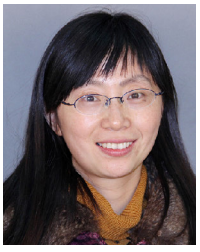
[75] J. Gu, K. Wang, H. Luo, C. Chen, W. Jiang, Y. Fang, S. Zhang, Y. You, and J. Zhao, "MSINet: Twins contrastive search of multi-scale interaction for object Reid," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19243–19253.

[76] G. Zhang, Y. Zhang, T. Zhang, B. Li, and S. Pu, "PHA: Patch-wise high-frequency augmentation for transformer-based person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14133–14142.

[77] 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.

[78] 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.

HUSHENG DONG received the M.S. and Ph.D. degrees from Soochow University, in 2008 and 2018, respectively. He is currently an Associate Professor with the School of Computer Engineering, Suzhou Vocational University. His research interests include computer vision, image processing, and deep learning.



PING LU received the B.Eng. and M.S. degrees from the School of Computer Science and Technology, Soochow University, in 2002 and 2005, respectively. She is currently an Associate Professor with Suzhou Institute of Trade and Commerce. Her research interests include digital image processing and pattern recognition.