Occluded Pedestrian Re-identification Method Based on Multi-scale Feature Fusion

Haitian Qin, Yang Xu

Abstract—**Pedestrian re-identification is a crucial research task in the fields of computer vision and video surveillance. The primary challenges include occlusion, illumination variation, and complex backgrounds, which significantly undermine the robustness and generalization capabilities of existing methods. To address these issues, a pedestrian re-identification method that effectively integrates multi-scale features and enhances the attention mechanism is required. This paper proposes an occluded pedestrian re-identification method based on multi-scale feature fusion. The method introduces an AAC (Add Noise and Concatenate) module, which injects noise into the central region of the input image to enhance the model's robustness and improve its generalization ability. The model employs EfficientNetB0 and DaViT_small as backbone networks. EfficientNetB0 processes the original input image, while DaViT_small handles the concatenated image with noise, incorporating a transposed convolution module for upsampling the feature maps to further extract high-level features and enhance spatial resolution. Additionally, a feature blocking and global fusion module is proposed, which splits the feature maps into multiple resolutions and uses different convolutional layers to further extract and fuse global features, ultimately generating a global feature vector. This design ensures the model can extract a rich variety of features from the images, thereby significantly improving the accuracy and reliability of the pedestrian re-identification task. Experimental results demonstrate that the proposed Occluded Pedestrian Re-identification method based on Multi-Scale Feature Fusion (OPR-MSFF) outperforms RNFPR (Relation Network for Person Re-Identification) with improvements of 0.7%, 0.8%, and 1.3% in rank-1 accuracy and increases of 0.9%, 1.2%, and 1.7% in mAP on the Market1501, DukeMTMC-reID, and Occluded-DukeMTMC datasets, respectively. These results validate the effectiveness of the proposed method in enhancing pedestrian re-identification performance.**

Index Terms—**person re-identification, deep learning, multi-scale feature fusion, convolutional neural networks**

I. Introduction

Pedestrian re-identification (Re-ID) is a critical research area within computer vision, focusing on the identification and tracking of the same individual across multiple images or video frames captured by different cameras. This technology has extensive applications in intelligent surveillance and urban safety, offering particular value in long-distance and cross-temporal pedestrian tracking and management in public spaces [1].

One of the key challenges in pedestrian re-identification (Re-ID) is the significant variation in appearance and the issue of occlusion. These variations arise from multiple factors, such as occlusions, lighting changes, complex backgrounds, and shifts in a pedestrian's pose [2]. These factors result in substantial visual differences in the same individual across different camera views, complicating the matching process. For instance, the appearance of a pedestrian can vary greatly depending on the viewing angle (e.g., front vs. back), which traditional feature extraction and comparison algorithms often struggle to manage effectively.

Hyunjong Park et al. [3] introduced the Relation Network for Person Re-Identification (RNFPR), which integrates single and multi-block features through a relation network. Building on this approach, we propose an enhanced method, the Occluded Pedestrian Re-Identification based on Multi-Scale Feature Fusion (OPR-MSFF). Our method incorporates an AAC (Add Noise and Concatenate) module, which strengthens the model's robustness to local image perturbations by injecting noise into the central region and concatenating it with the original image.

In contrast to RNFPR, which relies on ResNet50, OPR-MSFF employs both EfficientNetB0 [4] and DaViT_Small [5] for more diverse feature extraction. EfficientNetB0, known for its parameter and computational efficiency via compound scaling, processes the original image, while DaViT_Small, based on the Vision Transformer architecture, excels in capturing global context and understanding the relationships between pedestrians and their environments. It processes the concatenated noisy image, enhancing the model’s ability to interpret scene backgrounds and pedestrian interactions. By merging the outputs of both networks, our model leverages their respective strengths to improve generalization and robustness.

After feature extraction by DaViT_Small, we introduce a transposed convolution module to upsample the feature maps, facilitating the capture of features at various scales and improving recognition accuracy. Finally, a multi-scale feature fusion module integrates features from different scales into a unified representation, enhancing the model’s ability to capture both fine-grained and coarse-grained details, leading to superior performance in pedestrian re-identification.

Manuscript received September 03, 2024; revised November 12, 2024.

This work was supported by the National Natural Science Foundation of China (61775169), the Education Department of Liaoning Province (LJKZ0310).

H. T. Qin is a postgraduate student at the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China (e-mail: 1824234130@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: 705739580@qq.com).

II. Principle of the RNFPR

This paper introduces a novel relation network tailored for the pedestrian re-identification (Re-ID) task. The proposed method enhances the discriminative power of local features by accounting for the relationships between different body parts. The architecture comprises two key modules: the One-vs.-Rest Relation Module (ORM) and the Global Contrast Pooling (GCP) module.

The ORM module addresses one-to-one relationships between individual body parts, ensuring that each part-level feature not only captures information specific to that part but also incorporates contextual data from other body regions. The central concept of this module is to improve the discriminative capacity of local features by embedding relational information from surrounding body parts.

The GCP module, on the other hand, combines Global Average Pooling (GAP) [6] and Global Max Pooling (GMP) techniques to extract global feature mappings from the entire body. This module represents an innovative pooling approach that leverages contrastive features to identify differences and complementary information between pooling outcomes. By integrating residual learning, the GCP module amplifies the max-pooled features, thus enhancing overall model performance and accuracy.

The structure of the RNFPR algorithm is illustrated in Figure 1. The process begins with a pre-trained ResNet-50 model serving as the backbone to extract initial feature maps from the input images. These feature maps are then evenly divided into six horizontal strips, and GMP is applied to each strip. Subsequently, the ORM and GCP modules are employed to extract local relational features and global contrast features, respectively. This strategy strengthens the feature discrimination capabilities for the pedestrian re-identification task, ultimately improving the accuracy of pedestrian identification. The underlying principle of the RNFPR algorithm is depicted in Figure 1.

III. Improved Strategy

Hyunjong Park et al. proposed a relation network-based approach for pedestrian re-identification (RNFPR) that enhances the discriminative power of local features by considering relationships between different body parts. However, RNFPR has some limitations. First, while the One-vs.-Rest Relation Module (ORM) and Global Contrast Pooling (GCP) module integrate local and global features, the algorithm struggles to effectively extract complete pedestrian features in cases of severe occlusion. Second, RNFPR uses ResNet50 as its backbone. While ResNet50 is effective in many tasks, it may be limited in pedestrian re-identification, particularly in capturing fine-grained features in complex scenarios. Furthermore, RNFPR's focus on local feature discrimination through body part relationships may miss critical details under pose variations and occlusions, limiting its ability to fully capture a pedestrian's identity.

To address these issues and improve robustness and accuracy, this paper proposes an occluded pedestrian re-identification method based on multi-scale feature fusion (OPR-MSFF). The framework is shown in Figure 2: (1) The AAC (Add Noise and Concatenate) module injects Gaussian noise into the central region of the input image, encouraging the model to focus on a broader feature range and reducing reliance on background information. By concatenating the original image with the noisy image, the model captures diverse feature information, enhancing robustness and generalization. (2) OPR-MSFF replaces ResNet50 with EfficientNetB0 and DaViT_Small. EfficientNetB0, optimized for parameter and computational efficiency, processes the original image, while DaViT_Small, based on the Vision Transformer architecture, excels at capturing global context and processes the concatenated noisy image. This combination leverages the strengths of both networks, improving feature extraction and model robustness. (3) A transposed convolution module is introduced to upsample feature maps, facilitating fine-grained detail extraction and boosting recognition accuracy. (4) The feature maps are divided into strips for detailed local feature extraction, mitigating the effects of occlusion. Multi-resolution pooling applied to each strip captures features at various scales, further enhancing the model's discriminative power and robustness.

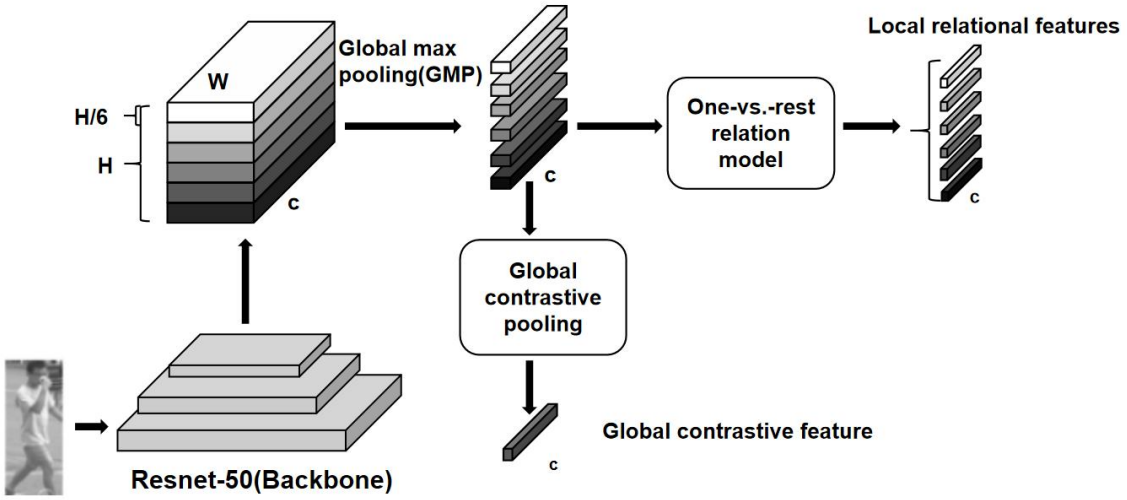


Fig. 1. Diagram of the RNFPR Algorithm Principles

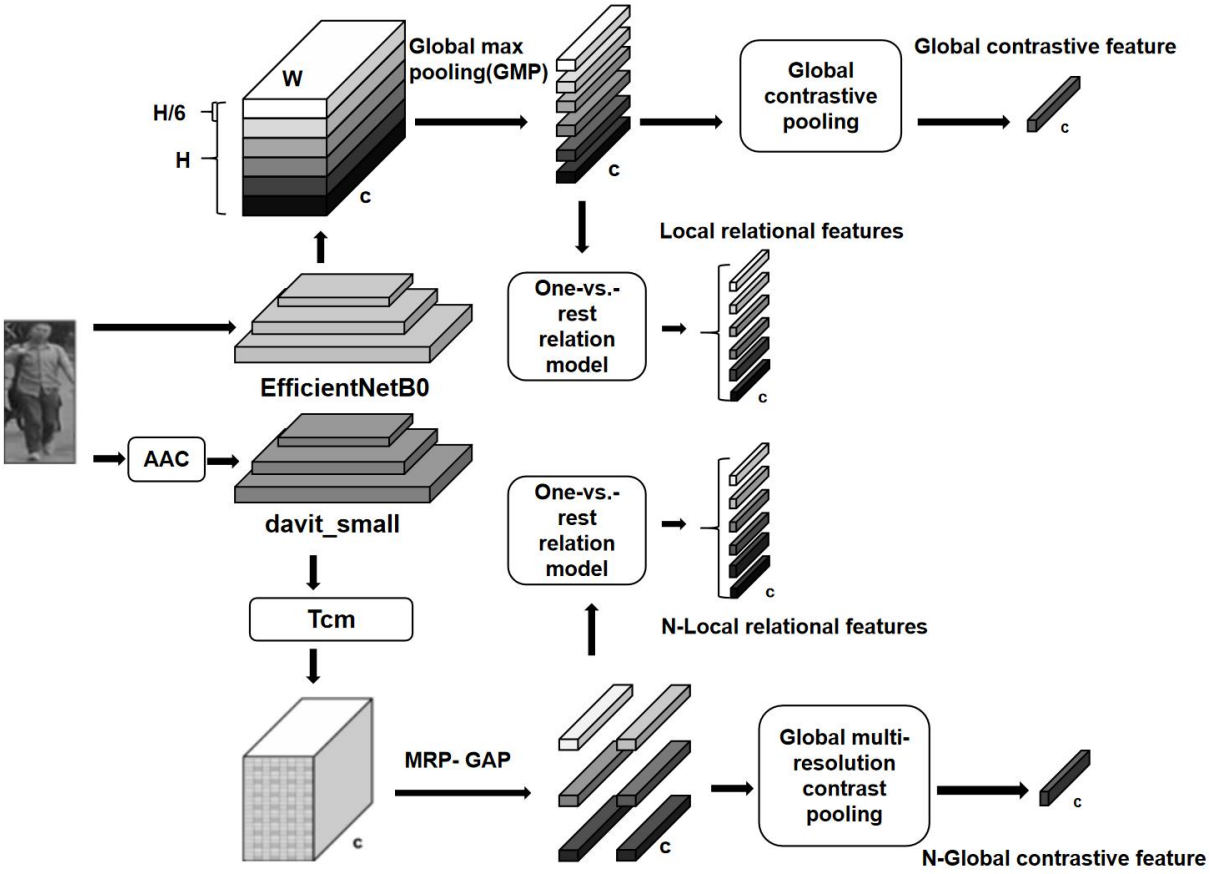


Fig. 2. Diagram of the OPR-MSFF Algorithm Principles

A. Optimize the Backbone Network

RNFPR utilizes ResNet50 as the feature extraction network. ResNet50, proposed by He et al. [7], is a variant of the deep residual network. This network employs a shortcut mechanism to merge residual units, facilitating residual learning and alleviating the degradation problem caused by increased network depth [8]. The network architecture is illustrated in Figure 3.

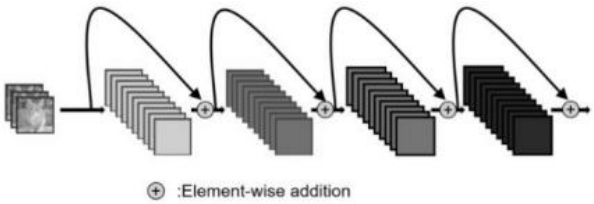


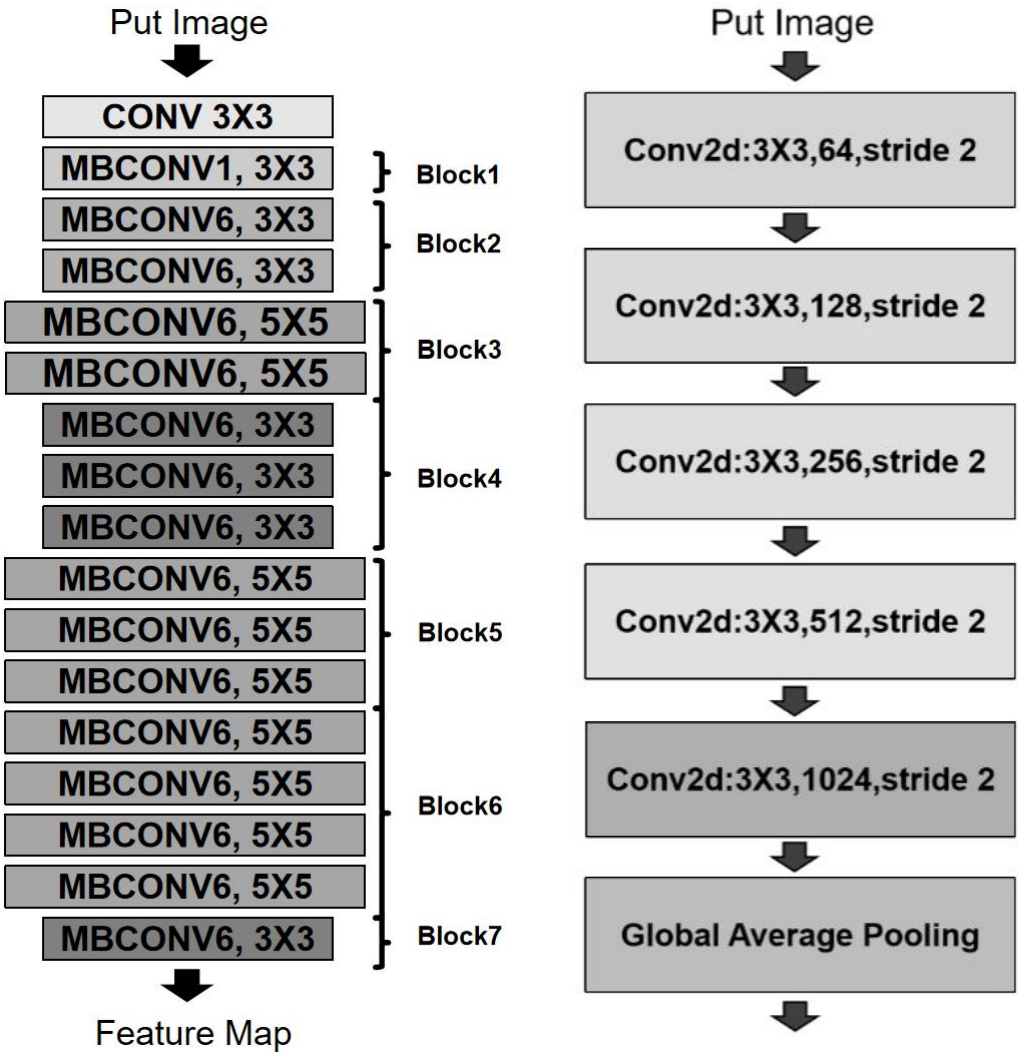
Fig. 3. Diagram of the ResNet50 Network Architecture

ResNet50 is effective in tasks such as image classification and object detection, but it exhibits limitations in pedestrian re-identification. While ResNet50 can extract rich local features, its representation of these features may fall short in capturing all critical details in pedestrian images, which often involve complex poses, diverse clothing, and occlusions. A single ResNet50 model may struggle to fully capture such complexities, particularly in scenarios with complex backgrounds and significant occlusions. Its convolutional layers are susceptible to background noise, potentially degrading feature discrimination. When large occlusions occur, the features extracted by ResNet50 may not sufficiently describe a pedestrian’s identity, highlighting its limitations in addressing occlusion challenges in pedestrian re-identification.

EfficientNet is a family of efficient convolutional neural networks that optimally balance network depth, width, and resolution through compound scaling. EfficientNetB0, the smallest variant, offers fewer parameters and lower computational complexity while maintaining a strong balance between accuracy and efficiency. It excels in extracting fine local features and, due to its compound scaling strategy, provides superior parameter and computational efficiency, achieving high performance with limited resources. The architecture of EfficientNetB0 is shown in Figure 4(a).

DaViT_Small is a compact, efficient model designed for visual tasks, leveraging the Vision Transformer architecture to capture global image features through self-attention mechanisms. Compared to traditional convolutional neural networks, Transformers excel at handling long-range dependencies and global context. The architecture of DaViT_Small is illustrated in Figure 4(b).

EfficientNetB0 and DaViT_Small are complementary in their feature extraction capabilities. EfficientNetB0 focuses on extracting fine local details, while DaViT_Small captures global contextual information. This combination provides a more comprehensive feature representation, which is particularly beneficial in pedestrian re-identification tasks where images often include complex backgrounds and varied poses. A single network may not fully capture this diversity, but the integration of both EfficientNetB0 and DaViT_Small improves robustness to background noise and occlusions. By fusing features from both architectures, the model learns richer, more diverse representations, thereby enhancing pedestrian re-identification performance.



（a）Diagram of the EfficientNetB0 Network Architecture　　（b）Diagram of the Davit_Small Network Architecture

Fig. 4. Diagram of the Backbone Feature Extraction Network Architecture

Let F_{ResNet} , $F_{EfficientNet}$, and F_{Davit} represent the feature representations extracted from ResNet50, EfficientNetB0, and Davit_Small, respectively. The model output using ResNet50 alone can be expressed as Equation 1:

$$O_{ResNet} = f\left(F_{ResNet}\right) \quad (1)$$

The model output utilizing both EfficientNetB0 and Davit_Small can be expressed as Equation 2:

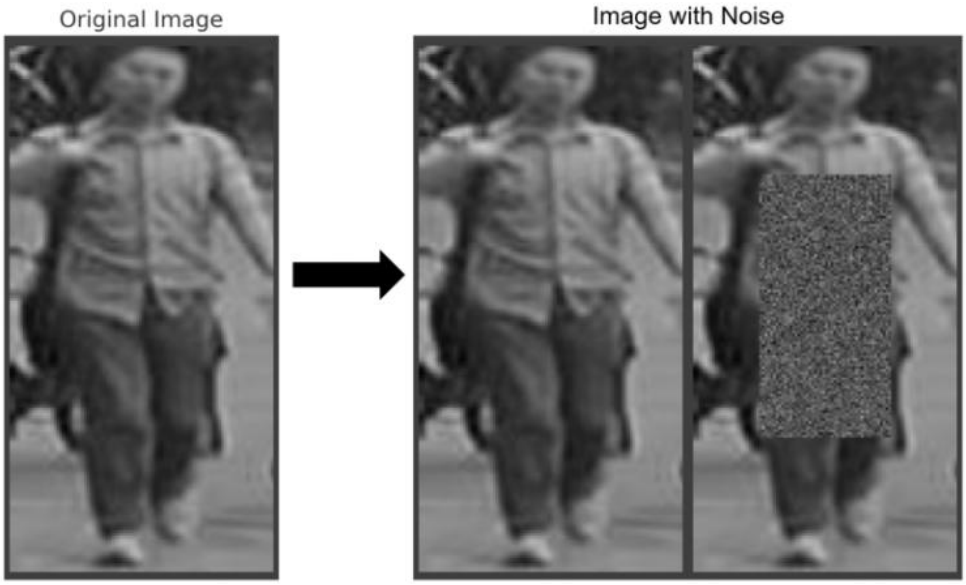
$$O_{Combined} = g\left(F_{EfficientNet} \oplus F_{Davit}\right) \quad (2)$$

where $\oplus$ denotes the concatenation operation, and f and g represent the mapping functions for using individual and combined feature extraction networks, respectively.

B. AAC Module

Pedestrian re-identification typically occurs in complex environments where the background may contain a significant amount of distracting information. These distractions can negatively impact the model's ability to extract and recognize pedestrian features. To mitigate the interference of redundant information, the AAC module adds noise to the central region of the image, forcing the model to focus on a broader area of the image and thereby reducing its reliance on background information. This approach enhances the model's robustness by encouraging it to concentrate more on the pedestrian's intrinsic features.

The primary concept of this module consists of adding noise and concatenating images. Initially, Gaussian noise, based on the mean and standard deviation of the central region, is added to the center of the input image. This technique exposes the model to greater variability during training, thereby enhancing its robustness. Subsequently, the noise-augmented image is concatenated with the original image along the width dimension, forming a new composite image. This concatenated image contains more diverse feature information, which aids the model in better learning both detailed and global features of the pedestrian. The structure of this module is illustrated in Figure 5.



（a）Input Image （b）Output Image

Fig.5. Schematic diagram of the AAC module

Let the input image be x, with dimensions (B, C, H, W), where B represents the batch size, C the number of channels, H the height, and W the width. Define the size of the noise in the central region as (N_H, N_W). The start and end positions of the central region are given by Equations 3 and 4:

$$start\_y = \frac{H - N_H}{2}, \quad end\_y = start\_y + N_H \quad (3)$$

$$start\_x = \frac{W - N_H}{2}, \quad end\_x = start\_x + N_W \quad (4)$$

Here, $start\_y$ and $start\_x$ represent the starting y -coordinate and x -coordinate of the noise region in the image, respectively, while end_y and end_x denote the ending y -coordinate and x-coordinate of the noise region. From the central region, features f_{center} are extracted, and their mean and standard deviation are calculated, as represented by Equation 5:

$$\mu = mean(f_{center}), \quad \sigma = std(f_{center}) \quad (5)$$

Here, μ represents the mean of the central region features, and σ represents the standard deviation of the central region features. Subsequently, Gaussian noise n is generated and applied to replace the central region, as shown in Equation 6:

$$n \sim N(\mu, \sigma^2)$$

$$x_{noisy} = x$$

$$x_{noisy}[:,:,start\_y:end\_y, start\_x:end\_x] = n \quad (6)$$

Here, x_{noisy} denotes the image tensor after the addition of noise. Finally, the original image and the noise-augmented image are concatenated, as represented by Equation 7:

$$x_{concat} = concat(x, x_{noisy}, \dim = 3) \quad (7)$$

By introducing noise during the training process, the model is able to extract more effective features in the presence of noise, thereby improving its accuracy in real-world noisy environments. Incorporating this module effectively achieves data augmentation without the need for additional data. The AAC module trains the model to recognize more diverse image features by altering the local properties of the image. This module enhances the model's adaptability to real-world challenges such as varying lighting conditions, occlusions, and environmental noise, thus significantly improving the robustness and accuracy of pedestrian re-identification.

C. Transposed Convolution Module

In pedestrian re-identification, features at different scales are crucial for achieving high recognition accuracy, as single-scale features often fail to comprehensively capture all the information in an image. To address this issue, the model combines transposed convolutions with multi-resolution pooling, enabling the extraction and fusion of features at various scales, thereby enhancing both the robustness and accuracy of recognition [9]. Furthermore, pedestrian re-identification tasks require capturing fine-grained details of individuals, such as clothing details and accessories. The transposed convolution module increases the resolution of feature maps through upsampling, allowing for more accurate capture of detailed information, which significantly improves recognition accuracy.

For standard convolution operations [10], the output size O can be expressed as Equation 8:

$$O = \left[\frac{I - K + 2P}{S}\right] + 1 \quad (8)$$

Here, I represents the size of the input feature map, K is the size of the convolution kernel, P denotes the padding size, and S indicates the stride. Since transposed convolution is the reverse operation of standard convolution, the output feature map size I' can be calculated using Equation 9:

$$I' = S \times (O - 1) + K - 2P \quad (9)$$

In this module, the parameters $K = 4$, $S = 2$, and $P = 1$ are set accordingly. Therefore, the size of the output feature map O is given by Equation 10:

$$O = 2 \times (I - 1) + 4 - 2 \times 1$$

$$O = 2I \tag{10}$$

The dimensions of the output feature map are twice that of the input feature map. A 4x4 convolution kernel is selected to enhance the smoothing and filling of the feature map during the upsampling process. Larger convolution kernels can capture a broader range of spatial information, contributing to smoother upsampling results. The stride is set to 2 to double the spatial dimensions of the feature map. This is a common practice in transposed convolution operations; by using a stride of 2, the width and height of the output feature map are doubled, achieving the desired upsampling.

Additionally, the padding is set to 1 to preserve edge information during the transposed convolution operation. Proper padding ensures that the output feature map's dimensions align with expectations while preventing the loss of edge information during convolution. By employing the transposed convolution module, the model can effectively utilize multi-scale features in pedestrian re-identification tasks, thereby enhancing both recognition accuracy and robustness.

D. Multi-Scale Feature Fusion Module

The pedestrian re-identification task faces challenges such as occlusion and illumination variations, making feature extraction and matching difficult. Existing methods often utilize either global or local features for matching, but a single type of feature description is usually insufficient to handle complex real-world scenarios. Zhao et al. proposed an attention-based model [11] that focuses on unoccluded parts to extract effective features. However, the attention mechanism has limited effectiveness when dealing with large occlusions. Wei et al. attempted to mitigate occlusion issues by dividing the image into multiple regions for separate feature extraction [12]. Nevertheless, this method struggles to effectively extract complete global features in complex occlusion scenarios.

The Multi-Scale Feature Fusion Module (MSFF) proposed in this paper extracts multi-scale features in local regions, partially compensating for information loss due to occlusion, and enhancing the model's recognition capability under partial occlusion. Additionally, by segmenting and fusing multi-scale features, the MSFF can capture feature information of pedestrians at different scales, thereby improving the model's robustness and recognition accuracy.

As shown in Figure 2, this module first divides the feature maps output by the transposed convolution module into several strips, each strip containing only a part of the pedestrian's region. This allows for more detailed feature extraction from these regions, facilitating the capture of local detail features. When facing local occlusion, only the feature extraction of certain strips is affected, while the overall feature map expression remains intact. By extracting features from the unoccluded parts of the strips, more effective information can be retained. Let NSh represent the number of strips in the height direction and NSw represent the number of strips in the width direction. The strip height Sh and strip width Sw can be expressed as Equations 11 and 12, respectively:

$$Sh = \frac{H}{NSh} \tag{11}$$

$$Sw = \frac{H}{NSw} \tag{12}$$

Let i and j denote the indices of the strips in the height and width directions, respectively. The feature map F is segmented to obtain local features Fl, as expressed in Equation 13:

$$Fl_{i,j} = F\left[:,:,i \cdot Sh:(i+1) \cdot Sh, j \cdot Sw:(j+1) \cdot Sw\right] \tag{13}$$

Next, each local feature strip undergoes multi-resolution pooling to extract features. By using pooling kernels of various sizes, the model can extract features at different scales. This diversity in feature extraction enhances the model's discriminative ability, enabling it to better differentiate pedestrians based on local details. Let $P_{x,y,k}$ represent the output feature of the multi-resolution max pooling for the x-th row and y-th column local feature strip using the k-th pooling kernel size. Let $L_{x,y}$ denote the x-th row and y-th column local feature strip. Let M represent the max pooling operation and A the adaptive average pooling operation. The multi-resolution max pooling for each local feature can be expressed as Equation 14:

$$P_{x,y,k} = A\left(M\left(L_{x,y},k\right),(1,1)\right) \tag{14}$$

The pooling kernel sizes for this module are set to (2×2), (3×3), and (5×5). The primary purpose of defining small-scale pooling kernels is to capture fine-grained local features. These fine-grained features are crucial for distinguishing pedestrians in local regions that appear similar but contain different details. Medium-scale pooling kernels are defined to capture broader features while maintaining attention to detail, which helps the model understand the global context and structure of the image. Large-scale pooling kernels are used to capture even broader contextual information and global features, providing the model with a comprehensive understanding of the pedestrian's overall shape and context.

Finally, global multi-resolution pooling is used to compare the pooled local features and global features, enhancing the model's ability to differentiate between local and global variations. Let $G_{x,y,k}$ represent the discrepancy feature of the x-th row and y-th column local feature strip using the k-th pooling kernel size. Let $C_{x,y,k}$ denote the global feature of the x-th row and y-th column local feature strip after adaptive average pooling using the k-th pooling kernel size. The output feature after global comparison pooling can be expressed as Equation 15:

$$C_{x,y,k} = G_{x,y,k} - A\left(P_{x,y,k},(H,W)\right) \tag{15}$$

This approach enhances the richness and diversity of feature representation, thereby improving the model's expressive capability. It also increases the model's robustness to variations such as deformation and occlusion, enabling better performance in complex scenarios. The schematic diagram of this method is illustrated in Figure 6.

It is recommended that footnotes be avoided (except for the unnumbered footnote with the receipt date on the first page). Instead, try to integrate the footnote information into the text and the reference part.

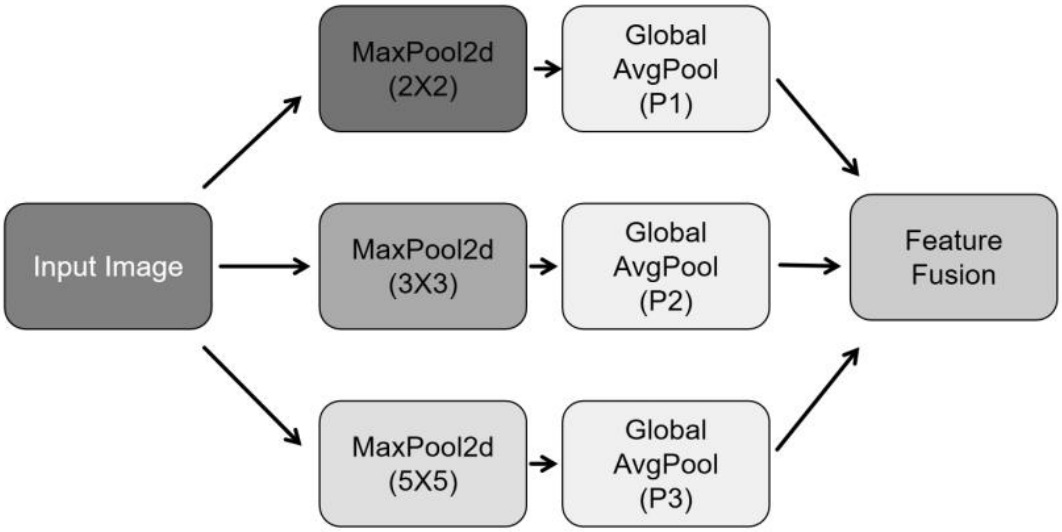


Fig.6. Global murti-resolution contrast pooling schematic

E. Training Loss

In the process of training deep learning models, the loss function is a crucial component used to measure the discrepancy between the model's predictions and the actual labels. To train the model effectively, this paper employs the cross-entropy loss L_{ce} and the triplet loss $L_{triplet}$. These are balanced using the parameter λ. The total loss function is represented by Equation 16:

$$L = L_{triplet} + \lambda L_{ce} \tag{16}$$

The cross-entropy loss is used to measure the difference between the predicted classification and the true classification. It is defined as Equation 17:

$$L_{ce} = -\sum_{n=1}^{N} \sum_{i} y_n \log \hat{y}_{ni} \tag{17}$$

where N represents the number of images in a mini-batch, y_n is the true identity label, and $\hat{y}_{ni}$ is the predicted identity label for each feature q_i. This is defined in Equation 18:

$$\hat{y}_{ni} = \arg\max_{c \in K} \frac{\exp\left(\left(w_c^i\right)^T q_i\right)}{\sum_{k=1}^{K} \exp\left(\left(w_k^i\right)^T q_i\right)} \tag{18}$$

where K represents the number of identity labels, and w_i^k is the classifier for feature q_i and label k. The classifier is implemented using a fully connected layer.

The triplet loss is used to enhance the discriminative ability of the model, ensuring that the features of the same class are closer to each other than those of different classes [13]. It is defined in Equation 19:

$$L = \sum_{i=1}^{N} [\| f(x_i^a) - f(x_i^p) \|_2^2 - \| f(x_i^a) - f(x_i^n) \|_2^2 + \alpha]_+ \tag{19}$$

Where x_i^a is the anchor sample in the i-th triplet, x_i^p is the positive sample in the i-th triplet, x_i^n is the negative sample in the i-th triplet, $f(x)$ represents the embedding function of sample x, $\|\cdot\|_2$ denotes the L2 norm, α is a hyperparameter, and J is used to control the margin between positive and negative sample distances.

IV. Experiments and Results Analysis

The experimental platform consists of both hardware and software components. On the hardware side, an Intel Core i7-11700 CPU and an NVIDIA GeForce GTX 3070 GPU were used. The software environment utilized the PyTorch 1.8-GPU deep learning framework and PyCharm Community IDE, facilitating the development and training of experiments. During the experiments, the batch size was set to 64 to fully leverage the GPU's computational resources, accelerating the training process without causing memory overflow or wasting computational resources. A larger batch size typically leads to more stable gradient updates, aiding faster model convergence. The number of strips for feature map segmentation was set to 6 to enhance the model's fine-grained feature extraction capability. The number of output channels for the local convolutional layer was set to 256, ensuring the model can capture detailed features while maintaining high computational efficiency.

A. Dataset Selection

This paper uses the Market-1501 [14], DukeMTMC-reID [15], and Occluded-DukeMTMC [16] datasets. The relevant information about these datasets is shown in Table I. The Market-1501 dataset is widely used for research and evaluation in pedestrian re-identification tasks, particularly for training and testing the performance of deep learning models. The DukeMTMC-reID dataset features complex imaging environments, encompassing various lighting conditions, occlusions, and pedestrian poses, which increase the difficulty and challenge of pedestrian re-identification. This dataset is primarily used for training and evaluating the performance of pedestrian re-identification models, allowing researchers to validate their models' performance across multiple cameras and in complex environments. The Occluded-DukeMTMC dataset is mainly used to assess the performance of pedestrian re-identification methods under occlusion conditions. The experimental results on this dataset demonstrate the superiority of new methods in occluded scenarios, and these methods do not require manual cropping during preprocessing.

TABLE I
COMPARED WITH DATASET

Dataset	Number of pedestrians	Training Images	Testing Images
Market-1501	1501	12936	19732
DukeMTMC-reID	1812	16522	19889
Occluded-DukeMTMC	1110	15618	2210

B. Experimental Evaluation Criteria

This study employs Mean Average Precision (mAP) and Rank-n as evaluation metrics. mAP is a key metric for assessing a model's performance in retrieval tasks, combining the advantages of Precision and Recall. It reflects the average precision (AP) over multiple queries, thereby offering a comprehensive measure of retrieval accuracy. Conversely, Rank-n is crucial for evaluating retrieval systems, indicating the likelihood of finding the correct result within the top n search results. A higher Rank-n value reflects the model's improved ability to identify the correct result among the top n candidates. The mathematical formula for mAP is provided in Equation 21:

$$mAP = \frac{\sum_{i=1}^{m} AP_i}{m} \tag{20}$$

Where m represents the total number of categories, and AP_i denotes the average precision of the i-th category.

C. Experimental Analysis

Figure 7 displays the loss curves before and after applying the proposed method. A comparison of the original and improved method reveals that the proposed approach achieves faster convergence, allowing the model to reach the optimal solution more quickly, thereby accelerating the training process. Moreover, the improved method exhibits greater stability, being less affected by fluctuations in the training data, and demonstrates enhanced robustness.

Figures 8 through 10 compare Rank-n and mAP between our model and the baseline model on the Market1501, DukeMTMC-reID, and Occluded-DukeMTMC datasets. The results show that our model consistently outperforms the baseline in both performance and ranking accuracy across all datasets. Additionally, our model identifies target individuals more quickly and accurately than the original model, with the improvement being particularly significant on the Occluded-DukeMTMC dataset. This suggests that our model is better suited for complex pedestrian re-identification tasks involving occlusions.

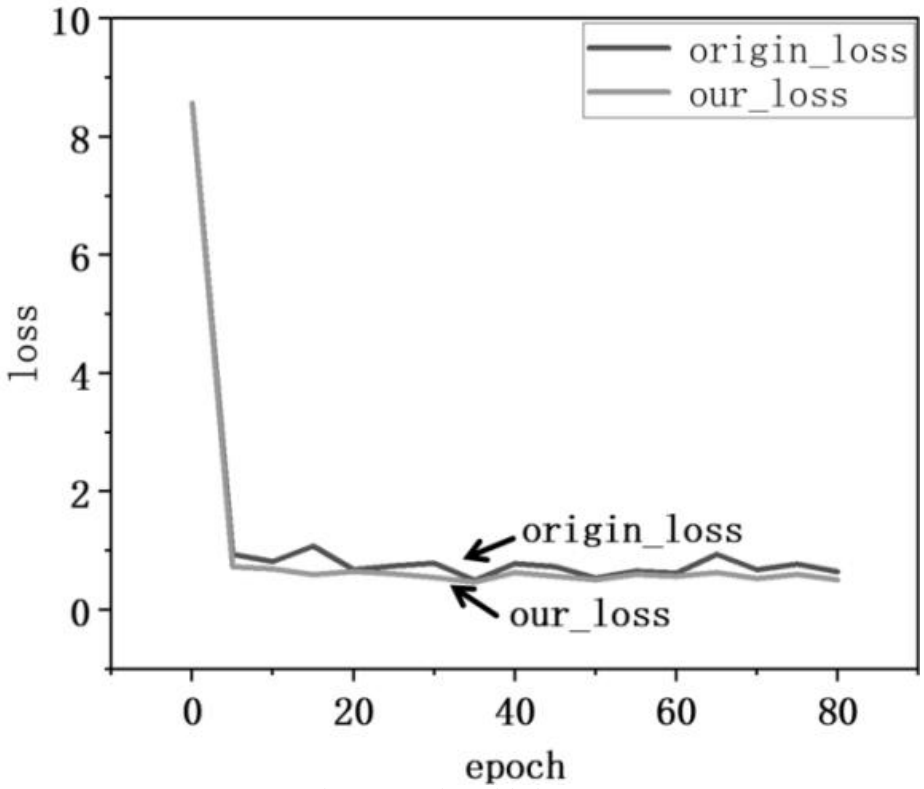


Fig.7. Comparison of the loss curves

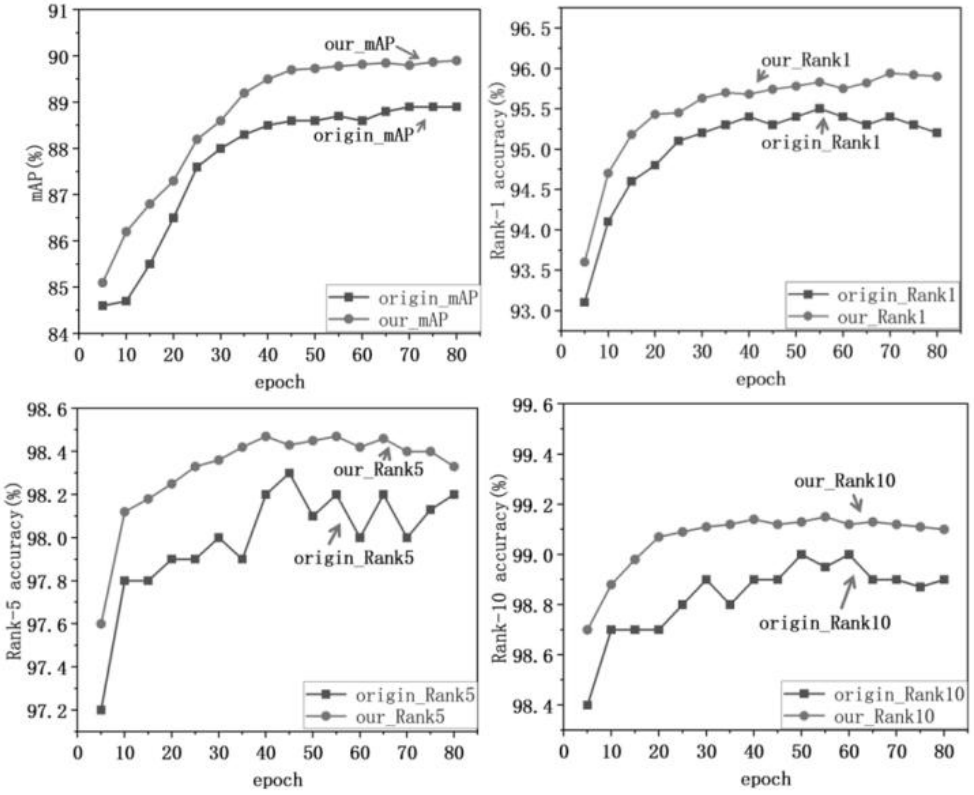


Fig.8. Point-line chart comparing mAP and Rank-n （Market-1501）

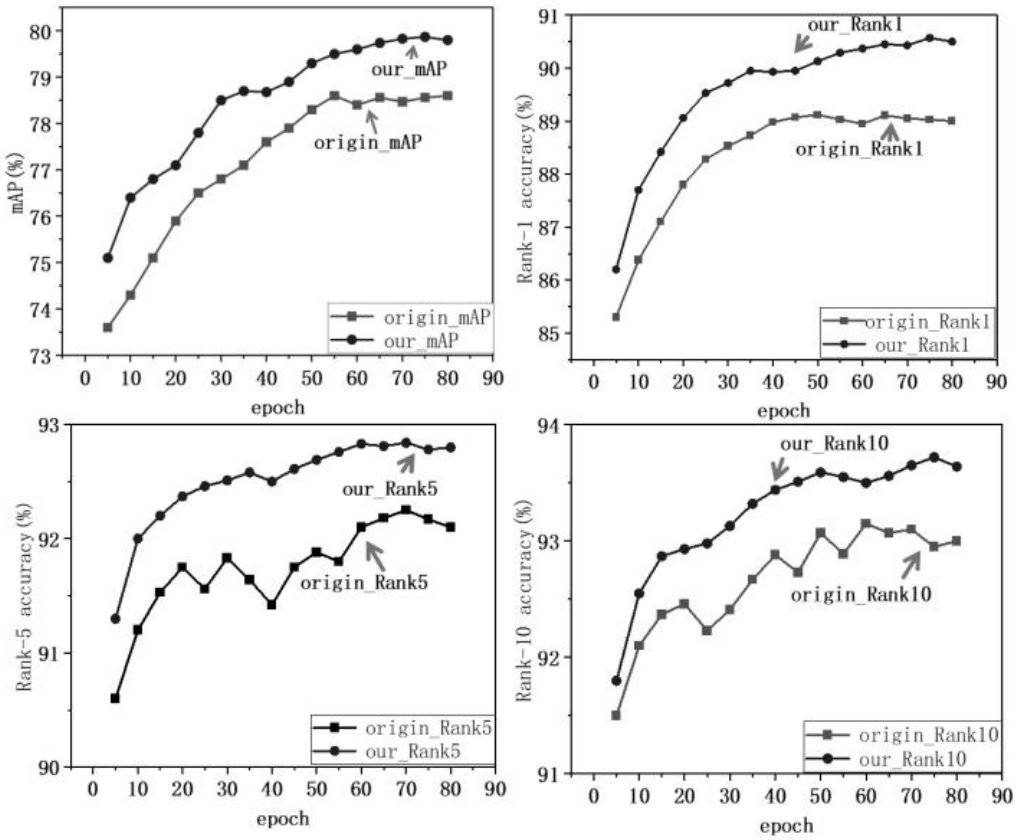


Fig.9. Point-line chart comparing mAP and Rank-n （DukeMTMC-reID）

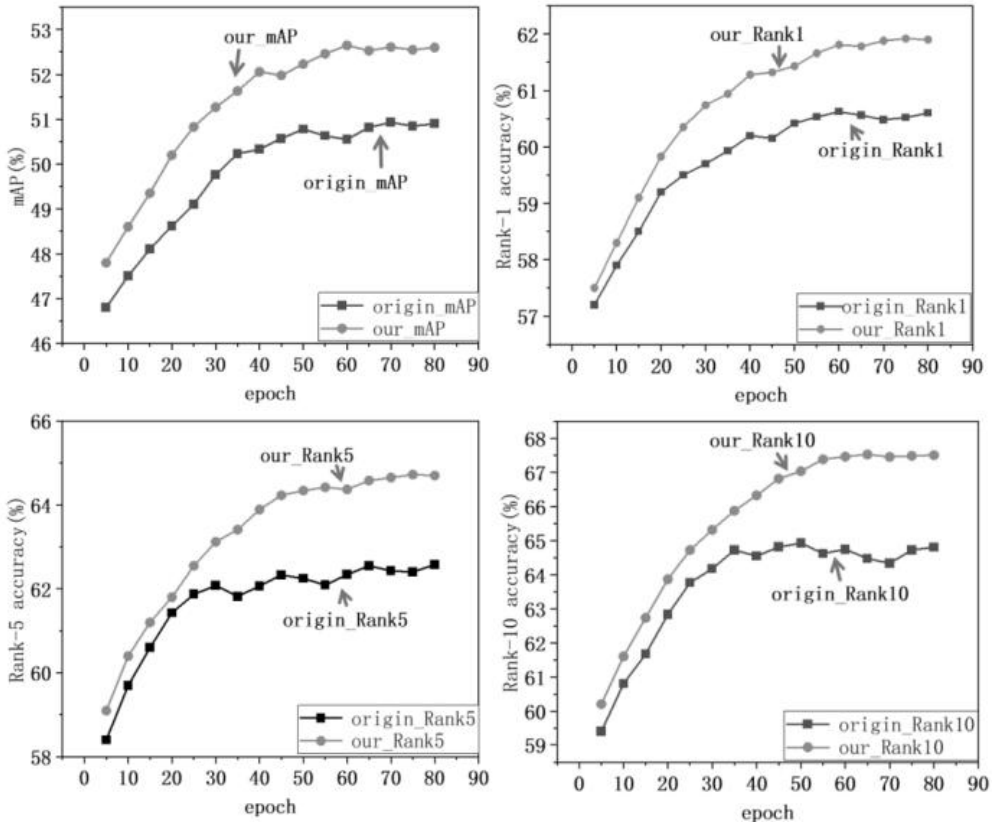


Fig.10. Point-line chart comparing mAP and Rank-n （Occluded-DukeMTMC）

D. Ablation experiment

To validate the effectiveness of the proposed improvements for pedestrian re-identification, a set of ablation experiments was designed to compare the following scenarios: (1) the original RNFPR model, (2) the RNFPR model with the AAC module applied, (3) the RNFPR model with the MSFF module applied, and (4) the RNFPR model with the OPR-MSFF method applied. These experiments were conducted under identical conditions on the DukeMTMC-reID dataset, and the specific performance results are shown in Table II.

As indicated in the table, merely adding the AAC module to the RNFPR model resulted in a slight increase of 0.3 percentage points in mAP and 0.1 percentage points in Rank-1. Incorporating the MSFF module into the RNFPR model led to a 0.9 percentage point increase in mAP and a 0.4 percentage point increase in Rank-1. These changes, while positive, were not highly significant individually. However, by applying the NEMFF method, there was a notable improvement with a 1.2 percentage point increase in mAP and a 0.8 percentage point increase in Rank-1, along with enhancements in Rank-5 and Rank-10.

E. Comparative Experiments

To thoroughly assess the detection performance of the proposed improved algorithm, we conducted a comparative analysis against several mainstream algorithms, including PCB and PGFA. The experiments were carried out using three widely recognized datasets: Market1501, DukeMTMC-reID, and Occluded-DukeMTMC. The results of this comparative evaluation are summarized in Table III, demonstrating the relative performance of each algorithm in pedestrian re-identification tasks under various conditions, including occlusion and complex backgrounds. These comparisons provide a clear indication of the effectiveness and robustness of our improved algorithm across different benchmark datasets.

TABLE II

ABLATION EXPERIMENT

Model	mAP(%)	Rank-1(%)	Rank-5(%)	Rank-10(%)
RNFPR	78.6	89.7	92.1	93.1
AAC + RNFPR	78.9	89.8	92.3	93.2
MSFF + RNFPR	79.5	90.1	92.5	93.4
OPR-MSFF + RNFPR	**79.8**	**90.5**	**92.8**	**93.6**

TABLE III
COMPARED WITH ADVANCED ALGORITHMS

Model	Market1501		DukeMTMC-reID		Occluded-DukeMTMC	
	rank-1	mAP	rank-1	mAP	rank-1	mAP
PCB [17]	92.1	77.3	81.8	66.1	47.6	34.7
PGFA [18]	91.3	76.7	82.6	65.5	53.4	38.4
Sun [19]	-	-	-	-	54.9	41.5
ABDNet [20]	95.4	88.2	88.7	78.6	-	-
PISNet [21]	95.6	87.1	88.8	78.7	-	-
CBN [22]	94.3	83.6	84.8	70.1	-	-
ISP [23]	95.3	88.6	89.6	80.0	-	-
HOReID [24]	93.9	84.5	86.9	75.6	55.1	43.8
OAMN [25]	93.5	80.1	86.2	72.6	60.2	50.5
BoT [26]	94.1	86.6	87.2	77.1	57.3	47.3
MGN[27]	95.7	88.6	88.7	77.2	58.4	47.2
IANet[28]	94.5	87.1	87.2	76.1	56.3	44.8
SAN[29]	95.3	88.1	87.9	76.5	57.9	46.5
JDGL[30]	94.8	87.4	87.5	75.8	57.1	45.7
RNFPR[3]	95.2	88.9	89.7	78.6	60.6	50.9
OPR-MSFF (ours)	**95.9**	**89.8**	**90.5**	**79.8**	**61.9**	**52.6**

As presented in the table, the proposed method consistently achieves superior Rank-n and mAP scores compared to other pedestrian re-identification algorithms. This performance advantage underscores the effectiveness of the enhancements introduced by our method, as it consistently outperforms existing algorithms. The results on the DukeMTMC-reID dataset further validate the method's efficacy, demonstrating its applicability and robustness across various pedestrian re-identification datasets. Notably, the strong performance on the Occluded-DukeMTMC dataset highlights the method's enhanced adaptability and generalization capabilities, particularly in handling pedestrian re-identification tasks involving occlusions.

These findings emphasize the considerable advantages of the proposed method in pedestrian re-identification, showcasing not only outstanding performance across multiple datasets but also demonstrating remarkable stability and consistency during the training process. The demonstrated effectiveness suggests that the proposed method holds significant potential and practical value within the domain of pedestrian re-identification.

To provide a more intuitive illustration of our method's effectiveness, retrieval experiments were conducted using the Occluded-DukeMTMC dataset. The results of these retrieval experiments are depicted in Figure 11. In the figure, standard borders represent correct retrieval outcomes, while bold black borders signify incorrect results. It is evident that the detection accuracy remains exceptionally high when pedestrian features are either unobstructed or only partially occluded, with misdetections predominantly occurring in scenarios where pedestrian features are heavily occluded.

Fig.11. Visualization of retrieval results

V. Conclusion

This paper presents an occluded pedestrian re-identification method based on multi-scale feature fusion, designed to address challenges such as occlusion, illumination variations, and complex backgrounds. The core elements of this approach include the AAC module, the integration of EfficientNetB0 and DaViT_Small networks, and the multi-scale feature fusion module. Together, these innovations enhance the model's robustness and generalization by effectively extracting and fusing multi-scale features, leading to improved overall performance.

The AAC module enhances robustness and generalization by injecting noise into the central region of the input image, encouraging the model to focus on a broader range of image features. EfficientNetB0 processes the original image, while DaViT_Small handles the concatenated noisy image, capturing high-level features and improving spatial resolution through transposed convolutions. The multi-scale feature fusion module further refines feature extraction by splitting and merging feature maps at multiple resolutions, ensuring the model captures a comprehensive set of feature information.

Experimental evaluations on the Market1501, DukeMTMC-reID, and Occluded-DukeMTMC datasets demonstrate that the proposed OPR-MSFF method improves Rank-1 accuracy by 0.7%, 0.8%, and 1.3%, and mAP by 0.9%, 1.2%, and 1.7%, respectively. These results highlight the effectiveness of the proposed method, particularly in handling re-identification scenarios involving significant occlusion.

Future work will focus on optimizing the algorithm to address challenges in small object detection and exploring its application in other domains that require robust feature extraction and fusion techniques.

References

[1] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person Re-identification: Past, Present and Future," arXiv preprint arXiv:1610.02984, 2016.
[2] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 2197-2206.
[3] H. Park and B. Ham, "Relation Network for Person Re-Identification," Proc. AAAI Conf. Artif. Intell., vol. 34, no. 07, pp. 11839-11847, 2020.
[4] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," ArXiv, abs/1905.11946, 2019.
[5] M. Ding, B. Xiao, N. C. F. Codella, P. Luo, J. Wang, and L. Yuan, "DaViT: Dual Attention Vision Transformers," in Proc. Eur. Conf. Comput. Vis., 2022.
[6] M. Lin, Q. Chen, and S. Yan, "Network In Network," CoRR, vol. abs/1312.4400, 2013.
[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in 2016 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770-778.
[8] J. Miao et al., "ResNet based on feature-inspired gating strategy," Multimed. Tools Appl., vol. 81, pp. 19283-19300, 2021.
[9] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning," arXiv preprint arXiv:1603.07285, 2016.
[10] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015.
[11] H. Zhao et al., "Spindle Net: Person Re-Identification with Human Body Region Guided Feature Decomposition and Fusion," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 1077-1085.
[12] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person Transfer GAN to Bridge Domain Gap for Person Re-Identification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 79-88.

[13] A. Hermans, L. Beyer, and B. Leibe, "In Defense of the Triplet Loss for Person Re-Identification," arXiv preprint arXiv:1703.07737, 2017.
[14] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable Person Re-identification: A Benchmark," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Santiago, Chile, 2015, pp. 1116-1124, doi: 10.1109/ICCV.2015.133.
[15] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in Vitro," in 2017 IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 3774-3782.
[16] X. Wang, Y. Jin, X. Li, C. Yan, and Y. Wang, "High-Order Information Matters: Learning Relation and Topology for Occluded Person Re-identification," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2024, pp. 1234-1243.
[17] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond Part Models: Person Retrieval with Refined Part Pooling," in Eur. Conf. Comput. Vis., 2017.
[18] J. X. Miao et al., "Pose-guided feature alignment for occluded person re-identification," in 2019 IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2019, pp. 542-551.
[19] Y. B. Sun and R. Wang, "Pedestrian re-identification method based on feature correlation and multi-loss fusion," China Sci. Pap., vol. 17, no. 3, pp. 233-239, 2022.
[20] K. Zeng, M. Ning, Y. Wang, and Y. Guo, "Hierarchical Clustering With Hard-Batch Triplet Loss for Person Re-Identification," in 2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, 2020, pp. 13654-13662, doi: 10.1109/CVPR42600.2020.01367.
[21] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable Person Re-identification: A Benchmark," in 2015 IEEE Int. Conf. Comput. Vis. (ICCV), Santiago, Chile, 2015, pp. 1116-1124, doi: 10.1109/ICCV.2015.133.
[22] E. Ristani, F. Solera, R. S. Zou, R. Cucchiara, and C. Tomasi, "Performance Measures and a Data Set for Multi-target, Multi-camera Tracking," in ECCV Workshops, 2016.
[23] Z. Zhuang et al., "Disassembling the Dataset: A Camera Alignment Mechanism for Multiple Tasks in Person Re-identification," ArXiv, abs/2001.08680, 2020.
[24] G. Wang et al., "High-Order Information Matters: Learning Relation and Topology for Occluded Person Re-Identification," in 2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 6448-6457.
[25] P. X. Chen et al., "Occlude them all: Occlusion-aware attention network for occluded person re-ID," in 2021 IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 11813-11822.
[26] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of Tricks and a Strong Baseline for Deep Person Re-Identification," in 2019 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2019, pp. 1487-1495.
[27] G. Wang, Y. Yuan, X. Chen, and J. Li, "Learning multi-granularity representations for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 1335-1344.
[28] R. Hou, H. Chang, B. Ma, S. Shan, and X. Chen, "Interaction-and-aggregation network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 9317-9326.
[29] L. Zhang, T. Xiang, and S. Gong, "Self-attention network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 988-997.
[30] T. Zhang, S. Li, Z. Zhang, and Z. Wang, "JDGL: Joint discriminative and generative learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 3377-3386.