Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-18882-w

Occluded person re-identification based on parallel triplet augmentation and parameter-free token spatial attention

Hangyu Li, Yu Zhu, Shengze Wang, Ziming Zhu, Jiongyao Ye, Xiaofeng Ling, et al.
[full author details at the end of the article]

Received: 11 December 2023 / Revised: 7 February 2024 / Accepted: 9 March 2024

Abstract
The task of occluded person Re-identification(Re-ID) is challenging because only local information can be used to make judgments. Also, occlusion may not be present in the training samples, leading to limited performance of the model in inference. Traditional data augmentation schemes that resize, flip, and erase the input image can alleviate this problem, but the serial approach still results in unbalanced samples. To overcome this problem, we propose Parallel Triplet Augmentation (PTA), which involves applying three different data augmentation schemes to a single image during the training phase, thereby robustly expanding the training data. At the same time, non-occluded critical regions of an image tend to provide more discriminative features, so Vision Transformer-based models that process images in chunks show significant advantages. Based on this, we design a parameter-free Token Spatial Attention (TSA) mechanism. TSA uses different schemes for different branches to calculate the weights of each image patch, and then fuses the information in all the patch embedding tokens with the classification head token, thus increasing the amount of spatial information in the classification head token. Using TransReID as a backbone, the experimental results on two occluded datasets (Occluded-Duke and Occluded-ReID) indicate that the proposed method is competitive compared to state-of-the-art methods, with a rank-1 accuracy 0.7% higher on Occluded-Duke. On two non-occluded datasets (Market-1501 and DukeMTMC-ReID) and one vehicle dataset (VeRi-776), the proposed method has also reached state-of-the-art methods, with a rank-1 accuracy 0.3% higher on the VeRi-776 dataset.

Keywords Occluded · Re-identification · Attention · Data augmentation

1 Introduction

Person Re-identification(Re-ID) is to identify the same target of different cameras, which is widely used in person tracking and other fields. The existing Re-ID methods pay attention

✉ Yu Zhu
zhuyu@ecust.edu.cn

Extended author information available on the last page of the article

to the overall and local details of the person. But the person in the cropped images is usually partially occluded.

Compared with the common person Re-ID, there are lots of challenges in the occluded person Re-ID [1]. As shown in Fig. 1(a), the person is often occluded by some objects (like trees or vehicles). Whether some key local details are occluded will affect the accuracy of the model. A data imbalance exists between the training set and the test set in occluded Re-ID datasets. The occluded images of persons in the training set will be less than in the test set. Many methods [2–5] will use data augmentation before training to make up for the balance of the data. Common data augmentations include resizing, erasing, and flipping. Erasing masks in some areas of the image to make them invisible. Each person in the occluded Re-ID dataset does not have the same number of occluded images, so the erasing is often used to generate occluded images from unmasked images during training. As shown in Fig. 1(b), the erasing only retains the local areas in the original image, which makes the erased image like the occluded image in the real world. Through erasing, the problem of insufficient occluded

Fig. 1 Examples of occluded persons(a) and data augmentation programs(b)(c)

images can be well solved. The network's ability to extract information from non-occluded parts can also be improved. As shown in Fig. 1(c), horizontal flipping performs symmetric flipping on the pictures in the horizontal direction. Due to the person walking in the opposite direction and the change of shooting angle between cameras, the person in different images tends to be close to horizontal symmetry. The images after flipping in Fig. 1(c) are like those taken by real cameras. It can be found that the horizontal flipping process is very effective for the Re-ID.

The traditional data augmentation process applies the above methods in series to an input image. In other words, only one output image can be obtained after the whole process. At the same time, different data augmentations are also used randomly, which means that the augmentation methods used by different pictures are different in the same training batch. This will result in an uneven augmentation. To address the limitations and problems in the traditional data augmentation process, we propose Parallel Triplet Augmentation (PTA). In simple terms, PTA includes three parallel branches, where different data augmentations are used in. Compared with the traditional data augmentation process, one training picture can get three different enhanced pictures after PTA, which expands the sample variation and expands the amount of data in the training process.

CNN's receptive field is limited to a small area by Gaussian distribution [6], so it is difficult to extract complete non-occluded features. The pooling and convolution also decrease the resolution of the feature map, which reduces the fine-grained feature information. Based on the different image patches, Vision Transformer (ViT) [3] can calculate the connection in the image to find the key areas in the overall picture. An image is divided into several patches in ViT. After rearranging into one-dimensional patch embedding tokens, they are concatenated with the classification([cls]) head token and sent to the transformer layers for learning. The final output image feature is the [cls] head token after multiple transformer layers. ViT does not lead to small receptive fields or low feature map resolution. But it uses the [cls] head token as the image feature will ignore all image tokens during training. Although self-attention is used in the ViT to allow the [cls] head token to interact with each image embedding token, the [cls] head token still can't contain all the spatial feature information in the entire image. In contrast, Global Average Pooling (GAP) and Maximum Pooling in CNN can integrate features from different spatial regions and preserve local texture features in the image. For occluded person images, the model needs to focus on non-occluded regions in the image, which provide many detailed discriminative features [1, 7]. Based on these considerations, we design a Token Spatial Attention (TSA) mechanism for ViT to remedy the above problems. Based on these considerations, we design a Token Spatial Attention (TSA) mechanism for ViT to remedy the above problems.

The main contributions of this paper are described as follows:

(1) During the training, this paper proposes a new Parallel Triplet Augmentation (PTA), which can solve the problem of unbalanced data samples in the occluded person Re-ID and improve the robustness of the overall network.
(2) At the output of the network, this paper proposes a novel parameter-less Token Spatial attention. TSA uses different methods to calculate attention weights according to different branches in the backbone network and compensates for the missing spatial information in the classification head vector.
(3) The accuracy of the Occluded-Duke [8] dataset has increased by 0.7% through using the designed PTA and TSA. We also test on two non-occluded datasets (Market-1501 [9], DukeMTMC-reID [10]) and the vehicle Re-ID dataset (VeRi-776 [11]) to prove the generalization of the model. Compared with the state-of-the-art methods, our result

shows that rank-1 achieves a 0.3% increase in the VeRi-776 and a close performance in two non-occluded datasets.

The overview of the paper is organized as follows. In the next section, we introduced the relevant work on Person and Vehicle Re-ID tasks including occlusion. In the Section 3, we provided a detailed introduction to the specific process of the proposed model. Afterwards, in the Section 4, we reported on the performance of the proposed model on different datasets. Finally, in Section 5, we evaluated the proposed model and provided prospects for future work.

2 Related work

2.1 Occluded Person ReID

Feature alignment, high-order semantic information, and region classification are three ways used in the existing occluded person Re-ID methods. In the approaches based on feature alignment, STNReID [12] is based on the pairwise spatial transformer, including an STN and ReID module. But simple one-on-one feature alignment is hard to achieve high robustness. The addition of high-order semantic information can solve this problem to a certain extent. With the use of pose information, Wang [13] designed a model with learning high-order relations and topology information for discriminative features and robust alignment. Its ADGC layer can suppress the message passing of meaningless features. When aligning different local features, the proposed CGEA layer can fully use alignment learned by suitable graph matching. Similarly, PVPM [14] extracts different pose information through the novel pose-guided attention module. PFD [15] uses different pose information to disentangle body components, which can selectively match non-occluded parts. In addition, some local regions can also provide lots of high-order local semantic information. It can be regarded as a supplement to overall semantic information. SCP-Net [16] extracts the high-order local spatial feature and achieves the fusion of global channel information and local spatial information. The methods of region classification perform classification and feature extraction on body regions. The additional external semantic cues are usually not applied during the region classification. Region classification attempts to classify body regions without using external cues. For example, VPM [17] classifies region features through a pre-trained human body partial model. It also uses shared regions of an image pair to suppress feature misalignment. VPM calculates the overall feature distance between images through region-level distance. Unlike VPM, ISP [18] uses pseudo-labels to classify body regions. It can obtain the local of both human body parts and personal belongings. Finally, only the features of visible parts are used in the verification phase of the ISP. This method focuses more on the region information and ignores pose information of the recognition subject. Therefore, compared to the pose based feature extraction, the region division and alignment effect in the recognition process is better, but the feature extraction effect for the recognition subject is relatively reduced. MVI2P [19] framework starting from the perspective of multiple perspectives, a multi view information integration module is proposed, including localization, quantification, and integration. DANet [20] utilizes attention mechanism to achieve diverse feature mining, which help the model automatically capture diverse discriminative features on a global scale. In general, attention and pose estimation are the more mainstream and representative methods for the occluded person Re-ID. Attribute annotation-based, clustering-based, figure convolution-based and regularisation-based methods, have received less attention [21].

2.2 Attention mechanisms

The attention mechanism uses a fixed calculation method to enhance the weight of specific regions in the network feature map. It can guide the network to focus on the critical regions of the input images and reduce the interference of irrelevant backgrounds. For example, SA [1] is a forward parameter-free spatial attention. Hu [22] considers the influence of different channels in the feature map and designs the SE attention. ECA [23] is based on the SE, which uses the faster adaptive 1D convolution to get the channel attention weights. CBAM [24] combines channel and spatial attention to achieve spatial and channel adaptive weight refinement. Non-local [25] obtains the association of weights by capturing the dependencies of long-distance regions. Chen [26] coordinated and optimized the selection of spaces and channels through the HAM module to enhance the feature expression ability of the network. ABDNet [27] uses complementary attention modules to focus on the weight distribution of channels and positions respectively and finally performs complementarity between different weights through orthogonal constraints.

2.3 Vehicle re-ID

Vehicle Re-ID is the application of the person Re-ID in vehicles. Among the Vehicle Re-ID methods, VANet [28] focuses on the differences between different viewpoints. This method has good recognition performance, but it requires prior input of the perspective relationship between image pairs. UMTS [29] uses a teacher-student model to extract integrated features in vehicle images. SAVER [30] improves recognition in cross-vehicle datasets through self-supervised learning. The above two methods improve model training performance through knowledge distillation and self supervised learning, thereby achieving better results for vehicle Re-ID. Transferred from person Re-ID methods, image segmentation and feature alignment methods are also widely used in vehicle Re-ID. PRReID [31] uses a detection branch to focus on local regions in vehicles. SAN [32] captures local texture information in local branches paired with different convolutional kernels and pooling layers. SPAN [33] and CFVMNet [34] mask different local regions to change the network's attention to local features. PGAN [35] increases the weight of local regions by local attention module. PVEN [36] cuts vehicle images into four parts for feature extraction. CLAMOR [37] adds local attention to its unsupervised network. MSINet-SAM [38] performs better supervision by TCM. This type of method eliminates the dependency on linear classifiers, thereby achieving unbinding of categories between the training and validation sets. The spatial alignment module within greatly enhances the generalization ability when facing images from different domains.

3 Methods

3.1 Network structure

This section describes the components of our proposed model for the Occluded Re-ID task [21]. As shown in Fig. 2, the proposed model consists of three modules: Parallel Triplet Augmentation, Backbone, and Token Spatial Attention. Firstly, when entering an image I_{input}, the PTA will output three different images named I_{Base}, I_{Erase} and $I_{Flip+Erase}$. After different images input, the backbone can extract the global features $f_*^g \{* \in (Base,\ Erase,\ Flip + Erase)\}$ and local features $f_*^{l_i} \{* \in (Base,\ Erase,\ Flip$

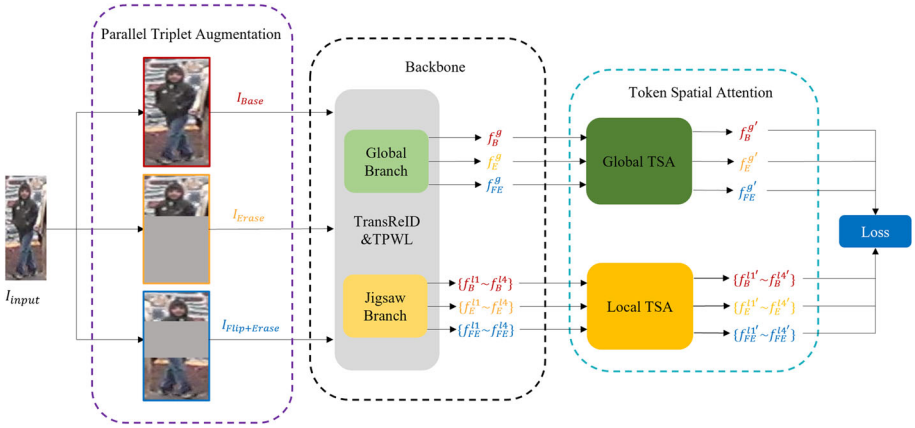

Fig. 2 General flowchart of the network

$+Erase)$, $i \in (1, 2, 3, 4)\}$. All global and local features are fed into the Token Spatial Attention module. The TSA module makes the network's attention focus on the critical regions of the image. As a result, the weight of the unobstructed part will increase, and the importance of the occluded part will decrease. Depending on different inputs, the weight in the TSA is calculated differently. The output of the global branch is fed into the global TSA, while the output of the Jigsaw branch is fed into the local TSA. Finally, all the features output from the TSA module are used to calculate different losses.

3.2 Parallel triplet augmentation

The traditional image data enhancement process is shown in Fig. 3(a), where data enhancement methods such as resizing, flipping, and erasing are used for the input image I_{input}. These enhancement methods are superimposed on the input image, so only one image can be output after data enhancement. Flipping and erasing are used in I^1_{input}, while only erasing is used in I^2_{input}, so some differences exist between the two images after enhancement. Different images in a batch may be enhanced by different means, leading to uneven sample data enhancement during training. In the occluded Re-ID dataset, the query set is almost all occluded images,

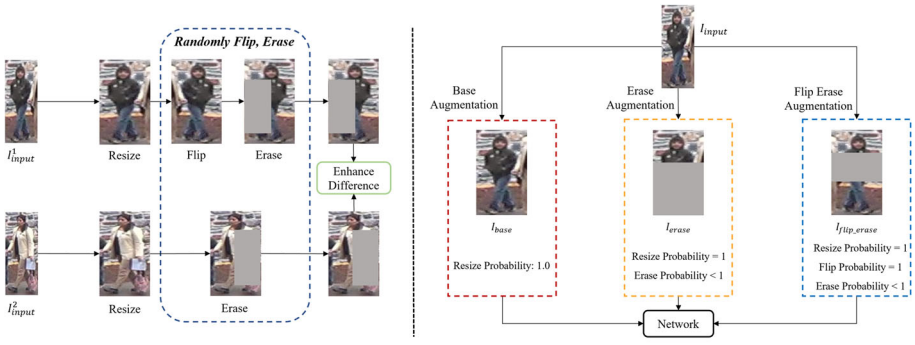

Fig. 3 Traditional augmentation and parallel triplet augmentation

and the gallery set is nearly all unoccluded. The imbalanced sample augmentation resulting from the traditional data augmentation during training introduces a problem where the network fails to effectively address the disparity between the query set and the gallery set in the testing phase, consequently causing a decline in the network's recognition accuracy. The designed Parallel Triplet Augmentation (PTA) can address the limitations in the traditional data augmentation process and further improve the robustness of the network.

The process of parallel ternary data enhancement is shown in Fig. 3(b). For each input image I_{input}, there are three parallel enhancement branches: base enhancement, erase enhancement, and flip-erase enhancement. Three common enhancement processes (resizing, erasing, and horizontal flipping) are used in three branches. In the base enhancement, the input image is enlarged so the network can learn the detailed information better. In the erase enhancement branch, each input image is first scaled up, then random regions of the image are masked. In the flip-erase enhancement branch, each original image is first enlarged, then horizontally symmetric flipped, and finally, random regions are masked. The occurrence probability of resizing and horizontal flipping in all branches is set to 1.0, while the occurrence probability of erasing is set to values less than 1. In other words, resizing and horizontal flipping are applied to each image in each training batch, while erasing occurs only in part of the images in each training batch. The reason for the erasing setting in this way is that the number of occluded images is different for each person in the occluded Re-ID dataset, and the proportion of the occluded area is also different. If erasing is performed on each image, there still exists a difference in the number of occluded images for each ID. Therefore, the erasure of randomly selected images in each batch can fit the difference in the number of occluded images between different IDs. The ablation experiment results prove this setting can improve the network. In contrast, resizing allows the model to capture detailed features better and images to fit better to the network. Horizontal flipping only changes the orientation of the images, allowing the network to extract key features of the same person in different orientations images. These two processes do not affect the balance of the enhanced samples compared to the erasure processing, so they can be used in each image in different branches during training. The normalization operation is used in all three enhancement branches, and the process is shown in (1), (2), and (3):

$$I_{Base} = Resize(I_{input}) \tag{1}$$

$$I_{erase} = Erase(Resize(I_{input})) \tag{2}$$

$$I_{flip_erase} = Erase(Flip(Resize(I_{input}))) \tag{3}$$

Compared with traditional data enhancement, which only gets one image after data enhancement, parallel ternary data enhancement can get three enhanced images I_{Base}, I_{erase} and I_{flip_erase} for one input image. These three images will be sent to the network for training, which expands the degree of image variation in the training process and improves the network's extraction ability for different types of person images. In Section 4.4.2, we discussed in detail the impact of erasure probability on model performance and the selection of the final erasure probability.

3.3 Backbone and optimization

In this model, we use the feature extractor in TransReID [3] as the backbone and implement the Token Pixel Weights Learning (TPWL). The backbone structure is shown in Fig. 4. The feature extractor uses the sliding convolution kernel to split the input image into different embedding tokens. However, the pixels in each token are not equally important. In this way,

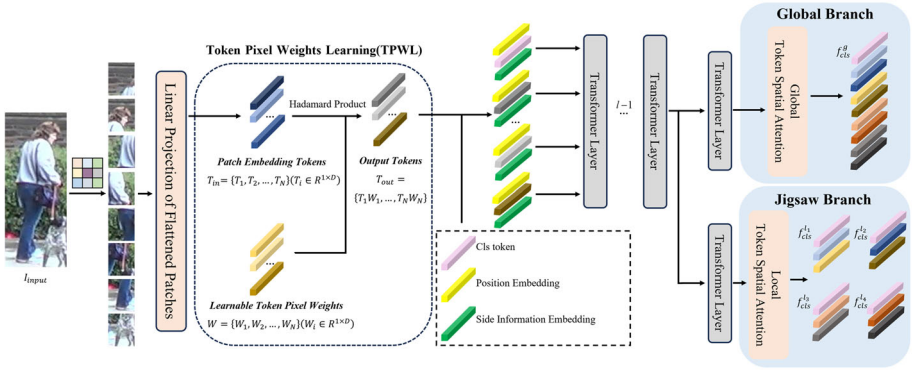

Fig. 4 The overall backbone

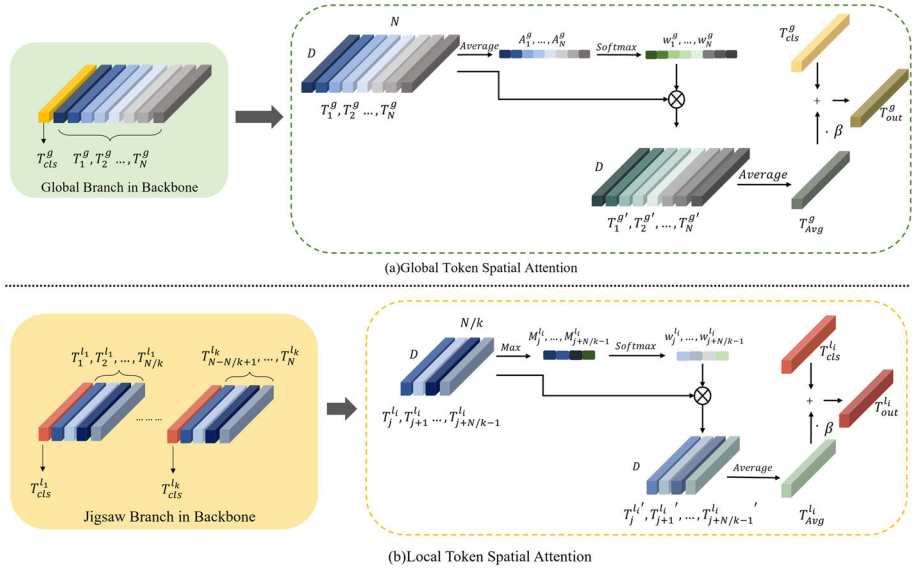

Fig. 5 Detailed diagram of token spatial attention

the network should have different attention to various pixels. The TPWL in our model can change the concentration of different pixels by the learnable token pixel weights.

The input image I_{input} (size of $H \times W \times C$) is firstly divided into N patches through the convolution kernel. Then all the blocks are transferred into two-dimensional patch embedding tokens, which are denoted as $T_{in} = [T_1, T_2, \ldots, T_N]\left(T_i \in R^{1\times D}\right)$. The learnable token pixel weights are denoted as $W = \{W_1, W_2, \ldots, W_N\}\left(W_i \in R^{1\times D}\right)$. The patch embedding tokens T_{in} and the learnable token pixel weights W are combined by the Hadamard product. The Hadamard product makes token pixel weights fully fuse into token pixels. After that, the network can learn the weight distribution of pixels in different regions and channels, which also improves the network's ability to extract personal semantic information. The final output is as shown in (4):

$$T_{out} = T_{in} \odot W = \{T_1 W_1, T_2 W_2, \ldots, T_N W_N\} \tag{4}$$

3.4 Token spatial attention

The Token Spatial Attention (TSA) obtains token weights by their average or maximum value, thus changing the network's attention for different image patches, as shown in Fig. 5. Besides, although the self-attention allows the [cls] embedding token to interact with each patch embedding token, the [cls] embedding token still lacks spatial information and translation invariance. The TSA can compensate for the missing spatial information in the [cls] embedding token. The TSA can be divided into global and local TSA depending on the different weight calculations on global or local patch embedding tokens. It is worth mentioning that the TSA is entirely parameter-free compared with other attention mechanisms. TSA will not bring too many extra calculations or parameters, which doesn't grow with the model's size.

3.4.1 Global token spatial attention

As shown in Fig. 5(a), the Global TSA's inputs are the patch embedding tokens $\left\{T_n^g\right\}_{n=1\ N}\left(T_n^g \in R^{1\times D}\right)$ and the [cls] embedding token T_{cls}^g, which all output from the global branch of the backbone. In the CNN model, the GAP (Global Average Pooling) is used to calculate the average of each region's pixels and maps all values to the output feature maps, which can retain the deep semantic information of the input image. Similar to GAP, global TSA computes the average of each token to preserve the deep information, which is denoted as $\left\{A_n^g\right\}_{n=1\ N}\left(A_n^g \in R^{1\times 1}\right)$. The average value represents the amount of semantic information contained in its token. Then the softmax is used to activate all the A_n^g so that they are distributed in the $(0,\ 1)$. The softmax can calculate the weight that each patch embedding token should have according to the amount of their information. All weights are denoted as $\left\{w_n^g\right\}_{n=1\ N}$. By multiplying w_n^g with the corresponding T_n^g , the network can change the attention distribution for different patches. After that, the new patch embedding token is denoted as $\left\{T_n^{g\prime}\right\}_{n=1\ N}$. Finally, all tokens and the [cls] head token are fused in the following way: firstly, calculating the average of all $T_n^{g\prime}$, which is denoted as T_{Avg}^g. Then the T_{Avg}^g is multiplied with the fusion coefficient β and summed with the [cls] head token T_{cls}^g, which outputs the final token T_{out}^g. The computation process of global TSA is shown in (5), (6), and (7):

$$T_n^{g\prime} = softmax\left(Avg\left(T_n^g\right)\right) \otimes T_n^g \tag{5}$$

$$T_{Avg}^g = \frac{\sum_{n=1}^{N} T_n^{g\prime}}{N} \tag{6}$$

$$T_{out}^g = T_{Avg}^g \cdot \beta + T_{cls}^g \tag{7}$$

3.4.2 Local token spatial attention

As shown in Fig. 5(b), in the Jigsaw branch of the backbone, all the image patch embedding tokens are divided into k groups. Each group and a corresponding local [cls] embedding token will send to the Transformer layer to get the local feature. All the local [cls] embedding token and token groups are denoted as $\left\{T_{cls}^{l_i}\right\}_{i=1\ k}\left(T_{cls}^{l_i} \in R^{1\times D}\right)$ and $\left\{T_j^{l_i}, \ldots, T_{j+N/k-1}^{l_i}\right\}_{i=1\ k, j=1\ N}\left(T_j^{l_i} \in R^{1\times D}\right)$. The max pooling in CNN calculates the

maximum of each region's pixels in the feature map. It can preserve more local texture information in the original image. Similar to the max pooling, the local TSA computes the max of each token in each group, which can amplify the texture information in all tokens. All the max values are denoted as $\left\{M_j^{l_i}\right\}_{i=1\ k,j=1\ N}\left(M_j^{l_i} \in R^{1\times 1}\right)$. Same as global TSA, the softmax is also used to perform the activation for all $M_j^{l_i}$. Softmax calculates the weight of each image patch embedding token based on the maximum value, and all the weights are denoted as $\left\{w_j^{l_i}\right\}_{i=1\ k,j=1\ N}$. Multiplying each $w_j^{l_i}$ and the corresponding token $T_j^{l_i}$ can change the attention of the network for each image patch, and the new patch embedding token is denoted as $\left\{T_j^{l_i\prime}\right\}_{i=1\ k,j=1\ N,}$. The $T_j^{l_i\prime}$ can further augment the information of $T_{cls}^{l_i}$ in the following way: firstly, calculating the average of $T_j^{l_i\prime}$, which is denoted as $T_{Avg}^{l_i}$. Then the $T_{Avg}^{l_i}$ is multiplied with the coefficient β and added with the [cls] head token $T_{cls}^{l_i}$, which outputs the final token $T_{out}^{l_i}$. The computation process of local TSA is shown in (8), (9), and (10).

$$T_j^{l_i\prime} = softmax\left(Max\left(T_j^{l_i}\right)\right) \otimes T_j^{l_i} \tag{8}$$

$$T_{Avg}^{l_i} = \frac{\sum T_j^{l_i\prime}}{N/k} \tag{9}$$

$$T_{out}^{l_i} = T_{Avg}^{l_i} \cdot \beta + T_{cls}^{l_i} \tag{10}$$

In Section 4.4.3, we conducted ablation experiments on attention factor β and determined the final values.

3.5 Loss function

We choose classification loss and triplet loss to train the network. All the global and local features in this network are used to calculate the above losses. The loss function is shown in (11):

$$L_{total} = L_{cls}\left(T_{out}^{g}\right) + L_{tri}\left(T_{out}^{g}\right) + \frac{1}{4}\sum_{i=1}^{4}\left(L_{cls}\left(T_{out}^{l_i}\right) + L_{tri}\left(T_{out}^{l_i}\right)\right) \tag{11}$$

In the process of inferencing, we concate the average of k local features and the global feature as the final inference feature $f_{inference}$, as shown in (12):

$$f_{inference} = Concat\left(T_{out}^{g}, \frac{\sum_{i=1}^{4} T_{out}^{l_i}}{4}\right) \tag{12}$$

4 Experiments

4.1 Datasets

Occluded-Duke Occluded-Duke [8] is a subset of the DukeMTMC-reID. The training set comprises 15618 images. The query consists of 2210 occluded images. The gallery, on the other hand, has 17661 images, some of which are occluded.

Occluded-ReID Occluded-ReID [1] contains 2000 images of 200 persons. Each person has 5 fully-body images and 5 occluded images. The query and gallery each contain

1000 images. The Occluded-ReID only provides the testing set, so we train our model the DukeMTMC-reID.

Market-1501 The people images in Market-1501 [9] are mainly collected by 6 static cameras (5 high-definition cameras and 1 low-definition camera) on the campus of Tsinghua University in the summer. Each person in the dataset is captured by at least two cameras, including a total of 32,217 images of 1,501 persons. The training set of Market-1501 is 12,936 images of 751 persons, and the test set is 23,100 pictures of 750 persons. During the test, the test set is divided into a query of 3368 images and a gallery of 19732 images.

DukeMTMC-reID The images in DukeMTMC-reID [10] are mainly collected by 8 static high-definition cameras at Duke University. Among them, 408 persons were only captured by one camera as interference data, and the remaining 1404 persons were at least Captured by two cameras, a total of 36,411 pictures of 1,812 persons were captured. The training set consists of 702 IDs randomly selected from all IDs, with a total of 16522 pictures. The test set is the other half of the IDs and the 408 IDs in the interference data, with a total of 19889 pictures. When testing, the 2228 pictures of the test set are used as the query set, and the 17661 pictures are used as the gallery.

VeRi-776 VeRi-776 [11] contains more than 50,000 images of 776 vehicles captured by 20 cameras, and each vehicle is captured by at least 2 cameras under different viewing angles, resolutions, and occlusions. Among them, 37778 images of 576 vehicles are used for training, and 11579 images of the remaining 200 vehicles are used for testing.

4.2 Experiment setting

4.2.1 Implementation details

The network is built through the Pytorch [45] and trained on four Nvidia Geforce GTX 1080Ti with 11 GB. The initial weights of the feature extractors in the backbone are pre-trained on ImageNet- 21K and then finetuned on ImageNet-1K [46]. In PTA, resizing adjusts all pictures to 256×128, and the probability of erasing is set to 0.4. In TSA, the attention factor β is set to 0.3. Due to the limitation of video memory, we use the traditional data augmentation process in the comparison experiment of vehicle Re-ID, which resizes the vehicle image to 256×256. 64 images were selected for each training batch during the training, including 4 images for each ID. The training parameters are consistent with the backbone (TransReID [3]), and the initial learning rate is set to 0.008 with cosine learning rate decay. The optimizer used is the SGD with a momentum of 0.9 and weight decay of 1e-4.

4.2.2 Evaluation metrics

Following conventions in the most person ReID papers, we evaluate all methods with Rank-K and the mean Average Precision (mAP)

4.3 Experimental results

4.3.1 Results on occluded Re-ID dataset

We evaluate the performance of the proposed network on the Occluded-Duke and Occluded-REID datasets. The compared methods can be divided into two categories depending on the backbone, as shown in Table 1. Among the methods using CNN as the backbone, PCB [2] horizontally slices the feature map into several pieces and extracts information individually

Table 1 Comparison with other methods on Occluded-Duke and Occluded-REID

Methods	Auxiliary Clues	Backbone	Occluded-Duke		Occluded-REID	
			Rank-1	mAP	Rank-1	mAP
PCB(2018ECCV) [2]	no	CNN	42.6%	33.7%	41.3%	38.9%
PGFA(ICCV2019) [8]	no	CNN	51.4%	37.3%	-	-
FPR(CVPR2019) [39]	yes	CNN	-	-	78.3%	68.0%
HOReID(CVPR2020) [13]	yes	CNN	55.1%	43.8%	80.3%	70.2%
MoS(AAAI2021) [40]	no	CNN	61.0%	49.2%	-	-
ISP(CVPR2020) [18]	no	CNN	62.8%	52.3%	-	-
PAT(CVPR2021) [41]	no	ViT	64.5%	53.6%	81.6%	72.1%
DRL-Net [42](2022)	no	ViT	65.0%	50.8%	-	-
TransReID(ICCV2021) [3]	no	ViT	66.4%	59.2%	70.2%	67.3%
FED(CVPR2022)(CVPR2022) [43]	no	ViT	68.1%	56.4%	86.3%	79.3%
PFD(AAAI2022) [15]	yes	ViT	69.5%	61.8%	81.5%	83.0%
PFT [44](2022)	no	ViT	69.8%	60.8%	83.0%	78.3%
Ours	no	ViT	70.5%	60.6%	81.1%	73.8%

before finally stitching them together as the overall feature. PGFA [8] uses pose labels to make the network focus on non-occluded regions. FPR [44] calculates the foreground probability of different regions to reduce interference from occluded regions during matching. HOReID [13] obtains important features through key human pose points in the image, and features of different images are also horizontally aligned. MoS [39] positions occluded person Re-ID as a set matching task without requiring spatial alignment. ISP [18] generates pseudo-labels for body localization. In the approaches using ViT as the backbone, PAT [40]'s encoder encodes the image as a whole based on the pixel background, and its decoder is based on different local regions to obtain multivariate local feature pairs. DRL-Net [41] inferred the local features in the image through a special transformer. FED [42] performs a random mask on the input images, and then a separate mask elimination module is used to enhance the quality of the features. PFD [15] uses the pose information to make the feature aggregation module extract and match features. PFT [43] improves the correlation between image patches.

On the Occluded-Duke, PFT is the best performance of all comparison methods. Our designed model improves Rank-1 by 0.7% to 70.5% compared to the best PFT, while mAP decreases slightly to 60.6%. On the Occluded-REID, FED performs best among all comparison models, with Rank-1 and mAP reaching 86.3% and 79.3%. Since the Occluded-REID does not contain a training set, we train the model on the DukeMTMC-reID. The test results achieved 81.10% and 73.80% for Rank-1 and mAP. Since the TransReID's feature extractor used in our model captures specific dataset perspective information, it causes performance degradation when the network is trained on one dataset and then transferred to other datasets for testing. Although our results are lower than FED, it improves by 10.9% and 1.9% on Rank-1 and mAP compared to the TransReID. Figure 6(a)(b)(c) shows the rank-5 performance of the model in the case of occlusion by cars, pedestrians, and objects, respectively. In Fig. 6(a), the wrong images in the adopted backbone (TransReID) selection all have the same occluder, while many of our selected graphs ignore the occlusion. In Fig. 6(b), the backbone does not select the correct image for the case containing occlusion, whereas our model selects the correct image even in the case of occlusion by a different person. As shown in Fig. 6(c), when occlusion is not present in the gallery, the backbone selects images with similar background stripes resulting in an error, which we avoid.

Fig. 6 Performance of rank-5 under different Occluded (The correct image is framed with green dashed box posts and the incorrect image is framed in red)

In addition, Table 1 lists whether additional auxiliary clues are used in the network. The additional auxiliary clues are categorized into body region parsing and key pose body points. FPR uses a foreground analyzer to parse human regions in the whole image, while HOReID and PFD extract human pose key points in images. The auxiliary clues can improve the overall network performance, but these results are still lower than some methods that do not use auxiliary clues. Besides, using additional auxiliary clues also imposes additional branches or parameter overhead on the overall network, which is why no additional clues are used in our method.

4.3.2 Results on non-occluded Re-ID

To demonstrate the generalization of the proposed model, we conduct experiments on two non-occluded holistic Re-ID datasets, as shown in Table 2. In addition to the abovementioned methods, we compared four other Re-ID methods in the non-occluded Re-ID datasets. Among them, SPReID [47] integrates human semantic information from images into the recognition process to improve the accuracy of the network. OSNet [4] connects deep point convolutional kernels of different sizes in the form of residuals, thus extracting spatial features of channels at different scales and finally fusing the features through aggregation gates. ABD-Net [27] activates channel attention by orthogonal constraints, and SCSN [48] merges the individual features into the final output by highlighting feature bootstrapping.

The experimental results in Table 2 show that SCSN and PFD perform better among all comparison methods. In the Market-150, our model gets 95.5% Rank-1 and 89.0% mAP. In the DukeMTMC, our model gets 91.0% Rank-1 and 81.6% mAP. Both results are close to the PFD and SCSN, proving that our network performs well on the occluded Re-ID dataset and has good generalization on the non-occluded Re-ID dataset.

Table 2 Comparison with other methods on Market-1501 and DukeMTMC

Methods	Auxiliary clues	Backbone	Market-1501		DukeMTMC	
			Rank-1	mAP	Rank-1	mAP
PGFA(ICCV2019) [8]	no	CNN	91.2%	76.8%	82.6%	65.5%
PCB(ECCV2018) [2]	no	CNN	92.3%	77.4%	81.8%	66.1%
SPReID(CVPR2018) [47]	yes	CNN	92.5%	81.3%	-	-
HOReID(CVPR2020) [13]	yes	CNN	94.2%	84.9%	86.9%	75.6%
OSNet(CVPR2019) [4]	no	CNN	94.8%	84.9%	88.6%	73.5%
ISP(ECCV2020) [18]	no	CNN	95.3%	88.6%	89.6%	80.0%
FPR(CVPR2019) [39]	yes	CNN	95.4%	86.6%	88.6%	78.4%
MoS(AAAI2021) [40]	no	CNN	95.4%	89.0%	90.6%	80.2%
ABDNet(ICCV2019) [27]	no	CNN	95.6%	88.3%	89.0%	78.6%
SCSN(CVPR2020) [48]	no	CNN	95.7%	88.5%	91.0%	79.0%
DRL-Net(2022) [42]	no	ViT	94.7%	86.9%	88.1%	76.6%
FED(CVPR2022) [15]	no	ViT	95.0%	86.3%	89.4%	78.0%
TransReID(ICCV2021) [3]	no	ViT	95.2%	88.9%	90.7%	82.0%
PFT(2022) [44]	no	ViT	95.3%	88.8%	90.7%	82.1%
PAT(CVPR2021) [41]	no	ViT	95.4%	88.0%	88.8%	78.2%
PFD(AAAI2022) [15]	yes	ViT	95.5%	89.7%	91.2%	83.2%
Ours	no	ViT	95.5%	89.0%	91.0%	81.6%

On the two non-occluded datasets, the results based on the ViT are also better than those based on the CNN, which can prove that the ViT is more suitable for the person Re-ID. In the comparison using auxiliary clues, SPReID, HOReID, FPR, and PFD all resolve body regions or use key point body features. Compared with the method without auxiliary clues, these four models can't get a massive improvement in Rank-1 or mAP and are worse than the best-performing SCSN on Market-1501, and only PFD is slightly better than SCSN on DukeMTMC. This also proves that using auxiliary clues may not get a considerable boost.

4.3.3 Results on vehicle ReID

We also experimented with VeRi-776 to verify the generalization of the proposed model. Table 3 lists all results and the methods, whether using the local information. Among all methods, only TransReID and our network use the ViT as the backbone; the rest of the methods use CNN as the backbone. From the results, it can be found that the performance of most methods using local features is better than that of networks using only global features. Compared with SAVER, the best TransReID also improves Rank-1 and mAP by 0.7% and 2.4%, which shows the importance of local features and ViT in vehicle Re-ID tasks. Our network obtains 97.4% Rank-1 and 81.7% mAP on the VeRi-776. Compared to the best-performing TransReID, it improves by 0.3% in Rank-1, which indicates that our model also has excellent generalization in the vehicle Re-ID dataset. Fig. 6(d) shows the rank-5 performance for the case where only cameras 008 and 009 are available in the gallery (containing only the occlusion of the trees), where our model selects the correct image with different occlusions, while backbone (TransReID) selects images that all have the same occlusion.

Table 3 Comparison with other methods on VeRi-776

Methods	Local feature	Rank-1	mAP
VANet(ICCV2019) [28]	no	89.8%	66.3%
UMTS(AAAI2020) [29]	no	95.8%	75.9%
SAVER(ECCV2020) [30]	no	96.4%	79.6%
PRReID(CVPR2019) [31]	yes	93.3%	72.5%
SAN(2020) [32]	yes	93.3%	72.5%
SPAN(ECCV2020) [33]	yes	94.0%	68.9%
CFVMNet(ACMMM2020) [34]	yes	95.3%	77.1%
PVEN(CVPR2020) [36]	yes	95.6%	79.5%
PGAN(2020) [35]	yes	96.5%	79.3%
GLAMOR(2020) [37]	yes	96.5%	80.3%
MSINet-SAM(CVPR2023) [38]	yes	96.8%	78.8%
TransReID(ICCV2021) [3]	yes	97.1%	82.0%
Ours	yes	97.4%	81.7%

4.4 Ablation study

We conducted ablation experiments on the overall network and parameters on the Occluded-Duke dataset, and the results are shown below.

4.4.1 Ablation of the overall network

The results of the ablation experiments of the overall network are listed in Table 4, where the probability of erasuring in PTA is set to 0.4. The parameter β in TSA is set to 0.3. We first integrate the different modules into the overall network separately for comparison. When only using PTA, the network gets 68.50% Rank-1 and 59.6% mAP. Then Rank-1 and mAP improved by 1.9% and 0.9% by exclusively using the TPWL. In contrast, only using the TSA can increase the Rank-1 and mAP by 1.8% and 0.5%. After that, we combine the proposed modules within the network. Incorporating TPWL and TSA leads to the 68.5% Rank-1 and 60.0% mAP. When adding the PTA and TPWL, we observe the Rank-1 and mAP increase

Table 4 Ablation experiment of the overall network

PTA	TPWL	TSA	Rank-1	mAP
no	no	no	66.4%	59.2%
yes	no	no	68.5% (+2.1%)	59.7% (+0.5%)
no	yes	no	68.3% (+1.9%)	60.1% (+0.9%)
no	no	yes	68.2% (+1.8%)	59.7% (+0.5%)
no	yes	yes	68.5% (+2.1%)	60.0% (+0.8%)
yes	no	yes	70.3% (+3.9%)	60.6% (+1.4%)
yes	yes	no	68.7% (+2.3%)	60.2% (+1.0%)
yes	yes	yes	70.5% (+4.1%)	60.6% (+1.4%)

Table 5 Ablation experiments on different enhanced branches in PTA

Index	B	E	FE	Rank-1	mAP
1	no	no	no	68.5%	60.0%
2	yes	yes	no	67.8%	58.4%
3	yes	no	yes	67.1%	57.7%
4	no	yes	yes	66.7%	58.7%
5	yes	yes	yes	70.5%	60.6%

by 2.3% and 1.0%. Compared to the above two results, employing the PTA and TSA yields the highest performance boost, with a 3.9% and 1.4% increase in Rank-1 and mAP. Finally, after all three modules are integrated into the network, the Rank-1 and mAP reach 70.5% and 60.6%.

4.4.2 Ablation of PTA

We perform ablation experiments for different enhancement branches in PTA, and the results are shown in Table 5. B, E, and FE denote the base enhancement, erase enhancement, and flip-erase enhancement. We maintained parallelism during the ablation experiments by retaining two parallel branches. This approach ensured consistency and comparability in this ablation experiment.

We can obtain 68.5% Rank-1 and 60.0% mAP when the network training with the traditional data enhancement. In the comparison of Rank-1, retaining the base enhancement yields better. It can lead Rank-1 to 67.8% or 67.15%. While using the erase enhancement and flip-erase enhancement delivers a lower 66.7% Rank-1. In comparing mAP, the experimental results tend to be higher for preserving erase enhancement. However, using the base enhancement and the flip enhancement obtains a lower result.

Observing the above results, it is evident that the effect of using two data enhancements in parallel is still lower than using traditional data enhancement. This is because using two data enhancements in parallel only preserves the parallelism but ignores the integrity. The absence of either branch may reduce the data diversity of the training images. The 70.5% Rank-1 and 60.6% mAP obtained after training with the complete PTA supports the above inference.

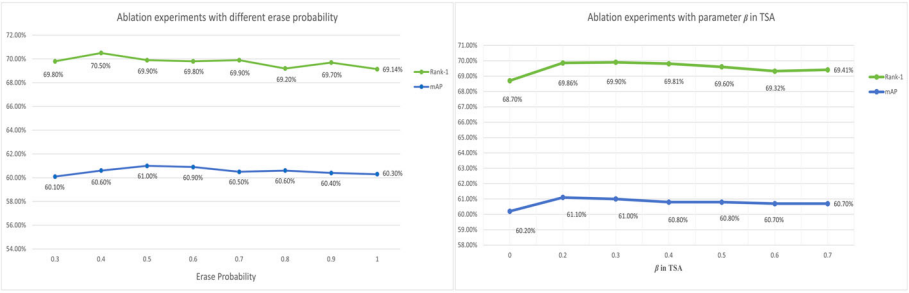

Fig. 7 Ablation of hyperparameters

Table 6 Ablation experiments of global and local TSA

Index	Global	Local	Rank-1	mAP
1	no	no	68.7%	60.2%
2	yes	no	69.4% (+0.7%)	60.8% (+0.6%)
3	no	yes	69.5% (+0.8%)	60.7% (+0.5%)
4	yes	yes	69.9% (+1.2%)	61.0% (+0.8%)

In addition, we took different probabilities of erase occurrence in PTA, as shown in Fig. 7(a). In comparing Rank-1, the network can get 70% results when setting the erase occurrence probability to 0.3∼ 0.7. At the same time, the Rank-1 will drop when the erasure occurrence probability is set to 0.8∼ 1. In comparing mAP, the network can achieve high mAP when the probability of erasure occurrence is set to 0.4∼ 0.6. After considering Rank-1 and mAP together, we finally chose 0.4 as the erasure occurrence rate used in the Occluded-Duke.

4.4.3 Ablation of TSA

We conducted ablation experiments for global or local TSA, where the erase probability in PTA is set to 0.5. Table 6 lists all the experimental results. When the network doesn't use the TSA, the Rank-1 and mAP reached 68.7% and 60.2%. When only the global or local TSA is added, the Rank-1 and mAP can be improved differently. Finally, after adding the global and local TSA, the Rank-1 and mAP further improved to 69.9% and 61.0%. This result indicates that using corresponding TSA for different branches of the backbone can make the network pay more attention to the critical global and local features.

We also investigated the impact of different weight calculation methods in TSA. Table 7 presents the experimental results. Max or Avg denotes using the max(·) or Avg(·) in TSA to compute the token weights. Both two approaches have improvements for the network. However, the global branch of the backbone extracts the overall semantic information of the image, so it is more balanced to use Avg(·) in the global TSA to calculate weights. Similarly, the Jigsaw branch of the backbone extracts the local information of different regions in the image, so using max(·) can give higher weights to local vital features. Finally, the Rank-1 and mAP are improved by 1.2% and 0.8% after using the appropriate calculation method.

The ablation experimental result of β is also presented in Fig. 7(b). When β is set from 0.2 to 0.4, the Rank-1 of the network is close to 70%. But if β continues to increase from 0.5 to 0.7, the Rank-1 gradually decreases to about 69.5%. The results of mAP of different β are all distributed around 61.0%. It can be found that the network can get better Rank-1 when the [cls] head token is supplemented with a small amount of spatial information.

Table 7 Different calculation methods in Trans Spatial attention

Weight calculation		Rank-1	mAP
Global TSA	Local TSA		
-	-	68.7%	60.2%
Max	Max	69.4% (+0.7%)	60.5% (+0.3%)
Avg	Avg	69.3% (+0.6%)	60.6% (+0.4%)
Avg	Max	69.9% (+1.2%)	61.0% (+0.8%)

5 Conclusion

Occluded person re-identification is challenging since it can only be judged by the information from the unoccluded part. Also, the training data for the model may not contain occlusion, and the data imbalance can lead to further performance degradation. In this paper, we propose a Parallel Triplet Augmentation (PTA) for occluded person re-identification. Unlike the traditional data enhancement process, PTA can compensate for the imbalance of each ID occlusion picture in the occluded Re-ID dataset, and parallel data enhancement can improve the robustness of the trained network. Meanwhile, based on the Vision Transformer, we design a parameter-free Token Spatial Attention (TSA) attention module. At the output end of the network, TSA calculates the weights according to the global or Jigsaw branch, respectively, and compensates for the lost spatial information in the classification head vector. This paper tests the proposed network on the occlusion Re-ID dataset, the overall Re-ID dataset, and the overall vehicle dataset. Compared with the current advanced methods, it is proved that PTA and TSA have excellent performance and generalization ability. In the ablation experiment stage, ablation experiments on different modules also demonstrate the importance of each module. At the same time, the ablation experiment results inside each module reflect logical rationality. The proposed model performs excellently in occluded person Re-ID and other Re-ID tasks. It is worth noting that we use random erasure in our enhancement scheme and do not make targeted strategies for different types of targets. For example, the proportion and parts of different types of vehicles to be erased during training should be considered more carefully. Designing more flexible erasure strategies will be part of our future work.

Acknowledgements This work was supported by the Shanghai Automotive Industry Science and Technology Development Foundation (2304)

Author Contributions Hangyu Li: Designed the framework and network architecture, carried out the implementation, performed the experiments and analysed the data. Yu Zhu: Experiment,Data curation, Writing-Original draft preparation, Software, Writing- Reviewing and Editing, Supervision. Shengze Wang: Experiment, Visualization, Investigation. Ziming Zhu: Experiment, Visualization, Investigation. Jiongyao Ye: Writing- Reviewing and Editing, Supervision. Xiaofeng Ling: Writing- Reviewing and Editing, Supervision.

Availability of data and materials The datasets analyzed during the current study are available in the repositories:
Occluded-Duke: https://github.com/lightas/Occluded-DukeMTMC-Dataset
Occluded-ReID: https://github.com/tinajia2012/ICME2018_Occluded-Person-Reidentification_datasets
Market-1501: https://github.com/DemacianPrince/Market1501Evaluation
DukeMTMC-reID: http://vision.cs.duke.edu/DukeMTMC/data/misc/DukeMTMC-reID.zip
VeRi-776: https://github.com/VehicleReId/VeRi

Declarations

Competing interests The author(s) declared no conflicts of interest with respect to the research, authorship, and publication of this paper.

Ethical statement Ethical and informed consent for data used.

Research involving human participants and/or animals Not involve.

Informed consent Not involve.

References

1. Zhuo J, Chen Z, Lai J, Wang G (2018) Occluded person re-identification. 2018 IEEE International Conference on Multimedia and Expo (ICME). pp 1–6. https://api.semanticscholar.org/CorpusID:4713514
2. Sun Y, Zheng L, Yang Y, Tian Q, Wang S (2017) Beyond part models: Person retrieval with refined part pooling. https://api.semanticscholar.org/CorpusID:10013306
3. He S et al. (2021) Transreid: Transformer-based object re-identification. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp 14993–15002. https://api.semanticscholar.org/CorpusID:231846818
4. Zhou K, Yang Y, Cavallaro A, Xiang T (2019) Omni-scale feature learning for person re-identification. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp 3701–3711. https://api.semanticscholar.org/CorpusID:145050804
5. Bhuyan, H. K., Vijayaraj, A. & Ravi, V. Development of secrete images in image transferring system. Multimedia Tools and Applications 82, 7529–7552 (2022). https://api.semanticscholar.org/CorpusID:251827392
6. Luo W, Li Y, Urtasun R, Zemel RS (2016) Understanding the effective receptive field in deep convolutional neural networks. https://api.semanticscholar.org/CorpusID:5665033
7. Zheng W et al. (2015) Partial person re-identification. 2015 IEEE International Conference on Computer Vision (ICCV). pp 4678–4686. https://api.semanticscholar.org/CorpusID:568909
8. Miao J, Wu Y, Liu P, Ding Y, Yang Y (2019) Pose-guided feature alignment for occluded person re-identification. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp 542–551. https://api.semanticscholar.org/CorpusID:207985433
9. Zheng L et al. (2015) Scalable person re-identification: A benchmark. 2015 IEEE International Conference on Computer Vision (ICCV). pp 1116–1124. https://api.semanticscholar.org/CorpusID:14991802
10. Ristani E, Solera F, Zou RS, Cucchiara R, Tomasi C (2016) Performance measures and a data set for multi-target, multi-camera tracking. https://api.semanticscholar.org/CorpusID:5584770
11. Liu X, Liu W, Ma H, Fu H (2016) Large-scale vehicle re-identification in urban surveillance videos. 2016 IEEE International Conference on Multimedia and Expo (ICME). pp 1–6. https://api.semanticscholar.org/CorpusID:662727
12. Luo H, Fan X, Zhang C, Jiang W (2019) Stnreid: Deep convolutional networks with pairwise spatial transformer networks for partial person re-identification. IEEE Transactions on Multimedia 22:2905–2913. https://api.semanticscholar.org/CorpusID:81978300
13. Wang G et al. (2020) High-order information matters: Learning relation and topology for occluded person re-identification. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp 6448–6457. https://api.semanticscholar.org/CorpusID:212747636
14. Gao S, Wang J, Lu H, Liu Z (2020) Pose-guided visible part matching for occluded person reid. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp 11741–11749. https://api.semanticscholar.org/CorpusID:214743196
15. Wang T, Liu H, Song P, Guo T, Shi W (2021) Pose-guided feature disentangling for occluded person re-identification based on transformer. https://api.semanticscholar.org/CorpusID:244909130
16. Fan X et al. (2018) Scpnet: Spatial-channel parallelism network for joint holistic and partial person re-identification. arXiv:1810.06996, https://api.semanticscholar.org/CorpusID:53114847
17. Sun Y et al. (2019) Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp 393–402. https://api.semanticscholar.org/CorpusID:90260003
18. Zhu K, Guo H, Liu Z, Tang M, Wang J (2020) Identity-guided human semantic parsing for person re-identification. arXiv:2007.13467, https://api.semanticscholar.org/CorpusID:220793215
19. Dong N, Yan S, Tang H, Tang J, Zhang L (2023) Multi-view information integration and propagation for occluded person re-identification. arXiv:2311.03828, https://api.semanticscholar.org/CorpusID:265043650
20. Ren T, Lian Q, Zhang D (2023) Constructing comprehensive and discriminative representations with diverse attention for occluded person re-identification. J Vis Commun Image Represent 97:103993. https://api.semanticscholar.org/CorpusID:265434225
21. Ning E, Wang C, Zhang H, Ning X, Tiwari P (2023) Occluded person re-identification with deep learning: A survey and perspectives. arXiv:2311.00603, https://api.semanticscholar.org/CorpusID:264832684
22. Hu J, Shen L, Albanie S, Sun G, Wu E (2017) Squeeze-and-excitation networks. 2018 IEEE/CVF conference on computer vision and pattern recognition. pp 7132–7141. https://api.semanticscholar.org/CorpusID:140309863

23. Wang Q et al. (2019) Eca-net: Efficient channel attention for deep convolutional neural networks. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp 11531–11539. https://api.semanticscholar.org/CorpusID:203902337
24. Woo S, Park J, Lee J-Y, Kweon I-S (2018) Cbam: Convolutional block attention module. arXiv:1807.06521, https://api.semanticscholar.org/CorpusID:49867180
25. Wang X, Girshick RB, Gupta AK, He K (2017) Non-local neural networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp 7794–7803. https://api.semanticscholar.org/CorpusID:4852647
26. Li W, Zhu X, Gong S (2018) Harmonious attention network for person re-identification. 2018 IEEE/CVF conference on computer vision and pattern recognition. pp 2285–2294. https://api.semanticscholar.org/CorpusID:3458516
27. Chen T et al. (2019) Abd-net: Attentive but diverse person re-identification. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp 8350–8360. https://api.semanticscholar.org/CorpusID:199442462
28. Chu R et al. (2019) Vehicle re-identification with viewpoint-aware metric learning. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp 8281–8290. https://api.semanticscholar.org/CorpusID:203951329
29. Jin X, Lan C, Zeng W, Chen Z (2020) Uncertainty-aware multi-shot knowledge distillation for image-based object re-identification. https://api.semanticscholar.org/CorpusID:210700928
30. Khorramshahi P, Peri N, Chen J-C, Chellappa R (2020) The devil is in the details: Self-supervised attention for vehicle re-identification. arXiv:2004.06271, https://api.semanticscholar.org/CorpusID:215754526
31. He B, Li J, Zhao Y, Tian Y (2019) Part-regularized near-duplicate vehicle re-identification. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp 3992–4000. https://api.semanticscholar.org/CorpusID:173174872
32. Qian J, Jiang W, Luo H, Yu H (2019) Stripe-based and attribute-aware network: a two-branch deep model for vehicle re-identification. Meas Sci Technol 31. https://api.semanticscholar.org/CorpusID:204512251
33. Chen T-S, Liu C-T, Wu C-W, Chien S-Y (2020) Orientation-aware vehicle re-identification with semantics-guided part attention network. arXiv:2008.11423, https://api.semanticscholar.org/CorpusID:221319661
34. Sun Z, Nie X, Xi X, Yin Y (2020) Cfvmnet: A multi-branch network for vehicle re-identification based on common field of view. Proceedings of the 28th ACM international conference on multimedia. https://api.semanticscholar.org/CorpusID:222278145
35. Zhang X et al. (2019) Part-guided attention learning for vehicle instance retrieval. IEEE Trans Intell Transp Syst 23:3048–3060. https://api.semanticscholar.org/CorpusID:221978047
36. Meng D et al. (2020) Parsing-based view-aware embedding network for vehicle re-identification. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp 7101–7110. https://api.semanticscholar.org/CorpusID:215737119
37. Suprem A, Pu C (2020) Looking glamorous: Vehicle re-id in heterogeneous cameras networks with global and local attention. arXiv:2002.02256, https://api.semanticscholar.org/CorpusID:211043643
38. Gu J et al. (2023) Msinet: Twins contrastive search of multi-scale interaction for object reid. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp 19243–19253. https://api.semanticscholar.org/CorpusID:257496331
39. He L et al. (2019) Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp 8449–8458. https://api.semanticscholar.org/CorpusID:118644956
40. Jia M et al. (2021) Matching on sets: Conquer occluded person re-identification without alignment. https://api.semanticscholar.org/CorpusID:235306331
41. Li Y et al. (2021) Diverse part discovery: Occluded person re-identification with part-aware transformer. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp 2897–2906. https://api.semanticscholar.org/CorpusID:235367907
42. Jia M, Cheng X, Lu S, Zhang J (2021) Learning disentangled representation implicitly via transformer for occluded person re-identification. IEEE Trans Multimed 25:1294–1305. https://api.semanticscholar.org/CorpusID:235742882
43. Wang Z et al. (2021) Feature erasing and diffusion network for occluded person re-identification. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp 4744–4753. https://api.semanticscholar.org/CorpusID:245218829
44. Zhao Y, Zhu S-C, Wang D, Liang Z (2022) Short range correlation transformer for occluded person re-identification. Neural Comput Appl 34:17633 – 17645. https://api.semanticscholar.org/CorpusID:245668804
45. Paszke A et al. (2019) Pytorch: An imperative style, high-performance deep learning library. https://api.semanticscholar.org/CorpusID:202786778

46. Deng J et al. (2009) Imagenet: A large-scale hierarchical image database. 2009 IEEE conference on computer vision and pattern recognition. pp 248–255. https://api.semanticscholar.org/CorpusID:57246310
47. Kalayeh MM, Basaran E, Gokmen M, Kamasak ME, Shah M (2018) Human semantic parsing for person re-identification. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp 1062–1071. https://api.semanticscholar.org/CorpusID:4564819
48. Chen X et al. (2020) Salience-guided cascaded suppression network for person re-identification. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp 3297–3307. https://api.semanticscholar.org/CorpusID:219630295

Authors and Affiliations

Hangyu Li[1] · Yu Zhu[1] · Shengze Wang[1] · Ziming Zhu[1] · Jiongyao Ye[1] · Xiaofeng Ling[1] ·

Hangyu Li
Y10220114@mail.ecust.edu.cn

Shengze Wang
y30230924@mail.ecust.edu.cn

Ziming Zhu
Y10220113@mail.ecust.edu.cn

Jiongyao Ye
yejy@ecust.edu.cn

Xiaofeng Ling
xfling@ecust.edu.cn

[1] The School of Information Science and Engineering, East China University of Science and Technolog, Lingyun, Shanghai 200237, Shanghai, China