Research on Occluded Person Re-identification Algorithm Based on Body Part Features

Zheng Li,Yang Xu

Abstract—**Occluded person re-identification (ReID) constitutes a critical challenge in computer vision, requiring accurate matching between occluded pedestrian images and their non-occluded counterparts. While part-based models have demonstrated potential through fine-grained feature representation for partial human body analysis, existing approaches exhibit limitations in two key aspects:(1) inadequate exploitation of discriminative local features under occlusion conditions, and (2) insufficient utilization of unannotated visual information. To address these challenges, we propose BPSEMA-ReID, a novel framework integrating two innovative components: a self-supervised body part attention module and a dynamic feature refinement strategy. Our architecture automatically generates discriminative part features without reliance on topological priors, employing a hierarchical feature learning approach that combines local semantic analysis with global context preservation. Specifically, the model decomposes input images into five semantic regions through adaptive feature partitioning, followed by spatial-channel attention fusion to construct comprehensive feature representations. Extensive experimental evaluations on Occluded-Duke and DukeMTMC-reID benchmarks demonstrate the framework's superior performance in occlusion scenarios. Notably, our method achieves state-of-the-art results on the Occluded-Duke dataset, with 68.55% Rank-1 accuracy and 54.2% mAP, outperforming existing part-based approaches by 1.85% and 0.1% respectively. The proposed solution effectively bridges the feature representation gap between occluded and holistic pedestrian images while maintaining computational efficiency.**

Index Terms—**Person re-identification; Computer vision; Fine grained feature representation; Attention**

I. Introduction

Pedestrian re-identification [1], [2] constitutes a pivotal function within the realm of computer vision, designed to retrieve images of an identical pedestrian from extensive image repositories that correspond to a specified query image. This technology finds extensive application in domains such as intelligent surveillance, public safety, and video analytics, particularly playing a critical role in the infrastructure of smart cities and security frameworks [3], [4]. Despite the myriad approaches attempted to address this challenge, pedestrian images are profoundly influenced by variables including viewpoint, illumination, pose variations, and occlusions, resulting in significant discrepancies and rendering pedestrian re-identification persistently daunting. Furthermore, the vast quantities of pedestrian images housed in databases necessitate the development of efficient and precise matching methodologies, a subject that has garnered considerable attention in contemporary research. Consequently, pedestrian re-identification mandates not only efficacious feature representation but also robust matching strategies to navigate the complexities inherent in real-world applications.

To address the aforementioned issues, the part-based approach [5-7] has been extensively employed in pedestrian re-identification. These methodologies segment pedestrian images into several distinct regions and independently extract features for each, thereby crafting a more nuanced and detailed representation. Nevertheless, this part-based representation technique encounters two primary challenges:

1. Insufficiency of Discriminative Local Attributes: Conventional pedestrian re-identification loss functions, such as identity loss or triplet loss, presuppose a pronounced disparity in the visual profiles of pedestrians with distinct identities, thereby necessitating divergent global feature vectors. Nevertheless, this premise falters when employing part-based feature vectors, as individuals of differing identities may exhibit strikingly similar traits in specific bodily regions. Given that local features often lack discriminative prowess, the application of traditional global feature learning losses to local feature learning proves problematic. Prior methodologies in part-based pedestrian re-identification have frequently neglected the unique challenges of local feature learning and their implications for loss function selection, a gap we address by proposing a novel solution. To this end, we utilize a training loss termed GiLet. GiLet adeptly mitigates the issues of occlusion and the dearth of discriminative local attributes, striving to cultivate a suite of local features wherein each element effectively characterizes its corresponding region while collectively upholding robust discriminative capability.

2. Insufficiency of Human Topology Annotations: Techniques predicated on body segments commonly depend on spatial attention maps to execute localized pooling from comprehensive feature maps, thereby extracting the target pedestrian's body part characteristics. Nevertheless, existing pedestrian re-identification datasets fall short in providing annotation data for localized region pooling. The creation of such annotations via auxiliary tools, such as pose estimation or part segmentation utilities, frequently encounters challenges in maintaining precision, attributable to inter-domain discrepancies and variances in image quality.Moreover, body part-centric feature pooling fundamentally diverges from pixel-level human parsing: spatial attention maps must not only pinpoint body segments within the image but also distill feature vectors that optimally

Manuscript received Mar 21, 2025; revised June 29, 2025. This research was funded bythe project of the Liaoning Provincial Department of Education (ProjectNumber: LJKZ0310).

Zheng Li is a master's student of University of Science and Technology Liaoning, Anshan, Liaoning, China. (corresponding author to provide phone: +086-150-0678-2683; e-mail: 15006782683@163.com).

Yang Xu is a Professor of University of Science and Technology Liaoning, Anshan, Liaoning, China. (corresponding author to provide phone: +086-138-8978-5726; e-mail: xuyang 198l@aliyun.com).

encapsulate the distinctive appearance of those segments. Consequently, the optimal attention map need not correspond precisely to a human segmentation outline. Previous pedestrian re-identification research endeavors have sought to leverage human parsing for generating part-based features, predominantly employing two methodologies:(1) acquiring local features through part discovery, independent of human topology priors [7-9]; (2) directly utilizing the outputs of pose estimation models as local attention masks, albeit without customization for the pedestrian re-identification task [10-12].

In this research, we introduce a body-part-centric attention module, refined via an innovative dual supervision mechanism. This module exemplifies the seamless integration of extrinsic human semantic data to produce refined body-part attention maps, specifically tailored for enhancing person re-identification processes.

Ultimately, we combine the body part attention module with the GiLet (Global-identity Local-Ema-triplet) loss and propose a body part-based pedestrian re-identification model, BPSEMA-ReID. This model successfully addresses the main challenges mentioned earlier.

II. Network structure

The network structure adopts the same backbone structure as the baseline network [13], namely the architecture used in Torchreid [14]. This network is built based on Torchreid and designed two main modules for the task of occluded person re-identification: the body part multi-attention module and the dynamic feature refinement module.

The Body Part Multi-Attention Module receives the feature map extracted by the backbone network as input, producing a suite of attention maps that accentuate the pedestrian target's body parts. This module incorporates a pixel-level part classifier and is refined via body part attention loss, grounded in coarse human parsing annotations. Given the model's end-to-end training regimen [7, 35-37], the module is not solely guided by body part prediction but also benefits from supplementary training cues through ReID loss. Consequently, the Body Part Attention Module integrates a dual supervision mechanism, concurrently optimizing both part prediction and pedestrian re-identification objectives. This dual oversight ensures that the module's generated attention maps are more effectively tailored to the pedestrian re-identification task.

The Dynamic Feature Refinement Module adeptly diminishes feature dimensions and enhances feature representation through an integrated application of convolution, batch normalization, ReLU activation, channel attention mechanisms, and residual connections. This is pivotal for the development of streamlined, high-precision deep learning architectures.

The overall work framework is shown in Fig. 1.

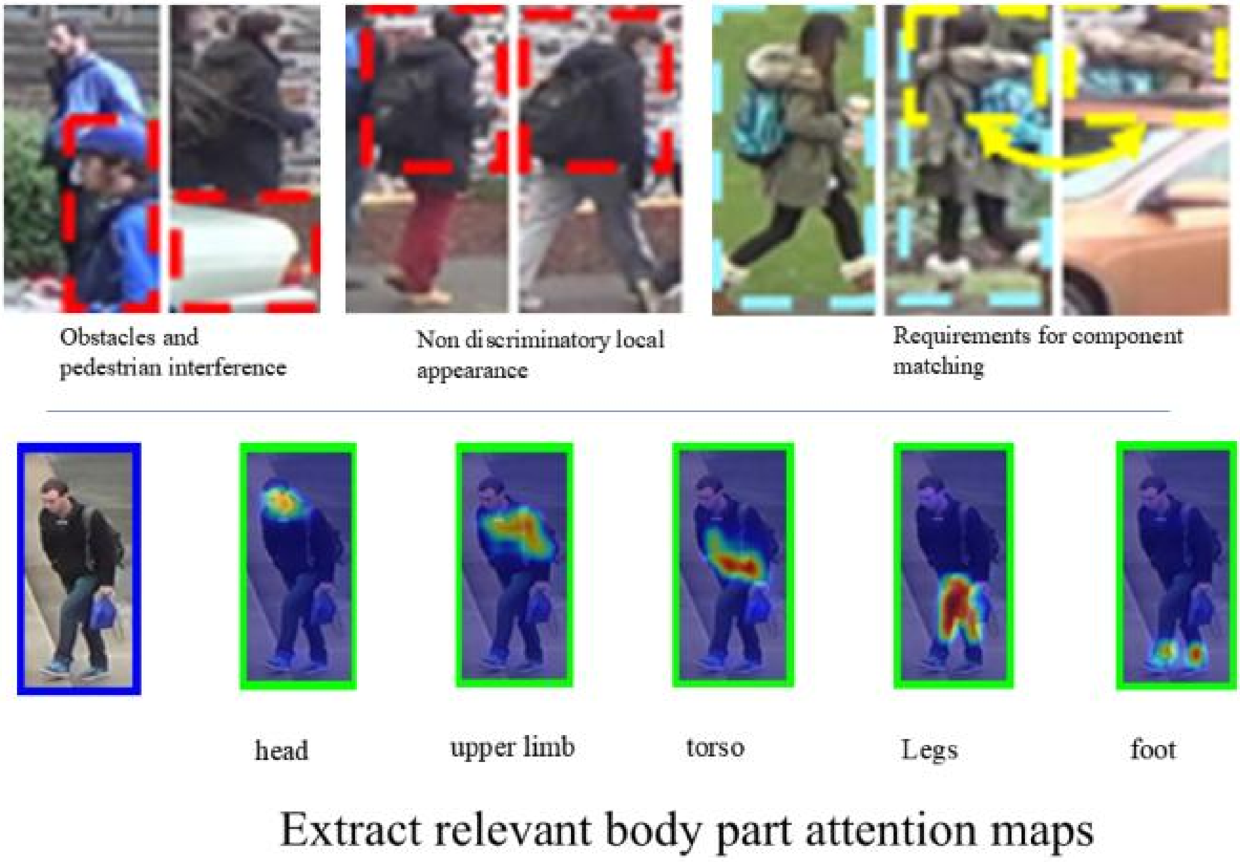
Obstacles and
pedestrian interference
Non di scriminatory local
appearance
Requirements for component
matching
head
upper limb
torso
Legs
foot
Extract relevant body part attention maps

Fig.1. Overall Framework Diagram

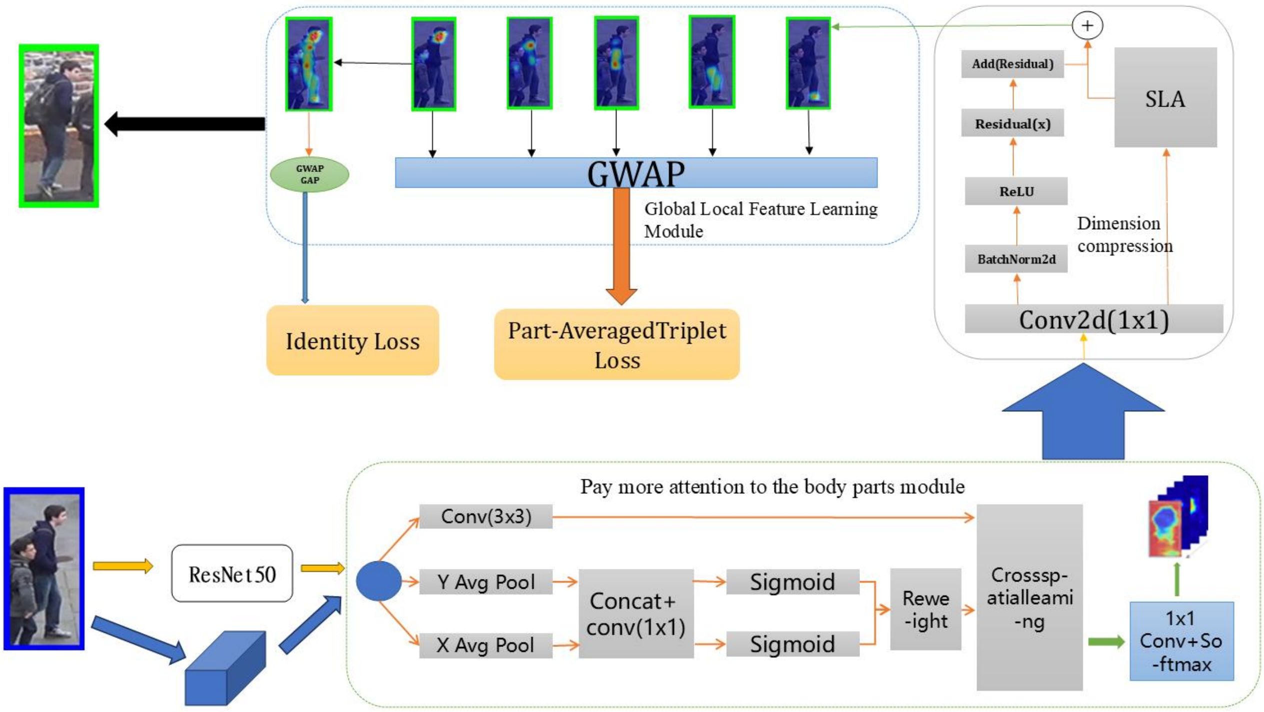

Fig.2. Overall Network Framework Diagram

A. *Overall Network Framework*

The overall framework of the network is shown in Fig.2.

The input data is a set of images containing multiple pedestrians $I=\{I_1, I_2, \cdots, I_n\}$, each image $Ii \in R^{C\times H\times W}$ For an RGB image with C channels, H height, and W width. Before inputting into the model, all images undergo normalization and data augmentation (such as cropping, scaling, flipping, etc.) to enhance the robustness of the model. ResNet-50 [33] is used as the feature extraction network, which employs residual connections to construct deep networks, thereby effectively learning image features. Let the backbone network be B, then the operation of extracting spatial features from image I is represented by formula (1):

$$F = B(I) \quad (1)$$

Among $F \in R^{N\times D\times Hf\times Wf}$ the feature tensor extracted by ResNet-50: D represents the output feature dimension of ResNet-50, typically 2048 (before the last fully connected layer of ResNet-50). Hf, Wf The height and width of the feature map, usually determined by the size of the input image and the convolution and pooling operations of the backbone network.After processing by the backbone network B, the output feature F is a feature map containing global spatial information.

The feature map Fi will undergo further processing through the dynamic attention module. Here, we introduce the EMA [34] attention mechanism.

The model processes the feature map through a foreground and background segmentation mechanism. It uses a pixel classifier to generate a probability distribution map for each component and performs local feature extraction for each component in the image. Then, it weightedly sums the feature map using the segmented foreground and background regions, followed by pooling operations on all features, including global pooling, foreground pooling, and component-level pooling.

The pooling results are concatenated into a large vector through a connection operation $g^{\text{concat-part}}$, i, The principle is formula (2):

$$g_{concat-part,i} = Concat(g_{part,i,1}, g_{part,i,2}, \ldots, g_{part,i,k}) \quad (2)$$

Ultimately, all features (including global features, foreground features, and component features) are input into multiple classifiers to obtain the category prediction for each feature.

These classification results will pass through the Softmax function, aggregating these local features into global features, and ultimately outputting the predicted identity of the pedestrian.

B. *Body Part Multi-Attention Module*

This module enhances the performance of pedestrian re-identification (ReID) by focusing on different parts of the image (such as the head, torso, legs, etc.). The key purpose of this module is to extract and weight the features of different parts, thereby enabling the model to "concentrate" on important areas for each part. As shown in Fig.3.

First, the model receives an input image and extracts the image features through a backbone network (such as ResNet-50 or HRNet). Then, a part classifier (pixel_classifier) is used to predict which part (head, torso, legs, etc.) each pixel in the image belongs to. This process generates a part mask or a probability map of part labels [41].

The original model applies a pixel-level part classifier in the body part attention module, as shown in formula (3):

$$\mathrm{M} = \mathrm{softmax}（\mathrm{GP^T}） \quad (3)$$

The body part attention module receives the appearance map G generated by the feature extractor as input, where the appearance map G is a tensor of dimensions ℝ^H×W×C. By applying a 1x1 convolutional layer with parameters P∈ℝ^(K+1)×C to G, followed by a softmax operation, M∈ℝ^H×W×(K+1) is obtained.

This paper's model, based on its original model, introduces the EMA [34] attention mechanism. The core idea of the EMA attention mechanism is to calculate a "smooth" version of the current feature map through weighted averaging. For the feature map Fi, the EMA operation can be represented as Formula (4):

$$\boldsymbol{F_i^{ema}} = \boldsymbol{\alpha} \cdot \boldsymbol{F_i} + (\boldsymbol{1} - \boldsymbol{\alpha}) \cdot \boldsymbol{F_i^{ema-prev}} \quad (4)$$

The text after the EMA module update is the feature map; α is the decay factor that controls the degree of smoothing, usually chosen as α ∈[0, 1]; is the EMA feature map from the previous round (which can be regarded as the cumulative information of historical feature maps). This allows for more focused feature extraction of the image on the original basis, resulting in more ideal final recognition results.

According to the output of the part classifier, the model generates masks for different parts. Each part's mask represents the spatial region of that part in the input image. Component Mask（Part Mask）：Through the output of the component classifier $p_{part}(x)$， Generate a mask for each component M_k, formula (5):

$$M_k(x) = \begin{cases} 1 & if\ \ p_{part}(x) = k \\ 0 & \quad\quad otherwise \end{cases} \quad (5)$$

Among, $M_k(x)$ The text is the mask of component k at position x.

Next, based on the mask of each component, generate a corresponding attention map. This map represents the influence weight of each component on the final feature representation. The generation of the attention map usually depends on the weighted operation of the feature map and the component mask. Attention Map: Use the component mask and the feature map (extracted from the backbone network) to generate an attention map for each component. This process can be implemented through convolutional layers or other learnable methods.

Using component masks and attention maps, the feature maps extracted by the backbone network are weighted. The weighted feature maps represent the model's attention to each component, thereby helping to highlight the features of important parts. Weighted Feature Map: For each component, the feature map is weighted based on its attention map and mask.

Then, the weighted features of each component are aggregated (e.g., pooling) to obtain the feature representation of each component. The aggregation methods can include Global Average Pooling (GAP), Max Pooling, etc. Finally, the features of each component are fed into a classifier (usually a fully connected layer) to generate the final classification result (i.e., the identity of the pedestrian). These component feature outputs are combined to achieve a local-to-global aggregation process, thereby obtaining the final pedestrian identity prediction. Combine the characteristics of all components f_1, f_2, ⋯, f_K, through a fully connected layer (FC) for identity classification, obtain the final pedestrian identity prediction, formula (6):

$$\hat{y} = FC(f_1, f_2, \ldots, f_k) \quad (6)$$

Among them, $\hat{y}$ the text is the output of the model, indicating the identity of the pedestrian.

C. Dynamic Feature Refinement Module

This module primarily processes the input feature maps and enhances their representation capabilities through a combination of convolutional layers, batch normalization layers (Batch Normalization), ReLU activation, SE attention mechanism [38], and residual connections. The following is a detailed workflow of this module, as shown in Fig. 4.

This module is mainly placed in the dimensionality reduction layer of the meta-model, and its main tasks are: first, convolution operation (Convolution): input feature map $x \in R^{B \times C^{in} \times H \times W}$, Where B is the batch size, C_{in} is the number of input channels H and W are the height and width of the input image, respectively. First, perform a 1x1 convolution operation on the input image, as shown in formula (7):

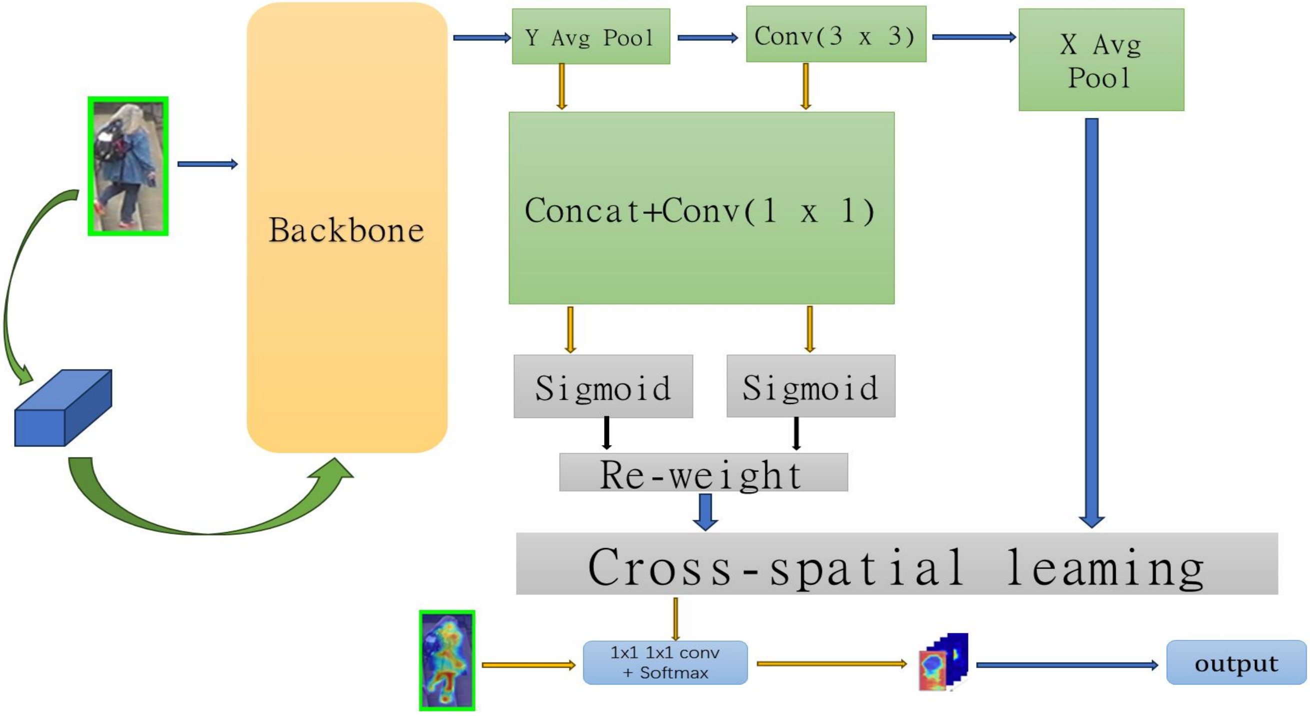

Fig.3. Multi-Attention Module for Body Parts

$$x^{'} = Conv(x) \ \ where \ \ x^{'} \in R^{B\times C_{out}\times H\times W} \tag{7}$$

Among them, the number of output channels for the convolution operation is C_{out}, The size of the convolution kernel is 1×1, without changing the spatial dimensions (stride=1, padding=0).

Next is Batch Normalization: Batch normalization is applied to the output of the convolutional layer, aiming to reduce internal covariate shift and enhance the stability of network training, as shown in formula (8):

$$BN\left(x^{'}\right) = \frac{x^{'}-\mu}{\sigma}\cdot\gamma+\beta \tag{8}$$

Among them, μ and σ are the mean and standard deviation on the batch dimension, respectively, and γ and β are the learned scaling factor and bias term.

Then perform ReLU activation (ReLU Activation): Apply the ReLU activation function to the normalized feature map to increase non-linearity, as shown in formula (9):

$$X^{''} = ReLU(\hat{X}) \tag{9}$$

ReLU activation function is defined as formula (10):

$$ReLU(z) = \max(0, z) \tag{10}$$

Where z is the normalized input.

Next, the SE attention mechanism is introduced: it enhances the expressive power of important channels by weighting the channels of the feature map. First, global average pooling is performed on each channel to generate a global description for each channel, as shown in equation (11):

$$z_c = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} x^{''}_{c,i,j} \tag{11}$$

Among them, $x^{''}_{c,i,j}$ Represents the feature value at the c-th channel, position (i, j).

Then, the channel description is passed into a small fully connected network (i.e., two FC layers), and the weighted coefficients of the channel are output, as shown in formula (12):

$$s_c = \sigma(W_2\cdot ReLU(W_1\cdot z_c)) \tag{12}$$

Among them, W_1 and W_2 are the weight matrix of the FC layer. $\sigma(\cdot)$ it is the Sigmoid activation function. $s_c \in [0,\ 1]$ the weight of the importance of channel c. Finally, apply the weight S_c to each channel, as shown in formula (13):

$$x_{SE} = x^{''}\cdot s_c \tag{13}$$

This step adjusts the contribution of each channel in the feature map by weighting the channels.

Finally, perform the residual connection: The residual connection adds the input feature map x to the feature map weighted by the SE mechanism, thereby preserving more original features in the information flow, as shown in formula (14):

$$x_{out} = x + x_{SE} \tag{14}$$

Among them, $x_{\mathrm{out}} \in R^{B\times C_{\mathrm{out}}\times H\times W}$ the final output feature map.

This module's main task is to further refine and enhance the extracted features, increasing the accuracy of image recognition.

D. Loss Function

This model generates global inter-sample distances by combining multi-part embedding vectors using a mean strategy [42], and is trained in conjunction with a hard triplet loss strategy (batch-hard triplet loss) [39], [40]. Through this method, the model can handle partial occlusion and optimize the feature representation of different parts.

First, calculate the inter-part contrast distance matrix: Each input sample (image) is represented by multiple parts (such as upper body, lower body, etc.), and these parts are separately encoded into embedding vectors. This method first calculates the pairwise distance matrix for each part, resulting in a matrix of shape [K, N, N], where K is the number of parts and N is the number of samples. Calculate the contrast distance matrix for each part, given the embedding matrix E $\in R^{K\times N\times C}$, as formula (15):

$$D_{ij} = ||E_i - E_j||^2 = |E_i|^2 - 2 < E_i, E_j > + |E_j|^2 \tag{15}$$

The Euclidean distance between sample i and sample j is represented by D_{ij}.

Secondly, there is also a distance matrix between combined parts: by merging the distance matrix of each part, the overall distance matrix for each pair of samples is calculated. The distance matrices of each part are combined into a global distance matrix.

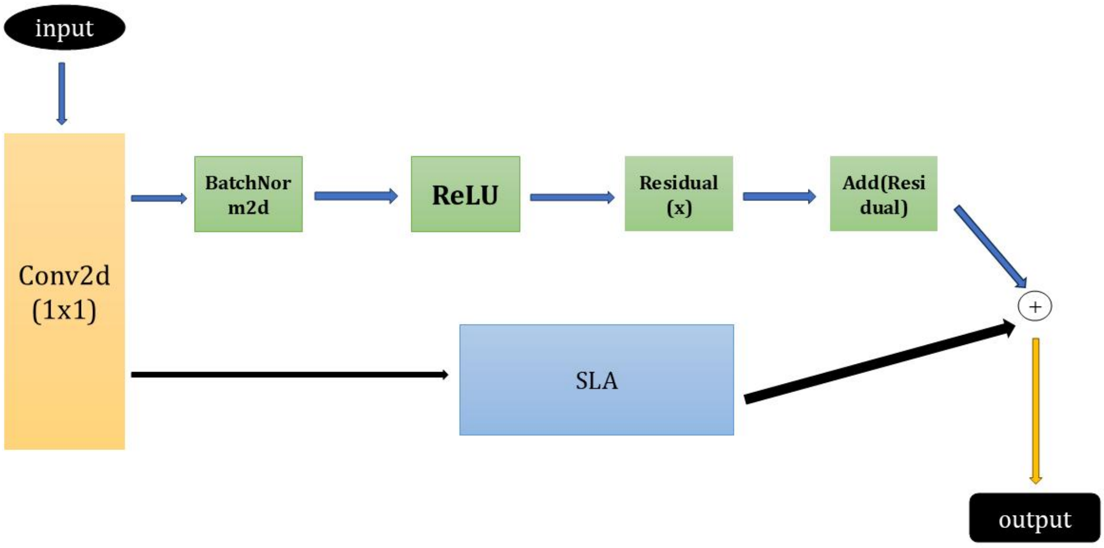

Fig.4. Dynamic Feature Refinement Module

Common merging methods include: calculating the average, maximum, or minimum value. In this method, the "mean" operation is used to calculate the distances between global samples. Additionally, the standard batch hard triplet loss is computed, using the merged pairwise distance matrix to calculate the standard triplet loss. The goal of triplet loss is: given an anchor sample, minimize the distance between it and the positive sample, while maximizing the distance between it and the negative sample, as shown in formula (16):

$$\mathcal{L}_{triplet} = \max\left(D_{positive} - D_{negative} + margin, 0\right) \quad (16)$$

Among them, margin is the given marginal threshold.

The calculation of soft margin triplet loss is also provided, as shown in Equation (17):

$$\mathcal{L}_{soft\,margin} = \log\left(1 + \exp\left(D_{negative} - D_{positive}\right)\right) \quad (17)$$

The soft margin loss function allows for a gradual transition between the distances of negative samples and positive samples, rather than rigidly satisfying the margin condition.

To obtain the batch hard triplet loss, the hardest positive and negative sample pairs must first be selected using the "hard mining" method. Additionally, it supports part occlusion information; if the visibility information of parts is provided, the distance matrix is adjusted based on the number of visible parts.

III. Experimental Results and Analysis

A. *Dataset and Evaluation Metrics*

The experiment is conducted based on three widely used, large-scale datasets, namely DukeMTMC-ReID [15], Occluded-DukeMTMC [16], and P-DukeMTMC [17]. Specifically, the DukeMTMC-reID (Duke Multi-Tracking Multi-Camera Re-Identification) dataset is a subset of DukeMTMC, used for image-based person re-identification. This dataset is created from high-resolution videos from 8 different cameras. It is one of the largest pedestrian image datasets, with images cropped by hand-drawn bounding boxes. The dataset contains 16522 training images of 702 identities, 2228 query images of another 702 identities, and 17661 gallery images; Occluded-DukeMTMC includes 15618 training images, 17661 gallery images, and 2210 occluded query images; P-DukeMTMC-reID is a modified version based on the DukeMTMC-reID dataset, with a training set of 12927 images (665 identities), a test set of 2163 images (634 identities), and a gallery set of 9053 images.

This experiment uses two standard ReID metrics: Cumulative Matching Characteristic (CMC) at Rank-1 and mean Average Precision (mAP). Performance evaluation does not require re-ranking in a single query setting [18]. CMC focuses on the probability of successful matching for the top K images, with the method primarily analyzing four scenarios: Rank-1, Rank-5, Rank-10, and Rank-20, which correspond to the probabilities of successful matching for the top 1, 5, 10, and 20 images, respectively. Additionally, the average precision for each query image is calculated, and based on this, the mean value of the average precision for all query images is summarized, which is the mAP value.

B. *Experimental parameter settings*

To save server resources, the model in this paper uses ResNet-50 as the main backbone feature extractor. The last fully connected layer and the global average pooling layer are removed, and the stride of the last convolutional layer is set to 1. To enable flexible training for different datasets, we also adopt optional backbone feature extractors, such as: ResNet-50-IBN (RI) [20] (e.g., [21-23]) and HRNet-W32 (HR) [24] backbone (e.g., [7]). For the overall dataset, the number of body parts K is set to 5, and for the occluded dataset, it is set to 8.

The training process is mainly based on the BoT [25] method. All images are resized to 384×128; the images are first enhanced by random cropping and 10-pixel padding, and then randomly erased with a 50% probability [26]. Each training batch contains 64 samples from 16 identities, with 4 images per identity. The model is trained end-to-end using the Adam optimizer on an NVIDIA Quadro RTX4060Ti GPU, with a total of 120 epochs. After the 10th epoch of training, the learning rate is linearly increased from 3.5×10^-5 to 3.5×10^-4, and decayed to 3.5×10^-5 and

TABLE I
THE COMPARISON BETWEEN OUR METHOD AND THE LATEST METHODS

Method	Holistic datasets		Occluded datasets			
	DukeMTMC-ReID		Occluded-DukeMTMC		P-DukeMTMC	
	R-1	mAP	R-1	mAP	R-1	mAP
BoT [25]	86.4	76.4	51.4	44.7	87	74.9
SGAM [27]	83.5	67.3	55.1	35.3	–	–
PGFA [28]	82.6	65.5	51.4	37.3	44.2	23.1
MHSA [29]	87.3	73.1	59.7	44.8	70.7	41.1
VGTri [30]	–	–	62.2	46.3	–	–
OAMN [31]	86.3	72.6	62.6	46.1	–	–
HG [32]	87.1	77.5	61.4	50.5	–	–
HOReID [11]	86.9	75.6	55.1	43.8	–	–
PAT [8]	88.8	78.2	64.5	53.6	–	–
BPBreID[13]	89.6	78.3	66.7	54.1	91.0	77.8
BPSEMA	**89.81**	**79.38**	**68.55**	**54.2**	**91.22**	**78.84**

3.5×10^-6 at the 40th and 70th epochs, respectively. The label smoothing regularization rate e is set to 0.1, and the margin for triplet loss is set to 0.3.

C. Comparison with the most advanced methods

On the DukeMTMC-ReID, Occluded-DukeMTMC, and P-DukeMTMC datasets, our method is compared with the state-of-the-art methods in recent years, and the comparison results are shown in Table I.

In Table I, all methods have chosen ResNet-50 as the backbone network. The last row of the table presents the experimental results of this paper, while the rest of the data are from the original model papers. To verify the effectiveness of the proposed method, we conducted comparative experiments with current mainstream pedestrian re-identification methods, mainly evaluating based on two metrics: mAP (mean average precision) and Rank-1 accuracy. To ensure the consistency of the experiments, we used the same dataset and evaluation protocol, and performed performance evaluations on the test set. Fig. 5 and 6 are line graphs comparing the experimental results of all methods in Table 1 on the DukeMTMC-ReID dataset and the Occluded-DukeMTMC dataset, clearly showing that the method proposed in this paper achieves more ideal performance.

The experimental results in Table I show that the proposed method significantly outperforms other comparison methods in terms of mAP and Rank-1 accuracy, especially demonstrating stronger advantages when human body parts are obscured. Additionally, our method maintains high efficiency in terms of computational complexity. Specifically, on the DukeMTMC-ReID dataset, compared to BPBreID (as shown in the performance comparison of various methods in Fig. 5), the proposed method improves Rank-1 accuracy by 0.21% and mAP by 1.08%; compared to the traditional BoT method, the proposed method improves Rank-1 accuracy by 3.41% and mAP by 2.98%. On the Occluded-DukeMTMC dataset (a subset with occlusion), compared to BPBreID (as shown in the performance comparison of various methods in Figure 4), the proposed method improves Rank-1 accuracy by 1.85% and mAP by 0.1%; compared to the traditional BoT method, the proposed method improves Rank-1 accuracy by 17.15% and mAP by 9.5%. On the P-DukeMTMC dataset (also an occlusion dataset), compared to BPBreID, the proposed method improves Rank-1 accuracy by 0.22% and mAP by 1.04%; compared to the traditional BoT method, the proposed method improves Rank-1 accuracy by 4.22% and mAP by 3.94%.

These results indicate that the proposed method has significant advantages in the extraction capabilities of both

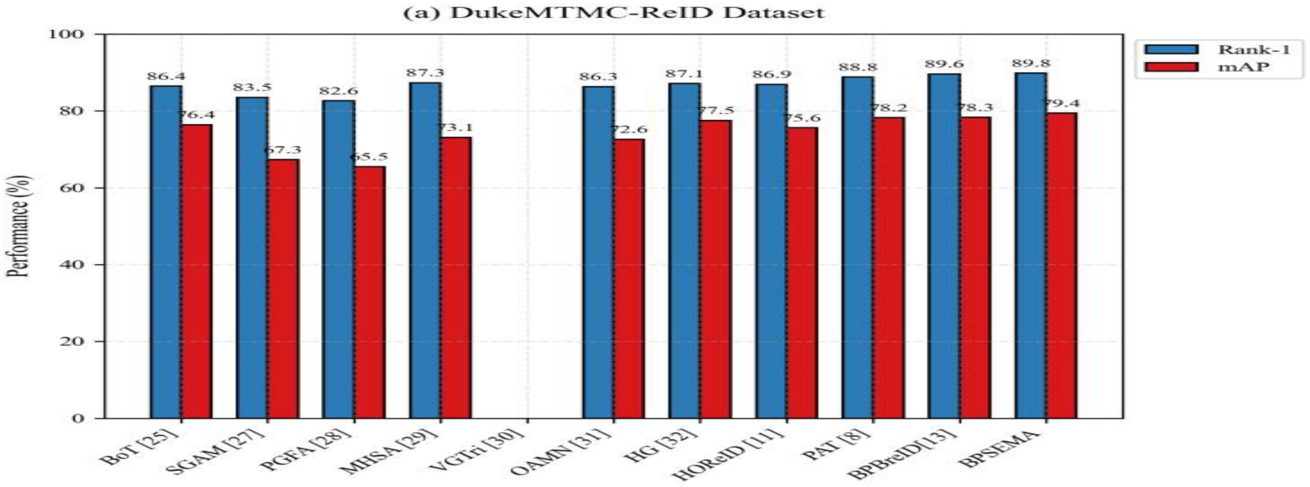

Fig.5. Performance Comparison on the DukeMTMC-ReID Dataset

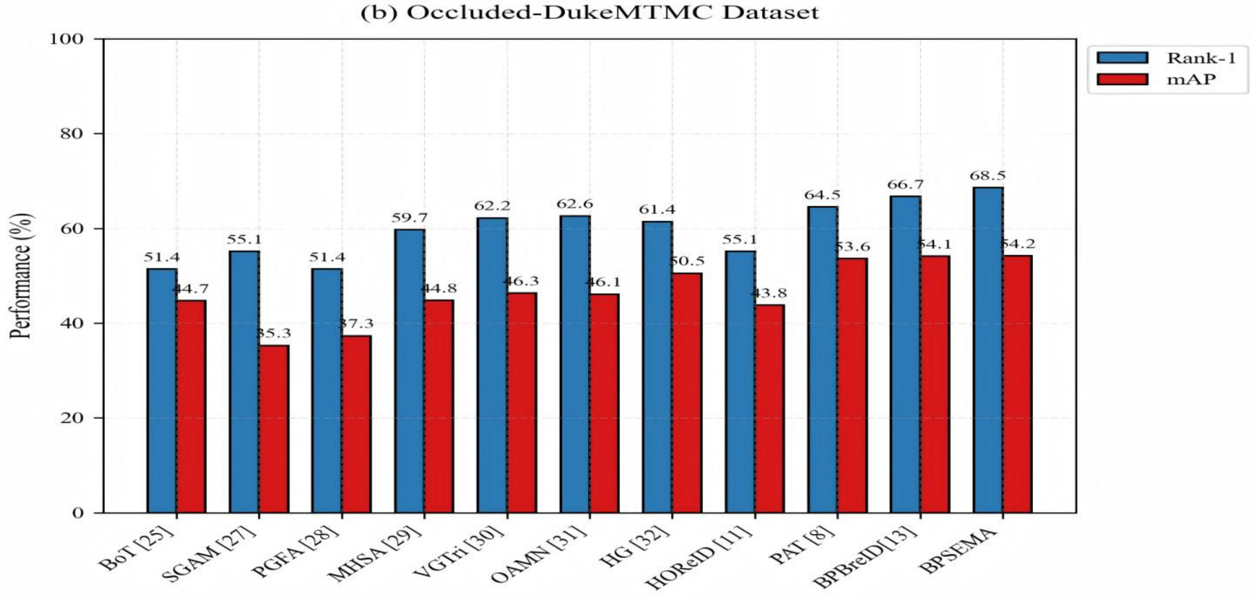

Fig.6. Performance Comparison on the Occluded-DukeMTMC Dataset

local features and spatial features, thereby significantly enhancing the overall performance of the pedestrian re-identification task.

D. Melting Experiment

On the DukeMTMC-ReI and Occluded-DukeMTMC datasets, we conducted a series of detailed ablation experiments aimed at verifying the actual effects of each key module in the designed network architecture. The experimental results are summarized in Table II, where "D1" represents the original results of the BPBreID method, "D2" represents the experimental results after introducing residual connections into the BPBreID method, "D3" represents the experimental results after introducing residual connections and adding the SE attention mechanism to the BPBreID method, and "D4" represents the experimental results after introducing residual connections, the SE attention mechanism, and incorporating the EMA attention module into the attention module in the BPBreID method.

Comparing the experimental results of the "D2" and "D1" models in Table2, the network using residual connections shows a significant performance improvement over the original BPBreID method in terms of Rank-1 cumulative matching features. On the DukeMTMC-ReID dataset, the mAP and Rank-1 metrics achieved improvements of 0.6% and 0.53%, respectively; whereas on the Occluded-DukeMTMC dataset, the mAP and Rank-1 metrics experienced a decrease of 1.11% and an increase of 0.27%, respectively. Further comparing the results of the "D4" model in the 4th row and the "D1" model in the 1st row of Table II, the network incorporating the dynamic feature refinement module (BPSEMA) demonstrates even more significant performance improvements over the BPBreID method. On the DukeMTMC-ReID dataset, the BPSEMA method improved the mAP and Rank-1 metrics by 1.08% and 0.21%, respectively; while on the Occluded-DukeMTMC dataset, the BPSEMA method improved the mAP and Rank-1 metrics by 0.1% and 1.85%, respectively. Thus, it can be seen that the proposed dynamic feature refinement module exhibits superior performance compared to the BPBreID method, validating the effectiveness and superiority of the module design. The specific performance comparison and trend changes on the DukeMTMC-ReID dataset and the Occluded-DukeMTMC dataset are shown in Fig. 7 and 8.

Based on these ablation experiments, we further conducted visual analysis based on parts and the whole, and the results showed that our method has higher accuracy and performance improvement. Included are the local recognition results of three parts of the character image and the local-global recognition results, where the green frame represents correct recognition and the red frame represents incorrect recognition. Obviously, our model based on the global and partial aspects has higher performance and accuracy. Specific information is shown in Fig. 9.

TABLE II
ABLATION EXPERIMENTS ON THE DUKEMTMC-REID AND OCCLUDED-DUKEMTMC DATASETS

Method	DukeMTMC-ReID		Occluded-DukeMTMC	
	R-1	mAP	R-1	mAP
D1	89.6	78.3	66.7	54.1
D2	90.13	78.9	66.97	52.99
D3	89.86	78.59	65.25	52.46
D4	89.81	79.38	68.55	54.2

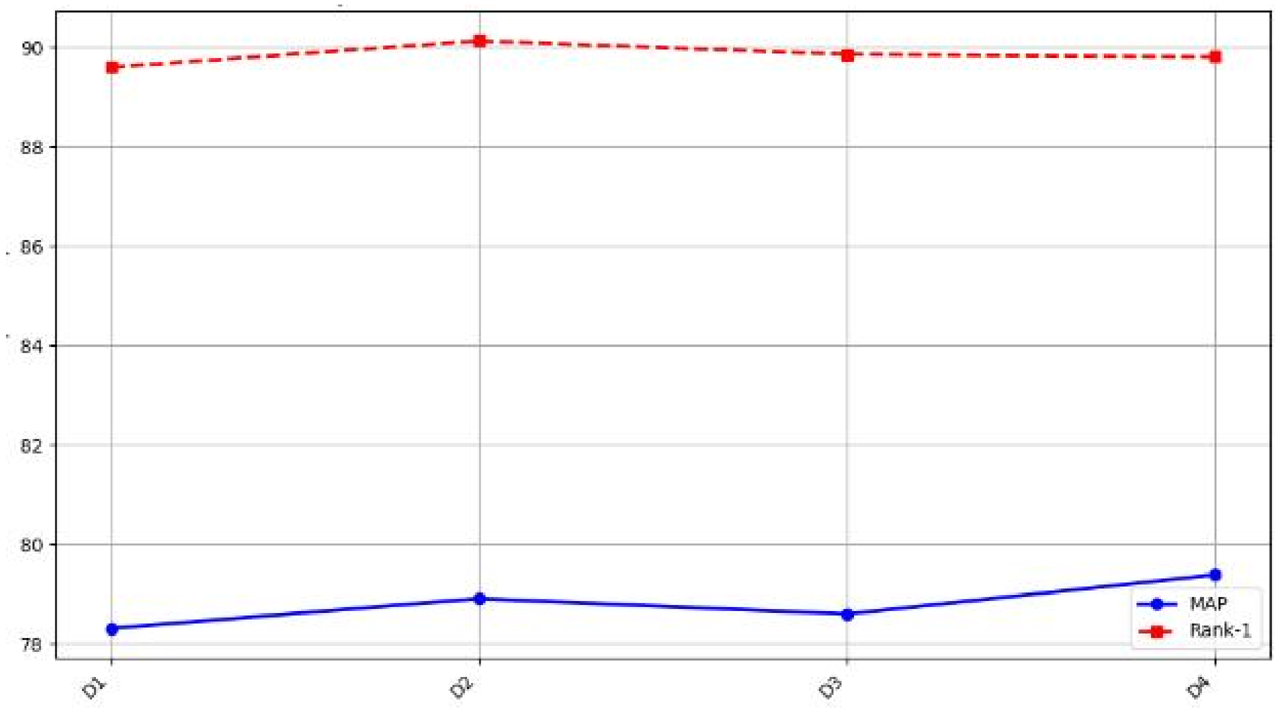

Fig.7. Ablation Study Comparison on the DukeMTMC-ReID Dataset

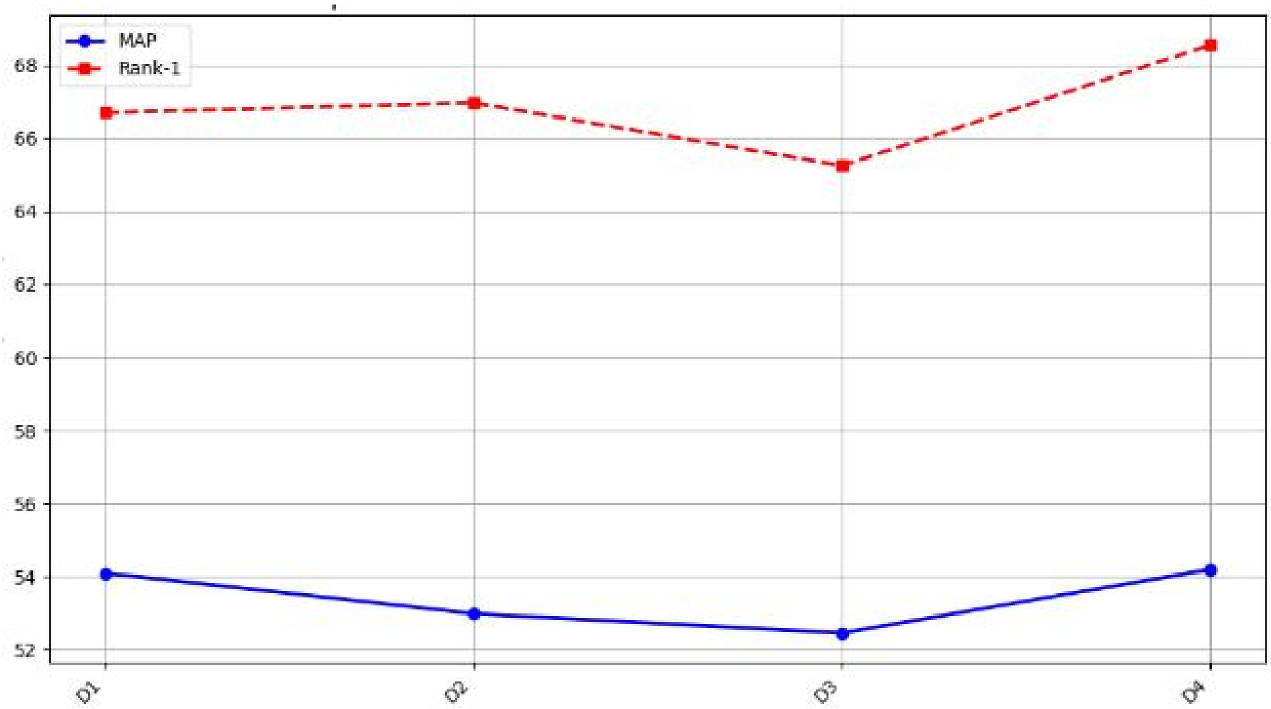

Fig.8. Ablation Study Comparison on the Occluded-DukeMTMC Dataset

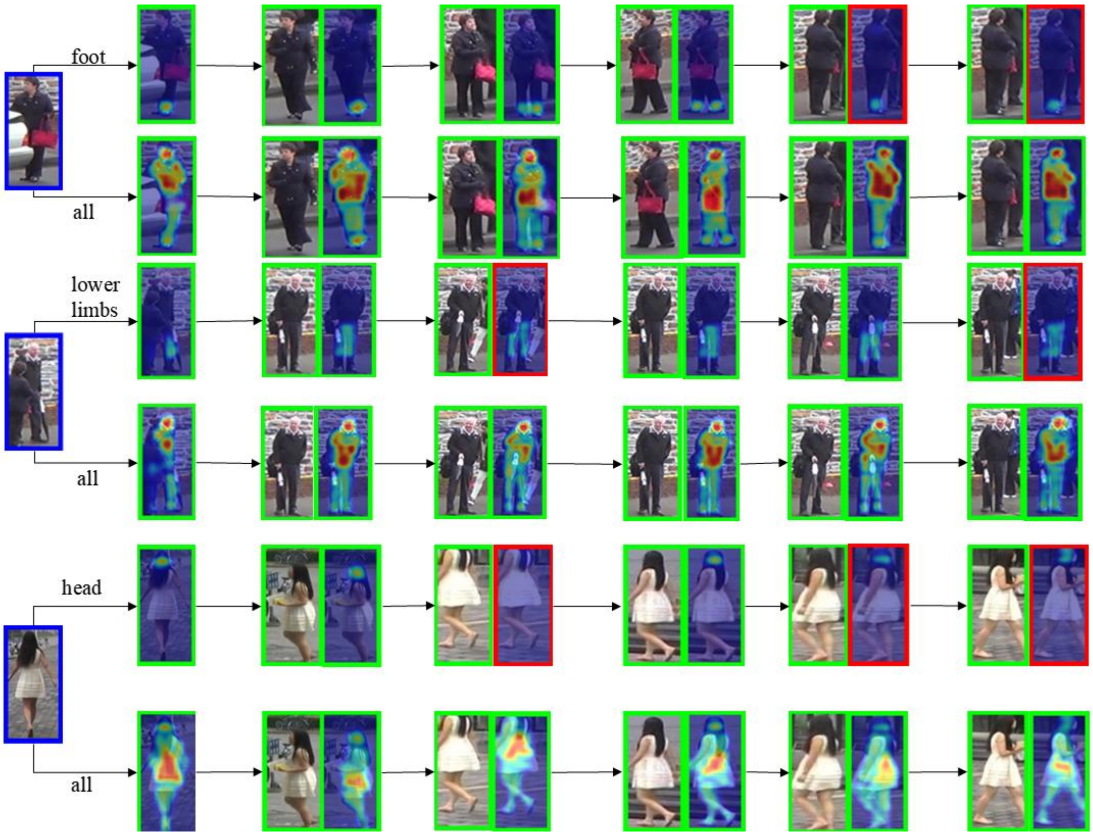

Fig.9. Visualization Analysis Comparison of Models

IV. Conclusion

This paper proposes an occluded pedestrian re-identification (ReID) algorithm based on body part features, addressing the issue of poor performance of existing pedestrian re-identification methods under occlusion conditions by introducing a new solution. By incorporating a multi-attention mechanism for body parts and a dynamic feature refinement module, we can effectively extract local features of pedestrians under various occlusion conditions and perform precise matching, ultimately aggregating them into global information. Experimental results show that compared to traditional global feature extraction methods, the local-global feature extraction method demonstrates significant advantages in handling occlusion scenes. In this study, our main contributions include: first, proposing a local-global feature extraction method based on a multi-attention mechanism for body parts, which can adaptively focus on the features of occluded areas; second, employing end-to-end joint training to effectively enhance the robustness of the model; and finally, through experimental validation on multiple standard datasets, our algorithm achieves an improvement of several percentage points in recognition accuracy under occlusion scenes compared to existing methods.

Despite our model's commendable performance in processing images of partially obscured pedestrians, it encounters several persistent challenges. Notably, in scenarios of substantial occlusion, the extraction of local features remains susceptible to disruption. Future investigations could delve into more sophisticated multimodal feature fusion techniques to augment the model's efficacy in intricate settings. Furthermore, the strategic integration of additional prior knowledge, such as camera perspective data and multi-scale feature information, represents a promising avenue for ongoing research.

In summary, the pedestrian re-identification method based on body part characteristics provides an effective solution to the occlusion problem, holding significant application value and research significance.

REFERENCES

[1] Feng Zhanxiang, Lai Jianhuang, Yuan Zang, Huang Yuli, Lai Peijie. Towards universal person re-identification: survey on the applications of large-scale self-supervised pre-trained model for person re-identification. Superintended by Chinese Academy of Sciences Sponsored by RADI, CSIG & IAPCN, Published Online: 21 August 2024.

[2] Zhang Guoqing, Yang Shan, Wang Hairui, Wang Zhun, Yang Yan, ZHOU Jieqiong. A Survey on Deep Learning-Based Multimodal Person Re-Identification. Journal.

[3] Xu Y, Guo X. Y, Rong L. L. A Survey on Unsupervised Learning Methods for Vehicle Re-Identification. Journal of Frontiers of Computer Science & Technology, 2023, Vol 17, Issue 5, p1017

[4] Jiang K. W, Wang J, Zhang L. Y, LU X. & LIU G. Q. (2023). Research on cross-modal person re-identification using local supervision. Application Research of Computers (Jisuanji Yingyong Yanjiu), 40 (4), 1226.

[5] Rongxin MI, Wenwen Yao, Binghao WU. Research on person re-identification algorithm based on multi-task learning [J]. Telecommunications Science, 2024, 40 (6): 127-136.

[6] Kuo Zhang, Xinyue Fan, Jiahui Li, Gan Zhang. Cross-Modal Person Re-Identification Based on Mask Reconstruction with Dynamic Attention [J]. Laser & Optoelectronics Progress, 2024, 61 (10): 1015001.

[7] Wu Xinyi; Deng Zhiliang; Liu Yunping; Dong Juan; Li Jiaqi. Multi-Feature Person Re-Identification Based on Cross-Attention Mechanism. Journal of Nanjing University of Information Science & Technology (Natural Science Edition) Nanjing Xinxi Gongcheng Daxue Xuebao (ziran kexue ban), 2024, Vol 16, Issue 4, p461

[8] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. Diverse Part Discovery: Occluded Person Re-identification with Part-Aware Transformer. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, number 2, pages 2897–2906, 6 2021.

[9] Deqiang Cheng, Guangkai JI, Haoxiang Zhang, He JIANG, Qiqi KOU. Cross-modality person re-identification based on multi-granularity fusion and cross-scale perception. Journal of Communications, 2025, 46 (1): 108-123.

[10] Shang Gao, Jingya Wang, Huchuan Lu, and Zimo Liu. Poseful visible part matching for occluded person ReID. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 11741–11749, 2020.

[11] Wang Guan'an, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. High-order information matters: Learning relation and topology for occluded person re-identification. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 6448 –6457, 2020.

[12] Chen Yu, Liu Hui, Liang Dongsheng, et al. Person re-identification based on pose estimation and Transformer model. Science Technology and Engineering, 2024, 24 (12): 5051-5058.

[13] Vladimir Somers, Christophe De Vleeschouwer, Alexandre Alahi.Body Part-Based Representation Learning for Occluded Person Re-Identification. Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV23), 2022.

[14] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, Tao Xiang. Omni-Scale Feature Learning for Person Re-Identification. Computer Vision and Pattern Recognition (cs. CV), 2019.

[15] Lin Wu, Yang Wang, Junbin Gao, Dacheng Tao. Deep Co-attention based Comparators For Relative Representation Learning in Person Re-identification. Computer Vision and Pattern Recognition (cs. CV), 2018.

[16] Miao, Jiaxu and Wu, Yu and Liu, Ping and Ding, Yuhang and Yang, Yi. Pose-guided feature alignment for occluded person re-identification. Proceedings of the IEEE International Conference on Computer Vision, 2019.

[17] Jiaxuan Zhuo, Zeyu Chen, Jianhuang Lai, Guangcong Wang. Occluded Person Re-identification. Computer Vision and Pattern Recognition (cs. CV); Artificial Intelligence (cs. AI); Multimedia (cs. MM), 2018.

[18] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Reranking person re-identification with k-reciprocal encoding. In Proceedings-30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, volume 2017-Janua, pages 3652 – 3661. Institute of Electrical and Electronics Engineers Inc., 1 2017.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 2016-Decem, pages 770 – 778. IEEE Computer Society, 12 2016.

[20] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at Once: Enhancing Learning and Generalization Capacities via IBN-Net. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11208 LNCS: 484–500, 7 2018.

[21] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, Tao Mei, and A I Research. FastReID: A Pytorch Toolbox for General Instance Re-identification. 6 2020.

[22] Xianghao Zang, Ge Li, Wei Gao, and Xiujun Shu. Learning to Disentangle Scenes for Person Re-identification.Image and Vision Computing, 116, 11 2021.

[23] Cheng Yan, Guansong Pang, Jile Jiao, Xiao Bai, Xuetao Feng, and Chunhua Shen. Occluded Person Re-Identification with Single-scale Global Representations. In ICCV 2021, pages 11875–11884, 2021.

[24] Wu H L, Liu H, Sun Y C. Vision Transformer-based pilot pose estimation. Journal of Beijing University of Aeronautics and Astronautics, 2024, 50 (10): 3100-3110 (in Chinese)

[25] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, volume 2019-June, pages 1487–1495, 2019.

[26] Jiang, W. T, Liu, Y. W & Zhang, S. C. Random Channel Perturbation-Based Image Data Augmentation Approach. Journal of Frontiers of Computer Science & Technology, 2024, Vol 18, Issue 11, p2980

[27] Qin Yang, Peizhi Wang, Zihan Fang, and Qiyong Lu. Focus on the visible regions: Semantic-guided alignment modelfor occluded person re-identification. Sensors (Switzerland), 20 (16): 1–15, 2020.

[28] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang DIng, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, volume 2019-Octob, pages 542–551, 2019.

[29] Hongchen Tan, Xiuping Liu, Baocai Yin, and Xin Li. MHSA-Net: Multihead Self-Attention Network for Occluded Person Re-Identification. IEEE Transactions on Neural Networks and Learning Systems, 8 2022.

[30] Jinrui Yang, Jiawei Zhang, Fufu Yu, Xinyang Jiang, Mengdan Zhang, Xing Sun, Yingcong Chen, and Wei-Shi Zheng. Learning to Know Where to See: A Visibility-Aware Approach for Occluded Person Re-identification. ICCV, pages 11885–11894, 2021.

[31] Peixian Chen, Wenfeng Liu, Pingyang Dai, Jianzhuang Liu, Qixiang Ye,Mingliang Xu, and Rongrong Ji. Occlude Them All: Occlusion-Aware Attention Network for Occluded Person Re-ID. Iccv, pages 11833–11842, 2021.

[32] Madhu Kiran, R Gnana Praveen, Le Thanh Nguyen-Meidine, Soufiane Belharbi, Louis-Antoine Blais-Morin, and Eric Granger. Holistic Guidance for Occluded Person ReIdentification. 4 2021.

[33] Li, G. D & Guo, L. J. Long-Term Pedestrian Re-Identification Based on Superpixel Random Erasure. Journal of Computer Engineering & Applications, 2023, Vol 59, Issue 10, p221.

[34] Fengming Lin, Yan Xia, Michael MacRaild, Yash Deo, Haoran Dou, Qiongyao Liu, Nina Cheng, Nishant Ravikumar, Alejandro F. Frangi. GS-EMA: Integrating Gradient Surgery Exponential Moving Average with Boundary-Aware Contrastive Learning for Enhanced Domain Generalization in Aneurysm Segmentation. Computer Vision and Pattern Recognition (cs. CV); Machine Learning (cs. LG), 2 2024.

[35] Zheng Aihua, Feng Mengya, Li Chenglong, Tang Jin, and Luo Bin. Bi-Directional Dynamic Interaction Network for Cross-Modality Person Re-Identification [J]. Journal of Computer-Aided Design & Computer Graphics, 2023, 35 (3): 371-382.

[36] Wen Rui; Kong Guangqian; Duan Xun. Unsupervised Domain Adaptive Person Re-Identification Based on Reliability Integration. Application Research of Computers / Jisuanji Yingyong Yanjiu, 2024, Vol 41, Issue 4, p1228

[37] Bochun Huang, Fan Li, Shujuan Wang. Cross-Classification-Based Sketch Person Re-Identification in Inconsistent Cross-Modal Identity Scenes [J]. Laser & Optoelectronics Progress, 2023, 60 (4): 0410006.

[38] Erik J Bekkers, Sharvaree Vadgama, Rob D Hesselink, Putri A van der Linden, David W Romero. Fast, Expressive SE (n) Equivariant Networks through Weight-Sharing in Position-Orientation Space. Machine Learning (cs. LG); Group Theory (math. GR). 10 2023

[39] Xu Guoliang, Hou Zhendong, Luo Jiangtao, Liu Yang, Liu Lizhu. Joint Reliable Instance Mining and Feature Optimization for Unsupervised Person Re-ID [J]. Journal of Computer-Aided Design & Computer Graphics, 2024, 36 (3): 368-378.

[40] Chen Fushi, Shen Yao, Zhou Chichun, Ding Meng, Li Juhao, Zhao Dongyue, Lei Yongsheng, Pan Yilun. Unsupervised Learning for Gait Recognition: A Review. Journal of Frontiers of Computer Science & Technology, 2024, Vol 18, Issue 8, p2014.

[41] Jiang Hailang, Liu Jianming. Fine-Grained Image Recognition via Multi-Part Learning . Journal of Computer-Aided Design & Computer Graphics, 2023, 35 (7): 1032-1039.

[42] Zheng, Jian, and Xin Yu. Parallel OPTICS Algorithm Using Mean Distance and Association Labeling. Journal of Computer Engineering & Applications, 2023, Vol 59, Issue 5, p232.