IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 28 March 2023, accepted 21 April 2023, date of publication 8 May 2023, date of current version 25 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3274473

RESEARCH ARTICLE

Neighboring-Part Dependency Mining and Feature Fusion Network for Person Re-Identification

CHUAN ZHU, WENJUN ZHOU, YINGJUN ZHU, AND JIANMIN MA
Department of Aeronautics and Astronautics, Fudan University, Shanghai 200201, China
Corresponding author: Jianmin Ma (jmma@fudan.edu.cn)

ABSTRACT Person re-identification (Re-ID) is a computer vision technique used to determine the presence of a specific pedestrian target in an image or video sequence. It is an important branch of image retrieval. With the advancements in deep learning, notable progress has been achieved in Re-ID research. However, existing methods primarily focus on the most prominent features in the image, ignoring other less obvious yet beneficial features and spatial interdependencies within the image. To address this issue, this paper proposes a neighboring-part dependency mining and feature fusion network (NDMF-Net). The network horizontally splits pedestrian features into multiple parts, using a part-level hybrid attention module (PHAM) to focus on the salient region of each part, and a neighboring-part dependency exploration module (NDEM) to extract the spatial dependencies between neighboring parts of the image. Eventually, different features are fused to form the final representation. We validate the NDMF-Net on mainstream datasets and the experimental results demonstrate that our method is effective and achieves state-of-the-art performance.

INDEX TERMS Person re-identification, deep learning, part-level hybrid attention, spatial dependencies, feature fusion.

I. INTRODUCTION

Person re-identification (Re-ID) is a typical sub-task of image retrieval, which uses computer vision, pattern recognition, machine learning, and other technologies to determine whether there is a specific person in the image or video sequence [1]. Person Re-ID has various applications in criminal investigation, video surveillance, intelligent commerce, and other fields [2], [3]. In recent years, it has received extensive attention from both industry and academia.

Before the emergence of deep learning in the field of person Re-ID, traditional methods based on manual features are mainstream. Traditional person Re-ID methods mainly rely on manually created features such as color, texture, shape, etc. This method is not only inefficient but also weak in robustness, especially when dealing with complex and variable scenes, it is difficult to extract features with sufficient discriminability [4], [5], [6]. In recent years, with the advancement of computer vision technology, deep learning-based person Re-ID has seen rapid development and achieved good performance [7], [8], [9], [10]. Compared with traditional methods, deep learning-based person Re-ID methods can not only perform end-to-end training, but also adaptively extract useful features from images, and avoid the subjectivity inherent in traditional methods. Furthermore, by jointly learning different features, the performance and robustness of the network have been greatly improved, surpassing that of traditional methods. Additionally, the current data explosion provides a large number of data samples, and deep learning-based methods can study the correlation between group samples, further improving the performance of person Re-ID. Nevertheless, in practical applications, posture changes, background interference, and partial occlusions still pose great challenges for feature recognition and extraction.

Convolutional neural networks (CNNs) are the most widely used deep learning networks in image processing, particularly for feature extraction in person Re-ID tasks. The prevalent features employed in person Re-ID tasks include

The associate editor coordinating the review of this manuscript and approving it for publication was Vivek Kumar Sehgal.

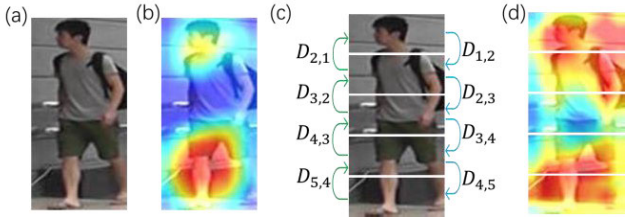


FIGURE 1. Motivation of the paper. (a) Original image. (b) Feature heat map by an attention-based method. (c) Dependencies between neighboring parts (taking the example of dividing the image into 5 parts), where D_{ij} represents the dependence of the ith part on the jth part. (d) Feature heat map by the NDMF-Net. The conventional attention module solely focuses on the most salient regions in the entire image while ignoring other less salient but beneficial features. Incorporating the PHAM and considering the dependencies between neighboring parts enables the NDMF-Net to extract more detailed and richer features, such as the backpack.

global features and local features. Global features are easy to recognize and represent, but they only focus on surface clues of the image and ignore specific information from local areas. This makes them unsuitable for identifying complex samples with background interference or partial occlusion [11]. On the other hand, local features provide a detailed account of specific regions in an image, encompassing richer and more refined features that can serve as a supplement to the limitations of global features. The current extensively applied approach for enhancing the performance and robustness of person Re-ID involves the joint extraction of global and local features to form the final representation features [12]. Typically, local features are extracted by using two methods: (1) utilizing attention mechanisms for obtaining information from prominent regions of the entire image as local features [13], [14], or (2) segmenting the image into different parts and using the global features of each part as local features [15], [16]. These two methods correspond to the attention-based and part-based branches of person Re-ID, respectively.

While both methods can extract richer features to some extent, they also possess certain drawbacks. (1) Attention-based methods typically only focus on the most prominent channels or regions, leading to the omission of other key information that may not be apparent at the overall pedestrian level but can contribute to identification. (2) Part-based methods not only fail to fully consider the relationships between different body parts, but also introduce background interference in parts when using global max pooling and global average pooling to extract the global features of parts as local features of the image. These factors may lead to the omission of some useful features and reduce the expressiveness of the extracted features. Therefore, to obtain more comprehensive discriminative features, it is necessary to fully consider the spatial dependencies between different parts of pedestrians while minimizing the interference from background information. Fig. 1 provides some intuitive representations of this concept.

To achieve this objective, we propose an NDMF-Net, which consists of a backbone network, two part-level hybrid attention modules (PHAM), and two neighboring-part dependency exploration modules (NDEM). The PHAM includes a part-level channel attention module and a spatial attention module. The network considers the interdependencies between neighboring parts of the pedestrian and employs a specially designed attention module to focus on salient channels and regions of different parts of the image. Specifically, the network first horizontally segments the pedestrian features extracted from the middle layer of the backbone network into several parts, then uses the PHAM to focus on salient regions in the feature maps to extract discriminative fine-grained features. Following that, the NDEM is employed to investigate the spatial dependencies between neighboring parts of the features. Finally, various features are fused into the final representation and sent to the classification layer for classification.

The main contributions of this work are as follows:

- We propose an end-to-end NDMF-Net, which can extract dependencies between neighboring parts in pedestrian images to improve the model's feature representation and recognition ability.
- We design a PHAM and an NDEM. The PHAM can guide the network to focus on salient regions at different positions in pedestrian images and extract corresponding global and local features. The NDEM can extract spatial dependencies between neighboring parts, enriching the semantic continuity of features.
- We evaluate our method on three mainstream datasets, i.e., Market-1501, DukeMTMC-ReID, and MSMT17, and the results show that our method achieves state-of-the-art performance.

This paper is organized as follows. Section II introduces related work. Section III provides a detailed description of the proposed NDMF-Net. Section IV presents the details and results of the experiments. Finally, Section V concludes the paper and discusses prospects.

II. RELATED WORK

In recent years, with the advancement of deep learning, researchers have proposed various deep learning-based methods to address the problem of person Re-ID and achieved remarkable performance. In the following sections, we will briefly introduce the related works on person Re-ID.

A. PART-BASED Re-ID

Current CNN-based person Re-ID models have achieved impressive results, especially when part-based feature extraction techniques are employed. These models have shown notable enhancements in feature representation and recognition, which have been verified through various studies [14], [17], [18], [19], [20]. To learn discriminative part features, researchers have proposed various part-based CNNs. The most common approach is to divide the input features into

several parts according to certain rules and then learn the features of each part separately. Reference [17] first proposed a strong part-based convolutional baseline (PCB), which divides the image vertically into six parts and then extracts the features of each part by global averaging pooling, resulting in a significant performance boost. Part misalignment is the problem that comes with part-based CNNs. To address this issue, [18] proposed a dynamic alignment method that uses dynamic programming to find the shortest possible path between local features for conducting alignment and matching tasks, leading to a significant improvement in performance. To extract individual part features, the paper utilizes a process of horizontal partitioning the features, which is then followed by a procedure of globally pooling each part in the horizontal direction. Reference [19] proposed a method for spatial feature reconstruction with the goal of achieving an alignment-free system. To overcome the obstacles posed by the local misalignment of individuals, a multi-scale block representation is proposed, where features at different scales are obtained via averaging pooling. Reference [20] proposed a multi-branch network to enhance feature representation by fusing global and local features. Local feature extraction is performed by horizontally segmenting input features into multiple strips, and each strip is then considered a local feature after implementing a global maximum pooling operation.

The aforementioned studies indicate that current part-based person Re-ID methods primarily enhance feature richness by jointly learning features of the whole body and body parts, thereby improving the performance of the network. However, the relationships between different body parts are not taken into full consideration, leading to a loss in feature expressiveness. To address this issue, an NDEM is constructed to extract the dependencies between neighboring parts, as depicted in Fig. 1(c). This internal dependency attribute based on single pedestrians is largely immune to misalignment and can further enrich the representation of pedestrian features.

B. ATTENTION-BASED METHODS

Attention mechanisms can adaptively find salient regions in complex scenes by simulating the human visual system. Many scholars have introduced attention mechanisms into person Re-ID and achieved remarkable results, making attention-based person Re-ID an important branch of this field. In recent times, researchers have proposed various attention models and enhanced the feature representation by collectively using multiple attention mechanisms to extract features at different levels.

Hu et al. [21] introduced the concept of channel attention and presented the SE-Net. Unlike conventional methods that completely pass on weights to the subsequent layer, the SE-Net establishes relationships among different channels and readjusts their weights based on their corresponding inter-correlations, ultimately exhibiting robust generalization abilities. Subsequently, this work was implemented in [22] and achieved remarkable performance. Chen proposed a High-Order Attention (HOA) [23], which models and utilizes complicated higher-order statistical information to differentiate pedestrian images. Cai [24] proposed a multi-scale body part mask-guided attention network (MMGA), that employs body part masks to steer corresponding attention training. However, how to extract the exact mask is a problem. Reference [25] designed a global pooling feature-pooling module to collect features from foreground, background, and spatial attention maps to generate a global descriptor. To extract more comprehensive local features, [26], [27] proposed a novel local attention model that utilizes the Squeeze-and-Excitation module to choose the most prominent channel features of the entire image as body part features. However, this approach may disregard features that are not globally prominent yet helpful for identification purposes.

The abovementioned attention-based methods typically focus on channels or regions that contain the most salient features. Nevertheless, the importance of these channels or regions may differ across different parts of the body, potentially leading to the omission of other essential information that may not be obvious at a general level but is helpful for identification. Additionally, segmenting the image into several distinct parts and treating the global features of each part as local features cannot prevent the interference of background information in these parts [15], [16]. To this end, we design a part-level hybrid attention module that adaptively selects important channels and regions for different body parts to extract as many features as possible, avoiding the omission of valuable features that may not be salient at the overall pedestrian level. Furthermore, all the features are filtered thoroughly at an overall level to retain the truly effective features, while minimizing the interference of background information in each body part. this is illustrated in Fig. 1(d).

III. PROPOSED METHOD

The proposed NDMF-Net aims to establish the internal dependencies of neighboring body parts by employing location information in a unified person Re-ID framework. The overall structure of the NDMF-Net, as shown in Fig. 2, consists of a backbone network, two CHAMs, and two NDEMs. The PHAM consists of a part-level channel attention module (PCAM) and a spatial attention module (SAM). In this paper, we use ResNet-50 as the backbone network, pre-trained on ImageNet, for extracting global features from pedestrian images. The PHAMs, embedded after Layer 1 and Layer 3, effectively suppress irrelevant noise in pedestrian images and enhance the expression of salient regions. The NDEMs are used to leverage the correlation of human spatial structure and extract dependencies between neighboring parts. Finally, the features extracted by different modules are fused to form the final discriminative features. A detailed introduction of the individual modules is provided in the following subsections.

A. BACKBONE

In this paper, we adopt ResNet-50, which is widely used in person Re-ID tasks and pre-trained on ImageNet, as the

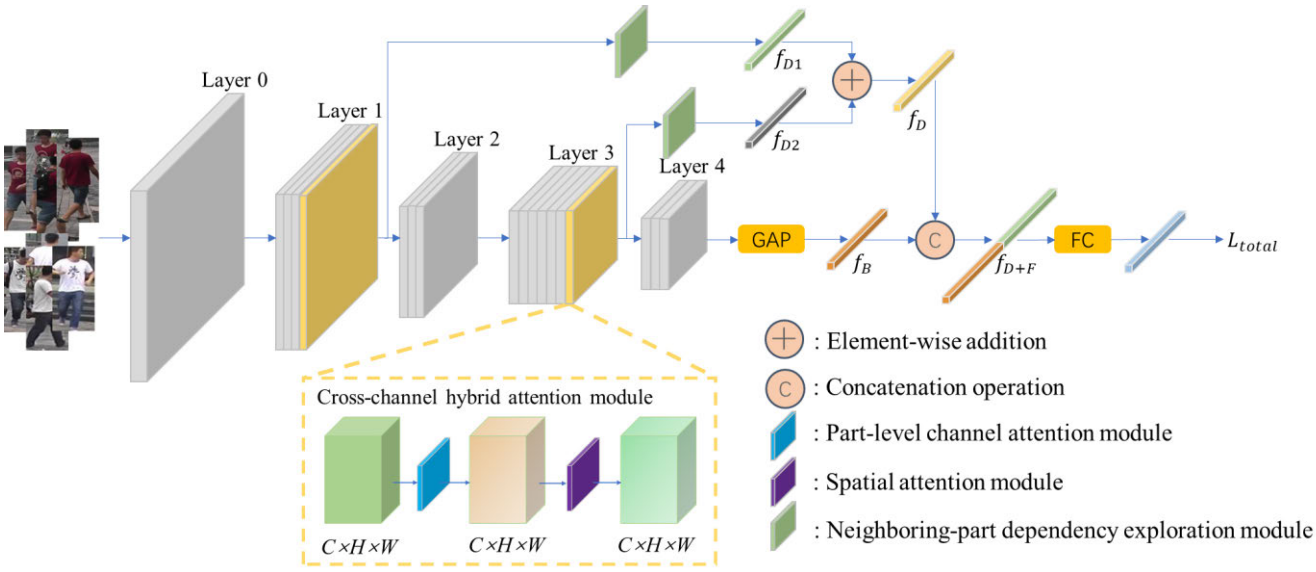


FIGURE 2. The overall structure of the NDMF-Net. The network consists of a ResNet-50 backbone network, with a CHAM and an NDEM embedded after Layer 1 and Layer 3. The CHAM consists of a part-level channel attention module and a spatial attention module. In addition, GAP stands for global average pooling, and FC stands for fully connected operation.

backbone network. ResNet-50 is a deep residual network proposed by He [28], which can effectively solve the problem of gradient disappearance in deep networks.

In the specific application, we made some small changes to the ResNet-50. Firstly, we set the stride of the *conv*2 layer and the *downsample* layer in Layer 4 to 1. Secondly, during the training process, we set the *out_feature* of the *fc* layer to the number of pedestrian IDs corresponding to different datasets.

B. PART-LEVEL HYBRID ATTENTION MODULE

The PHAM architecture comprises two modules: a part-level channel attention module (PCAM) and a spatial attention module (SAM). The PHAM aims to emphasize significant channels and regions in the extracted features. To achieve this, the PCAM uses a group of channel attention branches that selectively detect key channels for each part in an adaptive manner. On the other hand, the ASM globally adjusts and weights the features produced by the PCAM to extract genuinely useful features, while simultaneously reducing the influence of extraneous information from the surrounding background areas.

1) PART-LEVEL CHANNEL ATTENTION MODULE

The channel attention mechanism in a neural network is an additional network capable of explicitly modeling the correlation between different channels. This mechanism dynamically adjusts the weights of different channels via learning, increasing the weight of crucial channels and decreasing the weight of unimportant ones. This approach enhances significant features and suppresses unimportant ones. In the person Re-ID task, considering that features within the same channel are of different importance to different body parts of the pedestrian, we design a PCAM. The overall architecture of the PCAM is presented in Fig. 3.

As shown in Fig. 3, the proposed PCAM consists of a group of channel attention branches, each of which consists of two pooling layers, namely Global Average Pooling (GAP) and Global Max Pooling (GMP), a Fully Connected (FC) layer, and a Sigmoid function. For a given feature map $X \in \mathbb{R}^{C\times H\times W}$, where C is the channel numbers, H is the height, and W is the width, divide X along the height dimension into N parts and each part can be expressed as $X_i \in \mathbb{R}^{C\times H/N\times W}$ $(i = 1, 2, \ldots, N)$. Feed X_i into the corresponding channel attention branch, the generated branch channel attention map can be expressed as:

$$\boldsymbol{A_i^C = \sigma\left(W_i cat\left(X_i^{gap}, X_i^{gmp}\right)\right)} \tag{1}$$

where A_i^C is the branch channel attention map of X_i, σ denotes the sigmoid function, $W_i \in \mathbb{R}^{C\times 2C}$ denotes the parameters of the fully connected layer, *cat* $(\odot)$ denotes the concatenation operation along the channel dimension, X_i^{gap} and X_i^{gmp} denote the feature map generated by the GAP and the GMP.

After getting the branch channel attention map $A_i^C \in \mathbb{R}^{C\times 1\times 1}$, expand it to $\tilde{A}_i^C \in \mathbb{R}^{C\times H/N\times W}$ by broadcasting operation, and then embed $\tilde{A}_i^C$ into X_i via element-wise multiplication, a new feature map X_i^C can be obtained:

$$X_i^C = X_i \otimes \tilde{A}_i^C \tag{2}$$

where X_i^C is the output feature map of X_i, $\otimes$ denotes the element-wise multiplication.

Concatenate all new feature maps to obtain the final output feature map of the PCAM, which can be expressed as:

$$X_C = cat\left(X_i^C\right) \tag{3}$$

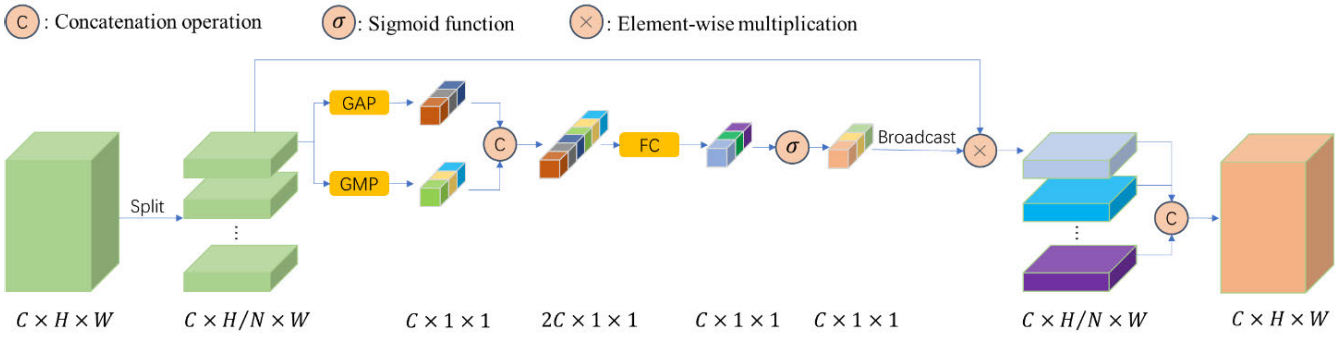


FIGURE 3. The architecture of the PCAM. The input feature map is horizontally divided into *N* parts, with each part being processed by an independent channel attention branch, thereby producing the corresponding new feature maps. These new feature maps are subsequently combined through concatenation to form the final feature map. GAP denotes global average pooling, GMP denotes global max pooling, and FC denotes fully connected operation.

where $X_C \in \mathbb{R}^{C\times H\times W}$ denotes the part-level channel attention map of X, $cat\ (\odot)$ denotes the concatenation operation along the height dimension, $X_i^C \in \mathbb{R}^{C\times H\times W}$.

2) SPATIAL ATTENTION MODULE

The spatial attention mechanism works by converting the original image into a new feature space and creating a weight mask for all locations in an adaptive manner. The weight mask is employed to modulate the output, resulting in an augmented representation of the target region of interest while concurrently attenuating the feature representation of irrelevant background regions.

In this paper, we design a SAM, the overall structure of which is shown in Fig. 4. For a given feature map $X \in \mathbb{R}^{C\times H\times W}$, where C is the channel number, H is the height, and W is the width, two spatial descriptors are generated after the cross-channel average pooling (CAP) and cross-channel max pooling (CMP) layers. These descriptors are subsequently concatenated to generate a new feature map. This new feature map is then processed using a convolutional layer and a sigmoid function to generate a more efficient descriptor $F \in \mathbb{R}^{1\times H\times W}$. The spatial attention map can be expressed as below:

$$A^S = \sigma\left(\varphi\left(cat\left(X^{cap}, X^{cmp}\right)\right)\right) \tag{4}$$

where $A^S \in \mathbb{R}^{1\times H\times W}$ denotes the spatial attention map generated by the SAM, $\sigma\ (\odot)$ denotes the Sigmoid function, $\varphi\ (\odot)$ denotes the convolution operation with a kernel size of 3×3, $cat\ (\odot)$ denotes the concatenation operation along the channel dimension, X^{cap} and X^{cmp} denote the spatial descriptors obtained by CAP and CMP.

After getting the spatial attention map $A^S \in \mathbb{R}^{1\times H\times W}$, expand it to $\tilde{A}^S \in \mathbb{R}^{C\times H\times W}$ by broadcasting operation, and then embed $\tilde{A}^S$ into X via element-wise multiplication. The output feature map of the SAM can be expressed as:

$$X_S = X \otimes \tilde{A}^S \tag{5}$$

where $X_S \in \mathbb{R}^{C\times H\times W}$ denotes the output feature map of the SAM, X denotes the input feature map, $\otimes$ denotes the element-wise multiplication, and $\tilde{A}^S$ denotes the expanded spatial attention map.

3) PART-LEVEL HYBRID ATTENTION MODULE

The PHAM is constructed by placing the PCAM and SAM in a sequential manner [29], as shown in Fig. 2. Given an input feature map $X \in \mathbb{R}^{C\times H\times W}$, where C is the channel numbers, H is the height, W is the width, the feature generated after passing through the CHAM can be expressed as:

$$X_{CHAM} = Att_S\left(Att_C\left(X\right)\right) \tag{6}$$

where $X_{CHAM} \in \mathbb{R}^{C\times H\times W}$ denotes the feature map generated by the PHAM, $Att_S\ (\odot)$ denotes the SAM, and $Att_C\ (\odot)$ denotes the PCAM.

C. NEIGHBORING-PART DEPENDENCY EXPLORATION MODULE

The Long Short-Term Memory (LSTM) network is a widely used variant of traditional recurrent neural networks developed by Hochreiter [30], capable of learning long-term dependencies in sequence modeling tasks. Unlike traditional RNNs, LSTM networks employ two modules to facilitate the learning of long-term dependencies, namely the memory module and the gate module. The gate module consists of an input gate, an output gate, and a forget gate. These gates enable the LSTM network to selectively add or remove information from the cell state and effectively capture long-term dependencies in sequence modeling tasks. The structure of the LSTM unit is illustrated in Fig. 5, and the relationship between the input and output is demonstrated as follows:

$$\begin{cases} f_t = \sigma\left(W_f\left[h_{t-1}, x_t\right] + b_f\right) \\ i_t = \sigma\left(W_i\left[h_{t-1}, x_t\right] + b_i\right) \\ \tilde{C}_t = tanh\left(W_C\left[h_{t-1}, x_t\right] + b_C\right) \\ C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \\ o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right) \\ h_t = o_t * tanh\left(C_t\right) \end{cases} \tag{7}$$

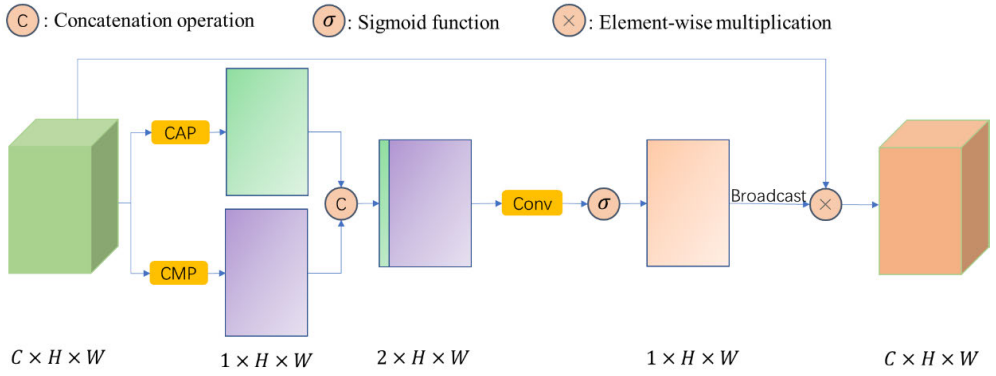


FIGURE 4. The architecture of SAM. CAP stands for cross-channel average pooling, CMP stands for cross-channel max pooling, and Conv stands for the convolution operation with a kernel size of 3 × 3.

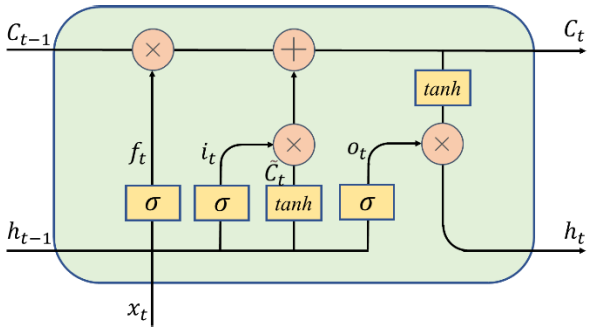


FIGURE 5. The structure of the LSTM unit.

where f_t, i_t, and o_t represent the input gate, the forget gate, and the output gate, respectively. h_t denotes the hidden state of the t step, x_t denotes the input of the t step, C_t denotes the cell state of the t step, $\tilde{C}_t$ is a vector of new candidate values, σ denotes the sigmoid function, W and b indicate the corresponding weights and biases, respectively.

The spatial information presented in pedestrian images includes not only rich global and local features but also interdependencies between neighboring parts. To effectively explore and establish these dependencies, an LSTM-based neighboring-part dependency exploration module (NDEM) is designed, as illustrated in Fig. 6, to uncover meaningful features in the image. For a given feature map $X \in \mathbb{R}^{C\times H\times W}$, where C is the channel number, H is the height, and W is the width. The corresponding feature vector $f_D \in \mathbb{R}^{1\times m}$ generated by the NDEM can be formulated as follows:

$$f_D = \sigma\left(Wcat\left(LSTM\left(\phi\left(X\right)\right)\right)\right) \tag{8}$$

where $\sigma(\odot)$ denotes the Sigmoid function, $W \in \mathbb{R}^{m\times 2NR}$ denotes the parameters of the FC layer, $cat(\odot)$ denotes the concatenation operation along the first dimension, $LSTM(\odot)$ denotes the LSTM network with two hidden layers, and the hidden size is R, $\phi(\odot)$ denotes the matrix of feature map processing operations that includes reshaping, permutation, and flattening operations. Through these matrix operations, we obtain features that incorporate rich spatial dependencies.

D. FINAL REPRESENTATION

As shown in Fig. 2, two PHAMs and two NDEMs are collectively embedded behind Layer 1 and Layer 3 of the backbone network. While the PHAMs guide the backbone network to generate the feature vector f_B, the NDEMs further generate the corresponding feature vectors f_{D1} and f_{D2} from the feature maps produced by each PHAM. Finally, f_{D1} and f_{D2} generate the feature vector f_D through element-wise addition, which can be written as:

$$f_D = f_{D1} \oplus f_{D2} \tag{9}$$

Then fuse the feature vector f_B and f_D to generate the final feature representation, which can be expressed as:

$$f_F = W \cdot cat\left([f_B, f_D]\right) \tag{10}$$

where W denotes the parameters of the FC layer, and $cat(\cdot)$ denotes the concatenation operation along the first dimension.

E. LOSS FUNCTION

To enhance feature distinctiveness, we utilize a joint loss framework to optimize the model parameters. The loss function comprises two primary components: cross-entropy loss and triple loss [31].

1) CROSS-ENTROPY LOSS

The cross-entropy loss function is capable of learning inter-class information and is commonly employed in image classification tasks. Additionally, to enhance the model's generalization ability, we adopt a cross-entropy loss with label smoothing, which can be mathematically represented as:

$$L_{ce} = -\sum_{i=1}^{N_0} q_n^k \cdot \log\left(p_n^k\right) \tag{11}$$

where N_0 represents the number of pedestrian images in a mini-batch, q_n^k represents the true label distribution of the k-th ID with label smoothing, which can be expressed as

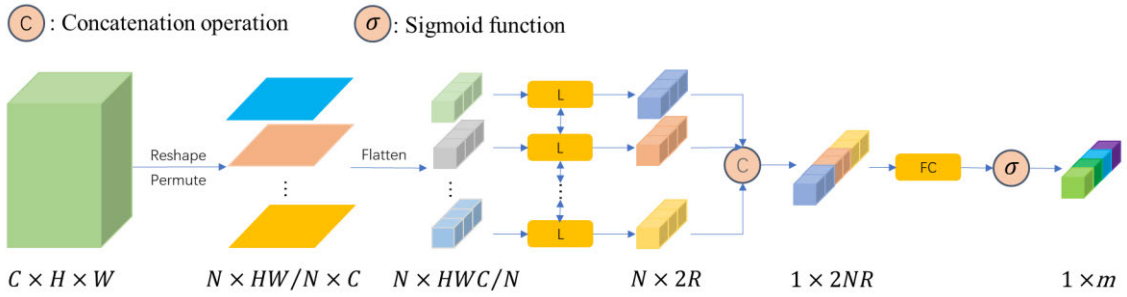


FIGURE 6. The architecture of NDEM. The input feature map is first split horizontally into *N* parts, then processed and fed into the LMST network, and finally obtains the feature vector. *L* stands for LSTM network, and FC stands for fully connected operation.

in Equation (12), and p_n^k represents the predicted probability distribution of the k-th ID.

$$q_n^k = \begin{cases} 1-\varepsilon, & y^n = k \\ \dfrac{\varepsilon}{K-1}, & y^n \neq k \end{cases} \tag{12}$$

where K represents the total number of pedestrian IDs, y^n represents the truth label of the n-th pedestrian image, and ε is the hyperparameter used for label smoothing.

2) TRIPLET LOSS

Triplet loss is a commonly used metric learning function that effectively minimizes the distance between positive sample pairs and maximizes the distance between negative sample pairs. To enhance the model's capability to identify pedestrians with similar appearances, we adopt a batch-hard triplet loss to optimize the model parameters, which can be mathematically expressed as:

$$L_{tri} = \frac{1}{N_0}\sum_{i=1}^{N_0}\left[\left\|f\left(x_i^a\right) - f\left(x_i^p\right)\right\|^2 - \left\|f\left(x_i^a\right) - f\left(x_i^n\right)\right\|^2 + m\right]_+ \tag{13}$$

where $f\left(x_i^a\right)$, $f\left(x_i^p\right)$, and $f\left(x_i^n\right)$ denote the feature map of the anchor, positive, and negative samples, respectively. $\|\cdot\|^2$ represents the distance between feature maps. m represents the hyperparameter that controls the inter-class and intra-class distance.

The final loss can be expressed as:

$$L_{total} = \alpha L_{ce} + \beta L_{tri} \tag{14}$$

where L_{total} denotes the final loss, α and β denote the loss balance parameters.

IV. EXPERIMENT AND ANALYSIS

To comprehensively evaluate the performance of the proposed NDMF-Net, a series of experiments were conducted. Firstly, the performance of NDMF-Net was tested on three mainstream datasets, Market-1501 [32], DukeMTMC-ReID [2], and MSMT17 [33], and the experimental results were compared with some current state-of-the-art person Re-ID methods. Secondly, ablation experiments were conducted on the Market-1501 dataset to evaluate the impact of each module in NDMF-Net. Besides, we also analyzed the effect of different split numbers, N, on the performance of NDMF-Net on the Market-1501 dataset.

A. EXPERIMENT

1) DATASETS

The Market-1501 dataset, gathered at Tsinghua University, includes 32,668 images of 1,501 pedestrians, captured by 5 high-resolution cameras and 1 low-resolution camera. The DukeMTMC-ReID dataset, collected at Duke University and released in 2017, encompasses 36,411 images of 1,812 pedestrians, captured by 8 cameras. The MSMT17 dataset, released in 2018, is a multi-scenario and multi-time dataset and has 126,441 images of 4,101 pedestrians, captured by 12 outdoor cameras and 3 indoor cameras. Detailed information can be found in Table 1.

2) EVALUATION METRIC

In this paper, we employ the cumulative matching characteristics (CMC) at Rank-1 and the mean average precision (mAP) as evaluation metrics for the model, without using any re-ranking technique during the evaluation process. The CMC at Rank-1 can be expressed as:

$$Rank - 1 = \frac{\sum_{i=1}^{m} x_i}{m} \tag{15}$$

where m is the total number of images, and x_i is an indicator variable. If the i-th probe and the top-ranked image in similarity ranking are the same, $x_i = 1$; otherwise, $x_i = 0$.

The mAP can be expressed as:

$$mAP = \frac{\sum_{i=1}^{k} AP_i}{k} \tag{16}$$

where k is the number of classes, AP_i is the average precision for class i, which is calculated by plotting a curve of precision versus recall and computing the area under the curve.

3) EXPERIMENTAL SETTINGS

During the model training process, we randomly select a batch consisting of P ($P = 16$) pedestrians, each with K

TABLE 1. Datasets details induction.

Dataset	Released	Training Set		Query Set		Gallery Set		Cameras
		Images	IDs	Images	IDs	Images	IDs	
Market-1501	2015	12,936	751	3368	750	16,364	750	6
DukeMTMC-reID	2017	16,522	702	2228	702	17,661	702	8
MSMT17	2018	32,621	4101	11659	3060	82,161	3060	15

($K = 4$) images, and pre-process all images before feeding them into the network. Firstly, we resize all images to 256×128 and then fill them with a value of 0 for 10 pixels around them. Subsequently, the resized images are cropped randomly and resized back to 256×128, followed by normalization. To augment the sample, we applied random horizontal flipping and random erasing operations. The Adam optimizer is used to optimize the model parameters, with the weight decay factor being set to 5e-4 and the momentum to 0.9, for a total of 200 epochs. The warm-up learning strategy is used to update the learning rate:

$$lr\,(ep)\begin{cases} 3.5\times10^{-5}\times\dfrac{e}{10}, & ep\le 10\\ 3.5\times10^{-4}, & 10<ep\le 40\\ 3.5\times10^{-5}, & 40<ep\le 70\\ 3.5\times10^{-6}, & 70<ep\le 200 \end{cases} \tag{17}$$

where ep is the current epoch.

Beyond that, we set the split part number N to 8, and set the loss balance parameters α and β to 0.3 and 0.7, respectively.

B. EXPERIMENTAL RESULTS AND ANALYSIS

To evaluate the performance of the NDMF-Net, we conduct experiments on three mainstream datasets and compare the results with several state-of-the-art person Re-ID methods. Among these methods, [34], [35], [46] use a re-ranking approach to improve the accuracy of the person Re-ID accuracy. Reference [36] investigate feature correlation by decomposing feature maps into multiple subspaces. To achieve a more generalized feature representation, [13], [41], [43], [44] propose a multi-scale feature fusion mechanism. References [7], [40], and [42] incorporate attribute information into the network to produce more discriminative feature representation. References [16], [19], [37], [38], and [39] improve person Re-ID accuracy by more effectively capturing local information within the image. Reference [45] propose a lightweight network that combines deep space and achieves impressive performance.

Table 2 to 4 shows the Rank-1 accuracy and mAP of our method and other state-of-the-art methods on three datasets, namely Market-1501, DukeMTMC-ReID, and MSMT17. The experimental results presented in Table 2 indicate that our method achieves 95.8% Rank-1 accuracy and 88.9% mAP on the Market-1501 dataset, demonstrating a superior performance compared with the second-best method with an increase of 0.7% in Rank-1 accuracy and 2.1% in mAP. Similarly, Table 3 presents the experimental results on the DukeMTMC-ReID dataset, where our method achieves 89.7% Rank-1 accuracy and 79.8% mAP. Compared with the second-best Rank-1 accuracy and mAP, achieved by the OSNet [43] and the RANGEv2 [45], respectively, our method outperforms them by 1.0% in Rank-1 accuracy and 1.6% in mAP. Furthermore, Table 4 shows the experimental results on the MSMT17 dataset, which show that our method achieves 78.4% Rank-1 accuracy and 54.5% mAP. The CDNet [44] achieves the best Rank-1 accuracy and mAP, with 78.0% and 54.7%, respectively, outperforming our method by 0.5% and 0.2%. Nonetheless, our method still outperforms other state-of-the-art methods and achieves the second-best performance.

TABLE 2. Comparison of NDMF-Net with the state-of-the-art methods on Market-1501. The best results are shown in red bold, and the second-best in blue bold.

Method	Rank-1 (%)	mAP (%)
Reranking [34] [CVPR'2017]	81.4	70.4
DaF [35] [BMVC'2017]	82.3	72.4
MLFN [36] [CVPR'2018]	90.0	74.3
HA-CNN [13] [CVPR'2018]	91.2	75.7
PCB [19] [ECCV'2018]	92.3	77.4
PCB+RPP [19] [ECCV'2018]	93.8	81.6
APR [7] [PR'2019]	87.0	66.9
PLNet [37] [TIP'2019]	88.2	69.3
AlignedReID++ [38] [PR'2019]	91.8	79.1
VPM64 [16] [CVPR'2019]	93.0	80.8
SFT [39] [ICCV'2019]	93.4	82.7
Deep-Person [40] [PR'2020]	92.3	79.6
CtF [41] [ECCV'2020]	93.7	84.9
APDR [42] [PR'2020]	93.1	80.1
HAN [43] [IJCV'2020]	91.6	76.7
OSNet [44] [TPAMI'2021]	94.8	86.7
CDNet [45] [CVPR 2021]	**95.1**	86.0
RANGEv2 [46] [PR'2022]	94.7	**86.8**
NDMF-Net (ours)	**95.8**	**88.9**

Compared with other state-of-the-art methods, the NDMF-Net focuses on salient regions of all parts of pedestrian images, avoiding the loss of detailed information caused by focusing only on globally salient regions and enriching the fine-grained features. Additionally, by exploring the dependencies between neighboring parts of pedestrian images, the semantic continuity of features is improved. Thanks to the above two aspects, the NDMF-Net achieves

TABLE 3. Comparison of NDMF-Net with the state-of-the-art methods on DukeMTMC-ReID. The best results are shown in red bold, and the second-best in blue bold.

Method	Rank-1 (%)	mAP (%)
MLFN [36] [CVPR'2018]	81.0	62.8
HA-CNN [13] [CVPR'2018]	80.5	63.8
PCB [19] [ECCV'2018]	81.8	66.1
PCB+RPP [19] [ECCV'2018]	83.3	69.2
APR [7] [PR'2019]	73.9	55.6
AlignedReID++ [38] [PR'2019]	82.1	69.7
VPM64 [16] [CVPR'2019]	83.6	72.6
SFT [39] [ICCV'2019]	86.9	73.2
Deep-Person [40] [PR'2020]	80.9	64.8
CtF [41] [ECCV'2020]	87.6	74.8
APDR [42] [PR'2020]	84.3	69.7
HAN [43] [IJCV'2020]	80.6	64.1
OSNet [44] [TPAMI'2021]	**88.7**	76.6
CDNet [45] [CVPR 2021]	88.6	76.8
RANGEv2 [46] [PR'2022]	87.0	**78.2**
NDMF-Net (ours)	**89.7**	**79.8**

TABLE 4. Comparison of NDMF-Net with the state-of-the-art methods on MSMT17. The best results are shown in red bold, and the second-best in blue bold.

Method	Rank-1 (%)	mAP (%)
MLFN [36] [CVPR'2018]	66.3	37.0
PCB [19] [ECCV'2018]	68.2	40.4
AlignedReID++ [38] [PR'2019]	69.8	43.7
SFT [39] [ICCV'2019]	73.6	47.6
HAN [43] [IJCV'2019]	66.2	46.2
OSNet [44] [TPAMI'2021]	78.7	52.9
CDNet [45] [CVPR 2021]	**78.9**	**54.7**
RANGEv2 [46] [PR'2022]	76.4	51.3
NDMF-Net (ours)	**78.4**	**54.5**

the best or second-best performance on the three mainstream datasets, Market-1501, DukeMTMC-ReID, and MSMT17, which demonstrates the effectiveness of our method for person Re-ID.

C. ABLATION STUDY AND HYPERPARAMETERS ANALYSIS

1) ABLATION STUDY

To verify the effectiveness of each component of the NDMF-Net, we conduct ablation experiments on the Market-1501 dataset. Specifically, we first use the backbone network as the baseline, and then add different components to the baseline one by one. Finally, the improvements in model performance are used to demonstrate the effect of each component. The results of the ablation experiments are shown in Table 5.

As can be seen from Table 5, the baseline achieves 92.1% Rank-1 accuracy and 84.6% mAP on the Market-1501 dataset without any additional components. The PCAM improves the model's Rank-1 accuracy and mAP to 94.2% and 87.1%,

TABLE 5. The ablation experimental results on the Market-1501 dataset.

Baseline	PCAM	SAM	NDEM	Rank-1 (%)	mAP (%)
√				92.1	84.6
√	√			94.2	87.1
√		√		93.9	86.9
√			√	94.8	87.6
√	√	√	√	95.8	88.9

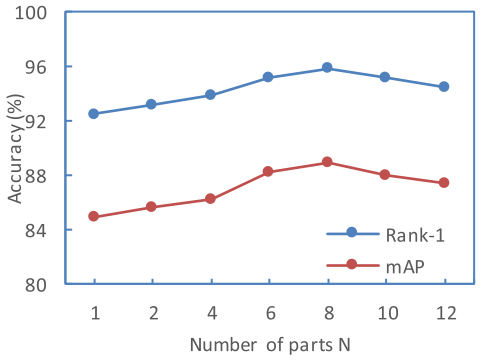


FIGURE 7. Experiment results of different *N* values on the Market-1501 dataset.

respectively, outperforming the baseline by 2.1% and 2.5%. Similarly, the SAM improves the model's Rank-1 accuracy and mAP to 93.9% and 86.9%, respectively, outperforming the baseline by 1.8% and 2.3 %. The NDEM improves the model's Rank-1 accuracy and mAP to 94.8% and 87.6%, respectively, outperforming the baseline by 2.7% and 3.0 %. By jointly utilizing the PCAM, the SAM, and the NDEM, the model achieves 95.8% Rank-1 accuracy and 88.9% mAP, outperforming the baseline by 3.7% and 4.3 %, respectively. The ablation experimental results demonstrate that our proposed components, the PCAM, the SAM, and the NDEM, are effective in facilitating the model to extract richer discriminative features, ultimately improving the performance of the model. This confirms the effectiveness of our proposed NDMF-Net.

2) HYPERPARAMETERS ANALYSIS

We conduct experiments on the Market-1501 dataset to investigate the effect of the number of parts N on model performance, and the results are shown in Fig. 7. The experimental results show that as N gradually increases from small to large, the model performance improves. When N is 8, the NDMF-Net performs the best, achieving 95.8% Rank-1 accuracy and 88.9% mAP. As N continues to increase, model performance decreases.

D. FEATURE MAPS VISUALIZATION

To visually demonstrate the effectiveness of the NDMF-Net, we extract feature maps from pedestrian images in the Market-1501 dataset and visualize them. Fig. 8 shows the visualization results, where Fig. 8(a) shows the original pedestrian images, Fig. 8(b) shows the feature map extracted

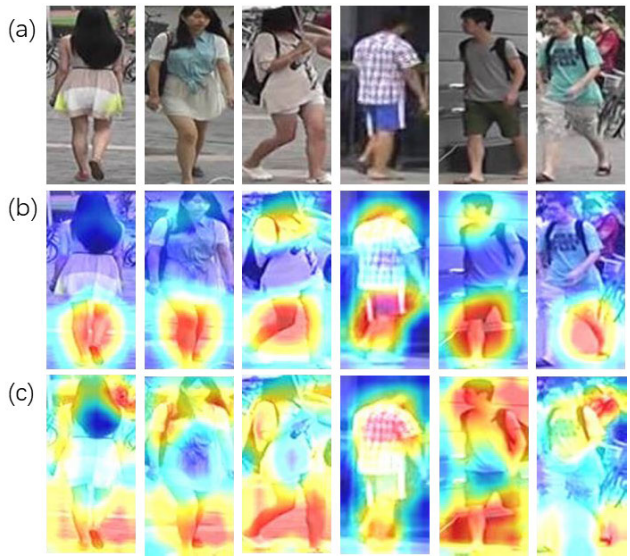


FIGURE 8. Feature maps visualization. (a) Original images. (b) Feature maps extracted by Baseline. (c) Feature maps extracted by NDMF-Net.

by the baseline, and Fig. 8(c) shows the feature map extracted by the NDMF-Net. The network's attention to the image is represented by different colors, with deep red indicating the most focused areas and deep blue indicating the least focused areas.

Fig. 8 illustrates that the Baseline selectively emphasizes the most salient regions of the image, disregarding other pertinent informative details, causing inadequate feature extraction. Contrarily, in comparison to the Baseline, the NDMF-Net generates salient features not only at the whole image level, but also at the part level by utilizing the CHAM to capture intricate details. Additionally, the NDEM directs the network to extract spatial dependency between neighboring parts of pedestrians, augmenting the semantic coherence of the features extracted. Consequently, the features extracted by our method are more exhaustive and representative.

E. VISUALIZATION OF RETRIEVAL RESULTS

We randomly select a certain number of images from the Market-1501 dataset query set and present their retrieval results in Fig. 9. The green rectangular box indicates a matching ID of the retrieved image with the probe, while the red rectangle indicates a mismatch. The retrieval results demonstrate that our method can retrieve the correct images in most cases.

F. FAILURE ANALYSIS

While our method generally produces correct identification results, we have noted that for pedestrian images with distinct IDs, the retrieval outcomes occasionally incorrectly associate them as the same person. This pattern is especially true when both the background and the pedestrians themselves are highly alike. The cause of this phenomenon is the extreme similarity in the dependency relationships among different

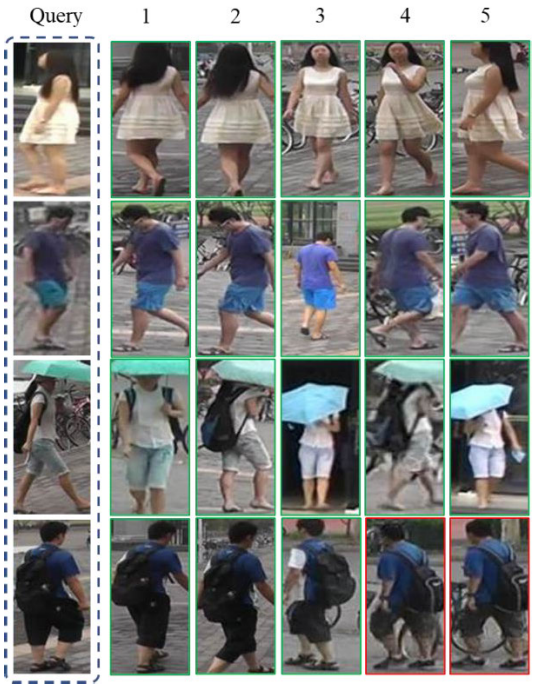


FIGURE 9. Visualization of retrieval results.

body parts of highly similar pedestrians, resulting in inaccurate identification outcomes. This pattern underlines the need for further investigation into methods for extracting distinct local features and dependency patterns in our approach, especially when dealing with highly similar individuals.

V. CONCLUSION

In this paper, we propose a novel NDMF-Net for person Re-ID, which consists of a backbone network, two PHAMs, and two NDEMs. The PHAM is composed of a PCAM and a SAM, which enables the network to focus on important features of different parts in pedestrian images, thus obtaining more discriminative fine-grained features. The NDEM can guide the network to extract dependencies between neighboring parts in pedestrian images, improving the semantic consistency of pedestrian features. Finally, by fusing different features, the final pedestrian representation is obtained. Thanks to these modules, the NDMF-Net can learn more comprehensive and discriminative features and ultimately improve the performance of person Re-ID. Experimental results have shown that our method can achieve better performance in person Re-ID tasks and outperform most existing methods.

Different parts of pedestrian images contain rich detailed information and have potential relationships. In the future, we will continue to investigate the information correlation of different parts and apply it to unsupervised person Re-ID tasks.

REFERENCES

[1] N. K. Singh, M. Khare, and H. B. Jethva, "A comprehensive survey on person re-identification approaches: Various aspects," *Multimedia Tools Appl.*, vol. 81, no. 11, pp. 15747–15791, Mar. 2022.

[2] 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. IEEE ECCV*, Nov. 2016, pp. 17–35.

[3] S. Zhang, L. Zhang, and A. G. Hauptmann, "Fuzzy least squares support vector machine with adaptive membership for object tracking," *IEEE Trans. Multimedia*, vol. 22, no. 8, pp. 1998–2011, Aug. 2020.

[4] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical Gaussian descriptor for person re-identification," in *Proc. IEEE CVPR*, Jun. 2016, pp. 1363–1372.

[5] G. Lisanti, I. Masi, A. D. Bagdanov, and A. D. Bimbo, "Person re-identification by iterative re-weighted sparse ranking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 8, pp. 1629–1642, Aug. 2015.

[6] R. Zhao, W. Oyang, and X. Wang, "Person re-identification by saliency learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 356–370, Feb. 2017.

[7] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang, "Improving person re-identification by attribute and identity learning," *Pattern Recognit.*, vol. 95, pp. 151–161, Nov. 2019.

[8] Z. Wang, J. Jiang, Y. Wu, M. Ye, X. Bai, and S. Satoh, "Learning sparse and identity-preserved hidden attributes for person re-identification," *IEEE Trans. Image Process.*, vol. 29, pp. 2013–2025, 2020.

[9] C. Wan, Y. Wu, X. Tian, J. Huang, and X.-S. Hua, "Concentrated local part discovery with fine-grained part representation for person re-identification," *IEEE Trans. Multimedia*, vol. 22, no. 6, pp. 1605–1618, Jun. 2020.

[10] H. J. Mohammed, S. Al-Fahdawi, A. S. Al-Waisy, D. A. Zebari, D. A. Ibrahim, M. A. Mohammed, S. Kadry, and J. Kim, "ReID-DeePNet: A hybrid deep learning system for person re-identification," *Mathematics*, vol. 10, no. 19, p. 3530, Sep. 2022.

[11] Z. Cao and H. J. Lee, "Learning multi-scale features and batch-normalized global features for person re-identification," *IEEE Access*, vol. 8, pp. 184644–184655, 2020.

[12] 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.

[13] L. Chen, H. Yang, Q. Xu, and Z. Gao, "Harmonious attention network for person re-identification via complementarity between groups and individuals," *Neurocomputing*, vol. 453, pp. 766–776, Sep. 2021.

[14] H. Liu, J. Cheng, S. Wang, and W. Wang, "Attention: A big surprise for cross-domain person re-identification," 2019, *arXiv:1905.12830*.

[15] Y. Hou, S. Lian, H. Hu, and D. Chen, "Part-relation-aware feature fusion network for person re-identification," *IEEE Signal Process. Lett.*, vol. 28, pp. 743–747, 2021.

[16] 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 CVPR*, Jun. 2019, pp. 393–402.

[17] 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. IEEE ECCV*, Sep. 2017, pp. 480–496.

[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*.

[19] L. He, J. Liang, H. Li, and Z. Sun, "Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach," in *Proc. IEEE/CVF CVPR*, Jun. 2018, pp. 7073–7082.

[20] 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 MM*, 2018, pp. 274–282.

[21] 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.

[22] T. Guo, D. Wang, Z. Jiang, A. Men, and Y. Zhou, "Deep network with spatial and channel attention for person re-identification," in *Proc. IEEE VCIP*, Dec. 2018, pp. 1–4.

[23] B. Chen, W. Deng, and J. Hu, "Mixed high-order attention network for person re-identification," in *Proc. IEEE/CVF ICCV*, Oct. 2019, pp. 371–381.

[24] H. Cai, Z. Wang, and J. Cheng, "Multi-scale body-part mask guided attention for person re-identification," in *Proc. IEEE/CVF CVPRW*, Jun. 2019, pp. 1555–1564.

[25] J. S. J. Rani and M. G. Augasta, "PoolNet deep feature based person re-identification," *Multimedia Tools Appl.*, vol. 82, pp. 1–23, Jan. 2023, doi: 10.1007/s11042-023-14364-7.

[26] 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.

[27] Y. Chen, H. Wang, X. Xun, 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.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, Jun. 2016, pp. 770–778.

[29] S. Woo, J. Park, J. Lee, and I. Kweon, "CBAM: Convolutional block attention module," in *Proc. ECCV*, 2018, pp. 3–19.

[30] S. Hochreiter and J. J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[31] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE CVPR*, Jun. 2015, pp. 815–823.

[32] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE ICCV*, Dec. 2015, pp. 1116–1124.

[33] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE CVPR*, Jun. 2014, pp. 152–159.

[34] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with *k*-reciprocal encoding," in *Proc. IEEE CVPR*, Jul. 2017, pp. 3652–3661.

[35] R. Yu, Z. Zhou, S. Bai, and X. Bai, "Divide and fuse: A re-ranking approach for person re-identification," in *Proc. BMVC*, 2017, pp. 1–13.

[36] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in *Proc. IEEE CVPR*, Jun. 2018, pp. 2109–2118.

[37] 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.

[38] 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.

[39] C. Luo, Y. Chen, N. Wang, and Z.-X. Zhang, "Spectral feature transformation for person re-identification," in *Proc. IEEE/CVF ICCV*, Oct. 2019, pp. 4975–4984.

[40] 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.

[41] G. Wang, S. Gong, J. Cheng, and Z. Hou, "Faster person re-identification," in *Proc. ECCV*, 2020, pp. 275–292.

[42] S. Li, H. Yu, and R. Hu, "Attributes-aided part detection and refinement for person re-identification," *Pattern Recognit.*, vol. 97, Jan. 2020, Art. no. 107016.

[43] W. Li, X. Zhu, and S. Gong, "Scalable person re-identification by harmonious attention," *Int. J. Comput. Vis.*, vol. 128, no. 6, pp. 1635–1653, Jun. 2020.

[44] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Learning generalisable omni-scale representations for person re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5056–5069, Sep. 2022.

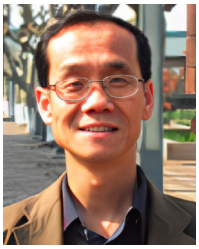
[45] H. Li, G. Wu, and W.-S. Zheng, "Combined depth space based architecture search for person re-identification," in *Proc. IEEE/CVF CVPR*, Jun. 2021, pp. 6725–6734.

[46] G. Wu, X. Zhu, and S. Gong, "Learning hybrid ranking representation for person re-identification," *Pattern Recognit.*, vol. 121, Jan. 2022, Art. no. 108239.

CHUAN ZHU received the B.Eng. degree from Harbin University, in 2015. He is currently pursuing the Ph.D. degree with Fudan University. His research interests include machine learning, computer vision, and natural language processing.

YINGJUN ZHU received the B.Eng. and B.BM. degrees, in 2020. He is currently pursuing the M.Sc. degree in general mechanics and mechanics with Fudan University. His research interests include speech emotion recognition and multimodal fusion.

WENJUN ZHOU received the B.Eng. degree in mechanical engineering from Donghua University, in 2018. She is currently pursuing the Ph.D. degree with Fudan University. Her research interests include machine learning and speaker recognition.



JIANMIN MA received the Ph.D. degree from Xi'an Jiaotong University, in 1998. He is currently a Professor with the Department of Aeronautics and Astronautics, Fudan University. His research interests include mechanical vibration and artificial intelligence.