A Multi-Dimensional Attention Feature Fusion Method for Pedestrian Re-identification

X. P. Chen, Y. Xu

Abstract—**Pedestrian re-identification aims to retrieve pedestrians across various cameras and scenes. However, the accuracy of re-identification is often affected by factors such as low-quality images of pedestrians and environmental conditions. Consequently, it is crucial for machine learning models to learn features from multiple dimensions. In response to these challenges, this paper proposes a Multi-Dimensional Attention Feature Fusion (MDAFF) method for pedestrian re-identification based on the NFormer approach. This method enables the model to learn and fuse pedestrian features from multiple dimensions, enriching the expressive power of the feature maps and improving the discrimination among pedestrians. By incorporating a PA module into the ResNeXt network for feature extraction, the model enhances its global perception and integrates pedestrian position information into the feature maps. This increases the model's sensitivity to pedestrian positions and reduces the impact of noise on re-identification accuracy. Furthermore, the method extracts channel and spatial correlations from the fused position feature maps and performs feature fusion, facilitating the fusion of multi-dimensional attention features. This alleviates the influence of varying scenarios and poses on re-identification, thereby enhancing the model's performance. Compared to the Res50+NFormer method, which directly models the relationships among different pedestrians after feature extraction, MDAFF integrates multi-dimensional features into the feature maps, improving the model's expressive power and capturing the relationships among different pedestrians more effectively. The proposed MDAFF method achieves a 1.3% increase in mAP and a 1.9% increase in Rank-1 on the Market1501 dataset, as well as a 1.7% increase in mAP and a 0.5% increase in Rank-1 on the DukeMTMC-reID dataset. Therefore, the MDAFF method effectively improves the accuracy of pedestrian re-identification.**

Index Terms—**Deep learning, Computer vision, Pedestrian re-identification, Multi-Dimensional Attention**

I. Introduction

Pedestrian re-identification is a significant research field in computer vision that aims to determine the presence of specific pedestrians in images or video sequences using computer vision techniques. It is considered a subproblem of image retrieval [1, 2]and is primarily used to overcome the limitations of camera perspectives. Due to variations in appearance caused by factors such as clothing, scale, occlusion, pose, and viewpoint, pedestrian re-identification has emerged as a valuable and challenging topic in the field of computer vision [3, 4].

In traditional methods, networks only focus on extracting representations from individual images, disregarding the potential correlations among images.

However, such correlations can enhance the representations of individual images. In the paper by Author et al. [5], NFormer was proposed, which employs Neighbor Transformer to model a large number of input images interactively, with the goal of obtaining enhanced image representations. To model the relationships between pedestrian images and extract more robust features while reducing computational complexity, the authors introduced two important modules: Landmark Agent Attention (LAA) and Reciprocal Neighbor Softmax (RNS). The LAA module processes long input sequences, capturing their internal local structures and dependencies. The RNS module reduces the length of long input sequences to improve computational efficiency, and it can be easily combined with existing methods to achieve performance improvements.

This paper proposes an improved MDAFF method based on NFormer. Instead of using the ResNet network structure mentioned in reference [5], the feature extraction is performed using ResNeXt [6], which has a deeper network structure but fewer parameters. This approach allows for widening and deepening the network while reducing the model's parameter count and operational costs. To further enhance the accuracy of pedestrian re-identification, a Position Attention (PA) module is added to the upstream position of ResNeXt, enabling the model to focus more on pedestrian position information and improve its understanding of effective features in other dimensions. Additionally, an attention feature fusion module is designed to fully exploit the information in the feature maps by effectively combining channel correlations and spatial position feature correlations. The LAA module captures internal relationships among sequences more efficiently, while the RNS module accurately preserves crucial information. These enhancements enable the model to better process long sequence inputs and further improve its accuracy.

II. Principle of the NFormer Algorithm

NFormer [5] proposes a deep learning method for pedestrian re-identification, aiming to improve the accuracy and robustness of this task.

Manuscript received April 28, 2023; revised August 12, 2023.This work was supported by the National Natural Science Foundation of China (61775169), the Education Department of Liaoning Province (LJKZ0310), the Excellent Young Talents Program of Liaoning University of Science and Technology (2021YQ04).

X. P. Chen is a postgraduate student at the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China (e-mail: 342014084@qq.com).

Y. Xu is a Professor at the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China (corresponding author, phone: 86-13889785726; e-mail: xuyang_1981@aliyun.com).

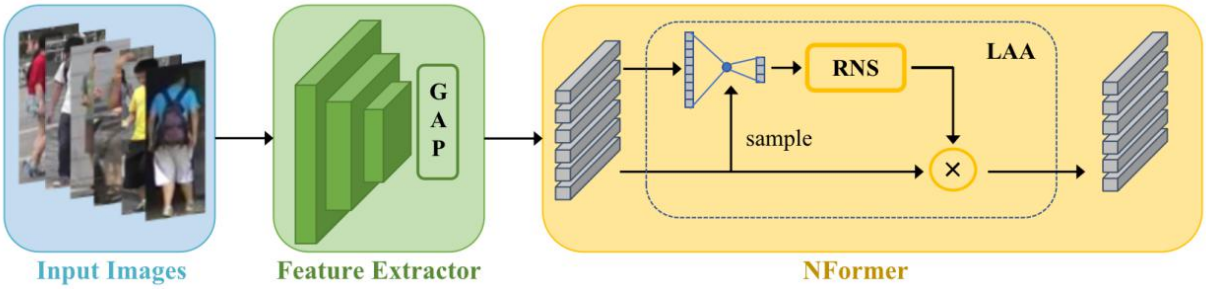

Fig. 1. Schematic Diagram of the NFormer Algorithm

Accuracy and robustness are crucial metrics in pedestrian re-identification, as they require the model to accurately identify pedestrian features and maintain stability in the face of varying scenes and poses. To address these challenges, NFormer introduces a novel deep learning approach. The architecture of the NFormer algorithm is depicted in Figure 1.In the NFormer method, the initial step involves using the ResNet50 network for feature extraction of input images. ResNet50 is a deep residual network that improves model performance by stacking multiple residual blocks. The output feature vectors from ResNet50 capture global information about pedestrians in the images, which are then used for modeling the neighborhood relationships. Next, the NFormer method employs a Transformer model to model the neighborhood relationships among pedestrians. During the neighborhood relationship modeling, NFormer adopts an adaptive multi-scale strategy to adapt to images of different scales. This strategy allows the model to better capture local features of pedestrians in different images, thereby enhancing the accuracy and robustness of pedestrian re-identification.

In the task of pedestrian re-identification, the relationships between pedestrians are often crucial, as different pedestrians may share similar features, such as similar clothing or hairstyles [7]. Therefore, modeling the neighborhood relationships between pedestrians can improve the model's ability to distinguish between pedestrians. In the neighborhood relationship modeling process, NFormer incorporates an adaptive multi-scale strategy. This strategy enables the model to adapt to images of different scales and capture the local features of pedestrians in different images. Specifically, the model automatically selects the appropriate scale to process the input images, leading to improved performance and stability of the model.

III. Improved Strategy

While NFormer utilizes attention mechanisms to handle long sequences, the model may not fully capture comprehensive local features of pedestrians during the process of modeling neighbor relationships. This limitation weakens the model's aggregation capability, resulting in decreased accuracy. To address this issue, we employ PA-ResNeXt for feature extraction from pedestrian images, enhancing the model's overall receptive field and incorporating pedestrian position features. Furthermore, we introduce an attention feature fusion module to improve the model's focus on both local and global features simultaneously. This module effectively aggregates critical global features without significantly increasing computational costs, facilitating the fusion of multi-dimensional pedestrian information within images.

As a result, the model can more effectively capture relationships among pedestrians in different images, thereby reducing the impact of various scenes or poses on recognition accuracy. By leveraging multi-dimensional features, the model achieves better discrimination of pedestrians, leading to a significant improvement in accuracy.

The MDAFF method proposed in this paper comprises two primary components. In the initial segment, we employ a ResNeXt network augmented with a fused Position Attention (PA) module as the foundational feature extraction framework. ResNeXt effectively mitigates the challenge of gradient vanishing through the integration of residual connections, while harnessing grouped convolutions to amplify the network's representational capacity. This design results in exceptional performance gains with reduced parameter count and computational overhead. ResNeXt achieves high performance while having fewer parameters and computational costs. By focusing on the positional features of pedestrians in high dimensions, the model reduces the impact of noise, enhances the global receptive field, and improves sensitivity to object positions, ultimately leading to higher model accuracy.

In the second part, we formulate an attention-driven feature fusion module tailored to accentuate multidimensional attributes within constrained dimensions, facilitating their amalgamation into a comprehensive multidimensional fused feature. Across varying depths, the network dynamically prioritizes distinct feature dimensions, ensuring adaptability and precision in feature extraction.

Finally, these features are integrated into the feature maps, resulting in a significant improvement in model accuracy. By combining the PA module and attention feature fusion module, the MDAFF method effectively captures both local and global features, leveraging multi-dimensional representations to enhance the model's accuracy.

A. Optimization of Backbone Networks

The depth of a neural network has a significant impact on the performance of machine learning models. Increasing the depth of a network enables it to extract more complex patterns, theoretically leading to better results. However, as the network depth increases, the model's accuracy may plateau or even decrease. To tackle this issue, He et al. [8] introduced the ResNet network, which incorporates residual units using a shortcut mechanism to facilitate residual learning and mitigate the problem of degradation stemming from network depth [9].

Compared to the ResNet network, the ResNeXt network enhances the network's expressive power by introducing

group convolutions [10]. When applied to image classification tasks, the main difference between ResNet and ResNeXt lies in their residual block structures. In ResNet, each residual block consists of a single branch, while in ResNeXt, each residual block contains multiple branches, as shown in Figures 2 and 3. These branches can learn different features in parallel and merge them to obtain more powerful feature representations. Specifically, the residual blocks in ResNeXt are designed as a base block plus multiple branches. Each branch consists of convolutional layers and batch normalization (BN) layers, with their inputs and outputs being the output of the base block. The outputs of each branch are concatenated and then dimensionality-reduced and dimensionality-increased through a 1×1 convolutional layer to obtain the final output of the residual block. This design allows ResNeXt to leverage multiple parallel pathways to learn diverse and rich features, enhancing its representation capacity and enabling it to capture intricate patterns in the data effectively.

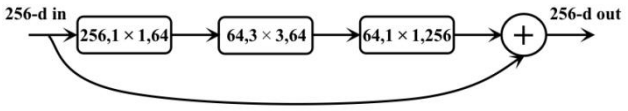

Fig. 2. Block structure of ResNet

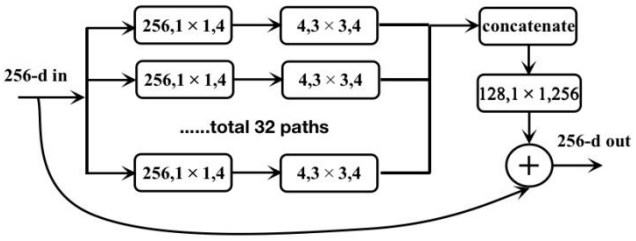

Fig. 3. Block structure of ResNext

To describe the structure of residual blocks in ResNet, mathematical formulas are employed. In ResNet, a residual block is represented by Equation (1):

$$y = F(x) + x \tag{1}$$

In this equation, x denotes the input feature map, while $F(x)$ represents the transformation function within the residual block. ResNeXt, an extension of ResNet, introduces parallel residual blocks, as shown in Equation (2):

$$y = \sum_{i=1}^{C} F_i(x) + x \tag{2}$$

In Equation (2), $\sum_{i=1}^{C} F_i(y)$ represents the output of the multi-branch residual network, where C is the number of parallel branches, C represents the input feature map, and $F_i(x)$ represents the transformation function of the i-th branch. Similar to ResNet, this formula also employs the concept of residual connections, where the input x is directly added to the network's output. This addition operation helps in better gradient propagation and avoids the issue of gradient vanishing.

By comparing the formulas above, we can observe that the primary distinction between ResNet and ResNeXt lies in the nature of the transformation function. $F(x)$. In ResNet, $F(x)$ is a simple convolutional neural network, while in ResNeXt, $F(x)$ is composed of multiple parallel branches. This design choice enables ResNeXt to leverage more model parameters, thus enhancing the network's representational capacity.

In sequence models, which treat input image sequences as temporal data, it is essential to consider both local and global features during feature extraction. ResNeXt50 exhibits greater expressive power than ResNet50 in capturing these features, rendering it a more appropriate choice for the feature extraction network in sequence models such as NFormer. Essentially, ResNeXt's distinctive architecture, characterized by its parallel branches, enables it to leverage a greater number of model parameters and attain a more comprehensive representation of features. This characteristic proves advantageous in sequence modeling tasks, where ResNeXt outperforms ResNet in capturing both local and global features, thereby leading to enhanced performance in applications such as NFormer.

In pedestrian re-identification datasets, images often appear blurry and contain substantial noise [11]. Moreover, pedestrian positions within images vary. To enhance pedestrian re-identification accuracy, machine learning models must concentrate on the pedestrian regions in images. By analyzing pedestrian positions along the x and y axes, the model can highlight relevant features, learn effective features more efficiently, and minimize the impact of noise. Consequently, this paper integrates the Position Attention (PA) module into the feature extraction network upstream. The PA module aims to reduce noise influence, expand the global receptive field, and heighten sensitivity to object positions. Expanding upon the Coordinate Attention (CA) module [12], the PA module enhances the focus on object position features in the upstream layers of the feature extraction network. In the ResNeXt50 network, this module is incorporated after the first convolutional layer of each block, allowing attention to be applied to higher-dimensional features. The structure of the PA-ResNeXt layer is depicted in Figure 4.

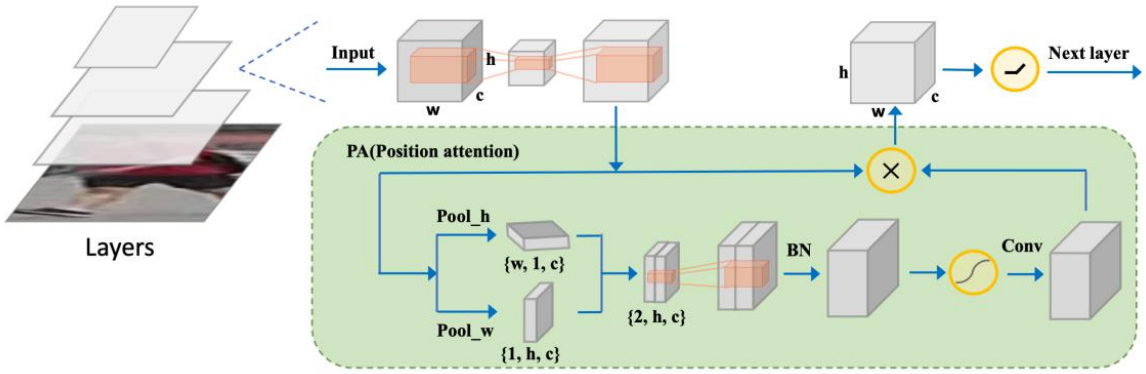

Fig. 4. Layer structure diagram of PA-ResNeXt

The PA (Position Attention) module aims to capture precise positional information features of objects. This module performs global pooling according to the following formula, which can be represented as Equation (3):

$$Z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i, j), \quad (3)$$

Specifically, given the input X, the PA module utilizes a pooling kernel of size $(H, 1)$ or $(1, W)$ to encode each channel along the vertical and horizontal coordinates. Thus, the output of the c-th channel with a height of h can be represented as Equation (4):

$$Z_c^h(h) = \frac{1}{W} \sum_{0 \le i < W} x_c(h, I), \quad (4)$$

Similarly, for the c-th channel with a width of w. It can be represented as Equation (5):

$$Z_c^w(w) = \frac{1}{H} \sum_{0 \le j < H} x_c(j, w) \quad (5)$$

After the aforementioned transformations, this part concatenates the previous transformations and applies a 1×1 convolutional transformation function F to it. This can be represented as Equation (6):

$$f = \delta(F([z^h, z^w])) \quad (6)$$

In Equation (6), the symbol $[\cdot, \cdot]$ represents the concatenate operation along the spatial dimension. δ denotes a non-linear activation function, and f represents the intermediate feature map that encodes spatial information in the horizontal and vertical directions. The feature map f is then decomposed into two separate tensors, $f^h \in R_r^{C \times H}$ and $f^w \in R_r^{C \times W}$, along the spatial dimension.

Additionally, two 1×1 convolutions, f_h and f_w, are applied to transform F_h and F_w, respectively, into tensors with the same number of channels as the input X. This can be represented as Equation (7) and Equation (8):

$$G^h = \sigma(F_h(f^h)) \quad (7)$$

$$G^w = \sigma(F_w(f^w)) \quad (8)$$

Here, σ represents the sigmoid activation function. Then, the outputs g^h and g^w are expanded and used as attention weights. Finally, the output can be represented as Equation (9):

$$Y(i, j) = x_c(i, j) \times g^h(i) \times g^w(j) \quad (9)$$

The attention dedicated to the detected pedestrian locations is crucial for the accuracy of pedestrian re-identification models. Normally, an attention module is applied on the feature map preceding the final convolutional layer to extract spatial and channel information for individual pixels. However, this usage may lead to inadequate learning of location features for the target objects, resulting in ineffective attention weights and potentially diminishing the model's representational capacity.

In ResNeXt, the convolutional layers in the basic blocks typically possess higher feature dimensions compared to the convolutional layers in the output layer. By applying the attention module to high-dimensional features, the model can more effectively capture spatial relationships among input features, leading to improved performance. Placing this module after the convolutional layer in the basic block enables direct connection between the module's output and the subsequent convolutional layer within the block, promoting information flow within the entire block and preventing information loss between layers. This can also help enhance the model's performance. By calculating the significance of different positions using the coordinate information from the feature map and adjusting the feature map accordingly, the modified feature map is utilized as the input for the next layer and subsequent operations. Ultimately, this process produces a feature map that emphasizes critical location information.

B. Attention Feature Fusion Module

In pedestrian re-identification tasks, techniques like SE attention [13] and CBAM [14] are effective in capturing vital image features such as pedestrian body parts, clothing colors and textures, background information, and spatial relationships among pedestrians, which play a crucial role in differentiating individuals. By incorporating SE attention and CBAM attention into pedestrian re-identification networks, the emphasis on these features can be heightened, leading to improved accuracy and robustness in pedestrian re-identification.

However, these attention modules mainly enhance the representational capacity of feature maps by assigning weights. When the weights of specific features approach zero, their contributions to the final feature representation are significantly weakened or even disregarded, resulting in information loss and error accumulation. Consequently, they fail to incorporate the correlations between different features, hindering a better understanding of the key object features. Thus, feature fusion techniques are crucial in the field of computer vision [15] as they enable the combination of channel features, spatial features, and object position features, effectively preserving the original information of diverse features and preventing important information from being neglected. This enhances the performance and robustness of deep learning models, allowing for more accurate identification and localization of various objects.

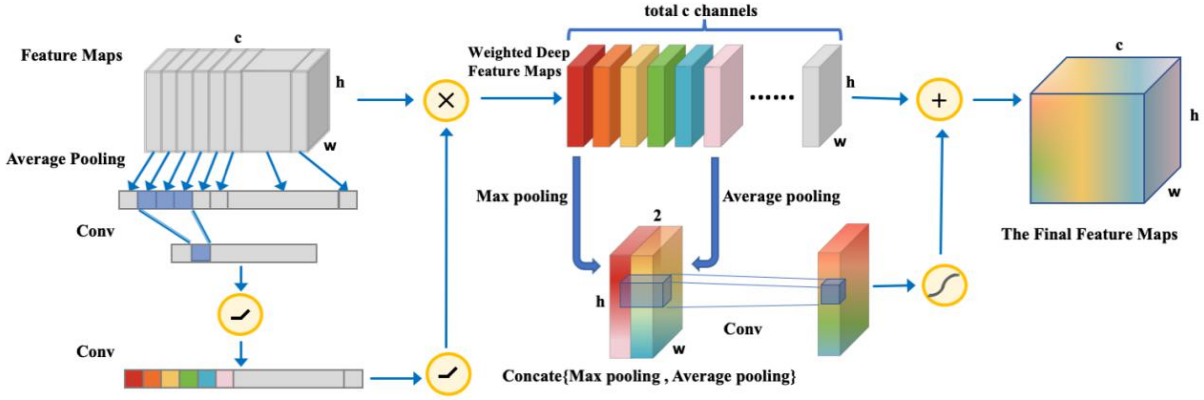

Fig. 5. Attention feature fusion module diagram

The attention feature fusion module proposed in this paper initially captures the crucial information by merging position features in the feature map, reducing the risk of overfitting and enabling the model to handle inputs from different positions more effectively, resulting in a feature map with weighted key position features. Subsequently, the module synergizes prominent position features from upstream, alongside pivotal channel characteristics and essential spatial attributes of the target object. By capitalizing on the amalgamation of object position features, the module exhibits heightened efficacy in assimilating channel and spatial attributes, thus adeptly capturing intricate object nuances and augmenting the model's feature representation.The attention feature fusion module is illustrated in Figure 5.

Initially, the feature map is dimensionally reduced via average pooling, reducing model parameters while retaining crucial features. Then, convolutional and activation operations are applied to the feature map, facilitating both linear and non-linear transformations to boost its expressive capacity. Employing the sigmoid activation function maps pixel values within the feature map to a probabilistic distribution, enabling a more informative representation of essential features. Subsequently, max pooling and average pooling operations are used to extract significant features, which are subsequently merged with the primary features, yielding a more comprehensive and expressive feature map. Finally, additional rounds of convolutional and activation operations further enhance the feature map's representation, ultimately improving the model's classification performance.

The attention feature fusion module extracts and combines channel and spatial features based on feature maps that contain positional information. Firstly, the extraction of channel features is performed by applying global average pooling to the input feature map, which can be represented as Equation (10):

$$Z_i = \frac{1}{H \times W} \sum_{j=1}^{H} \sum_{k=1}^{W} x_{I,j,k} \quad (10)$$

Where H and W represent the height and width of the input feature map respectively, and z has a dimension of C, representing the number of channels. Applying a convolution operation with ReLU activation function to the average pooling result z can be represented as Equation (11):

$$F = ReLU(Conv(z, K_1) + b_1) \quad (11)$$

In Equation (11), K_1 represents the convolutional kernel parameters, b_1 represents the bias parameters, and $Conv$ denotes the convolution operation. Applying another convolution operation with a sigmoid activation function to the output f can be represented as Equation (12):

$$A = sigmoid(Conv(f, K_2) + b_2) \quad (12)$$

Where K_2 represents the convolutional kernel parameters, b_2 represents the bias parameters. Multiplying the attention weights a obtained from the previous step with the input feature map x to obtain the weighted feature map can be represented as Equation (13):

$$M_{channel} = a_i \cdot x_{i,j,k} \quad (13)$$

$M_{channel}$ represents the feature map with channel attention, i represents the channel dimension, and j and k represent the height and width of the feature map, respectively. Then, the extraction of spatial features involves global average pooling and global max pooling operations on the obtained feature map, followed by concatenation. This can be represented as Equation (14):

$$F_2 = concat(AP(M_{channel}), MP(M_{channel})) \quad (14)$$

Where f_2 represents the feature map obtained by concatenating the global average pooling (AP) and global max pooling (MP) results from $M_{channel}$. Then, the feature map f_2 undergoes a convolutional operation followed by a sigmoid activation function. Finally, the result is added to the $M_{channel}$ feature map to obtain the final feature map, as shown in Equation (15):

$$M = sigmoid(Conv(f_2)) + M_{channel} \quad (15)$$

M represents the feature map obtained by integrating object position information and spatial information. This approach enables a more comprehensive utilization of the information contained in the feature map. It effectively combines the correlation between channel features and spatial position features, enabling the model to have a more holistic understanding of the input features. Furthermore, it reduces computational complexity in the spatial dimension, particularly for larger features. By concatenating and convolving salient features from different dimensions, the model converts the computations originally performed in the spatial dimension to the channel dimension, thereby reducing computational complexity and storage space requirements.

IV. Analysis of Experiments And Results

The experimental platform consists of two main components: hardware and software. The hardware platform comprises an Intel Xeon Platinum 8350C CPU and an RTX 3090 (24GB) graphics card, which provides computational power for the experiments. The software environment includes the Ubuntu 18.04.5 LTS operating system, PyTorch 1.8-GPU deep learning framework, and Pycharm Community IDE, which facilitate the development and execution of the experiments.

A. Dataset Selection

To accommodate the diverse research needs in pedestrian re-identification, multiple datasets have been proposed. For this experiment, the Market1501 dataset [16] will be used. Market1501 consists of pedestrian images captured from six cameras on the campus of Tsinghua University. It contains a total of 1,501 annotated pedestrians as shown in Figure 6. The dataset is divided into a training set with 751 annotated pedestrians and a testing set with 750 annotated pedestrians. There are no overlapping identities between the training and testing sets, meaning that the 751 pedestrians in the training set do not appear in the testing set. The bounding_box_train subset of Market1501, containing 12,936 images, will be used as the training set. The bounding_box_test subset, comprising 19,732 images, will be used as the testing set. The query set consists of manually labeled images, where one image per pedestrian from each of the six cameras in the testing set is selected, resulting in a query set of 3,368 images. Each pedestrian in the testing set can have up to six images, while the query set consists of 3,368 images.

Fig. 6. Pedestrian pictures in the Market1501 dataset

B. Experimental Evaluation Criteria

The evaluation metrics used in the experiment include the Rank-n and mAP (mean Average Precision). Rank-n refers to the probability of having correct results among the top n images in the search results based on their confidence scores. Precision (P) represents the percentage of correctly predicted positive samples out of all predicted positive samples. Recall (R) represents the percentage of correctly predicted positive samples out of the actual positive samples. AP (Average Precision) is the area under the Precision-Recall (P-R) curve. mAP is the average AP across all classes. These metrics can be expressed using equations 16-19.

$$P = \frac{TP}{TP+FP} \tag{16}$$

$$R = \frac{TP}{TP+FN} \tag{17}$$

$$AP = \int_0^1 PdR \tag{18}$$

$$mAP = \frac{\sum_{i=1}^{m} AP_i}{m} \tag{19}$$

Here TP, TN, FP and FN represent the following:

TP: True Positive, referring to the number of correctly classified positive samples.

TN: True Negative, indicating the number of correctly classified negative samples.

FP: False Positive, representing the number of negative samples incorrectly classified as positive.

FN: False Negative, indicating the number of positive samples incorrectly classified as negative.

Here m denotes the number of categories present in the sample.

C. Experimental Analysis

Figure 7 illustrates the loss and accuracy curves of the Res50+NFormer model before and after applying the proposed method. By comparing the curves of the original method and the improved method, it is evident that the improved method demonstrates faster loss reduction and accuracy improvement. This suggests that the proposed method facilitates a comprehensive understanding of the relevant features that the model needs to learn, leading to faster convergence speed and a smoother and more stable training process. Introducing an upstream position detector in the ResNeXt network restricts the model's focus solely to the position features of the upstream objects. After extracting features, the model highlights the channel and spatial features of the objects and then fuses them with multidimensional features. This enables the model to effectively learn detailed multidimensional features and achieve a more comprehensive representation. As a result, this method effectively enables the model to capture specific location features that require attention, leading to faster convergence and more stable training.

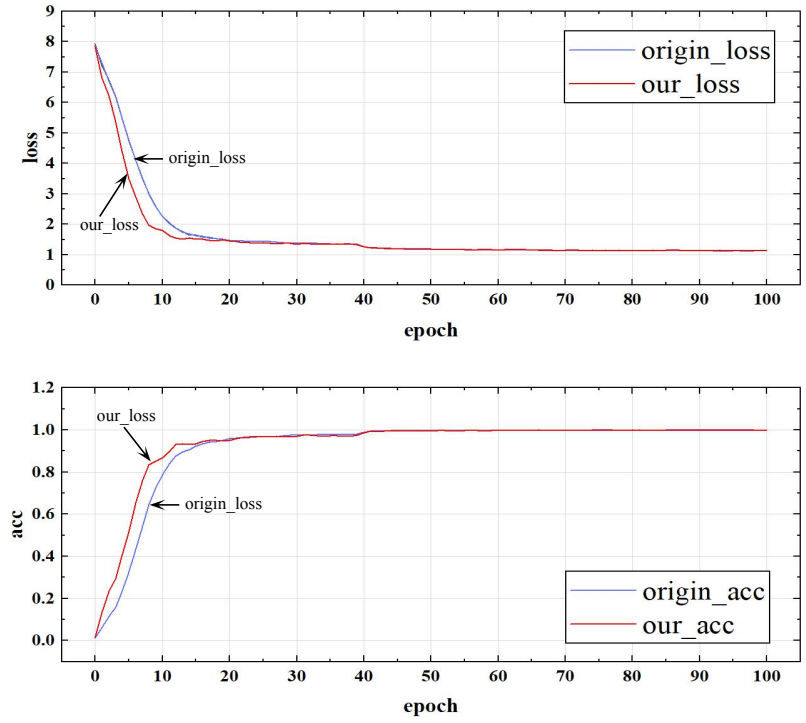

Fig. 7. Comparison curves of loss and accuracy between the original method and the improved method

1. Ablation experiment

In order to validate the effectiveness of the proposed improvements for pedestrian re-identification, a set of ablation experiments was designed to compare and analyze the following scenarios: (1) the original Res50+NFormer model, (2) using PA-ResNeXt as the feature extraction network, (3) adding the attention feature fusion module to Res50+NFormer, and (4) applying the MDAAF method to NFormer. Under the same experimental conditions, experiments were conducted on the Market1501 dataset, and the specific experimental performance can be found in Table I.The table shows that replacing ResNet with PA-ResNeXt as the feature extraction network led to a slight improvement of 0.3 percentage points in mAP and 0.5 percentage points in Rank-1. Likewise, incorporating only the attention feature fusion module to Res50+NFormer resulted in a 0.7 percentage point improvement in mAP and a minor 0.3 percentage point increase in Rank-1. The individual effects of these modifications were not very pronounced. However, by employing the MDAAF method with NFormer, there was a remarkable improvement of 1.3 percentage points in mAP and 0.9 percentage points in Rank-1.Additionally, there were also enhancements in Rank-5 and Rank-10.

To validate the effectiveness of the attention feature fusion module, this study compared it with other attention mechanisms, namely, SE [13] and CBAM [14], on the Market1501 dataset, as shown in Table II.

According to Table 2, the multi-dimensional attention feature fusion module utilized in this study demonstrated notable enhancements in mAP and Rank-1 compared to other attention mechanisms. More specifically, it surpassed SE attention by 0.6 percentage points in mAP and 0.7 percentage points in Rank-1. Although the improvement over CBAM was slightly smaller, with only a 0.3 percentage point increase in Rank-1, the feature fusion in the proposed method facilitated a more comprehensive representation, leading to a 0.4 percentage point improvement in mAP compared to CBAM.

2. Comparison experiment with mainstream algorithms

To validate the detection performance of the proposed improvement algorithm, it was compared with mainstream algorithms such as ABDNet, NFormer and so on The experimental results were obtained by conducting experiments on the Market1501 and DukeMTMC-reID [17] datasets, and the comparative results can be found in Table Ⅲ.

TABLE Ⅰ

ABLATION EXPERIMENT

Model	mAP(%)	Rank-1(%)	Rank-5(%)	Rank-10(%)
Res50+NFormer	91.1	94.6	97.7	98.7
PA-ResNeXt50+NFormer	91.4	95.1	97.9	98.6
Res50 + Attention Feature Fusion Module + NFormer	91.8	94.9	97.8	99.0
MDAFF+NFormer	**92.4**	**95.5**	**98.3**	**99.1**

TABLE Ⅱ

ADD DIFFERENT ATTENTION COMPARISON EXPERIMENTS

Model	mAP(%)	Rank-1(%)	Rank-5(%)	Rank-10(%)
Res50+SE+NFormer	91.2	94.2	97.5	98.8
Res50+CBAM+NFormer	91.4	94.6	97.8	98.6
Res50 + Attention Feature Fusion Module + NFormer	**91.8**	**94.9**	**97.8**	**99.0**

TABLE Ⅲ

COMPARED WITH ADVANCED ALGORITHMS

Model	Market1501		DukeMTMC-reID	
	Rank-1(%)	mAP(%)	Rank-1(%)	mAP(%)
ABDNet [18]	95.4	88.2	88.7	78.6
PISNet [19]	95.6	87.1	88.8	78.7
CBN [20]	94.3	83.6	84.8	70.1
ISP [21]	95.3	88.6	89.6	80.0
CBDB-Net [22]	94.4	85	87.7	74.3
CDNet [23]	95.1	86	88.6	76.8
PAT [24]	95.4	88	88.8	78.2
C2F [25]	94.8	87.7	87.4	74.9
Res50+NFormer [5]	94.6	91.1	89.4	83.5
BPBreID [26]	**95.7**	89.4	-	-
MSINet [27]	95.3	89.6	-	-
MDAFF(ours)	95.5	**92.4**	**89.9**	**85.2**

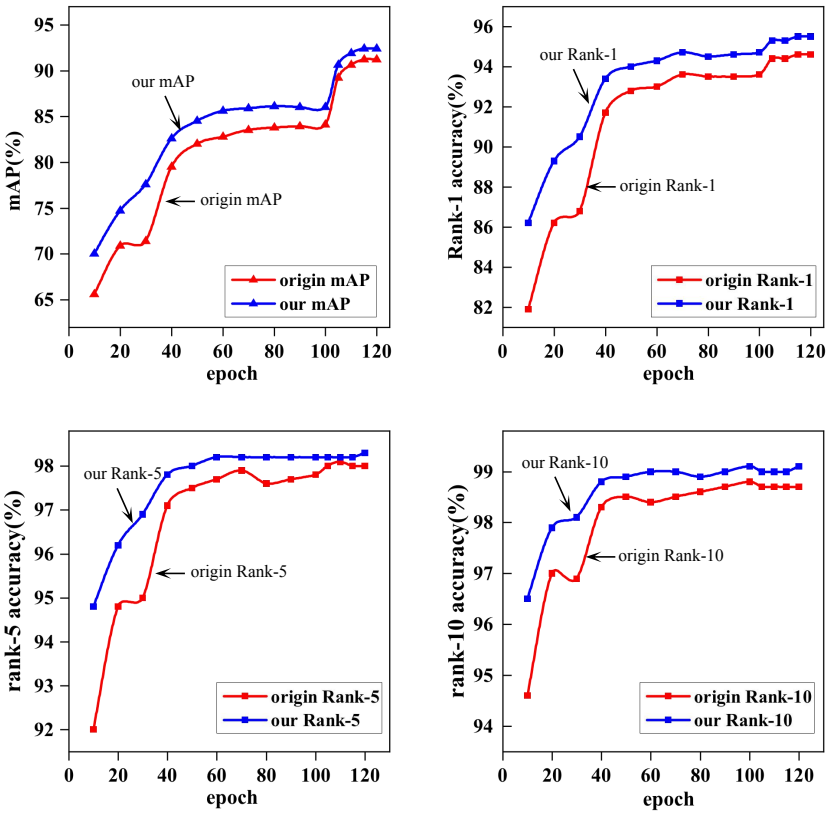

Fig. 8. Dot-line diagram of original algorithm and improved algorithm mAP and Rank-n change

Based on the data in the table, it is clear that the proposed method consistently achieves higher Rank-n and mAP scores compared to other pedestrian re-identification algorithms. This suggests that the results obtained using the proposed improvement method outperform other algorithms, confirming the effectiveness of the improvement algorithm. Furthermore, the results obtained on the DukeMTMC-reID dataset surpass those of other algorithms, indicating the effectiveness of the proposed method and its applicability to various pedestrian re-identification datasets.

Figure 8 illustrates the trends of average precision (mAP) and Rank-n for Res50+NFormer and the algorithm with the proposed method, plotted in a line graph. The results are obtained from 120 rounds of testing on the Market1501 dataset. It is noticeable that the improved algorithm surpasses the original model in terms of both the rate of improvement and stability. Furthermore, NFormer with the proposed method consistently obtains higher mAP and Rank-n scores compared to the original model. The learning process with the proposed method demonstrates increased stability, leading to smoother and more consistent performance.

V. Conclusion

In this paper, we propose an enhanced method for pedestrian re-identification based on the NFormer approach. Our objective is to address challenges such as high noise levels, low-quality images, and environmental factors that can hinder model learning and reduce sensitivity to detailed features. Our method utilizes the PA-ResNeXt network as the feature extraction backbone and incorporates a multi-dimensional attention feature fusion module. This allows the model to focus on specific features at different positions. By emphasizing pedestrian location feature learning during upstream feature extraction and capitalizing on multi-dimensional features, the model can more effectively capture fine-grained details. Ultimately, feature fusion boosts the expressive capacity of the feature maps. In comparison to the Res50+NFormer, our proposed method achieves a 1.3% increase in mAP and a 1.9% increase in Rank-1 on the Market1501 dataset. For the DukeMTMC-reID dataset, the method results in a 1.7% increase in mAP and a 0.5% increase in Rank-1. The experimental outcomes indicate that the enhanced model surpasses the original model in terms of accuracy and stability for pedestrian re-identification, exhibiting strong generalization capabilities across multiple datasets. Future research will concentrate on further optimizing the proposed method in light of the identified challenges.

References

[1] K. A. Shahrim, A. H. Abd Rahman, and S. Goudarzi, "Hazardous Human Activity Recognition in Hospital Environment Using Deep Learning," *IAENG International Journal of Applied Mathematics*, vol.52, no.3, pp. 748-753, 2022.

[2] X. Zhang, M. Hou, X. Deng, and Z. Feng, "Multi-cascaded attention and overlapping part features network for person re-identification," *Signal, Image and Video Processing*, vol.16, no.6, pp. 1525-1532, 2022.

[3] D. Wu, S.-J. Zheng, X.-P. Zhang, C.-A. Yuan, F. Cheng, Y. Zhao, et al., "Deep learning-based methods for person re-identification: A comprehensive review," *Neurocomputing*, vol.337, pp. 354-371, 2019.

[4] R. Zhu, Y. Xu, L. Wang, T. Sun, J. Yu, S. Ding, et al., "A Wide Range Multi-obstacle Detection Method Based on VIDAR and Active Binocular Vision," *IAENG International Journal of Applied Mathematics*, vol.53, no.1, pp. 381-392, 2023.

[5] H. Wang, J. Shen, Y. Liu, Y. Gao, and E. Gavves, "Nformer: Robust person re-identification with neighbor transformer," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7297-7307, 2022

[6] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," *Proceedings of*

the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1492-1500, 2017
[7] W. Qu, Z. Xu, B. Luo, H. Feng, and Z. Wan, "Pedestrian re-identification monitoring system based on deep convolutional neural network,"*IEEE Access*, vol.8, pp. 86162-86170, 2020.
[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016
[9] J. Miao, S. Xu, B. Zou, and Y. Qiao, "ResNet based on feature-inspired gating strategy,"*Multimedia Tools and Applications* pp. 1-18, 2022.
[10] T. Chen, B. Duan, Q. Sun, M. Zhang, G. Li, H. Geng, et al., "An efficient sharing grouped convolution via bayesian learning,"*IEEE Transactions on Neural Networks and Learning Systems*, vol.33, no.12, pp. 7367-7379, 2021.
[11] F. An, and J. Wang, "Pedestrian Re-Identification Algorithm Based on Multivariate Manifold Metric-Anti-Noise Manifold Space Learning,"*IEEE Transactions on Emerging Topics in Computational Intelligence*, vol.7, no.1, pp. 261-270, 2022.
[12] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13713-13722, 2021
[13] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132-7141, 2018
[14] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3-19, 2018
[15] M. Zhao, Q. Yue, D. Sun, and Y. Zhong, "Improved SwinTrack single target tracking algorithm based on spatio-temporal feature fusion,"*IET Image Processing* pp. 2410-2421, 2023.
[16] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1116-1124, 2015
[17] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," *European Conference on Computer Vision*, pp. 17-35, 2016
[18] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, et al., "Abd-net: Attentive but diverse person re-identification," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8351-8361, 2019
[19] S. Zhao, C. Gao, J. Zhang, H. Cheng, C. Han, X. Jiang, et al., "Do not disturb me: Person re-identification under the interference of other pedestrians," *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pp. 647-663, 2020
[20] Z. Zhuang, L. Wei, L. Xie, T. Zhang, H. Zhang, H. Wu, et al., "Rethinking the distribution gap of person re-identification with camera-based batch normalization," *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pp. 140-157, 2020
[21] K. Zhu, H. Guo, Z. Liu, M. Tang, and J. Wang, "Identity-guided human semantic parsing for person re-identification," *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pp. 346-363, 2020
[22] H. Tan, X. Liu, Y. Bian, H. Wang, and B. Yin, "Incomplete descriptor mining with elastic loss for person re-identification,"*IEEE Transactions on Circuits and Systems for Video Technology*, vol.32, no.1, pp. 160-171, 2021.
[23] H. Li, G. Wu, and W.-S. Zheng, "Combined depth space based architecture search for person re-identification," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6729-6738, 2021
[24] Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang, and F. Wu, "Diverse part discovery: Occluded person re-identification with part-aware transformer," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2898-2907, 2021
[25] 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," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 598-607, 2021
[26] V. Somers, C. De Vleeschouwer, and A. Alahi, "Body part-based representation learning for occluded person Re-Identification," *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1613-1623, 2023
[27] J. Gu, K. Wang, H. Luo, C. Chen, W. Jiang, Y. Fang, et al., "MSINet: Twins Contrastive Search of Multi-Scale Interaction for Object ReID," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19243-19253, 2023