Multimedia Systems (2024) 30:5
https://doi.org/10.1007/s00530-023-01235-2

REGULAR PAPER

Occluded pedestrian re-identification via Res-ViT double-branch hybrid network

Yunbin Zhao[1] · Songhao Zhu[1]

Received: 7 July 2023 / Accepted: 8 December 2023 / Published online: 12 January 2024

Abstract

Existing occluded pedestrian re-identification methods mainly utilize convolutional neural networks to realize the feature matching under different camera perspectives. Due to the complex occlusion situation, the accuracy of occluded pedestrian re-identification is not so satisfied where convolutional neural networks are utilized to extract local features. Convolutional neural network is unique in its ability to capture local features, but its global modeling ability is weak. In contrast, Vision Transformer (ViT) can efficiently extract global features from shallow layers with more spatial information and obtain intermediate features with high quality. To deal with the above issues, ViT is here introduced into the residual network to construct a dual-branch hybrid network of residual network and visual converter (DB-ResHViT), where the ViT branch is utilized to reduce training errors, while the residual-ViT branch is utilized to construct the global correlation of feature sequences extracted by the residual network. The proposed network proposes a novel data augmentation module, called partial image patch pre-convolution module (PPPC), which is utilized to input the extracted partial image patches into the pre-convolution network to replace the original image patches to achieve the goal of introducing local features into the ViT branch. In addition, the proposed network designs a novel module integrating residual and mobile vision transformer, called RMV Module, which is utilized to establish the global correlation of local features extracted by the residual network to achieve the goal of reducing the computational cost and improve the re-identification accuracy. Experimental results of a large number of occluded pedestrian re-identification datasets demonstrate that the performance of the proposed method is superior to other advanced methods.

Keywords Occluded pedestrian Re-identification · Double-branch hybrid structure · Image block partial pre-convolution enhancement · Residual-vision transformer module

1 Introduction

Pedestrian re-identification aims to identify target pedestrians appearing in different cameras and is widely used for safety monitoring in public places including schools, shopping malls, supermarkets, and train stations [1]. In recent years, scholars have proposed a variety of methods to solve the problem of pedestrian re-identification [2–6]. Most of these methods utilize convolutional neural network to extract discriminative feature representations and have achieved good recognition results on open experimental dataset. For example, image patch-based pedestrian recognition methods utilize the attention mechanism in reference [7], the multi-branch structure in reference [8], and other strategies to introduce local features of pedestrians to handle pedestrian re-identification problems. Fine-grained information-based pedestrian re-identification methods utilize various strategies including pose estimation and key points to extract fine-grained features of pedestrians to improve the performance of pedestrian re-identification. Generative adversarial network-based pedestrian re-identification methods generate pedestrian images under different angles and different illumination, which helps enrich training samples and improve the robustness of the network. However, it is difficult for most pedestrian re-identification methods to achieve satisfactory recognition accuracy in the case of illumination change, background clutter, pedestrian occlusion, etc.

Due to the Gaussian distribution, the receptive field of convolutional neural network is limited to a small region

✉ Songhao Zhu
zhush@njupt.edu.cn

[1] College of Automation and Artificial Intelligence, Nanjing University of Posts and Telecommunications, Nanjing, China

[9–12]. Under pedestrian occlusion, background clutter or other noise, it is easy for the smaller receptive field to receive incorrect feature information; in addition, due to the fact that down-sampling reduces the resolution of feature representation, it is easy to reduce the accuracy of occluded pedestrian re-identification methods with small receptive field [13–16]. Therefore, even if the strategy of feature alignment or attention mechanism in [17] is introduced, it is difficult to completely solve the challenge of using convolution neural network to solve the problem of occluded pedestrian re-identification.

Literature [18] has proved that visual transformer can be utilized for image classification, and its effect is not inferior to that of traditional convolution neural network. The reason is that vision transformer takes the multi-head self-attention mechanism as its core, and abandons convolution and down-sampling [19]. Specifically, firstly, vision transformer divides the original image into a series of image patches; then, vision transformer performs classification coding and position coding on these image patches; finally, visual transformer performs self-attention on these image patches.

In recent years, with the ability to capture global feature information and better self-attention mechanism, visual transformer has achieved good results in the field of pedestrian recognition [20, 21]. However, when most of the pedestrian's body is occluded or the background information is similar to the pedestrian's features, visual transformer is prone to making misjudgments because it is not good at capturing local features of the pedestrian, resulting in poor robustness. This article aims to simultaneously take into account the advantages of convolution neural network and vision transformer, and construct a novel network that can not only depict the correlation of the pedestrian’s global characteristics but also the pedestrian’s local characteristics, so as to further improve the accuracy of occluded pedestrian re-identification.

In this article, a dual-branch hybrid structure based on residual network and visual transformer is proposed to deal with the issue of occluded pedestrian re-identification. Firstly, an original input image is augmented by using the proposed partial patch pre-convolution module; secondly, the augmented images are input into the vision transformer branch to establish the global feature relationship of the image sequence; thirdly, the features extracted from the original input image are input into the residual-visual transformer branch, where the proposed residual-visual transformer module is utilized to extract local features; fourthly, the feature information extracted from these two branches is fused to obtain the pedestrian’s discriminative features; finally, the loss is iteratively calculated to complete the network training. Some representative results of class activation maps between the vision transformer-based method and the proposed DB-ResHViT method are given in the following figure. It can be seen from Fig. 1 that our DB-ResHViT pays more attention on local features than original vision transformer and improves the accuracy of occluded pedestrian re-identification.

The main contributions can be summarized as follows:

- We propose an effective occluded pedestrian re-identification network, which extracts local feature information and establishes global feature relationship through the proposed dual-branch hybrid structure of residual network and vision transformer.
- We propose a novel data augmentation module, which inputs randomly selected image patches into the pre-convolution module to replace those original image patches, thereby obtaining discriminative local features.
- We design a novel module integrating residual structure and visual transformer, which utilizes the translation invariance of convolution neural network to construct the relationship between global features, so as to reduce the computational cost and improve the occluded pedestrian re-identification accuracy.
- Extensive experiments have been conducted on public occluded pedestrian re-identification datasets, and experimental results demonstrate that the proposed network significantly improves the performance of occluded pedestrian re-identification.

2 Related work

Most studies mainly utilize the entire body information of a pedestrian to handle the pedestrian re-identification issue, with less consideration given to pedestrian re-identification under occlusion conditions. However, it is difficult to obtain the entire body information of a pedestrian in real life, especially in crowded scenes. Therefore, the issue of occluded pedestrian re-identification cannot be ignored. Next, we will review the research on occluded pedestrian re-identification in recent years.

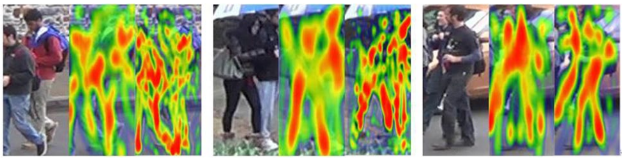

Fig. 1 Class attention maps of different methods:(1) A givenimage, (2) Original vision transformer-based method, (3) Proposed DB-ResHViT based method

2.1 Occluded pedestrian re-identification

Existing deep learning methods for occluded pedestrian re-identification mainly utilize convolution neural network to extract pedestrians' features. This category of method first aligns features or introduces higher-order semantic information (gesture information), then utilizes gesture model to estimate the position information of key points, and finally utilizes gesture information to complete pedestrian re-identification. Literature [22] proposes a robust soft matching feature alignment method, which utilizes hierarchical joint learning to obtain local features of a pedestrian's gestures to predict the similarity of different pedestrians. Literature [23] presents a local matching method based on gesture information, which designs a visible local feature predictor and utilizes the attention mechanism to achieve the representation of pedestrians' characteristics. Literature [24] utilizes a multi-granularity network based on gesture key points to extract multi-granularity features to eliminate the impact of occlusion. Although gesture information is beneficial for improving the performance of pedestrian re-identification, the key point estimation model makes the whole network slightly bloated, thereby reducing the network's running speed. In recent years, some scholars have attempted to solve the occlusion issue through local feature matching. Literature [25] first utilizes an object detection network. To segment each image into a sequence of image patches, then extracts multi-scale features, and finally realizes feature matching based on spatial similarity. Literature [26] proposes a dual-branch network, which improves the network robustness by extracting fine-grained multi-scale features.

2.2 Visual transformer

Visual transformer is a common deep learning model in the field of natural language processing. The multi-head attention mechanism proposed in literature [27] completely abandons network structures such as recurrent neural network and convolutional neural network, which is utilized to deal with the machine translation tasks and achieves good results. Google introduces the transformer model into the field of image classification and proposes the famous visual transformer model, which input the segmented image patch sequence into the transformer encoder, maximizing the preservation of the original structure of the transformer and achieving satisfactory results. Similar to convolutional neural network, visual transformer requires a large number of datasets to complete parameter training. Therefore, literature [28] proposes an efficient image data transformation framework through attention mechanism and optimizes the pedestrian re-identification problem using the teacher–student strategy. Literature [29] designs a pedestrian re-identification model based on visual transformer, which utilizes a mosaic module to classify the features of the last layer and calculates the losses separately, further enhancing the robustness of the model. However, the pedestrian re-identification model based on visual transformer still focuses on characterizing global features, while some issues such as local occlusion features and short-distance correlation have not been well addressed.

Compared with TransREID in [29] which only uses ViT, our method combines ViT with convolution to design a novel model that reduces the number of parameters and establishes the connection between local features. In addition, we enhance the local feature extraction capability by introducing residual network branch and improve the identification accuracy by introducing parallel dual branches.

3 Proposed method

This article proposes a dual-branch hybrid structure occluded pedestrian re-identification network integrating residual network and visual transformer, as shown in Fig. 2. The proposed network includes the following two branches: the visual transformer branch based on the partial image patch pre-convolution module, and the residual-visual transformer branch based on the residual-visual transformer module. The vision transformer branch is utilized to establish global feature correlation, and the residual-vision transformer branch is. Utilized to extract translation invariant features. The fused feature contains both translation invariant local feature and spatially relevant global feature, improving the distinguishability of extracted features and enhancing their generalization ability. Next, we will introduce the method proposed in this article in detail.

3.1 Visual transformer branch

Vision transformer branch utilizes the partial image patches pre-convolution module to extract shallow local features, which helps improve the ability of vision transformer to characterize local features. Next, we will introduce the image augmentation module designed in this article, namely, the partial image patch pre-convolution module.

Figure 3 gives the framework of partial image patch pre-convolution module (PPPC) designed in this article. Assuming an input image x (H, W, and C), where H, W, and C represent the height, width, and channel size of the input image x, respectively. Due to low resolution, illumination change, mutual occlusion, background clutter, and inconsistent feature distribution, deep learning neural network failed to well complete the task of occluded pedestrian re-identification and the re-identification accuracy is not so satisfactory. Therefore, it is necessary to perform image augmentation before training.

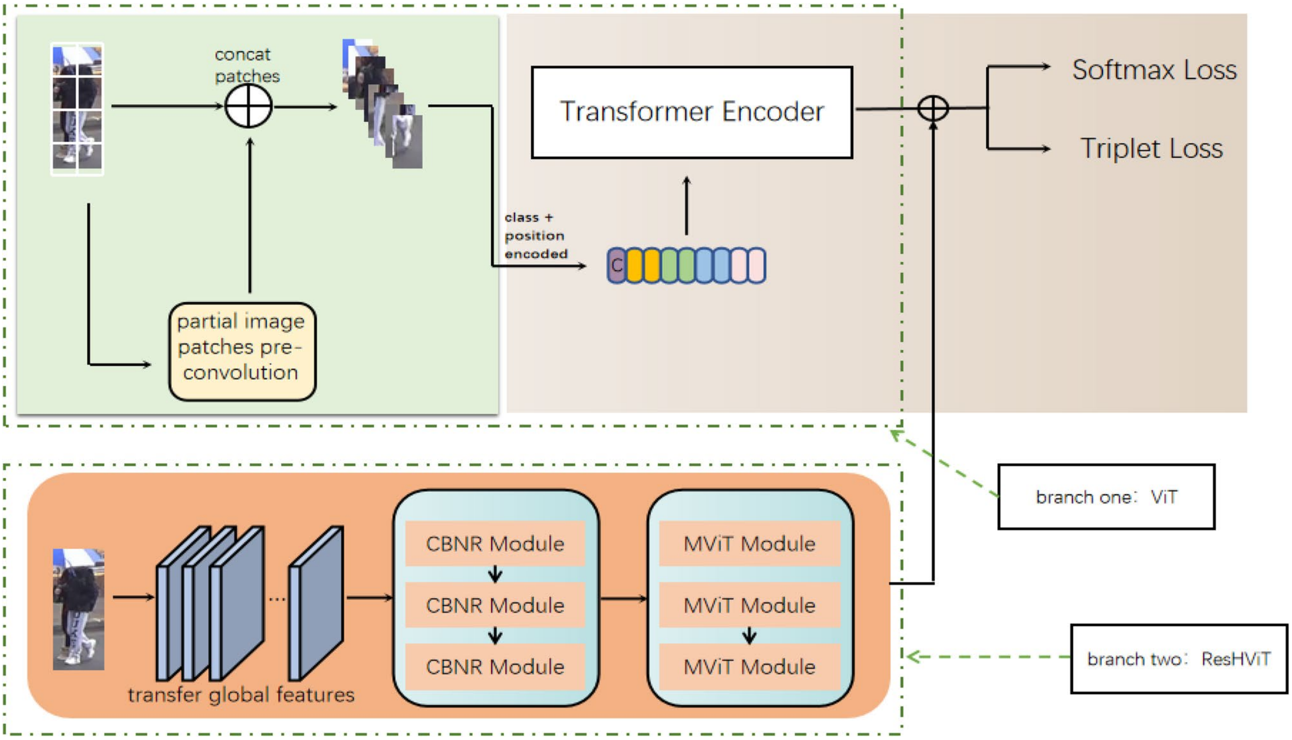

Fig. 2 Framework of the proposed dual-branch hybrid network for occluded pedestrian re-identification, where *branch one* is the vision transformer branch and branch two is the *residual-visual transformer* branch. The proposed partial image patch pre-convolution (PPPC) module and Convolution-Batch Normalization-Residual (CBNR) module are detailed in Sect. 3

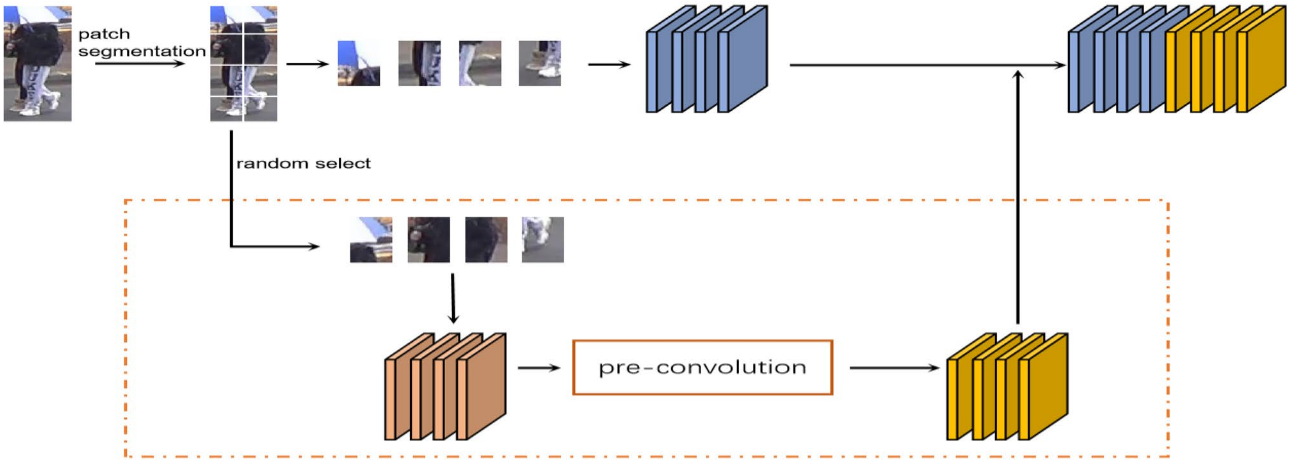

Fig. 3 Framework of the proposed partial image patch pre-convolutionmodule, where patch segmentation represents segmenting an image into multipleimage patches and random select represents selecting a portion of image patches for pre-convolution

Partial image patch pre-convolution module is utilized to implement image augmentation. Firstly, an input image is segmented into several image patches; secondly, some image patches are randomly selected and input into a convolutional neural network to extract shallow local features; thirdly, these extracted local features are input into the visual transformer branch. Therefore, the proposed PPPC module is utilized to introduce local features into global features extracted from the visual transformer branch.

During the process of randomly selecting image patches, all image patches are labeled with self-learning and a set of image patches is randomly selected to learn the discriminative feature distribution of occluded pedestrians for pre-convolution. Therefore, the robustness is guaranteed.

In this article, partial image patch pre-convolution module is added into the vision transformer branch, and the corresponding pseudocode of partial image patch pre-convolution module is shown in Table 1. Firstly, the input image is segmented into several image patches; secondly, a random percentage is set to select the corresponding proportion of segmented image patches for pre-convolution; then, a convolutional neural network is selected as the backbone network to complete the pre-convolution task, and ResNet50 is here selected as the backbone network through comparative experiments; finally, the convolutional features of randomly selected image patches are spliced with the features of original image patches in the extraction order, which helps ensure that the original feature distribution from these randomly selected image patches is not affected.

The algorithm of the partial image patch pre-convolution module is described as follows:

Step 1: Obtain the length of initial image patch sequence.

Step 2: Utilize the *Random* library function to randomly select *length*Percent* image patches, where *Select* represents the index list of image patches.

Step 3: Utilize the obtained index list *Select* to obtain the image patches sequence *SelectPatch* for subsequent feature extraction.

Step 4: Subtract *SelectPatch* from original image patches sequence *OriginalPatch* to obtain the rest image patches sequence *RestPatch*.

Step 5: Input *SelectPatch* into the pre-convolution network of *Resnet50* to obtain the pre-convolution image patch feature sequence *PreConvPatch*.

Step 6: Splice the pre-convolution image patch feature sequence *PreConvPatch* with the rest image patch sequence *RestPatch* to obtain the partial image patch pre-convolution feature sequence *OutputPatch*.

Step 7: End.

Due to the consistency between the selection order of partial image patches and the splice order of corresponding image patches, the feature distribution of the original input image will not change. In addition, due to the strong representation ability and low model parameters of the ResNet50 network, it can ensure that the convolutional features of randomly selected image blocks have strong discriminability. That is to say, thereby introducing local features into global features extracted by the visual transformer branch helps improve the robustness of the proposed model for occluded pedestrian re-identification.

Table 1 Pseudocode of Partial Image Patch Pre-Convolution

```
Algorithm: Partial Image Patches Pre-Convolution
Input: OriginalPatch and Percent
Output: OutputPatch
length = Length (OriginalPatch)
Select = Random(length*Percent)
SeletcPatch = OriginalPatch[Select]
RestPatch = OriginalPatch - SelectPatch
PreConvPatch = Resnet50(SelectPatch)
OutputPatch = PreConvPatch + RestPatch
End
```

3.2 Residual-visual transformer branch

Residual networks are often utilized to solve the issue of occluded pedestrian re-identification, where the reason is that residual networks can effectively deal with the issue of gradient disappearance. However, there exists the following two problems for residual networks: (1) With the increase in the number of network layers, there still exists many redundant parameters in deep residual network; (2) Compared with other convolutional neural networks, residual networks enlarge the receptive field to a certain extent; however, the enlarged receptive field is still limited to a certain region, which affects the performance of occluded pedestrian re-identification based on local receptive field.

Recent research has shown that vision transformer also has excellent performance in the field of computer vision. For small-scale and medium-scale datasets, the performance of visual transformer is inferior to that of residual network, because residual network has built-in biases and translation invariant. In other words, for large-scale datasets, the performance of visual transformer is much better than that of residual network, where the reason is that the advantages of visual transformer can be fully demonstrated through the training on large-scale datasets. Compared with residual network, visual transformer also has many disadvantages, such as multiple model parameters, slow computational speed, and weak local feature extraction ability. Therefore, this article attempts to combine the advantages of these two models to design a residual-vision transformer branch, as shown in Fig. 4. Such a branch design is not only conducive to extracting discriminative local features using residual network but also conducive to constructing the correlation between local features using visual transformer.

The training process of the residual-visual transformer branch is described as follows: firstly, the feature of an occluded pedestrian is extracted through convolution, and the preliminary feature is extracted through batch normalization, ReLu activation function, and maximum pooling; secondly, the extracted preliminary feature is input into the convolution-batch normalization-residual module (CBNR) as shown in Fig. 5 to extract deep feature; thirdly, the extracted deep feature is input into the mobile visual transformer module as shown in Fig. 5 to extract discriminative

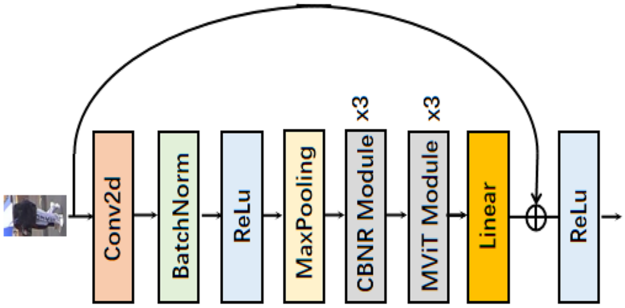

Fig. 4 Framework of the proposed residual-vision transformer branch, which consists of convolution, bath normalization, nonlinearization, max pooling, convolution-batch normalization-residual (CBNR) andlinearization

feature; fourthly, the extracted discriminative feature is input into the linear layer to adjust the number of feature channels, and a short-cut structure is constructed with original input image feature; finally, the residual-visual transformer feature is the output result of ReLu activation function.

The residual-visual transformer branch leverages the advantages of residual network and visual transformer: (1) Visual transformer can not only represent global features but also reduce the computational cost of the network; (2) Residual network can not only effectively prevent over-fitting but also add the spatial accumulation bias missing in visual transformer, which optimizes the network structure and improves the network accuracy.

Next, we will introduce the residual-visual transformer module proposed in this article in detail.

3.3 Residual-vision transformer module

The design idea of the convolution-batch normalization-residual module is described as below. As shown in the dotted box on the left of Fig. 5, the convolution-batch normalization-residual module utilizes the short-cut structure in residual networks to extract shallow features. Such a CBNR module can deepen the number of feature channels and reduce the disappearance of feature gradients. Theoretically, such a CBNR module can be stacked countless times, and experimental results demonstrate that optimal performance can be achieved when the module is stacked three times.

Due to the large number of model parameters and high-complexity of floating-point calculations in the visualization transformer, a mobile visual transformer module (MViT) is here utilized to reduce network complexity and improve computational speed. The reasons for such a processing are described as follows: (1) The mobile vision transformer requires fewer parameters to model the local and global features of the input tensor. (2) The accumulated bias generated during the convolution process is introduced into vision transformer to improve the stability and robustness of the model.

As shown in the dotted box on the right of Fig. 5, the mobile visual transformer module first transforms the input feature tensor into a feature sequence through convolution and inputs the feature sequence into visual transformer, which corresponds to the structure from *Conv* to *Dropout*; then, the self-attention mechanism is added into the feature sequence, which corresponds to the structure from *Linear* to *Dropout*; finally, the feature sequence is transformed back to the original dimension through convolution, which corresponds to the structure from *Conv* to *Conv*. Theoretically, such a mobile visual transformer module can be stacked countless times, and experimental results demonstrate that optimal performance can be achieved when the module is stacked three times.

4 Experiments

A large number of experiments are conducted on the proposed dual-branch hybrid network combining residual network and visual transformer to verify the effectiveness of

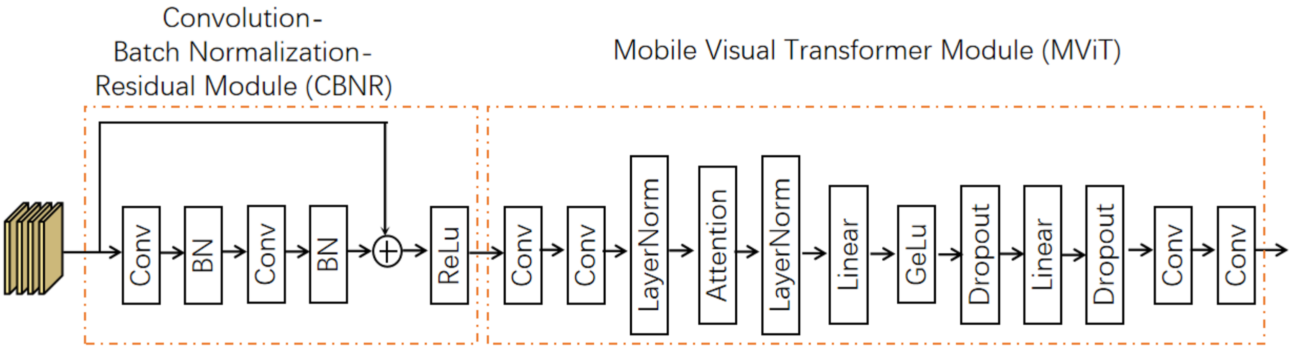

Fig. 5 Framework of the proposed residual-vision transformer module, which consists of convolution, batch normalization, residual, attention, linearization, and nonlinearization

the proposed network in solving the problem of occluded pedestrian re-identification and to verify whether the proposed network fully leverages the advantages of these two models to achieve better performance than residual network.

4.1 Experimental dataset

The performance of the dual-branch hybrid network proposed in this article is evaluated on six public datasets, including the Occluded-REID dataset in literature [30], the Occluded-Duke dataset in literature [31], the Market-1501 dataset in literature [32], the DukeMTMC-REID dataset in literature [33], the Partial-REID dataset in literature [34], and the Partial-iLIDS dataset in literature [35].

Occluded-REID dataset: This dataset contains 2000 images from 200 pedestrians, including 5 full-body images and 5 occluded images for each pedestrian.

Occluded-Duke dataset: This is the largest occluded pedestrian re-identification dataset so far, which contains 35,489 images from 1110 pedestrians, including 15,618 training images, 17,661 validation images, and 2210 testing images.

Market-1501 dataset: This dataset contains 32,668 images from 1501 pedestrians, including 12,936 training images and 19,732 testing images.

DukeMTMC-REID dataset: This dataset contains 36,411 images from 1812 pedestrians, including 16,522 randomly selected training images, 2228 randomly selected validation images, and 17,661 randomly selected testing images.

Partial-REID dataset: This is the first pedestrian re-identification dataset, which contains 900 images from 60 pedestrians, including 5 full-body images, 5 local images, and 5 occluded images for each pedestrian.

Partial-iLIDS dataset: This dataset is a pedestrian re-identification dataset based on iLIDS, which contains 476 images from 119 pedestrians, including an average of 4 images for each pedestrian.

4.2 Experimental setting

Backbone Network: This article utilizes a self-designed backbone network with visual transformer branch and residual-visual transformer branch, where the visual transformer branch is composed of data enhancement module and visual transformer module, and the residual-visual transformer branch is composed of the residual module and residual-visual transformer module.

Training Detail: PyTorch is here utilized to build the network framework: firstly, the size of each input image is uniformly adjusted to 256 × 128, and each input image is augmented through flipping, filling, random horizontal, random cutting, and random erasing; secondly, the batch size is set to be 96, and the momentum of the Adam optimizer is set to be 0.9; thirdly, the weight attenuation parameter is set to be $1e^{-4}$, and the initial learning rate is set to be 0.01; finally, the first ranking accuracy (Rank-1), the top 5 ranking accuracy (Rank-5), the top 10 ranking accuracy (Rank-10), and the mean average accuracy (mAP) are utilized to evaluate the performance of the proposed network, and all experiments are performed under a single query condition.

4.3 Comparison method

Four categories of pedestrian re-identification methods are here selected for performance comparison, including the common pedestrian re-identification methods (Part Aligned [36]、PCB [37]、Adver occluded [38]), external information or semantic information based pedestrian re-identification methods on (PGFA [31]、PVPM [23]、HONet [24]、PAFM [25]、Part Bilinear [39]、FD-GAN [40]), local feature based pedestrian re-recognition methods (MFM + HFR [26]、FGMFN [27]、DSR [35]、SFR [41]、MoS [42]、FPR [43]、VPM [44]、MGCAM [45]), and occluded pedestrian re-recognition methods (TransReID [29]、MAT [46]、PAT [47]、DRL-Net [51]、PFD [53]、FRT [54]).

4.4 Experimental result

- Performance Verification on Occluded-Duke Dataset

According to the experimental results on Occluded-Duke dataset as shown in Table 2, compared with the HONet method based on the gesture information, the proposed DB-ResHViT method has increased by 15.8 percent of Rank-1 and 18.3 percent of mAP; at the same time, compared with other occluded pedestrian re-recognition method, the retrieval accuracy of the proposed DB-ResHViT method is also greatly improved. The above experimental results demonstrate that the combination of vision transformer and residual network is more conducive to solving the problem of occluded pedestrian re-recognition.

As the first method to introduce visual transformer into the field of pedestrian re-recognition, TransReID has refreshed the best performance of many pedestrian re-recognition, as well as in the field of occluded pedestrian re-recognition. According to the experimental results as shown in Table 2, compared with 66.4 percent of Rank-1 and 59.2 percent of mAP of the TransReID method, the Rank-1, and mAP of the proposed DB-ResHViT method increased by 4.5 percent and 2.9 percent, respectively. This proves that the performance of visual transformer has been further improved after incorporating the local feature receptive field of convolution.

Table 2 Experimental results of different methods on Occluded-Duke dataset

Methods	Occluded-Duke			
	Rank-1	Rank-5	Rank-10	mAP
Part Aligned [36]	28.8	44.6	51.0	20.2
PCB [37]	42.6	57.1	62.9	33.7
Part Bilinear [39]	36.9	–	–	–
FD-GAN [40]	40.8	–	–	–
HONet [23]	55.1	–	–	43.8
PAFM [24]	55.1	–	74.5	42.3
DSR [35]	40.8	58.2	65.2	30.4
SFR [41]	42.3	60.3	67.3	32.0
MoS [42]	61.0	74.4	79.1	49.2
FGMFN [26]	65.5	80.4	84.4	55.8
MFM + HFR [25]	48.9	59.1	65.1	41.9
PGFA [31]	51.4	–	83.6	37.3
Ad Occluded [38]	44.5	–	–	32.2
TransReID [29]	66.4	–	–	59.2
MAT [47]	66.2	–	–	58.8
PFD [53]	69.5	–	–	61.8
FRT [54]	70.7	–	–	61.3
DB-ResHViT(Ours)	**70.9**	**83.0**	**87.2**	**62.1**

Bold value represents the maximum in the corresponding column

Table 3 Experimental results of different methods on Occluded-REID dataset and Partial-REID dataset

Methods	Occluded-REID		Partial-REID	
	Rank-1	mAP	Rank-1	Rank-1
PCB [37]	41.3	38.9	66.3	63.8
Part Bilinear [39]	54.9	50.3	57.7	59.3
DSR [40]	72.8	62.8	43.0	–
FPR [43]	78.3	68.0	81.0	–
HONet [23]	80.3	70.2	85.3	–
PAFM [24]	76.4	68.0	–	–
PVPM [22]	70.4	61.2	78.3	–
PGFA [31]	–	–	68.0	–
PAT [38]	81.6	72.1	88.0	81.6
FGMFN [26]	84.4	79.5	82.5	72.8
MFM + HFR [25]	–	–	68.2	64.9
PFD [53]	81.5	79.3	-	–
FRT [54]	80.4	71.0	88.2	–
DB-ResHViT(Ours)	**84.8**	**79.7**	88.5	**78.9**

Bold value represents the maximum in the corresponding column

- Performance Verification on Occluded-REID and Partial-REID Datasets

Here, Market1501 dataset is selected as the pre-training dataset of the Partial-REID dataset, and the Occluded-Duke dataset is selected as the pre-training dataset of the Occluded-REID dataset. This is because that the Occluded-REID dataset can be considered as an occlusion dataset, so using occlusion dataset for pre-training is easier to achieve satisfactory results. Experimental results as shown in Table 3 verify such hypothesis.

It can be seen from the experimental results as shown in Table 3 that the DB-ResHViT method proposed in this article achieves the best Rank-1 and mAP on the Occluded-REID dataset and Partial-ReID dataset. Specifically, (1) compared with 80.3 percent of Rank-1 obtained by convolution neural network-based HONet method on the Occluded-REID dataset, the proposed DB-ResHViT method achieves 84.8 percent of Rank-1; (2) compared with 81.6 percent of Rank-1 and 72.1 percent of mAP obtained by vision transformer-based PAT method on the Occluded-REID dataset, the proposed DB-ResHViT method increases 3.2 percent of Rank-1 and 7.6 percent of mAP, respectively; (3) compared with 85.3 percent of Rank-1 obtained by convolution neural network-based HONet method on the Partial-ReID dataset, the proposed DB-ResHViT method achieves 88.5 percent of Rank-1; and (4) compared with the 88.0 percent of Rank-1 obtained by visual transformer-based PAT method on the Partial-ReID dataset, the proposed DB-ResHViT method improves 0.5 percent.

According to the experimental results as shown in Table 2 and Table 3, the DB-ResHViT method proposed in this article achieves good performance and high robustness on occluded pedestrian re-identification dataset. Next, the performance of the DB-ResHViT method proposed in this article will be verified on non-occlusion pedestrian re-identification dataset.

- Performance Verification on Market-1501 and Duke-MTMC Datasets

It can be seen from the experimental results as shown in Table 4, for the Market-1501 dataset, ISP method achieves the best performance among the convolution neural network-based methods, and PFD method achieves the best performance among the vision transformer-based methods. Specifically, (1) the Rank-1 and mAP of ISP method reach 95.3% and 88.6% respectively, and the Rank-1 and mAP of PFD method reach 95.5% and 89.7%, respectively, which demonstrates that the recognition performance of convolutional neural network and visual transformer is very close; (2) the Rank-1 and mAP of the proposed DB-ResHViT method reach 95.7% and 89.8% respectively, which shows that the performance of the proposed method is superior to that of ISP and PFD methods; (3) experimental results validate our original assumption that the proposed method can better play the respective advantages of convolutional neural network and visual transformer, namely, effective combination

Table 4 Experimental results of different methods on Market-1501 dataset and DukeMTMC dataset

Methods	Market-1501		DukeMTMC	
	Rank-1	mAP	Rank-1	Rank-1
PCB [37]	92.3	77.4	81.8	66.1
PGFA [31]	91.2	76.8	82.6	65.5
VPM [44]	93.0	80.8	83.6	72.6
MGCAM [45]	83.8	74.3	46.7	46.0
SPReID [38]	92.5	81.3	–	–
OSNet [49]	91.3	84.9	88.6	73.5
HONet [23]	94.2	84.9	86.9]	75.6
PAFM [24]	95.1	88.2	90.6	78.6
ISP [50]	95.3	88.6	89.6	80.0
FGMFN [26]	90.6	85.0	88.5	80.4
TransReID [29]	95.2	88.9	90.7	82.0
DRL-Net [51]	94.7	86.9	88.1	76.6
PAT [47]	95.4	88.0	88.8	78.2
PFD [53]	95.5	89.7	91.2	83.2
FRT [54]	95.5	88.1	90.5	81.7
DB-ResHViT (Ours)	**95.7**	**89.8**	**91.4**	**83.5**

Bold value represents the maximum in the corresponding column

Table 5 Experimental results of different methods on Partial-iLIDS dataset

Methods	Partial-iLIDS	
	Rank-1	mAP
AMC + SWM [34]	21.0	32.8
DSR [35]	58.8	67.2
SFR [41]	63.9	74.8
FPR [43]	68.1	–
PGFA [31]	69.1	80.9
MFM + HFR [25]	68.1	64.5]
HOReID [23]	72.6	86.4
MHSA-Net [52]	74.9	87.2
TransReID [29]	71.4	87.4
DB-ResHViT(Ours)	**75.2**	**87.6**

Bold value represents the maximum in the corresponding column

of convolution neural network and visual transformer can achieve performance beyond individual models.

- Performance Verification on Partial-iLIDS Dataset

Because the Partial-iLIDS dataset contains few images, the Occluded-Duke dataset is selected as the training set, and the Partial-iLIDS dataset is selected as the testing set. According to the experimental results shown in Table 5, the Rank-1 of the DB-ResHViT method proposed in this article is 75.2%, which is 2.6% and 3.8% higher than the HOReID method and TransReID method, respectively.

Since the feature distribution of the Partial-ILIDS dataset is relatively inconsistent, the experimental results in Table 5 also proves that proposed DB-ResMViT can solve this kind of problem well through data enhancement and robust dual-branch network structure in the case of such micro-dataset and inconsistent feature distribution.

4.5 Ablations and analysis

This section studies the dual-branch structure of the DB-ResHViT method proposed in this article and the effectiveness of each proposed module. Based on the visual transformer branch, the ablation experiments of each module and the dual-branch are performed in turn. Table 6 shows ablation experiment results on the Occluded-Duke dataset, which verifies the effectiveness of each module and double-branch structure in the process of occluded pedestrian re-identification.

Effectiveness of partial image patches pre-convolution module: According to the experimental results of index 2 in Table 6, the Rank-1 of partial image patch pre-convolution module is improved by 2.3% compared with visual transformer. This verifies two purposes of the partial image patch pre-convolution module designed in this article: (1) This module really achieves the purpose of image augmentation and effectively improves the accuracy of occluded pedestrian re-identification. (2) This module is utilized to extract shallow local features, thus improving the ability of feature representation and effectively compensating for

Table 6 Ablation experimental results on Occluded-Duke dataset, where ViT represents the branch network of visual transformer, PPPC represents the partial image patch pre-convolution module, ResDB represents the double-branch network containing the residual structure, RViTDB represents the double-branch network containing the residual-visual transformer network module

Index	ViT	PPPC	ResDB	RViTDB	Rank-1	Rank-5	Rank-10	mAP
1	√				63.6	77.1	81.9	55.1
2	√	√			65.9	79.0	82.7	56.8
3	√	√	√		67.9	80.6	84.9	58.5
4	√		√	√	69.1	81.1	85.2	60.5
5	√	√	√	√	**70.9**	**83.0**	**87.2**	**62.1**

Bold value represents the maximum in the corresponding column the shortcoming of visual transformer in extracting local features.

Effectiveness of the dual-branch structure of residual-vision transformer: According to the comparative experimental results of index 2 and index 3 in Table 6, after adding the dual-branch structure of residual-vision transformer, the Rank-1 and mAP of the network model are improved by 2.0% and 1.7% respectively, which shows that the performance is obviously improved. This verifies the two purposes of the design of the dual-branch structure of residual-vision transformer: (1) The dual-branch structure of residual-vision transformer can effectively establish the correlation of global feature sequences and can effectively extract shallow local features with high discrimination. (2) The dual-branch structure of residual-vision transformer can well maintain the consistency of feature distribution, thus making the training results more convergent and accurate.

Effectiveness of the residual-vision transformer module: According to the comparative experimental results of index 3 and index 4 in Table 6, compared with the simple dual-branch structure, after introducing the residual-vision transformer module in the dual-branch residual branch, the Rank-1 and mAP of the network model reached 69.1% and 60.5%, respectively, increasing by 1.2% and 2.0%. This verifies the two purposes of the design of the residual-vision transformer module in this article: (1) Such a module can adjust the convolution feature distribution to be as consistent as possible with the visual transformer feature distribution before the convolution feature is fused with the visual transformer feature. (2) Such a module is a lightweight module, which can approach the performance of the vision transformer branch by only stacking a few layers.

4.6 Computational complexity

Table 7 gives the complexity of the proposed network and compares it with existing works.

5 Conclusion

In this article, an occluded pedestrian re-identification method based on the dual-branch hybrid framework is proposed, including the designed dual-branch hybrid framework integrating residual network and visual transformer, the designed partial image patch pre-convolution augmentation method and the designed residual-vision transformer module. The dual-branch hybrid framework integrating residual network and visual transformer takes visual transformer and residual network as two parallel branches, which is conducive to extracting more robust features, integrating the correlation of local features and the distinguishability of global features. The partial image patch pre-convolution augmentation method introduces local feature information through the convolution of partial image patches, thus realizing image augmentation. The residual-vision transformer module integrates global features and local features, thus establishing the global correlation of local features. Experimental results demonstrate that the DB-ResHViT method proposed in this article has achieved good results in occluded pedestrian re-identification.

Table 7 Computational complexity of different methods

Method	Parameters(M)	FLOPs(G)
AMC + SWM	78.4	113.4
DSR	53.58	86.97
SFR	52.2	105.3
FPR	65.2	150.7
PGFA	71.8	208.4
MFM + HFR	62.0	124.7
HOReID	146.7	148.1
MHSA-Net	51.20	90.31
TransReID	78.4	242.9
DB-ResHViT	43.7	83.6G

Author contributions SZ and YZ wrote the main manuscript text, and ZL proofread the language description. All authors reviewed the manuscript.

Data availability This declaration is not applicable.

Declarations

Conflict of interests This declaration is not applicable.

Ethical approval This declaration is not applicable.

References

1. Zheng, L., Yang, Yi., Hauptmann, A.G.: Person re-identification: past, present and future. CoRR **16**(10), 1–20 (2016)
2. He Li, Mang Ye, Cong Wang, and Bo Du. Pyramidal Transformer with Conv-Patchify for Person Re-identification. ACM International Conference on Multimedia, 2022: 7317–7326.
3. Chen, C., Ye, M., Qi, M., Jingjing, Wu., Jiang, J., Lin, C.: Structure-aware positional transformer for visible-infrared person re-Identification. IEEE Trans. Image Process. **31**, 2352–2364 (2022)
4. Tao, H., Duan, Q., An, J.: An adaptive interference removal framework for video person re-identification. IEEE Trans. Circuits Syst. Video Technol. **33**(9), 5148–5159 (2023)
5. Tao, H., Bao, W., Duan, Q., Zhenwu, Hu., An, J., Xie, C.: An improved interaction and aggregation network for person re-identification. Multimedia Tools Applications **82**(28), 44053–44069 (2023)
6. Duan, Q., Zhenwu, Hu., Minghao, Lu., Tao, H.: Learning discriminative features for person re-identification via multi-spectral

channel attention. Signal Image Video Process **7**(6), 3019–3026 (2023)

7. Huang, P., Zhu, S., Liang, Z.: Cross-modal pedestrian recognition based on triple attention feature aggregation. J Nanjing Univer Posts Telecommun **41**(5), 101–112 (2021)
8. Xiaofu, Wu., Yin, Z., Song, Y., Zhang, L., Xie, B., Zhao, S., Zhang, S.: Progress in the construction of multi-branch deep neural network for pedestrian recognition diversity feature mining. J Nanjing Univer Posts Telecommun **41**(1), 78–85 (2021)
9. Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard S. Zemel:. Understanding the Effective Receptive Field in Deep Convolutional Neural Networks[C]. Advances in Neural Information Processing Systems, 2016: 4898–4906.
10. Shijie Wang, Zhihui Wang, Haojie Li, Jianlong Chang, Wanli Ouyang, and Qi Tian. Accurate Fine-grained Object Recognition with Structure-driven Relation Graph Networks. International Journal of Computer Vision, 2023.
11. Shijie Wang, Jianlong Chang, Zhihui Wang, Haojie Li, Wanli Ouyang, and Qi Tian. Fine-Grained Retrieval Prompt Tuning. AAAI Conference on Artificial Intelligence, 2023: 2644–2652.
12. Shijie Wang, Jianlong Chang, Haojie Li, Zhihui Wang, Wanli Ouyang, and Qi Tian. Open-Set Fine-Grained Retrieval via Prompting Vision-Language Evaluator. IEEE Conference on Computer Vision and Pattern Recognition, 2023: 19381–19391.
13. Tao Wang, Hong Liu, Pinhao Song, Tianyu Guo, and Wei Shi. Pose-Guided Feature Disentangling for Occluded Person Re-Identification Based on Transformer. AAAI Conference on Artificial Intelligence, 2022: 2540–2549.
14. Boqiang, Xu., He, L., Liang, J., Sun, Z.: Learning feature recovery transformer for occluded person re-identification. IEEE Trans. Image Process. **31**, 4651–4662 (2022)
15. Wenfeng Liu, Xudong Wang, Lei Tan, Yan Zhang, Pingyang Dai, Yongjian Wu, and Rongrong Ji. Learning Occlusion Disentanglement with Fine-grained Localization for Occluded Person Re-identification. ACM International Conference on Multimedia, 2023: 6462–6471.
16. Zhikang Wang, Feng Zhu, Shixiang Tang, Rui Zhao, Lihuo He, and Jiangning Song. Feature Erasing and Diffusion Network for Occluded Person Re-Identification. IEEE Conference on Computer Vision and Pattern Recognition, 2022: 4744–4753.
17. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 2017: 5998–6008.
18. Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person reidentification[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2020: 3186–3195.
19. Xuesong Chen, Canmiao Fu, Yong Zhao, Feng Zheng, Jingkuan Song, Rongrong Ji, and Yi Yang. Salience-guided cascaded suppression network for person re-identification. IEEE Conference on Computer Vision and Pattern Recognition, 2020: 3300–3310.
20. Guiwei Zhang, Yongfei Zhang, Tianyu Zhang, Bo Li, and Shiliang Pu. PHA: Patch-Wise High-Frequency Augmentation for Transformer-Based Person Re-Identification. IEEE Conference on Computer Vision and Pattern Recognition, 2023: 14133–14142.
21. Haocong Rao and Chunyan Miao. TranSG: Transformer-Based Skeleton Graph Prototype Contrastive Learning with Structure Trajectory Prompted Reconstruction for Person Re-Identification. IEEE Conference on Computer Vision and Pattern Recognition, 2023: 22118–22128.
22. Shang Gao, Jingya Wang, Huchuan Lu, and Zimo Liu. Pose-guided visible part matching for occluded person ReID. IEEE Conference on Computer Vision and Pattern Recognition, 2020: 11744–11752.
23. Guanan Wang, 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. IEEE Conference on Computer Vision and Pattern Recognition, 2020: 6449–6458.
24. Yang, J., Zhang, C., Tang, Y., Li, Z.: PAFM: pose-drive attention fusion mechanism for occluded person re-identification. Neural Comput. Appl. **34**(10), 8241–8252 (2022)
25. Li, Y., Yang, Z., Chen, Y., Yang, D., Liu, R., Jiao, L.: Occluded person re-identification method based on multiscale features and human feature reconstruction. IEEE Access **10**, 98584–98592 (2022)
26. Zhang, G., Chen, C., Chen, Y., Zhang, H., Zheng, Y.: Fine-grained-based multi-feature fusion for occluded person re-identification. J. Vis. Commun. Image Represent. **87**, 103581 (2022)
27. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations, 2021: 1–12.
28. Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers and distillation through attention. International Conference on Machine Learning, 2021: 10347–10357.
29. Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. IEEE Conference on Computer Vision, 2021: 14993–15002.
30. Jiaxuan Zhuo, Zeyu Chen, Jianhuang Lai, and Guangcong Wang. Occluded Person Re-identification. IEEE Conference on Multimedia and Expo, 2018: 1–6.
31. Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-Guided Feature Alignment for Occluded Person Re-Identification. IEEE International Conference on Computer Vision, 2019: 542–551.
32. Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. IEEE Conference on Computer Vision, 2015: 1116–1124.
33. Ergys Ristani, Francesco Solera, Roger S. Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. European conference on computer vision, 2016: 17–35.
34. Weishi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shaogang Gong. Partial person re-identification[C]. IEEE Conference on Computer Vision, 2015: 4678–4686.
35. Lingxiao He, Jian Liang, Haiqing Li, and Zhenan Sun. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7073–7082.
36. Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. IEEE Conference on Computer Vision, 2017: 3219–3228.
37. Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)[C]. European conference on computer vision, 2018: 480–496.
38. Houjing Huang, Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. Adversarially occluded samples for person re-identification. IEEE Conference on Computer Vision and Pattern Recognition, 2018: 5098–5107.
39. Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. European Conference on Computer Vision, 2018: 402–419.

40. Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, and Hongsheng Li. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. Annual Conference on Neural Information Processing Systems, 2018: 1230–1241.
41. He, L., Sun, Z., Zhu, Y., Wang, Y.: Recognizing partial biometric patterns[J]. CoRR **18**(10), 1–13 (2018)
42. Mengxi Jia, Xinhua Cheng, Yunpeng Zhai, Shijian Lu, Siwei Ma, Yonghong Tian, and Jian Zhang. Matching on sets: Conquer occluded person re-identification without alignment. AAAI Conference on Artificial Intelligence, 2021: 1673–1681.
43. Lingxiao He, Yinggang Wang, Wu Liu, He Zhao, Zhenan Sun, and Jiashi Feng. Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification. IEEE Conference on Computer Vision, 2019: 8450–8459.
44. Yifan Sun, Qin Xu, Yali Li, Chi Zhang, Yikang Li, Shengjin Wang, and Jian Sun. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. IEEE Conference on Computer Vision and Pattern Recognition, 2019: 393–402.
45. Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. IEEE Conference on Computer Vision and Pattern Recognition, 2018: 1179–1188.
46. Zhou, Mi., Liu, H., Lv, Z., Hong, W., Chen, X.: Motion-aware transformer for occluded person re-identification. CoRR **22**(10), 1–20 (2022)
47. Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. Diverse Part Discovery: Occluded Person Re-Identification With Part-Aware Transformer. IEEE Conference on Computer Vision and Pattern Recognition, 2021: 2898–2907.
48. Mahdi M. Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E. Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. IEEE Conference on Computer Vision and Pattern Recognition, 2018: 1062–1071.
49. Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. IEEE Conference on Computer Vision, 2019: 3702–3712.
50. Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. Identity-guided human semantic parsing for person re-identification. European Conference on Computer Vision, 2020: 346–363.
51. Jia, M., Cheng, X., Shijian, Lu., Zhang, J.: Learning disentangled representation implicitly via transformer for occluded person re-identification. CoRR **21**(7), 1–10 (2021)
52. Tan, H., Liu, X., Tian, S., Yin, B., Li, X.: MHSA-net: multi-head self-attention network for occluded person re-identification. CoRR **20**(8), 1–11 (2020)
53. Tao Wang, Hong Liu, Pinhao Song, Tianyu Guo, and Wei Shi. Pose-guided feature disentangling for occluded person re-identification based on transformer. AAAI Conference on Artificial Intelligence. 2022: 2540–2549.
54. Boqiang, Xu., He, L., Liang, J., Sun, Z.: Learning feature recovery transformer for occluded person re-identification. CoRR **23**(1), 1–11 (2023)

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