The Journal of Supercomputing (2025) 81:363
https://doi.org/10.1007/s11227-024-06831-1

Research on occlusion pedestrian re-identification based on ViT model

Yuepeng Guo[1] · ZhenPing Lan[1] · Yanguo Sun[2] · Yuheng Sun[1] · Xinxin Li[1] · Yuru Wang[1] · Bo Li[1]

Accepted: 17 December 2024
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract

Pedestrian re-identification (Re-ID) is an important task in intelligent surveillance and public safety. Traditional pedestrian re-identification methods show obvious limitations when facing the occlusion problem, leading to a significant decrease in re-identification accuracy. For this reason, we design an occlusion perceptual attention module (OPAM), which seeks to improve the model's capacity to grasp both local and global contextual information. Secondly, we bring in an improved feature fusion module FtF (Feature to Feature), which aims to fully utilize the rich information of convolutional features to enhance the feature representation capability of the visual transformer model. Finally, this paper constructs a comprehensive loss function robust triplet loss (RTL), which combines the triad loss and occlusion perception loss to enhance the performance and efficiency of pedestrian re-recognition. We conduct experiments on two recognized occluded pedestrian datasets, with Rank-1 of 73.6% and mAP of 63.2% on the Occluded-Duke dataset, which is an increase of 2.6% and 2.2% from baseline, respectively; and Rank-1 of 90.8% and mAP of 82.4% on the DukeMTMC-reID dataset. The good performance compared with state-of-the-art methods fully validates its validity and generalizability.

Keywords Occlusion · Pedestrian re-recognition · Feature fusion · Attention mechanism

1 Introduction

Occluded Person's re-identification (OPReID) is a significant research focus in the areas of pattern recognition and computer vision, aiming at solving the problem of difficult pedestrian recognition caused by occlusion, change of viewing angle and difference of lighting conditions in real scenes through feature extraction and

ZhenPing Lan and Yanguo Sun contributed equally to this work.

Extended author information available on the last page of the article

Published online: 03 January 2025

matching of pedestrian images [1]. With the wide deployment of surveillance cameras, pedestrian re-identification technology has become particularly important in applications such as intelligent surveillance, public safety and intelligent transportation. However, the occlusion problem poses a great challenge to pedestrian re-recognition, and existing methods often perform poorly in dealing with the occlusion problem.

Existing pedestrian re-recognition methods mainly focus on feature matching and extraction of full-body images. Still, in practical applications, pedestrian images are often subject to occlusion, e.g., by other pedestrians, vehicles, trees, etc., which results in part of the body parts not being visible. Such occlusion not only leads to the lack of feature information but also introduces a large number of noise features, which reduces the accuracy of re-recognition. To cope with this problem, recently, scholars have proposed various occlusion-based pedestrian re-recognition methods, which can be mainly divided into two major groups: methods based on traditional methods and methods utilizing deep learning.

Traditional methods, on the other hand, rely on hand-designed features [2, 3], such as color histograms and texture features, or use image processing and analysis techniques, such as image segmentation and keypoint detection, to cope with the occlusion problem; yet, they are difficult to adapt to complex and changing real-world scenarios. With advancements in deep learning technology, considerable progress has been achieved in deep learning-based pedestrian re-recognition methods [4, 5]. In occluded pedestrian re-recognition, vision transformer (ViT) [6] is able to better deal with feature reuse and local information loss through global context modeling and adaptive attention capability, thus improving the model's performance in complex scenes. Therefore, this article conducts research based on the ViT model. Traditional feature fusion methods may not be able to fully utilize multi-level features, resulting in information loss or insufficient feature expressiveness [7]. In this article, we improve a novel feature fusion module, FtF, for enhancing the feature expression capability of the visual transformer model. The design of this module is inspired by the need for multi-level feature fusion and aims to fully utilize the rich information of convolutional features while maintaining feature integrity. With this module, we are able to effectively fuse multi-level convolutional features, thus enhancing the feature expression ability and the model's ultimate classification performance. Similarly, the traditional self-attention mechanism may be insufficient in capturing local features [8]. In this article, we put forward an occlusion perceptual attention module (OPAM module), OPAM module is an improved attention mechanism designed to improve the model's capacity to capture both local and global contextual information. The module extracts local features by introducing convolutional operations and combines them with the traditional self-attention mechanism to capture global features. The traditional ternary loss not only relies strongly on sample selection, requiring manual selection of positive and negative samples, which is inefficient, but also lacks occlusion awareness, and the model performs poorly in occlusion situations and is susceptible to occlusion [9]. For this reason, we construct a comprehensive loss function robust triplet loss (RTL), which introduces a dynamic selection mechanism in the traditional triad loss to screen positive and negative samples according to the degree of occlusion. Different weights can be assigned to different triplets when calculating the loss. To enhance the concentrate on features of different scales, ternary loss for hierarchical features can be introduced in the loss calculation.

In summary, the contributions of this article can be summarized as follows:

- A new feature fusion module, FtF, is proposed to better handle feature reuse and local information loss, thereby enhancing the model's performance in complex scenes.
- An occlusion perceptual attention module (OPAM) is designed to improve the model's capability to capture both local and global contextual information.
- A comprehensive loss function (RTL) is constructed with an innovative ternary loss function using dynamic selection, weighting mechanism, and multi-scale feature fusion, which better captures features in complex occlusion scenarios and improves the overall efficiency of the model.
- Comprehensive experiments are performed on well-established pedestrian re-identification datasets, and the effectiveness of the method is demonstrated through comparison with other approaches.

2 Related work

2.1 Pedestrian re-recognition methods in conventional environments

Pedestrian re-identification in traditional environments is in unobstructed conditions. As deep learning progresses, an increasing number of researchers have put forward deep learning-based methods for ReID and proved its effectiveness. The proposed methods are roughly categorized into two types, which are based on the metric-based learning and representation learning. The network architecture based on representation learning proposed by Wu et al. [10] fully extracts the information on the global features. Luo et al. [11] proposed a BN-Neck structure in the CNN-based network architecture to fully extract the global information. However, only global features are obtained and local variables of specific environments cannot be recognized, which is less robust, so the method of using local features to generate fine-grained information on images has been widely studied and used by scholars. Zheng et al. [12] used the pyramid method to horizontally segment the input images or feature maps into several parts for fine-grained information representation. Wei et al. [13] suggested a technique grounded in the generative adversarial network for pedestrian re-identification, which reduces the gap between different data domains and enhances the accuracy of pedestrian re-identification through the use of a generative adversarial network. Xiao et al. [14] proposed a deep metric learning framework that uses deep neural networks to learn the similarity of pedestrian features and developed a novel loss function that integrates the contrastive loss and the ternary loss in order to refine the feature space so that the same identity samples are as close as possible in the feature space, while samples with different identities are as far away as possible.

2.2 Pedestrian re-recognition methods in occluded environments

In real life, the movement of people and objects is specifically uncertain, so occlusion is very common, when the retrieved person is occluded by an object, the useful information is only a section of the body, then the traditional pedestrian re-recognition method that sees the person as a whole is no longer applicable. Occlusion pedestrian re-recognition should be centered on extracting human features and separating the human body from the background. Earlier scholars have worked on removing the effect of end-to-end obstacles in the framework to address this issue and to make the revealed part derived into a global feature representation. Zhuo et al. [1] introduced an application of binary to differentiate between the occluded region of the feature map and the overall region. Although this work is easier to implement, it is highly susceptible to the interference of external noise and has poor robustness. Gao et al. [15] utilized the HONet method and introduced a key point model, where more effort is put into extracting local features to match each part of the body. Miao et al. [2], through the PGFA method, can effectively mitigate the impact of occlusion on feature extraction by aligning the pedestrian poses in different images and thus improve the recognition accuracy. Zheng et al. [16] design a novel feature mapping strategy to obtain feature representations that are unaffected by occlusion and pose changes. Focusing on how to minimize the impact of pedestrian's pose changes on the recognition process, a pose estimation model is constructed to capture the pedestrian's features from different viewpoints. Yan et al. [17] effectively improve the pedestrian re-recognition performance under occlusion by introducing semantic information, designing an occlusion processing module, integrating multilayered features, and using a joint loss function. Han et al. [18] improve the pedestrian re-recognition performance under occlusion conditions by means of an attribute-driven attentional mechanism to improve the pedestrian re-recognition performance under occlusion conditions. Luo et al. [19] no longer extracted local features but introduced vision transformer (ViT) into the field of re-recognition and obtained good performance. However, ViT is good at capturing global features, but in some scenarios, local features (e.g., detailed textures or features of small objects) may be neglected, which may affect the model's ability to recognize complex targets. For this reason, we construct a new network framework to capture global and local features more rationally and enhance the generalization ability of the model.

2.3 Attention mechanism

Pedestrian re-identification (ReID) methods based on attention mechanisms improve recognition performance by automatically focusing on key areas of the image. The attention mechanism can effectively enhance the model's perception of critical information, especially when dealing with occlusion or complex backgrounds. Liu et al. [20] suggested a holistic approach comparative attention network that improves the precision and reliability of pedestrian re-identification in challenging scenarios using deep feature extraction and adaptive attention mechanism. Woo et al. [21] enhanced the feature representation capability of convolutional neural networks through the integration of spatial attention and channel attention mechanisms, thereby improving the performance of mandates including target detection and image categorization. Song et al. [22] proposed a mask-guided comparative attention model, which enhances the feature extraction capability of occluded pedestrians by combining mask information and a comparison learning mechanism, thereby improving the pedestrian re-recognition accuracy and robustness. In this article, the occlusion perceptual attention module (OPAM) is proposed to improve the model's capacity to capture both global and local contextual details. By skillfully combining the advantages of convolution and self-attention, it effectively enhances the local feature extraction capability while maintaining the powerful modeling capability of the transformer, which is a promising mechanism for improving attention.

3 Method

3.1 Overall framework

As shown in Fig. 1, the framework of this paper uses ViT model to extract image features, segment the image into patches and add positional coding. Features are processed through self-attention and MLP layers, and layer normalization and residual concatenation are applied. Multiple transformer block stacking builds high level

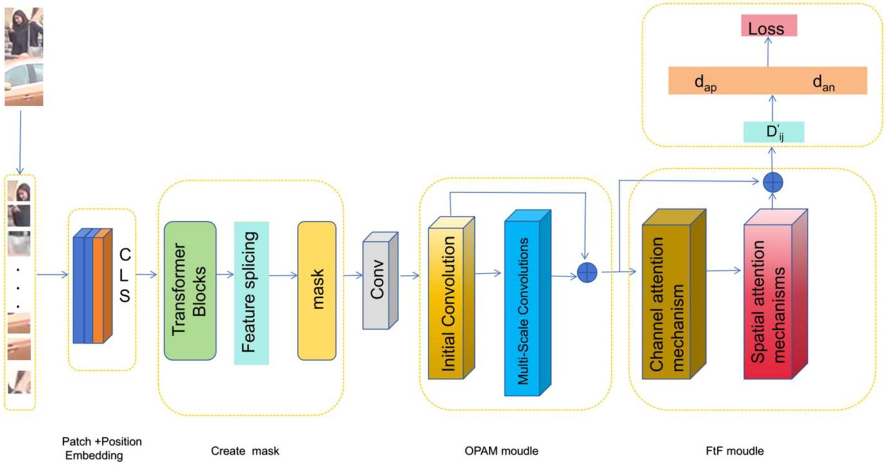

Fig. 1 The model first splits the input image into small chunks and embeds them into a high-dimensional space, which is then processed through a series of transformer blocks to generate feature representations and masks. Next, the convolutional features are combined through the attention, feature fusion module, and finally classified through the fully connected layer. In the loss function, the matrix distance D'_{ij} between samples is computed, then the distance between each anchor point and its hardest positive and negative samples is extracted from the distance matrix, denoted as dap and dan, respectively, and finally, the loss function is computed through the RTL class, which optimizes the model by adjusting the value of d_{ap} and d_{an}, so that the anchor points have smaller distances from the positive samples, and greater distances from the negative samples. Thus, a more discriminative feature representation is learned

features [6]. A mask is generated from class token, which is continuously valued and indicates the level of importance of each region [23], and its features are passed through OPAM and FtF modules, respectively. OPAM and FtF are two modules used for feature processing. OPAM first processes the input features through a shared initial convolutional layer and then generates an attention map using multi-scale convolution in the attention branch, after adaptive mean pooling and Softmax normalization, and then weighted fusion with the features generated in the value branch. FtF, on the other hand, generates the attention weights through the channel and spatial attention mechanisms, respectively, in combination with the feature transforms and residual concatenation. These two modules enhance the representation of input features through different attention mechanisms and feature fusion strategies. The final result is output through the loss function.

3.2 Occlusion perceptual attention module

As shown in Fig. 2 for the occlusion perceptual attention module module proposed in this paper, the core idea is to combine convolutional operations and attention mechanisms to capture both local and global information. By using grouped convolution and dimensionality reduction operations, it improves computational efficiency while maintaining strong modeling capabilities. This design enables OPAM to excel in numerous computer vision tasks, especially in scenarios that require simultaneous taking into account global context and local details. The OPAM module employs an innovative approach to feature processing and fusion. First, the input features are passed through a shared initial convolutional layer that is utilized to obtain the base features. These features are then split into two branches: an attention branch and a value branch. In the attention branch, the features undergo a multi-scale convolutional operation to capture spatial information at different scales, followed by compression of the spatial dimensions through an adaptive pooling layer. Meanwhile, the value branch uses pointwise convolution for feature transformation. Next,

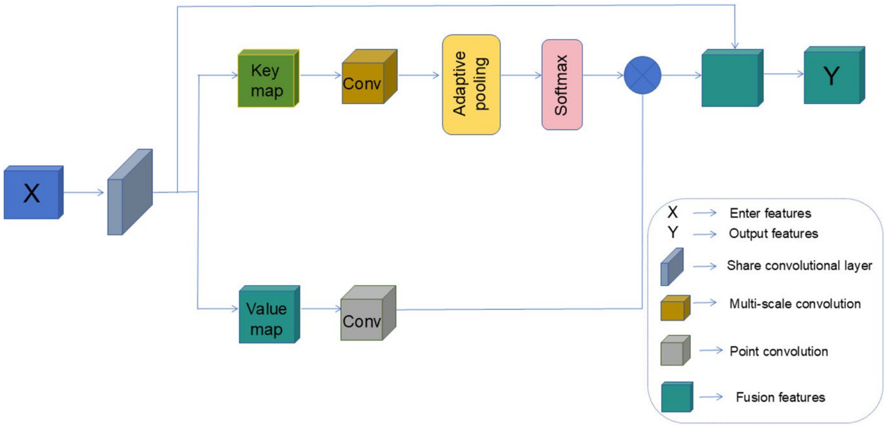

Fig. 2 Detailed internal structure of the OPAM module

the output of the attention branch is softmax normalized to generate an attention graph. This attention map is then multiplied element by element with the output of the value branch to realize weighted feature fusion. Finally, the fused features are summed with the original input via residual concatenation to form the final output. This design skillfully balances the processing of local and global information while enabling adaptive feature selection through the attention mechanism. This structure enables the module to effectively handle complex spatial dependencies and adapt to the needs of various visual tasks, especially excelling in scenes that require simultaneous consideration of details and context. Where the pointwise convolution formula is:

$$Output(c,x,y) = \sum_{i=1}^{i} (Input(i,x,y) \times Weight(c,i)) \tag{1}$$

Where c is the output channel index, x, y are spatial locations, and i traverses all input channels. This means that for each spatial location (x, y), the output channel c's value is the products of the values summed up of all input channels at that location and the corresponding weights.

For each convolutional kernel size k, this can be expressed as:

$$Output_k(c,x,y) = \sum_{i=1}^{i} \sum_{dx=1}^{c} \sum_{dy=1}^{c} \left(Input(i, x + \mathrm{d}x, y + \mathrm{d}y) \times Weight_k(c, i, \mathrm{d}x, \mathrm{d}y)\right) \tag{2}$$

Where: k is the convolution kernel size, x, y are the output spatial locations, i iterates over all input channels, dx, dy iterates over the spatial offsets of the convolution kernel.

The final result of a multi-scale convolution is the sum of all the different kernel size results:

$$FinalOutput(c,x,y) = \sum_{k=1}^{k} Output_k(c,x,y) \tag{3}$$

Where k iterates over all the convolution kernel sizes used.

In order to avoid the challenge that this module may face in practical applications with different degrees of occlusion and the inability to automatically adjust the balance between local and global information, we introduce an adaptive mechanism in the model. This is done as follows:

We introduce an adaptive mechanism in the OPAM module to handle different occlusion scenarios, which is realized by dynamically adjusting the weights of local and global information. We can do this by introducing a learnable parameter that is adjusted to the input data during the forward propagation process. We set this parameter as α, which is an adaptively adjustable parameter for weight balancing between local and global information during the training process. The initial value is set to 0.5, which indicates that the weights of local and global information are equal at the initial time. In the forward method, α is used to adjust the weights between the fused features and the original input x to achieve adaptive feature fusion. In this way, the model can automatically learn how to adjust the balance of local and global information under different occlusion conditions during the training process.

3.3 Feature fusion module FtF

In deep learning models, especially in transformer-based architectures, various levels of features encompass different layers of information. Shallow features typically include more lower level information (e.g., edges and texture), while deep features contain more high level semantic information. Traditional model architectures usually utilize only the last layer of features, which can lead to information loss and model performance limitations. Our proposed feature fusion module draws on and extends the design concept of CBAM [21]. Specifically, we retain the core idea of CBAM's channel and spatial attention mechanisms but add an extra feature transformation step and explicit residual linking. We further introduce a residual learning component for more stable training and information retention. With residual connectivity, input features can be passed directly to the output layer to avoid losing important information during feature transformation. In addition, we introduce a learnable parameter β that enables the module to dynamically adjust the weights of feature transformations and residual connections. This dynamic fusion strategy provides greater flexibility and adaptability for the model to adaptively choose the optimal feature fusion strategy under different training stages or data distributions. This improvement not only retains the advantages of CBAM but also enhances the performance of the model through an innovative fusion approach. As shown in Fig. 3, the input features are first passed through an adaptive average pooling layer that compresses the spatial dimensions to 1x1, then through two consecutive convolutional layers (to form a multilayer perceptron similar to a bottleneck structure), and finally through a sigmoid activation function to produce the channel weights. The weighting formula is as follows:

$$W_C = \sigma\left(Conv2d_2\left(\mathrm{Re}LU\left(Conv2d_1\left(AdaptiveAvgPool2_d(X)\right)\right)\right)\right) \tag{4}$$

Where W_C denotes the channel weights and X denotes the input feature map.

This mechanism captures the interdependencies between different channels and highlights important feature channels. Next, the channel attention processed

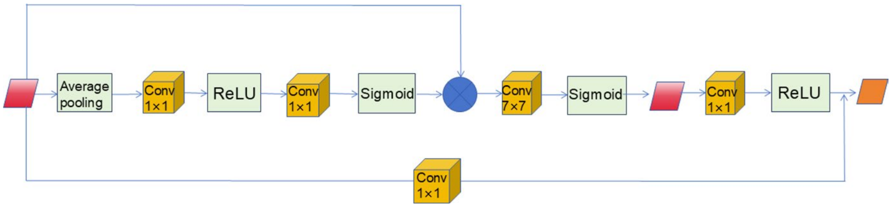

Fig. 3 Detailed flowchart within the FtF module

features are then passed through a 7x7 convolutional layer and sigmoid activation function to produce a spatial attention map, where the spatial weights are formulated as follows:

$$W_S = \sigma\left(Conv2d_3\left(X_C\right)\right) \tag{5}$$

Where W_S represents the spatial weights and X_c represents the feature map after channel attention processing.

This step helps the model to focus on important spatial regions in the input feature map and enhances the perception of key regions. The attention-enhanced features are then nonlinearly transformed by a 1x1 convolutional layer, batch normalization and ReLU activation function. This step adjusts the dimensionality of the features and introduces nonlinearities that increase the expressive power of the model.The formula is as follows:

$$T = ReLU\left(BatchNorm2d\left(Conv2d_4\left(X_S\right)\right)\right) \tag{6}$$

Where T is the transformed feature and X_S represents the feature map after spatial attention processing.

At the same time, the original input features are passed through an optional 1x1 convolutional layer (used when the number of input and output channels are different) to form a residual path. The formula is as follows:

$$R = Conv2d_5\left(X_S\right) \tag{7}$$

Where R stands for the residual feature.

This residual connection helps preserve the original information and facilitates the backpropagation of the gradient, thus mitigating the issue of gradient depletion in deep networks. Finally, the transformed features are summed with the features of the residual path to achieve the final feature fusion. The formula is as follows:

$$O = \beta T + (1 - \beta)R \tag{8}$$

Where O is the final output feature and β represents the dynamic fusion weights.

This fusion retains the information of the original features and introduces new features enhanced by the attention mechanism and nonlinear transformation.

Through this series of well-designed steps, the module is able to adaptively emphasize important features while achieving feature enhancement while maintaining the original information. This multi-stage processing allows the module to better capture and integrate different levels of feature information, thereby enhancing the model's expressiveness and ability to generalize in various computer vision tasks. In addition, the module is designed to be flexible enough to be further adapted and optimized according to specific task requirements.

In order to improve the flexibility and generalization ability of FtF, the introduction of a dynamic feature fusion strategy can be considered. A learnable parameter β is also added to dynamically adjust the fusion ratio of transforme and residual. The initial value is set to 0.5, indicating that both are initially weighted equally. In the forward method, β is used to weight the transforme and residual to achieve dynamic feature fusion. In this way, the model can automatically adjust the proportion of feature fusion during the training process to adapt to different scenarios and data distributions, thus improving the generalization ability.

3.4 Loss function

We propose a comprehensive loss function (RTL) that combines dynamic difficult sample mining, occlusion detection and multi-scale feature extraction. This approach significantly enhances the model's robustness and effectiveness in complex scenarios. First, we introduce a masking detection mechanism, which generates masking scores for each sample via a masking detection function. These scores are used to dynamically adjust the distance between samples such that occluded samples are given reduced weight for positive sample selection and increased weight for negative sample selection. This approach effectively reduces the mismatching problem due to occlusion. Specifically, for samples i and j, the adjusted distance D'_{ij} is defined as:

$$D'_{ij} = \begin{cases} D_{ij}(1 - O_i), y_i = y_j \\ D_{ij}(1 + O_i), y_i \neq y_j \end{cases} \tag{9}$$

Where D'_{ij} is the adjusted distance, D_{ij} is the original distance, O_i is the occlusion score of sample $i(0 \leq O_i \leq 1)$, and y_i, y_j are the labels of samples i and j, respectively.

Second, we implement a dynamic difficult sample mining strategy that takes into account the masking factor. We dynamically adjust the distance matrix based on the masking score to ensure that the most challenging and representative pairs of positive and negative samples are selected. For each anchor sample a, we select:

$$\begin{aligned} d_{ap} &= \max\{D'_{aj} | y_a = y_j\} \\ d_{an} &= \min\left\{D'_{ak} | y_a = y_k\right\} \end{aligned} \tag{10}$$

Where: d_{ap} is the distance from anchor sample a to the most difficult positive sample, d_{an} is the distance from anchor sample a to the most difficult negative sample, and y_a, y_j, y_k are the labels of samples a, j, and k, respectively.

Furthermore, we utilize a multi-scale feature extraction method to compute the loss at different scales. This approach boosts the model's capability to detect visual features at different scales and improves the robustness of the feature representation. Finally, we introduced a hardness factor for further adjusting the spacing between pairs of positive and negative samples to increase the difficulty of training. The adjusted distance is:

$$\begin{aligned} d'_{ap} &= d_{ap}\,(1 + h) \\ d'_{an} &= d_{an}(1 - h) \end{aligned} \tag{11}$$

Where: d'_{ap} is the adjusted positive sample distance, d'_{an} is the adjusted negative sample distance, and h is the hardness factor. At the same time, we use the masking score to weight the loss contribution of each sample to ensure that the model focuses more on the reliable samples that are not masked.

The final loss function is defined as:

$$L = \sum_{i=1}^{N} \max\left(0, m + {d'}_{ap} - {d'}_{an}\right) \tag{12}$$

Where N is the number of samples, L is the total loss, and m is the boundary parameter.

Experimental results show that this comprehensive ternary loss function we propose performs well in dealing with complex scenes, occluded objects and multi-scale features and significantly enhances the performance and model's capacity for generalization.

Due to the diversity of occlusion types, the occlusion scores in RTL may not be stable in practical applications, so we pick experiments under different occlusion conditions to verify the validity of the scores in different scenarios. The details are as follows: First, we prepare an image dataset containing multiple occlusion conditions. Each sample in the dataset is manually labeled to mark the occlusion regions. Second, we integrate a mask R-CNN model to detect occluded regions in the images. The model is able to identify and label the occluded part of the image, which provides the basis for subsequent feature extraction and processing. Then, we weighted the features according to the results of occlusion detection. Specifically, for occluded features, we reduced their weights to minimize their impact on the final score. We grouped the datasets by occlusion condition and calculated the loss and accuracy for each group separately. We define a metric for the degree of occlusion and analyze the impact of the degree of occlusion on the model performance.

By comparing the model performance under different occlusion conditions, we find that the occlusion detection method significantly improves the robustness of the model in complex scenes. As shown in Table 1, specifically, the accuracy of the model improves by about 3% under conditions with a higher degree of occlusion. This suggests that by implementing a more refined occlusion detection method, the model is able to cope with different occlusion scenarios more effectively. Overall, the accuracy of the model improves under different occlusion conditions, especially under moderate and heavy occlusion conditions. This indicates that the improved occlusion detection method effectively enhances the robustness of the model in complex scenes.

Table 1 Comparison of loss functions for different occlusion scenarios

Sheltering situation	Triplet loss (mAP)	RTL (mAP)
Not obstructed	87.3%	88.4%
Mildly obscure	80.6%	82.4%
Medium shade	77.2%	79.2%
Heavy masking	68.3%	71.3%

4 Experimental

4.1 Dataset and experimental platform setup

4.1.1 DukeMTMC-reID

The DukeMTMC-reID dataset [24] consists of 36,411 images of 1,812 pedestrians from eight static cameras, where the training set contains 702 pedestrians providing 16,522 images; the query image contains 2,228 images, and the gallery contains 17,661 images.

4.1.2 Occluded-Duke

The Occluded-Duke dataset [2] is constructed based on the DukeMTMC-reID dataset and focuses on dealing with the problem of recognizing pedestrians in occluded situations, including 15,618 training images, of which occluded images account for 9%, 17,661 gallery images, of which occluded images account for 10%, and 2,210 query images, of which occluded images account for 100%.

4.1.3 Experimental platform setup

In this paper, the PyTorch framework is used to conduct the experiments and trained and tested on NVIDIA 4060ti GPUs. Existing methods were followed, using the pre-trained ViT model on ImageNet [25] as the backbone. All image sizes were set to 256 × 128 and common methods were applied to augment their image data, such as horizontal flipping, random cropping, padding, and random erasure. The initial learning rate is 0.008, and the learning rate decay is cosine. The optimizer is used as SGD, where the epoch of Occluded-Duke dataset and DukeMTMC-reID dataset is 150 times.

To ensure a fair comparison, this paper adopts mean average precision (mAP) and cumulative matching feature (CMC) as evaluation indexes on the basis of previous research. Note: The results of cumulative matching features are used to rank the pedestrian images to assess the degree of matching between the reference image and the candidate images. The process was repeated 10 times to obtain this average value from two datasets.

4.2 Comparison with other advanced methods

In this section, to demonstrate the performance of the proposed model in this article, it is evaluated against the leading methods from the past 3 years, as shown in Table 2, and it can be seen that our model obtains an excellent performance of 73.6% Rank-1 and 63.2% mAP in the most challenging dataset of occluded pedestrians re-recognition, as shown in Table 3. We want to demonstrate that the method in this paper is not only applicable to the occlusion environment, but also to the

Table 2 Comparison with advanced methods on the Occluded-Duke dataset

	Occluded-Duke	
Method	Rank-1	mAP
PCB[26]	42.6%	33.7%
OAMN[27]	62.6%	46.1%
RFCNet[28]	63.9%	54.5%
FRT[29]	70.7%	61.3%
PFT[30]	69.8%	60.8%
OCNet[31]	64.3%	54.4%
DRL-Net[32]	65.8%	53.9%
PFD[5]	67.7%	60.1%
FED[33]	68.1%	56.4%
ETNDNet[34]	68.1%	57.9%
TransReID[19]	66.4%	59.2%
PRE-Net[35]	68.3%	55.2%
OURS	73.6%	63.2%

Table 3 Comparison with advanced methods on the DukeMTMC-reID dataset

	DukeMTMC-reID	
Method	Rank-1	mAP
PCB[26]	81.8%	66.1%
OAMN[27]	86.3%	72.6%
RFCNet[28]	90.7%	80.7%
FRT[29]	90.5%	81.7%
PFT[30]	90.7%	82.1%
OCNet[31]	90.5%	80.2%
DRL-Net[32]	88.1%	76.6%
PFD[5]	90.6%	82.2%
FED[33]	89.4%	78.0%
ETNDNet[34]	82.7%	58.0%
TransReID[19]	90.7%	82.0%
PRE-Net[35]	89.3%	77.8%
OURS	90.8%	82.4%

overall dataset. Therefore, we choose the representative DukeMTMC dataset for our experiments, and we can see that it also outperforms the other methods, achieving 90.8% Rank-1 and 82.4% mAP. Our improved method mainly involves two modules, OPAM and FtF, which have significant advantages in handling occluded pedestrian recognition and avoiding trajectory fragmentation. Among them, OPAM implements the attention mechanism through multi-scale convolution and adaptive pooling, which can effectively capture both local and global information. With learnable balancing parameters, the module is able to dynamically adjust the fusion of local and global information to improve the recognition of occluded pedestrians. FtF combines the channel attention and spatial attention mechanisms, which enhances the expression ability of features. By dynamically fusing the weights, the module is able to perform effective feature fusion between feature transformations and residual connections to reduce the occurrence of trajectory fragmentation. The modules are designed to improve the robustness and accuracy of the model in complex scenarios, especially in dealing with occlusion and trajectory continuity. With these innovations, our approach outperforms existing leading methods in recognizing occluded pedestrians and avoiding trajectory fragmentation. Not only but also our proposed RTL loss function significantly improves the recognition of occluded pedestrians and effectively reduces the occurrence of trajectory fragmentation by combining dynamic hard-case mining, multi-scale feature extraction and occlusion-weighted loss, thus showing better robustness and accuracy than the existing methods in complex scenes. It is clear that the approach presented in this article delivers superior performance in both the occlusion environment and the overall environment. That is, it is a comprehensive network framework that mainly solves the problem of pedestrian re-recognition in occluded environments but also solves pedestrian re-recognition in traditional environments.

4.3 Visualization and analysis

The heat map results show the areas that the model focuses on when processing the input image. The change in color allows us to visualize which parts have the most influence on the model's decisions. Typically, red areas indicate parts that the model pays high attention to, while blue or green areas indicate parts that are paid less attention to. This visualization helps to understand the behavior of the model, verifying whether it focuses on key features in the image or whether there are noisy regions that mislead the model. As shown in Fig. 4, two images are in a group, and in the same group, the later image is the heat map presented by the earlier image, respectively, and we can observe that the color of the highly focused part can be almost accurately localized to the unobstructed part of the human body, which proves the validity of our suggested approach.

4.4 Ablation study

The ablation experiments of the method proposed in this article are conducted based on the Occluded-Duke dataset, which contains the baseline, the feature fusion module FtF, the occlusion perceptual attention module OPAM and the integrated loss function RTL in order to demonstrate the effectiveness of our method. Our baseline

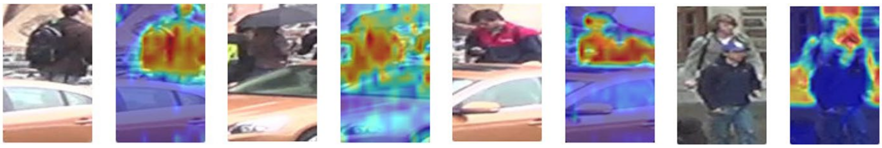

Fig. 4 Heat map visualization and analysis of occluded pedestrian re-identification for the model proposed in this article

is [36]. In order to obtain accurate and stable data, we conducted 10 separate experiments to obtain the average value. To prove the superiority of the proposed modules, they are added to the network model one by one. As shown in Table 4, the addition of OPAM module improves the two key metrics criteria rank1 and mAP by 1.0 and 0.6%, respectively. The OPAM module improves the overall performance of the model through several mechanisms. First, the module employs multi-scale convolution, which enables the model to extract features from different receptive fields and enhances feature representation. Second, the attention mechanism dynamically adjusts the importance of features by weighted summation of the outputs of the multi-scale convolution. The value branch extracts features through pointwise convolution and performs element-level multiplication with the output of the attention branch to achieve weighted feature fusion, combining local and global information. With these designs, OPAM contributes significantly in improving the model's multi-scale feature fusion capability, enhancing detailed feature capture and improving the overall performance, which is an integral part of the model. The FtF module improves rank1 and mAP by 0.9 and 0.9%, respectively, compared to baseline. The FtF boosting performance is analyzed as follows: First, the module employs a channel attention mechanism that dynamically adjusts the importance of each channel through adaptive average pooling and a series of convolution operations. Second, the spatial attention mechanism generates a spatial weight map through convolutional operations. The feature transform further extracts and enhances the features. Residual concatenation, on the other hand, ensures that the original information of the input features is preserved, enhancing the stability and robustness of the model. With these designs, FtF contributes significantly in enhancing the feature selection capability of the model, increasing spatial and channel attention and improving the overall performance, and is an indispensable part of the model. RTL improves rank1 and mAP by 1.1 and 0.5%, respectively. RTL optimizes the feature space distribution of the model by calculating the difference in distance between the anchor samples and the positive samples (similar) and the negative samples (dissimilar), thus enhancing the discriminative ability of the model. Second, the loss function introduces an optional margin parameter to control the minimum difference between the distances of the positive and negative samples, which further improves the robustness of the model. In addition, a feature normalization option is also supported to

Table 4 Evaluation of components on the Occluded-Duke dataset

Baseline	OPAM	FtF	RTL	mAP	Rank-1	epoch
✓				61.0%	71.0%	150
✓	✓			61.6%	72.0%	150
✓		✓		61.9%	71.9%	150
✓			✓	61.5%	72.1%	150
✓	✓	✓		62.8%	73.1%	150
✓	✓		✓	62.7%	72.9%	150
✓		✓	✓	61.8%	71.8%	150
✓	✓	✓	✓	63.2%	73.6%	150

ensure the stability of the distance calculation. With these designs, RTL contributes significantly in optimizing the feature discrimination ability of the model, enhancing the sample differentiation and improving the overall performance. When the OPAM module acts together with the FtF module, compared to baseline, rank1 and mAP are improved by 2.1 and 1.8%, respectively. Finally, when OPAM, FtF and RTL are all added, rank1 and mAP are improved by 2.6 and 2.2% compared to baseline. The results show that the added modules as well as the improved loss function are meaningful, and these components make an excellent contribution to an effective framework that achieves good performance.

5 Conclusion

In this article, we propose a ViT-based occluded pedestrian re-identification model and make the generated mask pass through the occlusion perceptual attention module and the feature fusion module in turn, which effectively captures the local and global features, improves the model's expressiveness and recognition accuracy and performs the feature enhancement, and then make it pass through the improved ternary loss function in this paper - the composite loss function, which enhances the stability and accuracy of the model in dealing with difficult samples by introducing a regularization term and a dynamic weight adjustment mechanism, making similar samples closer and dissimilar samples more separated. Through experimental validation on several public datasets, our method achieves good performance. Future work will further explore more attention mechanisms and feature fusion strategies with the aim of achieving better performance on larger datasets.

Author's contribution YG designed and conducted experiments, data analysis and manuscript editing. ZL, YS and BL provide theoretical guidance for this paper. YS and XL organize the charts. All authors reviewed the manuscripts.

Funding This work was supported by the Project of Liaoning Provincial Education Department (LJKZ0515), Cross-modal Pedestrian Re-identification Based on Deep Learning and Corporate Co-operation Project (LJKFZ20220206) and Personnel and Vehicle Monitoring System for Logistics Parks.

Data availability The data supporting the findings of this study are available from the corresponding author upon request.

Declarations

Conflict of interest The authors have no conflict of interest to declare.

References

1. Zhuo J, Chen Z, Lai J, Wang G (2018) Occluded Person Re-identification. In: 2018 IEEE International Conference on Multimedia and Expo (ICME), pp 1–6 . IEEE

2. Miao J, Wu Y, Liu P, Ding Y, Yang Y (2019) Pose-Guided Feature Alignment for Occluded Person Re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 542–551
3. Gao S, Wang J, Lu H, Liu Z (2020) Pose-Guided Visible Part Matching for Occluded Person Reid. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 11744–11752
4. Jia M, Cheng X, Lu S, Zhang J (2022) Learning disentangled representation implicitly via transformer for occluded person re-identification. IEEE Trans Multimed 25:1294–1305
5. Wang T, Liu H, Song P, Guo T, Shi W (2022) Pose-guided feature disentangling for occluded person re-identification based on transformer. Proc AAAI Conf Artif Intell 36:2540–2549
6. Dosovitskiy A (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929
7. Li Y, Zhang Y, Gao Y, Xu B, Liu X (2024) Fcl: pedestrian re-identification algorithm based on feature fusion contrastive learning. Electronics 13(12):2368
8. Tan H, Liu X, Yin B, Li X (2022) Mhsa-net: multihead self-attention network for occluded person re-identification. IEEE Trans Neural Netw Learn Syst 34(11):8210–8224
9. Zhou X, Zhong Y, Cheng Z, Liang F, Ma L (2023) Adaptive Sparse Pairwise Loss for Object Re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 19691–19701
10. Wu L, Shen C, Hengel A v d (2016) Personnet: person re-identification with deep convolutional neural networks. arXiv preprint arXiv:1601.07255
11. Luo H, Gu Y, Liao X, Lai S, Jiang W (2019) Bag of Tricks and a Strong Baseline for Deep Person Re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp 0–0
12. Zheng F, Deng C, Sun X, Jiang X, Guo X, Yu Z, Huang F, Ji R (2019) Pyramidal Person Re-identification Via Multi-loss Dynamic Training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 8514–8522
13. Wei L, Zhang S, Gao W, Tian Q (2018) Person Transfer Gan to Bridge Domain Gap for Person Re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 79–88
14. Yi D, Lei Z, Liao S, Li S Z (2014) Deep Metric Learning for Person Re-identification. In: 2014 22nd International Conference on Pattern Recognition, pp 34–39. IEEE
15. Wang G, Yang S, Liu H, Wang Z, Yang Y, Wang S, Yu G, Zhou E, Sun J (2020) High-Order Information Matters: Learning Relation and Topology for Occluded Person Re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 6449–6458
16. Zheng L, Huang Y, Lu H, Yang Y (2019) Pose-invariant embedding for deep person re-identification. IEEE Trans Image Process 28(9):4500–4509
17. Zhang X, Yan Y, Xue J-H, Hua Y, Wang H (2020) Semantic-aware occlusion-robust network for occluded person re-identification. IEEE Trans Circuits Syst Video Technol 31(7):2764–2778
18. Jin H, Lai S, Qian X (2021) Occlusion-sensitive person re-identification via attribute-based shift attention. IEEE Trans Circuits Syst Video Technol 32(4):2170–2185
19. He S, Luo H, Wang P, Wang F, Li H, Jiang W (2021) Transreid: Transformer-Based Object Re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 15013–15022
20. Liu H, Feng J, Qi M, Jiang J, Yan S (2017) End-to-end comparative attention networks for person re-identification. IEEE Trans Image Process 26(7):3492–3506
21. Woo S, Park J, Lee J -Y, Kweon IS (2018) Cbam: Convolutional Block Attention Module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 3–19
22. Song C, Huang Y, Ouyang W, Wang L (2018) Mask-Guided Contrastive Attention Model for Person Re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1179–1188
23. Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-Cam: Visual Explanations from Deep Networks Via Gradient-Based Localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp 618–626
24. Zheng Z, Zheng L, Yang Y (2017) Unlabeled Samples Generated by Gan Improve the Person re-Identification Baseline in Vitro. In: Proceedings of the IEEE International Conference on Computer Vision, pp 3754–3762

25. He K, Zhang X, Ren S, Sun J (2016) Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778
26. Sun Y, Zheng L, Yang Y, Tian Q, Wang S (2018) Beyond Part Models: Person Retrieval with Refined Part Pooling (and a Strong Convolutional Baseline). In: Proceedings of the European Conference on Computer Vision (ECCV), pp 480–496
27. Chen P, Liu W, Dai P, Liu J, Ye Q, Xu M, Chen Q, Ji R (2021) Occlude Them All: Occlusion-Aware Attention Network for Occluded Person Re-id. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 11833–11842
28. Hou R, Ma B, Chang H, Gu X, Shan S, Chen X (2021) Feature completion for occluded person re-identification. IEEE Trans Pattern Anal Mach Intell 44(9):4894–4912
29. Xu B, He L, Liang J, Sun Z (2022) Learning feature recovery transformer for occluded person re-identification. IEEE Trans Image Process 31:4651–4662
30. Zhao Y, Zhu S, Wang D, Liang Z (2022) Short range correlation transformer for occluded person re-identification. Neural Comput Appl 34(20):17633–17645
31. Kim M, Cho M, Lee H, Cho S, Lee S (2022) Occluded Person Re-identification Via Relational Adaptive Feature Correction Learning. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 2719–2723 . IEEE
32. Jia M, Cheng X, Lu S, Zhang J (2022) Learning disentangled representation implicitly via transformer for occluded person re-identification. IEEE Trans Multimed 25:1294–1305
33. Wang Z, Zhu F, Tang S, Zhao R, He L, Song J (2022) Feature Erasing and Diffusion Network for Occluded Person Re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 4754–4763
34. Dong N, Zhang L, Yan S, Tang H, Tang J (2023) Erasing, transforming, and noising defense network for occluded person re-identification. IEEE Trans Circ Syst Video Technol
35. Yan G, Wang Z, Geng S, Yu Y, Guo Y (2023) Part-based representation enhancement for occluded person re-identification. IEEE Trans Circ Syst Video Technol 33(8):4217–4231
36. Tan L, Dai P, Ji R, Wu Y (2022) Dynamic Prototype Mask for Occluded Person Re-identification. In: Proceedings of the 30th ACM International Conference on Multimedia, pp 531–540

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Yuepeng Guo[1] · ZhenPing Lan[1] · Yanguo Sun[2] · Yuheng Sun[1] · Xinxin Li[1] · Yuru Wang[1] · Bo Li[1]

✉ ZhenPing Lan
lanzp@dlpu.edu.cn

✉ Yanguo Sun
563083592@qq.com

Yuepeng Guo
230520854000584@xy.dlpu.edu.cn

Yuheng Sun
230520854000563@xy.dlpu.edu.cn

Xinxin Li
230510811000507@xy.dlpu.edu.cn

Yuru Wang
wangyr@dlpu.edu.cn

Bo Li
libo_15@dlpu.edu.cn

[1] Information Science and Engineering, Dalian Polytechnic University, Dalian 116038, China

[2] Informat Ctr Second Hosp, Dalian Medical University, Dalian 116034, China