



# A multi-branch attention and alignment network for person re-identification

Chunyan Lyu<sup>1</sup> · Wu Ning<sup>1</sup> · Chenhui Wang<sup>2</sup> · Kejun Wang<sup>1</sup> 

Accepted: 30 September 2021 / Published online: 18 January 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

## Abstract

Person re-identification plays a critical role in video surveillance and has a variety of applications. However, the body misalignment caused by detectors or pose changes sometimes makes it challenging to match features extracted from different images. To address the issues above, we propose a multi-branch attention and alignment network (MAAN). This approach is based on a deep network with three main branches. One branch is used for global feature representations. Another branch implements a multi-attention process based on keypoints, filters the practical information in the image, and then horizontally partitions the image to extract local features. For the last branch, we create a method based on part feature alignment. We obtain 17 keypoints from a pretrained pose estimation model, and nine local regions from the corresponding feature map are extracted for alignment. Experimental results on various popular datasets demonstrate that our method can produce competitive results under posture changes and body misalignment.

**Keywords** Person re-identification · Multi-branch deep network · Keypoints · Feature fusion

## 1 Introduction

Person re-identification (Re-ID) aims to detect and return images containing the same person from an image gallery. Re-ID is critical in intelligent surveillance systems and has essential research influence and practical significance because it is important in the field of public safety due to the increasing number of surveillance cameras. Since the scene complexity of images from surveillance videos is high, Re-ID's primary challenge comes from considerable changes in the image's subject, such as posture, viewpoint,

occlusion, clothing, background clutter, detection/tracking errors, and illumination changes, among others. These factors contribute to the difficulty of identifying unique individuals from among extensive galleries.

Among these, the crucial factor that affects the Re-ID accuracy is misalignment. First, pedestrians will naturally adopt a variety of postures. Changes in posture mean that the body part's location inside the bounding box is uncertain. Detection errors may also cause misalignment. Pedestrians can appear in various locations throughout the image at varying scales. In addition, different camera viewpoints may cause misalignment due to factors such as different clothing characteristics. As shown in Fig. 1, we present some misaligned images that demonstrate the above factors from three popular Re-ID datasets.

In general, previous Re-ID methods extract features from the entire image and use them for retrieval. These methods either directly use global character features [1, 2] or combine local features from various parts [3–6]. However, when pedestrians are not appropriately aligned, the Re-ID accuracy may be significantly reduced. A typical Re-ID practice, for example, is to divide the bounding box into horizontal stripes [3, 7–10]. This approach assumes minor vertical misalignment. However, when vertical misalignment occurs, a person's head may match the background of another picture. As a result, when extreme misalignment occurs, the horizontal partition's advantages

---

✉ Kejun Wang  
wangkejun@hrbeu.edu.cn

Chunyan Lyu  
lyuchunyan@hrbeu.edu.cn

Wu Ning  
ningwu@hrbeu.edu.cn

Chenhui Wang  
harbin0451@g.ucla.edu

<sup>1</sup> College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin 150001, China

<sup>2</sup> Department of Statistics, University of California, Los Angeles, CA 90095-1554, USA

**Fig. 1** Examples of detected pedestrian images from three popular Re-ID datasets (from left to right are Market-1501 [8], CUHK03 [7], and DukeMTMC-reID [20, 21]). The green bounding box represents the actual location of the pedestrian in the image. The first row shows the misalignment caused by posture and camera viewpoint changes, and the second row shows the misalignment from detection errors



may be diminished. In another case, the feature extractor may incorrectly weigh the background under various pedestrian poses, impairing subsequent matching accuracy.

To our knowledge, several previous works [11–19] explicitly consider the misalignment problem. [11, 12] utilize the image structure (pictorial structure), and the extracted features are affected by noise and produce specific errors. [14, 17, 19] segment the human body structure into blocks using a more precise pose estimation network and employ a particular procedure for reorganization or feature fusion. However, its network structure is relatively simple, and the fusion process results in some information loss. In [87], although the PAP module is used for part alignment, its performance primarily comes from the image’s segmentation constraints in the target domain, which requires impractical and complex pseudo label generation. [13] introduced a four-stream network that acquires global and part features based on the head, upper body, and lower body simultaneously. It then combines multiple features to produce a GLAD descriptor. It lacks, however, strong robustness to changes in posture and viewpoint. In summary, [15, 16, 18, 19] add attention factors, [16, 18] are based on posture keypoints, and [15, 19] are derived from similarity calculations. These factors guide the model’s attention to the critical portion of the input image that contributes to alignment. Although relying solely on attention for alignment is not ideal, the robustness is insufficient when attention modules are not used. Unlike the above methods, the network we built incorporates multiple branches based on attention and alignment. Diverse attention mechanisms and feature partitioning are used in various branches, with global feature representations for complementation, contributing to better feature alignment and improving the model’s performance.

Considering the problems mentioned above and the limitations of existing methods, we propose a multi-branch feature fusion strategy based on pose-guided multi-attention and feature alignment. The attention mechanism in deep learning causes the network to pay more attention to feature extraction of areas practical for the Re-ID task, and the keypoints have important guiding significance for alignment in high-level semantics. We propose a concise but efficient multi-branch attention and alignment network that combines the ideas mentioned above. It is divided into three parts. Before processing the input images, they are routed via a backbone network (ResNet-50). Following this process, the global features are obtained and sent to three distinct components. Part 1 is used for global feature representations. In Part 2, we introduce a multiple attention mechanism including spatial attention, channel attention, and keypoint attention; then, we set a horizontal feature partition for local feature mining. In Part 3, we introduce an alignment method based on part features. We use a pretrained pose estimation model to obtain 17 keypoints; relying on these keypoints, we divide the input feature map into nine parts to achieve alignment. In addition, and considering the errors of attention blocks and pose estimation, both Part 2 and Part 3 introduce a sub-branch for feature information complementation. Notably, our feature extraction is based on the backbone, and only extra attention and region divisions are used in subsequent branches to further complete feature extraction. This design minimizes the number of network parameters and makes it easy to train. Simultaneously, we calculate the loss function for each minimum branch’s features. Finally, these features are concatenated to create the final representation of the input images; this process assists in developing the extracted features’ discrimination and robustness, thus improving the Re-ID accuracy.

This paper's main contributions are:

- We propose a simple yet effective Re-ID pipeline called the multi-branch attention and alignment network (MAAN). This method can simultaneously learn local features using the attention mechanism while finding part features partitioned according to keypoints. In addition, to construct a global feature representation branch, we create a separate sub-branch to maintain the global features as supplementary before each main branch's feature partitioning process. We achieve feature alignment and increase the network's robustness by combining features from multiple levels and different critical locations.
- We use keypoints to partition the global features into nine specialized parts and connect a classification loss function to each part feature in the part alignment branch. This process enables finer-grained mining of part features, mitigates the effects of real factors such as sample noise, and leads to improved alignment.
- Using MAAN, we report competitive Re-ID accuracy on the Market-1501 [8], CUHK03 [7], and DukeMTMC-reID [20, 21] datasets.

## 2 Related work

### 2.1 Attention mechanism

In the last few years, the attention mechanism has been widely used in computer vision as a method that may enhance convolutional neural networks (CNNs). Its primary objective is to choose the most critical pieces of information from a large amount of data. SENet [22] proposed a squeeze and exception network based on the relationship between feature channels. The interdependence between feature channels is modelled explicitly, and the weight of each feature channel is automatically obtained by learning. Then, according to the weight, the model enhances the valuable features and suppresses the useless features for the current task. SKNet [23] inspired cortical neurons to dynamically adjust their receptive fields according to different stimuli, used multiscale feature information to guide the distribution, and focused on which kernel representation to use. To emphasize the meaningful features in the two region and channel dimensions, CBAM [24] combines the two channel and spatial attention modules to achieve better feature representation.

### 2.2 Pose estimation

The research on pose estimation has diverged from classical methods [11, 12] to deep learning [25–27]. In general,

this problem can be divided into four tasks: single-person skeleton estimation, multi-person pose estimation, video pose tracking, and 3D skeleton estimation. For single-person skeleton estimation, a picture of the cropped person is used as input, and then the keypoints needed in the body area, such as the head, left hand, and right knee, are predicted. Keypoints can be used to indicate the position of the human body and can assist with a variety of visual issues. In this paper, we use OpenPose [26] to produce keypoints, including the position of keypoints on the face, hand, and joints of the human body, since it is a multistage pose estimator with continuous pose prediction.

### 2.3 Person re-identification

Person re-identification addresses the problem of matching pedestrian images across disjoint cameras. The key challenge lies in the large intragroup and small intergroup differences caused by different views, illumination, occlusion, and poses. Existing techniques can be classified as follows: hand-crafted descriptors [10, 30], metric learning methods [32–34], and deep learning algorithms [2–10, 35]. The aforementioned Re-ID approaches are not robust to changes in human pose and camera viewpoint, restricting their applicability in real-world surveillance scenarios.

### 2.4 Part-based person re-identification

The part-based Re-ID methods use local descriptors from different regions to enhance the discrimination and robustness of the feature representation. Part-based deep feature extraction methods can be divided into two groups. The first group turns to the predicted keypoints, which require the help of pretrained pose estimators. [36] suggested a novel pose-based attention perception synthesis network. In addition, part of the visibility is also incorporated into the final feature representation. [37] suggested combining the person's fine and coarse posture information to learn the discriminative embedding, directly splicing the confidence map of 14 keypoints, and the model automatically learns alignment. In [87], under the guidance of pose estimation and semantic segmentation, part aligned pooling and part segmentation constraints were proposed to improve the cross-domain Re-ID behaviour. The second group does not require keypoints or segmentation information. A simple method is to divide the person image or feature map into uniform partitions. [3] divided the feature map into  $p$  horizontal stripes and trained each embedding part independently using a non-shared classifier. Additionally, one can extract the local features using pose-driven RoI extraction [14], human parsing results [38], or learning attention regions based on appearance features [5, 6, 39]. For instance, [14] proposed

using posture detection to generate a local area through a manually crafted cutting method and then gradually fusing part features. [38] extracted features of body parts from human semantic analysis results. In [5, 6, 39], they attempted to exploit local information using appearance-based attention maps.

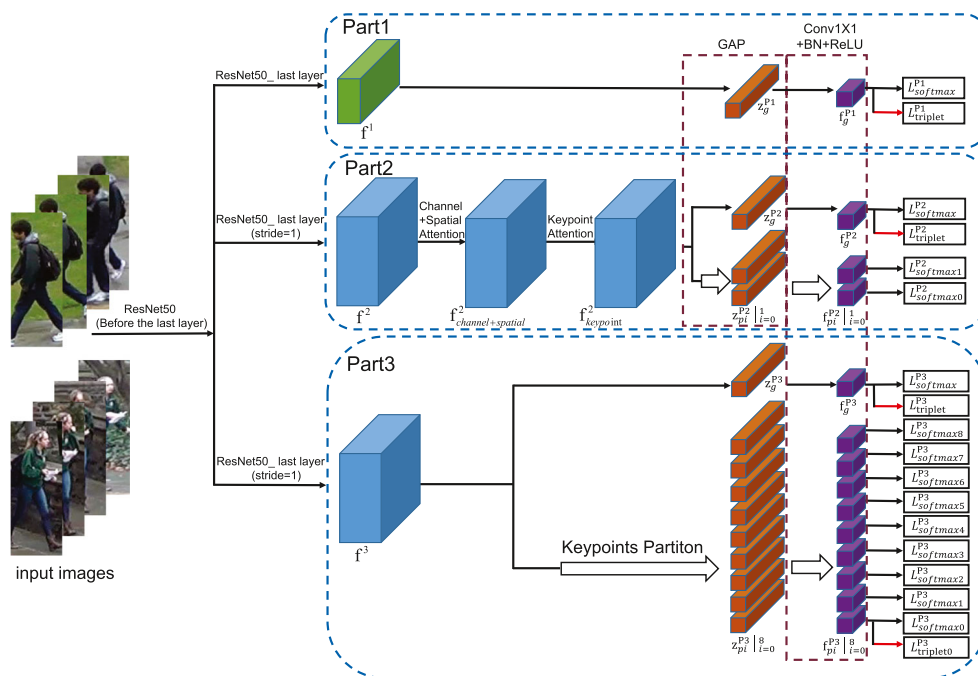
Compared with the above related Re-ID methods, we creatively construct a more complete and robust fine-grained feature extraction and alignment framework based on multi-branch deep networks and multi-task learning. We simultaneously introduced a multi-attention mechanism and alignment method. In the former, we combine spatial attention, channel attention, and keypoint attention while carrying out multiple feature segmentation operations at the appropriate scale to strengthen the information mining of local features. In the alignment branch, we propose a new horizontal segmentation method with reused hierarchical information, which is different from previous work. For example, 'upper leg' features will appear in the 'upper leg', 'lower body', and 'whole body', and they are further divided into three sub-branches. This design is more reasonable because each branch can provide better gradient flow during training and alleviate the problem of the uneven gradient of each classification loss function. When calculating feature similarity, the feature robustness can also be improved, and better results can be achieved.

### 3 Method

This paper proposes a multi-branch Re-ID network consisting of a global feature extraction network; a multi-attention mechanism that incorporates channel, spatial, and keypoint attention; and a pose-guided part feature alignment network. We outline the proposed method's overall framework in Section 3.1, and the design of the global feature extraction network is shown in Section 3.2. The pose-guided attention mechanism and the part alignment method are discussed separately in Sections 3.2 and 3.4. Section 3.5 summarizes the overall structure.

#### 3.1 The overall framework

In a Re-ID system, an input pedestrian image's global features can be used to achieve a reasonable Re-ID effect. However, by learning more refined local features, the effect is improved when compared to using only global features. Traditional methods usually use uniform partitioning and do not pay attention to the region around keypoints. As a result, the change in pedestrian pose and camera viewpoint can substantially influence the network's performance. To solve this problem, this paper proposes a MAAN module. The overall structure is shown in Fig. 2.



**Fig. 2** Overview of our multi-branch attention and alignment network (MAAN). ResNet50 is used as the backbone for global feature extraction, whereas different branches adopt different pooling strides at the last layer. Pose information is included using a keypoint attention mechanism as well as a keypoint partition operation. The MAAN

consists of three main branches, Part 1, Part 2, and Part 3. Part 1: global feature extraction network from all input images. Part 2: multi-attention mechanism including spatial attention, channel attention, and keypoint attention. Part 3: part feature alignment network based on a pose-guided feature partition operation

### 3.2 Global feature extraction

As shown in Fig. 2, Part1 is a global feature extraction branch for input images. To increase the scale of extracted features in the network, the input image is resized to  $384 \times 128$ . We adopt ResNet-50 as the backbone to extract feature  $f^1$  with a size of  $2048 \times 12 \times 4$ . Part1 learns global information using global average pooling (GAP),  $1 \times 1$  convolution(Conv1×1), BatchNorm(BN) and the ReLU activation function. We use ResNet-50 to extract features since it can converge quickly and reduce the number of parameters. In addition, it can make the model easier to train, which not only prevents the gradient from disappearing but also prevents the loss from diverging.

The global feature  $f_g^{P1}$  is extracted from the input image using this branch, and the size of  $f_g^{P1}$  is  $256 \times 1 \times 1$ . Through dimension reduction,  $f^1$  can be reduced to 256-dim from 2048-dim, which is more effective for feature calculation. The 256-dim feature  $f_g^{P1}$  is used to simultaneously calculate the softmax loss  $L_{softmax}^{P1}$  and hard triplet loss  $L_{triplet}^{P1}$ . Both losses are summed for backpropagation. For feature  $f_i$ , the softmax loss is formulated as follows:

$$L_{softmax} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(W_{y_i}^T f_i)}{\sum_{k=1}^C \exp(W_k^T f_i)} \quad (1)$$

$W_k^T$  is the weight vector for class  $k$ , where  $N$  denotes the number of mini-batches in the training period,  $C$  denotes the number of classes in the training dataset, and  $W_{y_i}^T$  corresponds to a weight vector when input sample  $i$  is predicted for the correct class  $y_i$ . To improve the ranking performance, the global feature  $f_g^{P1}$  is trained using a hard triplet loss that consists of an anchor sample, a positive sample, and a negative sample. The anchor and positive samples are the most dissimilar positive sample pairs, while the anchor and negative samples are the most similar negative sample pairs. The following formula is used to express the hard triplet loss function.

$$L_{triplet} = \sum_i^N \left[ \max \left\| f_a^i - f_p^i \right\|_2 - \min \left\| f_a^i - f_n^i \right\|_2 + margin \right]_+ \quad (2)$$

$f_a^i, f_p^i, f_n^i$  are features extracted from an anchor, a positive sample and a negative sample respectively, and the margin is used to control the inter-class distance.  $\left\| f_a^i - f_p^i \right\|_2$  is the Euclidean distance between the anchor and positive sample, and  $\left\| f_a^i - f_n^i \right\|_2$  is the Euclidean distance between the anchor and negative sample. The global feature has good performance on Re-ID tasks; however, it introduces some

interference factors such as background noises, so we use it as a supplement to the overall feature and combine it with other branches.

### 3.3 Multi-attention mechanism

The attention mechanism is an important tool in computer vision tasks, and it causes the network to pay more attention to the effective part of the input image. Therefore, in the second part of the network (Part 2), we introduced a multi-attention mechanism combined with classic horizontal segmentation to complete local feature extraction. The principle of channel attention is to use different channels' coefficient weights, and that of spatial attention is to use different regions' coefficient weights. Neither of the two, however, recognize the significance of human pose variation. Thus, following the channel and spatial attention blocks, we introduce a keypoint attention block. This combination causes the network to focus on valuable parts of input images while diminishing the importance of insignificant or even harmful regions.

As shown in Part 2 of Fig. 2, the last layer of ResNet-50 cancels the downsampling operation that extracts feature  $f^2$  with a size of  $2048 \times 24 \times 8$ . The last stride is changed from 2 to 1, so the size of feature  $f^2$  is twice as large as that of  $f^1$ . A larger feature size indicates that more information can be obtained, which is more helpful when learning details.

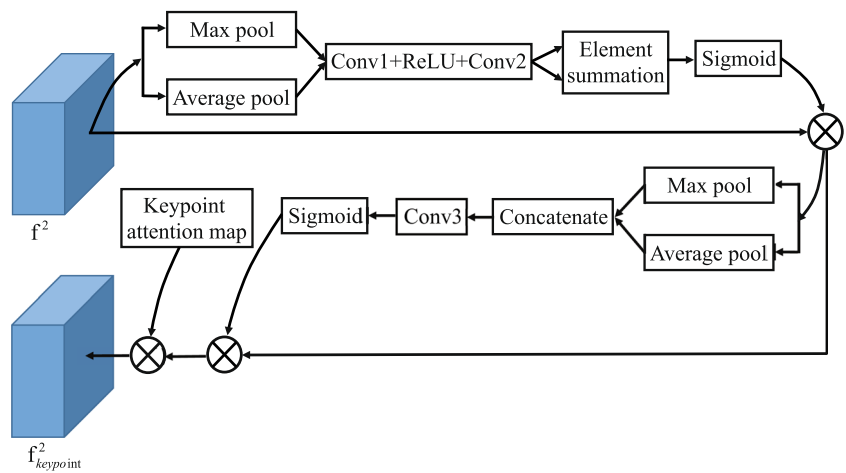
We detail the attention mechanism in Fig. 3. First,  $f^2$  passes through a global max pooling and GAP block. Both features are then forwarded to another network, which conducts channel attention. This network is composed of two convolution layers and a ReLU activation function. After adding both output features, which are based on each element and activated by the sigmoid function, this subsequent network generates the final channel attention map. The channel attention map  $f_{channel\_map}$  is computed as:

$$f_{channel\_map} = \sigma \left( W_1 \left( ReLU \left( W_0 avgpool \left( f^2 \right) \right) \right) + W_1 \left( ReLU \left( W_0 maxpool \left( f^2 \right) \right) \right) \right) \quad (3)$$

where  $\sigma$  denotes the sigmoid function, the convolution weights  $W_0$  and  $W_1$  are shared by both inputs, and the ReLU activation function is followed by  $W_0$ .

The channel attention map  $f_{channel\_map}$  and the feature  $f^2$  are multiplied element-wise to generate feature  $f_{channel}^2$ . Then, feature  $f_{channel}^2$  is used as the input for the spatial attention module. First,  $f_{channel}^2$  passes a global max pooling layer and a GAP layer separately, and then we concatenate these two features on the channel dimension. The dimension is reduced to one channel using another convolution layer. After that, the spatial attention map

**Fig. 3** Illustration of our proposed multi-attention mechanism



$f_{spatial\_map}$  is generated using a sigmoid function, and  $f_{spatial\_map}$  is computed as:

$$f_{spatial\_map} = \sigma \left( W_2 \left[ avgpool \left( f_{channel}^2 \right), maxpool \left( f_{channel}^2 \right) \right] \right) \tag{4}$$

where  $\sigma$  denotes the sigmoid function, and  $W_2$  is the convolution weight. Then, feature  $f_{channel}^2$  is multiplied by  $f_{spatial\_map}$  to obtain the input feature for the keypoint attention module.

The channel and spatial attention blocks produce the feature  $f_{channel+spatial}^2$ , which is then used to process the keypoint attention. Specifically, we use OpenPose[26] pretrained on the MS COCO2017 dataset[41] to obtain the coordinates of 18 keypoints in the original image, including 'nose', 'neck', 'right shoulder', 'right elbow', 'right wrist', 'left shoulder', 'left elbow', 'left wrist', 'right hip', 'right knee', 'right ankle', 'left hip', 'left knee', 'left ankle', 'right eye', 'left eye', 'right ear', and 'left ear'. According to our experience, when applying the pretrained model to Re-ID datasets, the estimation of 'neck' has a larger deviation. Therefore, we removed 'neck' and used the remaining 17 keypoints for the attention calculation in Part 2 and the partition of the global feature in Part 3. The keypoint coordinates are modified using Gauss's transformation and binarization.

$$f_{keypoint\_map} [m \geq 0.8] = 1$$

$$f_{keypoint\_map} [m < 0.8] = 0 \tag{5}$$

where  $m$  denotes the confidence values on the keypoint attention map  $f_{keypoint\_map}$ . The elements on this attention map matrix that exceed 0.8 are set to 1, as we regard these elements as belonging to the region near the keypoints, and the remaining elements are set to 0. As shown in Fig. 4, bright areas represent the regions near the keypoints. The keypoint attention map focuses only on the areas

surrounding 17 keypoints, omitting other areas. After multiplying by the corresponding features, it is possible to emphasize significant characteristics near the keypoints and extract more practical features for pedestrian identity representation.

Then,  $f_{channel+spatial}^2$  is multiplied by the keypoint attention map to obtain  $f_{keypoint}^2$ , which has a size of  $2048 \times 24 \times 8$ . Next, it is sent to two branches, one of which learns global information using GAP, a  $1 \times 1$  convolution(Conv1 $\times$ 1), BatchNorm(BN) and the ReLU activation function. The global feature  $f_g^{P2}$  is extracted from the input image using this sub-branch, and the size of



**Fig. 4** Visualization of keypoint attention maps on four different pedestrian images

$f_g^{P2}$  is  $256 \times 1 \times 1$ . We retain this global feature following the upper attention module as a complement. In another sub-branch, feature  $f_{keypoint}^2$  passes through GAP and a  $1 \times 1$  convolution layer, to obtain features with a size of  $256 \times 2 \times 1$ . Using horizontal partitioning, the feature is divided into two uniform parts,  $f_{pi}^{P2} |_{i=0}^1$ , which have the same size of  $256 \times 1 \times 1$ . We can obtain better local features by mining after this division operation.

The softmax loss  $L_{softmax}^{P2}$ ,  $L_{softmax0}^{P2}$ ,  $L_{softmax1}^{P2}$  and the hard triplet loss  $L_{triplet}^{P2}$  are calculated after the calculation of the above branches. All the losses are added together for backpropagation.

In this section, only softmax loss is used for the two local features  $f_{pi}^{P2} |_{i=0}^1$ , because if hard triplet loss is used for the local feature mining branches, the results will be less satisfactory. When an image is divided into two parts from top to bottom, it is not clear whether the upper part represents the upper body or the lower part represents the lower body. In reality, the upper part may include the background, while the lower part may represent the whole body. If we choose the hard triplet loss, the distance between the background and the upper body is meaningless, and the training data will cause the model to learn the wrong prediction.

### 3.4 Pose-guided part feature alignment

In this section, a part feature alignment network is proposed. When matching the local features from two pedestrian images, the identified accuracy based on the same body part is better. Therefore, according to human parsing, we could extract nine local regions using 17 estimated keypoints, as mentioned in Sec. 3.2. By aligning the corresponding part features, we can solve the offsets caused by misalignment, thus enhancing the discrimination of extracted features and the robustness of the entire network.

Part 3 of Fig. 2 includes one sub-branch for global information extraction and nine sub-branches for part information mining. The input image first passes through the same backbone used in Part 2, and the last layer of ResNet-50 cancels the downsampling operation to extract the feature  $f^3$ , which has a size of  $2048 \times 24 \times 8$ , the same as  $f^2$ .

After that, we distribute the network into two branches. One learns global information using GAP and a  $1 \times 1$  convolution to obtain the global feature  $f_g^{P3}$ , and the other obtains the features  $z_{pi}^{P3} |_{i=0}^8$  using pose-guided partitioning, as shown in Fig. 5. As mentioned in Sec. 3.2, we use OpenPose[26], which is pretrained on the MS COCO2017



**Fig. 5** Examples with 17 keypoints and nine part partitions. We used OpenPose [26], which was pretrained on the MS COCO2017 dataset [41], to obtain the coordinates of 18 keypoints in the original image, including 'nose', 'neck', 'right shoulder', 'right elbow', 'right wrist', 'left shoulder', 'left elbow', 'left wrist', 'right hip', 'right knee', 'right ankle', 'left hip', 'left knee', 'left ankle', 'right eye', 'left eye', 'right ear', and 'left ear'. According to our experience, when applying the pretrained model to Re-ID datasets, the estimation of 'neck' has a large deviation. Therefore, we removed 'neck' but retained the remaining 17 keypoints. The yellow dots in each image represent the 17

keypoints. Nine rectangular boxes with different colours divide the body into different parts, including the head between the top of the skull and shoulders, the upper torso between the shoulders and chest, the lower torso between the chest and waist, the upper leg between the waist and knees, the lower leg between the knees and ankles, the feet between the ankles and soles of the feet, the upper body between the head and waist, the lower body between the waist and feet, and the entire body between the head and feet. In addition, none of the nine regions contain ineffective background information above the head or under the feet

dataset [41], to obtain 17 keypoints from the total 18 keypoints for feature partitioning; we discard the neck keypoint. Specifically, we can obtain the coordinates of 17 keypoints in the original image and map to the feature map using the size ratio of the original image as well as the feature map. The vertical coordinates' maximums for the left eye and right eye were selected as the vertical coordinate  $Y1$  for the eyes, and the vertical coordinates' minimum for the left ankle and right ankle were selected as the vertical coordinate  $Y2$  for the ankles. According to the human body's proportions, 120% of the difference  $M$  between the vertical eye and mouth coordinates is calculated as the distance from the eyes to the top of the head. The vertical coordinate  $Y1$  for the eyes plus this distance  $M$  is used as the vertical head coordinate  $Y3$ , and the vertical coordinate  $Y2$  for the ankles minus this distance  $M$  is used as the vertical coordinate  $Y4$  for the feet. Therefore, a person's height is the difference  $|Y3 - Y4|$  between the vertical coordinates for the head and the feet, and the person image is segmented into six stripes, as shown in Fig. 5. The average position of the symmetrical keypoints is used as the partition boundary. The average position of the left and right shoulders is used as the boundary between the head and upper torso, and the area between the shoulder and waist is divided equally into two parts: the upper torso and the lower torso. The average position of the left and right waist is used as the boundary between the upper body and the lower body. The average position of the left and right knees is used as the boundary between the upper leg and the lower leg. The average position of the left and right ankles is used as the boundary between the lower leg and feet. In this way, pedestrians can be divided into six parts. In addition to the above operations, and considering the general information, the image can be divided into three parts. The upper body is composed of the head, upper torso, and lower torso. The lower body is composed of the upper leg, lower torso, and feet, and the whole body is composed of the head to the feet. Therefore, the feature  $z_{pi}^{P3} \mid_{i=0}^8$  is finally divided into nine different parts.

This division method causes the network to pay attention to specific human body regions; thus, the extracted features are more specific, and the entire image is free from interference such as background noise. For example, the network can focus on details about the local information, such as the logo on the clothes.

The division method also modifies a uniform partition, such as PCB, to realize partial feature alignment. Additionally,  $1 \times 1$  convolution(Conv1 $\times$ 1), BatchNorm(BN) and the ReLU activation function are used to obtain features  $f_{pi}^{P3} \mid_{i=0}^8$ , and each has a size of  $256 \times 1 \times 1$ .

The global feature  $f_g^{P3}$  and the features  $z_{pi}^{P3} \mid_{i=0}^8$  from the other branch are used to learn each part

feature of an input image, and these subnetworks share the same weights. Instead of training ten subnetworks separately, they are trained to share weights in the convolution layer to avoid overfitting. The softmax losses,  $L_{softmax}^{P3}, L_{softmax0}^{P3}, L_{softmax1}^{P3}, L_{softmax2}^{P3}, L_{softmax3}^{P3}, L_{softmax4}^{P3}, L_{softmax5}^{P3}, L_{softmax6}^{P3}, L_{softmax7}^{P3}, L_{softmax8}^{P3}$ , and the hard triplet losses,  $L_{triplet}$  and  $L_{triplet0}$ , are calculated. Eventually, all losses are added together for backpropagation. As  $f_{pi}^{P3} \mid_{i=0}$  means that all of the body features belong to the global feature, features  $\{f_g^{P3}, f_{pi}^{P3} \mid_{i=0}\}$  are trained using the hard triplet loss to improve the network's performance.

In the training stage, the global feature  $f_g^{P3}$  and nine part features  $f_{pi}^{P3} \mid_{i=0}^8$  are calculated separately and are then concatenated to form the entire identity representation during the test stage. Every branch in Part 3 shares parameters during training to enhance the model's performance, as the concatenation of global features and each part feature with a total size of  $10 \times 256$  is used in the test stage. The shared convolution kernels are forced to learn both the global and local features, and more samples are used during training to avoid overfitting.

This pose-guided feature partitioning method can effectively focus on the critical body parts and mine the corresponding information while suppressing the misalignment caused by background noise and detection errors. Compared with the existing partitioning methods [13, 14, 17, 19], the features we extracted consider the representation at different granularities; they are more informative and complete. We also prove that this method can achieve good results in the experiment section.

### 3.5 Training and inference

The entire network structure is composed of Part 1, Part 2, and Part 3. These branches include both cooperation and division of labour. The weights of the lower layers are shared, and those of the higher layers are independent. Global features are the overall common representation, and then the multi-attention mechanism and pose-guided feature alignment focus on local features at different levels. Combining the global and local features as the final identity representation could strengthen the network's discrimination. Thus, in the training stage, the total loss function is formulated as follows:

$$L = L_{softmax}^{P1} + L_{softmax}^{P2} + L_{softmax}^{P3} + L_{triplet}^{P1} + L_{triplet}^{P2} + L_{triplet}^{P3} + L_{triplet0}^{P3} + \sum_{i=0}^1 L_{softmaxi}^{P2} + \sum_{i=0}^8 L_{softmaxi}^{P3} \quad (6)$$



**Table 1** Descriptions of the Market-1501, DukeMTMC-reID, and CUHK03 (detected) datasets

Dataset	Market-1501	DukeMTMC-reID	CUHK03(detected)
Release time	2015	2017	2014
Shooting location	Tsinghua University	Duke University	The Chinese University of Hong Kong
Identities	1501	1404	1467
Identities in training set	751	702	767
Images in training set	12936	16522	7365
Identities in test set	750	702	700
Images in test set	19732	17661	5332
Cameras	6	8	2

our method realizes end-to-end learning by integrating global information, multi-attention-based local features, and pose-guided part feature alignment. In retrieval, as shown in Fig. 2, there are 18 purple blocks of 256-dim vectors from top to bottom at the end of the network. We concatenate these to form an identity feature with the size of  $18 \times 256$ . The Euclidean distance was used for similarity calculation.

## 4 Experiments

The experiments are described in nine sections. Sections 4.1 and 4.2 introduce the datasets, evaluation protocols, experimental environments, and implementation details. In Section 4.3, we verify the effectiveness of the global feature extraction network. Section 4.4 demonstrates that the accuracy is further improved by the multi-attention mechanism. Section 4.5 proves the effectiveness of pose-guided part feature alignment. Section 4.6 shows the superiority of the overall network. In Section 4.7, we conduct an additional cross-domain experiment to demonstrate the generalization ability of this network. Section 4.8 provides a discussion of the time and space complexity. Finally, Section 4.9 conducts a synthesis comparison between our method and several state-of-the-art methods.

### 4.1 Dataset and evaluation protocol

We conduct experiments on three popular Re-ID datasets: Market-1501 [8], DukeMTMC-reID (Duke) [20, 21], and CUHK03 [7]. The Market-1501 dataset was collected on the Tsinghua University campus, and the images are from six non-overlapping cameras, including one with low resolution. The images are automatically detected and cut by the detector. This dataset includes pose changes, illumination variations, and occlusion, which are similar to an actual scene. The DukeMTMC-reID dataset is the largest person Re-ID dataset and was collected at Duke University. It also provides the annotation of pedestrian attributes such

as gender and sleeve length, among others. The CUHK03 dataset was collected at the Chinese University of Hong Kong and contains some detection errors. The details of these datasets are described in Table 1.

In our experiments, to evaluate the performances of Re-ID methods, we report the cumulative matching characteristics (CMC) at Rank-1, Rank-3, Rank-5, Rank-10 and the mean average precision (mAP) on all the candidate datasets.

### 4.2 Experimental environment and implementation details

Our model is trained on PyTorch, and the details of the experimental environment are shown in Table 2.

During the training process, the backbone network adopts the ResNet-50 model, which was pretrained on ImageNet. Pretraining was been completed before network initialization, to reduce the training time. To increase the scale of feature extraction, the input image is resized to  $384 \times 128$ . We use random flipping and random erasing to augment the data. The ADAM optimizer and smooth label are used to train the network for 500 epochs, and the first-order momentum coefficient and second-order momentum coefficient are 0.9 and 0.999, respectively. The weight decay is set to  $5e-4$ . The initial learning rate is set to  $2e-4$ , which decays by 0.1 at epochs 300 and 400 to avoid overfitting. The mini-batch is composed of 16 samples, and the batch size for every sample is set to 4. The margin in the triplet loss is set to 1.2.

**Table 2** Experimental environment

Configuration	Parameter
Operating System	Ubuntu 16.04
CPU	Intel(R) Core(TM) i5-6500 CPU@3.20GHz
GPU	Nvidia GeForce RTX 2080Ti
Cudnn	7.1.4
Software Platform	Pytorch 1.4.0

**Table 3** Results of the methods "ResNet-50" and "Part1" on the Market-1501, DukeMTMC-reID and CUHK03 (detected) datasets

Dataset	Method	Rank1	Rank3	Rank5	Rank10	mAP
Market-1501	ResNet-50	87.6	94.99	96.71	98.23	71.02
	Part1	90.56	95.75	96.91	98.07	77.6
DukeMTMC-reID	ResNet-50	81.04	88.89	91.05	94.06	64.63
	Part1	82.63	89.45	90.98	93.31	68.4
CUHK03	ResNet-50	45.14	60.07	68.14	77.29	43.3
	Part1	63.36	75.64	80.86	88.29	59.07

### 4.3 Effectiveness of the global feature extraction branch

As shown in Part 1 of Fig. 2, the global feature extraction branch is used separately for Re-ID research. The softmax loss and hard triple loss are introduced for training. This method is represented by "Part1". We continue to experiment on the ResNet-50 network and only use the softmax loss to calculate the loss. This method is represented by "ResNet-50", and we compare Part 1's performance with ResNet-50. The results of "Part1" and "ResNet-50" are recorded in Table 3.

Table 3 shows that after adding the global feature, when compared with ResNet-50, the Rank1/mAP rises from 87.6%/71.02% to 90.56%/77.6% (+2.96%/6.58%) on Market-1501, from 81.04%/64.63% to 82.63%/68.4% (+1.59%/3.77%) on DukeMTMC-reID, and from 45.14%/43.3% to 63.36%/59.07% (+18.22%/15.77%) on CUHK03. The global network leads to better Re-ID accuracy. In addition, from Table 4, we can see that "Part1" is superior to other popular methods based on the global feature. Nevertheless, the global feature does not focus on detailed information. Combining the local features from Part2 and Part3 can improve the network.

### 4.4 Effectiveness of multi-attention mechanism branch

As shown in Part2 of Fig. 2, in this section, we use this branch separately for Re-ID research. Specifically, the

softmax and hard triple loss are used for training, and we concatenate the output features from the three sub-branches of Part 2 to create the final representation of the input images during retrieval. This method is denoted by "Part2" in the table. For this multi-attention mechanism branch, Fig. 6 shows the feature response maps of two query images. The feature response map can filter background information and focus on the valuable information in the pedestrian images, and highlights only the fields surrounding the head, shoulders, waist, limbs, and feet near 17 keypoints.

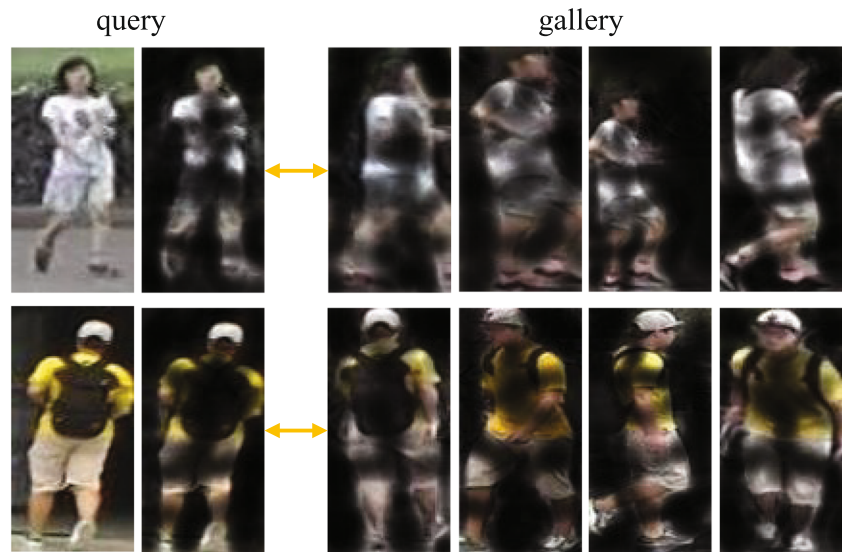
To verify the effect of the multi-attention mechanism, we designed three other network structures for comparison. The first removes all attention modules, but the network structure is the same as Part 2 and includes the same feature segmentation method. This method is denoted as "Base". The second network structure uses only the spatial and channel attention modules. This method is denoted as "Spatial+Channel". The third network structure uses only the keypoint attention modules. This method is denoted as "Keypoint". The results of this comparative experiment are shown in Table 5.

Table 5 shows that after adding the channel, spatial, and keypoint attention blocks, the model performance is significantly improved. Compared with "Base", the results of Rank1/mAP increase from 91.95%/77.59% to 92.67%/82.99% (+0.72%/5.4%) on Market-1501, from 83.17%/68.93% to 86.0%/73.52% (+2.83%/4.59%) on DukeMTMC-reID, and from 54.36%/51.5% to 73.57%/70.61% (+19.21%/19.11%) on CUHK03. In addition, combining spatial and channel attention with keypoint

**Table 4** Comparison with several global-feature-based SOTA methods on the Market-1501 and DukeMTMC-reID datasets

Method	Market-1501		DukeMTMC-reID	
	Rank1	mAP	Rank1	mAP
IDE [40]	79.5	59.9	-	-
SVDNet [1]	82.3	62.1	76.7	56.8
TriNet [41]	84.9	69.1	-	-
AWTL [42]	89.5	75.7	79.8	63.4
Part1(Ours)	<b>90.56</b>	<b>77.6</b>	<b>82.63</b>	<b>68.4</b>

**Fig. 6** Feature response maps from the pose-guided attention network in the query set and gallery set. The left side of the arrow denotes an original pedestrian image from the query set and its feature response map from the pose-guided attention network; the right side of the arrow denotes the feature response maps produced by the pose-guided attention network for four pedestrian images from the gallery set, in which the ID is the same as the query image on the left. The bright fields indicate the positions near the keypoints of the human pose, which are more valuable than other areas



attention results in better performance than when they are used separately. Different attention mechanisms affect the model’s performance from different aspects. Channel attention represents global information, spatial attention represents local information, and keypoint attention focuses on the information near human body keypoints. In addition, following the multi-attention mechanism, Part 2 is divided into a global feature and two local features so that the network can better learn the local features. Thus, in Table 6, we find that the multi-attention mechanism branch achieves better performance than some common attention-based SOTA methods.

### 4.5 Effectiveness of the pose-guided part feature alignment branch

As shown in Part 3 of Fig. 2, the pose-guided part feature alignment branch partitions the global features into nine different parts according to the 17 keypoints and then outputs features including the global feature  $f_g^{P3}$ , the head  $f_{pi}^{P3} |_{i=1}$ , the upper torso  $f_{pi}^{P3} |_{i=2}$ , the lower torso  $f_{pi}^{P3} |_{i=3}$ , the upper leg  $f_{pi}^{P3} |_{i=4}$ , the lower leg  $f_{pi}^{P3} |_{i=5}$ , the feet  $f_{pi}^{P3} |_{i=6}$ , the upper body  $f_{pi}^{P3} |_{i=7}$ , the lower body  $f_{pi}^{P3} |_{i=8}$ , and the entire body  $f_{pi}^{P3} |_{i=0}$ .

**Table 5** Results with different Part 2 settings on the Market-1501, DukeMTMC-reID and CUHK03 (detected) datasets

Dataset	Method	Rank1	Rank3	Rank5	Rank10	mAP
Market-1501	base	91.95	96.23	97.09	98.22	77.59
	spatial+channel	92.58	96.24	97.12	98.27	80.09
	keypoint	92.43	96.00	97.19	98.12	81.82
	Part2	92.67	96.17	97.24	98.19	82.99
DukeMTMC-reID	base	83.17	89.18	91.74	93.9	68.93
	spatial+channel	84.76	90.38	92.31	94.68	71.50
	keypoint	85.28	90.40	92.42	94.86	72.55
	Part2	86.0	90.57	92.55	94.66	73.52
CUHK03	base	54.36	67.29	74.21	82.86	51.5
	spatial+channel	71.79	82.14	86.86	92.14	69.58
	keypoint	72.93	82.71	87.21	92.36	69.95
	Part2	73.57	82.79	88.14	92.43	70.61

”Base” refers to the single network that removes all the attention blocks from Part 2. ”Spatial+Channel” refers to using only the spatial attention and channel attention modules from Part 2 while removing the keypoint attention block. ”Keypoint” refers to using only the keypoint attention module from Part 2 while removing the other attention blocks. ”Part2” refers to this main branch’s complete structure

**Table 6** Comparison with several attention-based SOTA Re-ID methods on the Market-1501, DukeMTMC-reID and CUHK03 (detected) datasets

Method	Market-1501		DukeMTMC-reID		CUHK03(detected)	
	Rank1	mAP	Rank1	mAP	Rank1	mAP
DLPAP [43]	81.0	63.4	-	-	-	-
HA-CNN [39]	91.2	75.7	80.5	63.8	41.7	38.6
Mancs [44]	<b>93.1</b>	82.3	84.9	71.8	65.5	60.5
PAN [45]	82.8	63.4	71.6	51.5	36.3	34.0
PPA [46]	92.3	79.3	85.1	70.2	64.6	62.0
DuATM [47]	91.4	76.6	81.8	64.6	41.7	38.6
Part2(Ours)	92.67	<b>82.99</b>	<b>86.0</b>	<b>73.52</b>	<b>73.57</b>	<b>70.61</b>

In this section, we use this branch separately for Re-ID research and conduct experiments using various feature fusion and training strategies on Market-1501. The detailed experimental results are recorded in Table 7.

As seen from Fig. 7, we divided the global feature of the corresponding position in the input image into the following parts using the keypoints: the upper body, the lower body, the head, the upper torso, the lower torso, the upper leg, the lower leg, and the feet. These corresponding parts may be located in different positions in different images, and the original sizes vary; however, the last features extracted represent the same semantic information. Therefore, through this partition operation,

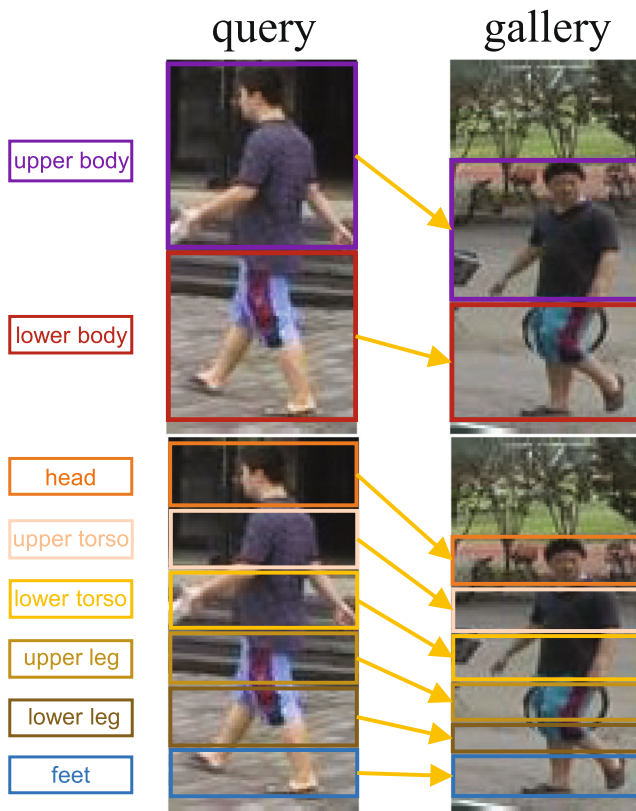
when we calculate the distance between different features, the results are more accurate and can align the same features in different images.

In Table 5, the effect of the "Part3" fusion strategy is the best. The network performance with weight sharing is better than that which does not share parameters because sharing convolution kernels results in better training and avoids overfitting. The performance of the global feature is better than that of the local features because the global feature contains complete information, which results in better discrimination. Because the head, upper torso, lower torso, upper leg, lower leg, and feet are hard to recognize, the performance of the fusion strategy "head

**Table 7** Results of different feature fusion strategies on Market-1501

Training Strategy	Fusion Strategy	Rank1	mAP
No Weight Sharing	global	90.11	77.25
	upper body	85.15	68.01
	lower body	83.79	65.63
	whole body	90.54	77.48
	head + upper torso + lower torso + upper leg + lower leg + feet	82.4	63.38
	global + head + upper torso + lower torso + upper leg + lower leg + feet	91.33	75.51
	global + upper body	91.48	75.82
	global + lower body	91.26	79.62
	global + whole body	91.46	79.43
	Part3	92.16	81.07
Weight Sharing	head + upper torso + lower torso + upper leg + lower leg + feet	85.38	68.26
	global + head + upper torso + lower torso + upper leg + lower leg + feet	92.19	81.36
	global + upper body	92.56	81.62
	global + lower body	92.24	81.83
	global + whole body	92.97	82.35
	Part3	93.3	83.23

Note that the fusion strategy "head + upper torso + lower torso + upper leg + lower leg + feet + global + upper body + lower body + whole body" is denoted by "Part 3". "Global" denotes using only the global feature extraction branch and discarding the pose-guided feature alignment parts. "No weight sharing" means that all of the last convolution layers in Part 3 are trained separately, "Weight sharing" refers to sharing parameters between the last convolution layers (including the sub-branch for extracting global feature  $f_g^{P3}$ ). Please refer to the schematic in Fig. 7 for details about other descriptors, such as "upper body" and "lower body"



**Fig. 7** Schematic of part feature alignment with the same ID between a query image and a gallery image. The left side of the arrow denotes the parts of a query image; the right side of the arrow denotes the corresponding parts of the gallery image. The parts of the human body enclosed by the different-coloured bounding boxes are illustrated in boxes of the same colour on the left side. The same part features can be matched one by one using a part alignment network that also calculates the distance between different input images

+ upper torso + lower torso + upper leg + lower leg + feet” is worse than that of the fusion strategy “upper body”, “lower body” and “whole body”. However, these six parts can still focus on detailed information. The performance of “Part3”, which adds these parts, is the best and rises from Rank1/mAP=91.11%/77.25% to 93.3%/83.23%(+2.19%/5.98%). When comparing this result to “global” on Market-1501, it is clear that the head, upper torso, lower torso, upper leg, lower leg, and feet features are helpful for Re-ID because they are easy to align across images.

Therefore, we conduct experiments using “Part3” on the Market1501, DukeMTMC-reID, and CUHK03 datasets. The results are recorded in Table 8. In Table 9, we further compare “Part3” with other popular SOTA methods that aim to solve the alignment problem to prove its effectiveness. We can see that “Part3” demonstrates the best performance, which verifies its alignment effectiveness.

### 4.6 Effectiveness of MAAN

This section combines the global feature extraction branch, the multi-attention mechanism branch, and the pose-guided part feature alignment branch for Re-ID. The experiments are conducted on the Market-1501, DukeMTMC-reID, and CUHK03 datasets. “Part2+Part3” denotes that the network consists of only Part 2 and Part 3, and “MAAN” denotes that the network consists of all sub-branches, i.e., the complete method proposed in this paper. In addition, we include the results of “MAAN (rerank)”, which denotes using MAAN as well as the reranking tricks for Re-ID. The results are recorded in Table 10.

By reviewing the data in Table 10, it is clear that when the global feature, the pose-guided attention mechanism, and the part feature alignment network are combined, the Re-ID accuracy is increased in comparison with “Part2+Part3” as well as the results in Sections 4.3, 4.4, 4.5, demonstrating the effectiveness of our proposed MAAN method. Rank1 and the mAP for the “Part2 + Part3” method are 96.07% and 87.44%, respectively, on Market-1501, and the effect is good. In addition, the MAAN structure is more reasonable, with Rank1 and mAP values of 96.97% and 88.38%, respectively, on Market-1501. This result is obtained because different branches encourage each other or share valuable information. Hence, the MAAN effect is better than the “Part2 + Part3” method. When rearranging the test results, Rank1 is 97.16%, and mAP is 94.66% on the Market-1501 dataset. The method “MAAN (rerank)” uses the information from the test set itself. In a practical sense, the test set is likely to be open. Thus, there is no way to use the information from the test set, and this method therefore cannot be achieved, making the original Rank1 and mAP are more valuable. However, for some known test sets, rerank technology can be used to improve network accuracy. In our experiment, mAP is improved from 88.38% to 94.66% on Market-1501 using “rerank”, which is very significant.

**Table 8** Results of “Part3” method on the Market-1501, DukeMTMC-reID, and CUHK03 (detected) datasets

Dataset	Method	Rank1	Rank3	Rank5	Rank10	mAP
Market-1501	Part3	93.3	96.48	97.87	99.03	83.23
DukeMTMC-reID	Part3	86.63	91.82	93.98	95.82	73.82
CUHK03	Part3	74.11	84.07	89.29	92.51	71.34

**Table 9** Comparison with several alignment-oriented SOTA methods on the Market-1501, DukeMTMC-reID, and CUHK03 (detected) datasets

Method	Market-1501		DukeMTMC-reID		CUHK03(detected)	
	Rank1	mAP	Rank1	mAP	Rank1	mAP
Spindle [14]	76.9	-	-	-	-	-
SVDNet [1]	82.3	62.1	76.7	56.8	41.5	37.3
PDC [17]	84.1	63.4	-	-	-	-
GLAD [7]	89.9	73.9	-	-	-	-
Pose Transfer [48]	87.7	68.9	68.6	48.1	30.1	28.2
PN-GAN [49]	89.4	72.6	73.6	53.2	-	-
PSE [37]	87.7	69.0	79.8	62.0	30.2	27.3
MSCAN [50]	80.8	57.5	-	-	-	-
PIE [19]	78.7	53.9	-	-	-	-
Part-Aligned [51]	91.7	79.6	84.4	69.3	-	-
SPReID [38]	92.5	81.3	84.4	71.0	-	-
Part3(Ours)	<b>93.3</b>	<b>83.23</b>	<b>86.63</b>	<b>73.82</b>	<b>74.11</b>	<b>71.34</b>

To briefly summarize, the proposed MAAN learns the feature information from different perspectives. The global branch has a larger receptive field, which captures full and sparse features. The multi-attention branch focuses on the critical location of features while suppressing useless feature information such as background noise, and the pose-guided part alignment branch causes the model to compare the same semantic parts from different images, leading to alignment. What different branches have learned from different perspectives can complement each other. Under the same network settings, the synthesis MAAN method behaves much better than any single method.

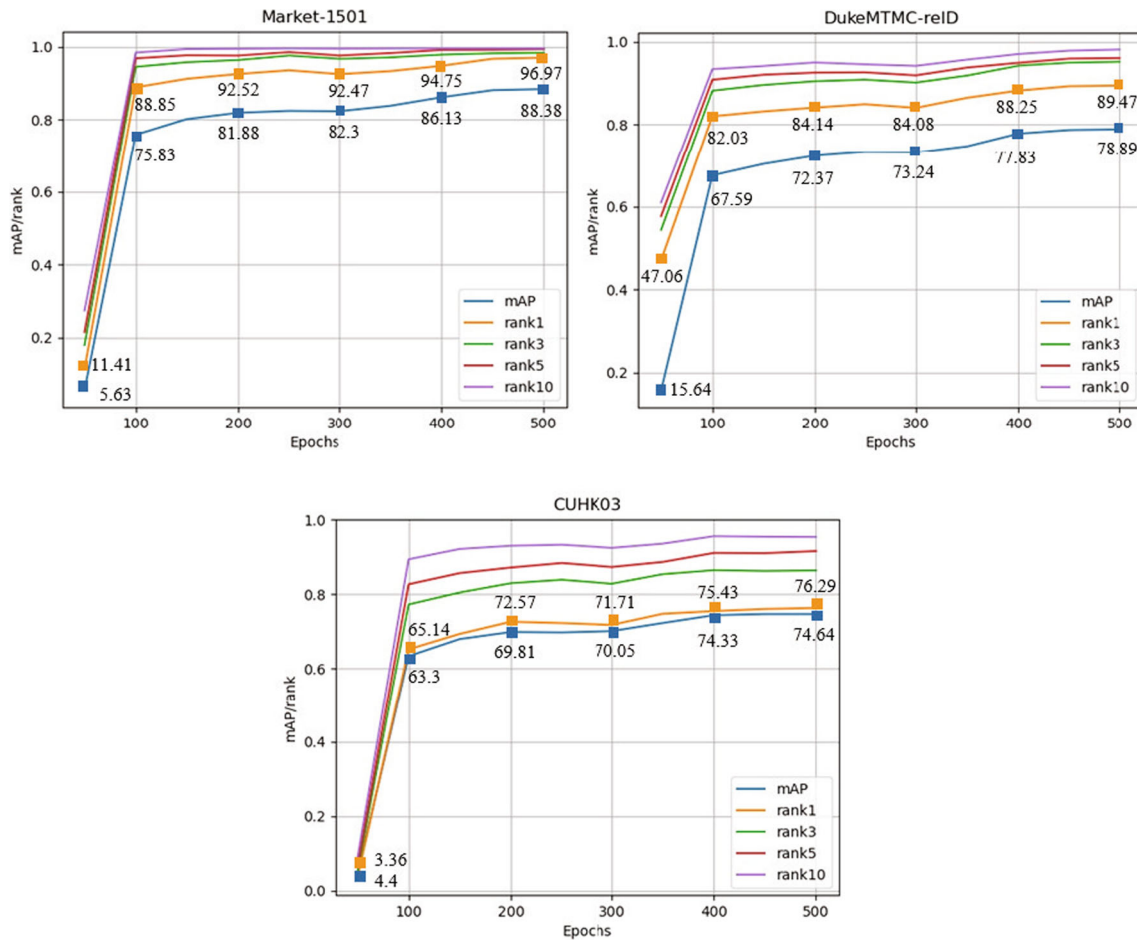
To analyse the experimental results, we draw the Cumulative Match Characteristic (CMC) curve for the proposed MAAN, as shown in Fig. 8. This experiment has a duration of 500 training epochs and tests every 50 epochs. The network is jointly trained by three branches, and every branch is responsible for different tasks, including global feature extraction, the multi-attention mechanism and part

feature extraction. Therefore, the branches have many different loss functions. Owing to the reasonable settings, the network is not difficult to train and can converge to a good effect after 100 epochs. Due to the richness of the feature extraction, the gradients of the losses can maintain their updates after 100 epochs, and the network can continue to converge better and obtain a better effect of feature extraction. From Fig. 8, it can be seen that mAP and Rankn are further improved at 350 epochs because the model jumps out of the local minimum under continuous training. The CMC curve gradually becomes stable between 400 and 500 epochs, and the best convergence effect is obtained at 500 epochs. After 500 epochs, the CMC curve no longer fluctuates. Therefore, the network's training potential is relatively large. By training for 500 epochs, the network's different losses and gradients can be better matched.

Figure 9 shows the retrieval results of four different query images. The first two query images are all identified accurately whether the pedestrians are sideways or shown

**Table 10** Results of the methods "Part2+Part3", "MAAN", and "MAAN (rerank)" on the Market-1501, DukeMTMC-reID, and CUHK03 (detected) datasets

Dataset	Method	Rank1	Rank3	Rank5	Rank10	mAP
Market-1501	Part2+Part3	96.07	97.67	98.94	99.5	87.44
	MAAN	96.97	98.31	99.29	99.5	88.38
	MAAN(rerank)	97.16	98.47	99.31	99.5	94.66
DukeMTMC-reID	Part2+Part3	88.52	94.26	95.92	97.74	78.19
	MAAN	89.47	95.22	96.1	98.17	78.89
	MAAN(rerank)	91.11	96.02	97.15	98.63	88.92
CUHK03	Part2+Part3	75.0	86.14	90.29	94.79	73.52
	MAAN	76.29	86.36	91.57	95.36	74.64
	MAAN(rerank)	77.64	87.93	92.26	95.64	85.62



**Fig. 8** CMC curve of MAAN on the Market-1501, DukeMTMC-reID and CUHK03 (detected) datasets. The abscissa represents epochs, and the ordinate represents mAP/Rank-n. Five curves with different colours in each image, from top to bottom, represent mAP, Rank1, Rank2, Rank3, Rank5, and Rank10

from the back. The first query pedestrian has his back to the camera, and his front face is found in the retrieval result. There is considerable background information above these gallery images corresponding to the first query image, which can still be identified accurately using the pose-guided part feature alignment branch of MAAN. The fifth and tenth vague gallery images corresponding to the second query pedestrian were also identified accurately, reflecting the robustness of MAAN. For the seventh gallery image with a red dotted box corresponding to the third query pedestrian, the person's clothes and looks are nearly identical to the correct one, which leads to errors. For the ninth gallery image with a red dotted bounding box corresponding to the fourth query image, the head information is missing due to an extreme detection error, while its other appearance is nearly the same as the query's, leading to a mistake.

#### 4.7 Cross-Domain Re-ID

This section conducts cross-domain Re-ID experiments on the "Part2+ Part3" and "MAAN" methods. We first use the model trained on Market1501 to test on DukeMTMC-reID. In addition, we also use the model trained on DukeMTMC-reID to test on Market-1501. The results are recorded in Table 11.

From the results of the cross-domain experiment, it is clear that the MAAN method can achieve competitive accuracy when compared with the state-of-the-art unsupervised domain adaptation (UDA) methods, even though it does not use any unlabelled source data. The MAAN performance on the Duke target dataset is better than most UDA methods in Table 11, and on target dataset Market-1501, MAAN's performance is better than SPGAN, TJ-AIDL, and ATNet, which verifies its generalization ability. Some of the com-



**Fig. 9** Top 10 ranking list from four query images from the Market-1501 dataset by MAAN. The retrieved images are all from in the gallery set. The images with green numbers belong to the same identity, and those with red numbers belong to a different identity

**Table 11** Cross-domain Re-ID results

Method	Source	Target	Rank1	Rank5	Rank10	mAP
MMFA [52]	Market+ Duke (U)	Duke	45.3	59.8	66.3	24.7
SPGAN [53]	Market+ Duke (U)	Duke	46.4	62.3	68.0	26.2
CamStyle [54]	Market+ Duke (U)	Duke	48.4	62.5	68.9	25.1
TJ-AIDL [55]	Market+ Duke (U)	Duke	44.3	59.6	65.0	23.0
ATNet [56]	Market+ Duke (U)	Duke	45.1	59.5	64.2	24.9
HHL [57]	Market+ Duke (U)	Duke	46.9	61.0	66.7	27.2
ECN [58]	Market+ Duke (U)	Duke	63.3	75.8	80.4	40.4
Part2+Part3	Market	Duke	45.47	60.0	65.17	25.31
MAAN(ours)	Market	Duke	46.16	61.31	66.40	25.85
MMFA [52]	Duke + Market (U)	Market	56.7	75.0	81.8	27.4
SPGAN [53]	Duke + Market (U)	Market	57.7	75.8	82.4	26.7
CamStyle [54]	Duke + Market (U)	Market	58.8	78.2	84.3	27.4
TJ-AIDL [55]	Duke + Market (U)	Market	58.2	74.8	81.1	26.5
ATNet [56]	Duke + Market (U)	Market	55.7	73.2	79.4	25.6
HHL [57]	Duke + Market (U)	Market	62.2	78.8	84.0	31.4
ECN [58]	Duke + Market (U)	Market	75.1	87.6	91.6	43.0
Part2+Part3	Duke	Market	57.06	75.3	81.96	26.02
MAAN(ours)	Duke	Market	57.72	75.96	82.51	26.73

"Market" denotes the Market-1501 dataset, and "Duke" denotes the DukeMTMC-reID dataset. The MAAN method (highlighted rows) does not use any unlabelled data from the target domain. (U: Unlabelled)



**Table 12** Time complexity, space complexity and top-1 accuracy (%) on Market-1501

Method	Params	MACs	Top-1
ResNet-50	25.56M	4.028G	87.6
Part1	24.23M	4.026G	90.56
Part2	26.19M	6.108G	92.67
Part3	30.69M	6.112G	93.3
MAAN	33.17M	8.856G	96.97

M: Million; G: Gillion

**Table 13** Comparison with state-of-the-art methods on the Market-1501 dataset

Type	Method	Rank1	mAP	Reference
Stripe-based	AlignedReID [15]	90.6	77.7	Arxiv2017
	Deep-Person [59]	92.3	79.6	PR2017
	SCPNet [60]	91.2	75.2	ACCV2018
	BFE [61]	94.5	85.0	Arxiv2018
	PCB+RPP [3]	93.8	81.6	ECCV2018
	MGN [62]	95.7	86.9	CVPR2018
Attention-based	HA-CNN [39]	91.2	75.7	CVPR2018
	Mancs [44]	93.1	82.3	ECCV2018
	PAN [45]	82.8	63.4	TCSVT2018
	PPA [46]	92.3	79.3	CVIU2021
	DuATM [47]	91.4	76.6	CVPR2018
	Attention-Driven [63]	95.0	86.5	PR2019
	ABD-Net [64]	95.6	88.3	ICCV2019
	PAT [29]	95.4	88.0	CVPR2021
Pose-guided	MSCAN [50]	80.8	57.5	CVPR17
	SVDNet [1]	82.3	62.1	ICCV2017
	PDC [17]	84.1	63.4	ICCV17
	Pose Transfer [48]	87.7	68.9	CVPR2018
	PN-GAN [49]	89.4	72.6	ECCV2018
	PSE [37]	87.7	69.0	CVPR2018
	Part-Aligned [51]	91.7	79.6	ECCV2018
	SPReID [38]	92.5	81.3	CVPR2018
	PGFA [69]	91.2	76.8	ICCV2019
	HOReID [28]	94.2	84.9	CVPR2020
Others	HAP2S [65]	84.59	69.43	ECCV2018
	PABR [66]	90.2	76	ECCV2018
	KPM [67]	90.1	75.3	CVPR2018
	VPM [68]	91.3	77.8	CVPR2019
	DistributionNet [70]	87.26	70.82	ICCV2019
	DRAL [71]	84.2	66.26	ICCV2019
	PL-Net [72]	88.2	69.3	TIP2019
	APR [73]	87.04	66.89	PR2019
	MuDeep (SL) [31]	95.34	84.66	TPAMI2019
	SMGAN [74]	92.13	76.48	Appl Intell2020
	DUNet [75]	91.6	75.9	Appl Intell2020
	Ensemble [76]	91.9	79.5	PR2020
	Multi-Scale & Multi-Patch [77]	93.7	81.2	Neurocomputing2020
	UnityStyle [78]	91.8	76.5	CVPR2020
	IRN+ARN [79]	92.8	79.5	Appl Intell2020
Ours	MAAN	<b>96.97</b>	<b>88.38</b>	

pared methods in the cross-domain experiment are better than the proposed MAAN method, such as ECN. These are used for domain adaptation based on UDA methods and not general architectures. The proposed method is a supervised and general architecture, not a network specially used for migration. In addition, the proposed network extracts detailed features based on three branches, so the fusion effect can achieve competitive migration performance. The proposed MAAN method can achieve better results in both the original domain and the cross domain, while the other methods are only better in the cross domain. Their accuracy in the original domain is not as good as our method.

Therefore, in terms of overall performance, the MAAN method is better than other methods and has a comparable cross-domain ability. Without the global feature branch, the performance of "Part2 + Part3" is slightly worse. In summary, our proposed MAAN can be used as a backbone model for cross-domain Re-ID.

#### 4.8 Complexity analysis

The proposed approach is based on a deep network with three branches, but the time and space complexity are not high. Notably, our feature extraction is based on

**Table 14** Comparison with state-of-the-art methods on the DukeMTMC-reID dataset

Type	Method	Rank1	mAP	Reference
Stripe-based	AlignedReID [15]	81.2	67.4	Arxiv2017
	Deep-Person [59]	80.9	64.8	PR2017
	SCPNet [60]	80.3	62.6	ACCV2018
	BFE [61]	88.7	75.8	Arxiv2018
	PCB+RPP[3]	83.3	69.2	ECCV2018
	MGN [62]	88.7	78.4	CVPR2018
Attention-based	HA-CNN [39]	80.5	63.8	CVPR2018
	Mancs [44]	84.9	71.8	ECCV2018
	PAN [45]	71.6	51.5	TCSVT2018
	PPA [46]	85.1	70.2	CVIU2021
	DuATM [47]	81.8	64.6	CVPR2018
	Attention-Driven [63]	86.0	74.6	PR2019
	ABD-Net [64]	89.0	78.6	ICCV2019
	PAT [29]	88.8	78.2	CVPR2021
Pose-guided	SVDNet[1]	76.7	56.8	ICCV2017
	Pose Transfer [48]	68.6	48.1	CVPR2018
	PN-GAN [49]	73.6	53.2	ECCV2018
	PSE [37]	79.8	62.0	CVPR2018
	Part-Aligned [51]	84.4	69.3	ECCV2018
	SPReID [38]	84.4	71.0	CVPR2018
	PGFA [69]	82.6	65.5	ICCV2019
	HOReID [28]	86.9	75.6	CVPR2020
	Others	HAP2S [65]	75.94	60.64
KPM [67]		80.3	63.2	CVPR2018
DistributionNet [70]		74.73	55.98	ICCV2019
DRAL [71]		74.28	56	ICCV2019
APR [73]		73.92	55.56	PR2019
MuDeep (SL) [31]		88.19	75.63	TPAMI2019
SMGAN [74]		84.82	64.89	Appl Intell2020
DUNet [75]		82.1	66.5	Appl Intell2020
Ensemble [76]		84.8	70.3	PR2020
Multi-Scale & Multi-Patch [77]		84.4	70.4	Neurocomputing2020
UnityStyle [78]		82.1	65.2	CVPR2020
IRN+ARN [79]		82.7	66.4	Appl Intell2020
Ours		MAAN	<b>89.47</b>	<b>78.89</b>

**Table 15** Comparison with state-of-the-art methods on the CUHK03(detected) dataset

Type	Method	Rank1	mAP	Reference
Stripe-based	PCB+RPP [3]	63.7	57.5	ECCV2018
	MGN [62]	66.8	66.0	CVPR2018
Attention-based	HA-CNN [39]	41.7	38.6	CVPR2018
	Mancs [44]	65.5	60.5	ECCV2018
	PAN [45]	36.3	34.0	TCSVT2018
	PPA [46]	64.6	62.0	CVIU2021
Pose-guided	SVDNet [1]	41.5	37.3	ICCV2017
	Pose Transfer [48]	30.1	28.2	CVPR2018
	PSE [37]	30.2	27.3	CVPR2018
Others	MultiScale [80]	40.7	37.0	ICCV2017
	MLFN [82]	52.8	47.8	CVPR2018
	MuDeep (SL) [31]	71.93	67.21	TPAMI2019
	TriNet+Era [81]	55.5	50.7	AAAI2020
	SMGAN [74]	50.16	47.12	Appl Intell2020
	DUNet [75]	51.6	49.9	Appl Intell2020
	Ensemble [76]	62.9	57.6	PR2020
	Multi-Scale & Multi-Patch [77]	70.1	67.2	Neurocomputing2020
Ours	MAAN	<b>76.29</b>	<b>74.64</b>	

the backbone (ResNet-50), and only extra attention and region divisions are used in subsequent branches to further complete feature extraction. The function of the multi-branch network is to perform post-processing without introducing too many complex operations. Specifically, Part 1 extracts the global feature. For Part 2, the parameters used for the channel and spatial attention are small, and the keypoint attention does not require parameters that do not occupy space. For Part 3, the part feature is calculated based on keypoints, so the amount of calculation is small, and no parameters are introduced. After every branch, GAP does not introduce parameters, and the parameters for Conv1×1 are small. Although the network has multiple branches, we introduce fewer parameters and computations. As shown in Table 12, Macs and Params denote the time complexity. Compared with ResNet-50, the network does not substantially increase the computational burden, and the accuracy is greatly improved. Therefore, the computational efficiency is excellent. This design minimizes the number of network parameters and makes it easy to train.

#### 4.9 Comparison with state-of-the-art methods

We compare the MAAN method with several state-of-the-art methods on the Market-1501, DukeMTMC-reID, and CUHK03 datasets to verify its performance. These methods are divided into four classes: stripe-based, attention-based, pose-guided, and others. MAAN belongs to the Pose-guided class. The detailed results for different datasets are provided in Tables 13, 14, and 15.

From Tables 13, 14, and 15, we find that MAAN's Rank1 and mAP are 96.97% and 88.38% on Market-1501, 89.47% and 78.89% on DukeMTMC-reID, and 76.29% and 74.64% on CUHK03, respectively, which are much higher than other methods such as MGN. MGN is a powerful Re-ID method based on multi-branch and horizontal feature partitioning, and it is the nearest competitor on the Market-1501 and CUHK03 datasets. MAAN's Rank1 and mAP surpass MGN by 1.27% and 1.48% on Market-1501, and 9.49% and 8.64% on CUHK03(detected). In addition, ABD-Net is the nearest competitor on the DukeMTMC-reID dataset. MAAN's Rank1 and mAP surpass ABD-Net by 0.47% and 0.29% on DukeMTMC-reID. Comparing the above results demonstrates that MAAN has a comparative performance to state-of-the-art Re-ID methods on several commonly used person Re-ID benchmarks.

The proposed MAAN can achieve excellent performance because the global feature extraction branch, the multi-attention mechanism branch, and the pose-guided part feature alignment branch mine more valuable and more substantial information from complete perspectives, thus broadly enhancing the discrimination and robustness of the final feature representation.

## 5 Conclusion

In this paper, we propose a robust person re-identification algorithm called MAAN. By integrating different functional branches, MAAN can complete mining of both the global

features and valuable local features while performing part feature alignment. Specifically, MAAN extracts complete and detailed identity information using global feature extraction and a multi-attention mechanism. MAAN also suppresses the misalignment caused by pose changes and camera viewpoint changes using pose-guided part feature partitioning. Extensive experiments have shown that our method outperforms several state-of-the-art methods on three mainstream Re-ID benchmarks and has good discrimination, robustness, and generalization ability. In the future, we will also extend our idea and network to other intelligent computing, such as 3D mesh simplification with feature preservation [83], social representation learning [84], training detection networks [85] and deep residual convolutional dehazing networks [86].

**Acknowledgements** This work is supported by the National Natural Science Foundation of China (61573114). We also gratefully acknowledge the support of the College of Intelligent Systems Science and Engineering, Harbin Engineering University.

#### Compliance with Ethical Standards

**Conflict of Interests** The authors declare that they have no conflict of interest.

## References

- Sun Y, Zheng L, Deng W, Wang S (2017) Svdnet for pedestrian retrieval. In: Proceedings of the IEEE International Conference on Computer Vision, pp 3800–3808
- Hermans A, Beyer L, Leibe B (2017) In defense of the triplet loss for person re-identification. arXiv:1703.07737
- Sun Y, Zheng L, Yang Y, Tian Q, Wang S (2018) Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European conference on computer vision (ECCV), pp 480–496
- Fu Y, Wei Y, Zhou Y, Shi H, Huang G, Wang X, Huang T (2019, July) Horizontal pyramid matching for person re-identification. Proc AAAI Conf Artif Intell 33(01):8295–8302
- Liu X, Zhao H, Tian M, Sheng L, Shao J, Yi S, Wang X (2017) Hydraplus-net: Attentive deep features for pedestrian analysis. In: Proceedings of the IEEE international conference on computer vision, pp 350–359
- Zhao L, Li X, Zhuang Y, Wang J (2017) Deeply-learned part-aligned representations for person re-identification. In: Proceedings of the IEEE international conference on computer vision, pp 3219–3228
- Li W, Zhao R, Xiao T, Wang X (2014) Deepreid: Deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 152–159
- Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2015) Scalable person re-identification: a benchmark. In: Proceedings of the IEEE international conference on computer vision, pp 1116–1124
- Ahmed E, Jones M, Marks TK (2015) An improved deep learning architecture for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3908–3916
- Liao S, Hu Y, Zhu X, Li S (2015) Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2197–2206
- Dong SC, Cristani M, Stoppa M, Bazzani L, Murino V (2011) Custom pictorial structures for re-identification. In: British Machine Vision Conference, vol 6
- Cheng DS, Cristani M (2014) Person re-identification by articulated appearance matching. In: Person Re-Identification. Springer, London, pp. 139–160
- Wei L, Zhang S, Yao H, Gao W, Tian Q (2018) GLAD: Global–Local-alignment descriptor for scalable person re-identification. IEEE Transactions on Multimedia 21(4):986–999
- Zhao H, Tian M, Sun S, Shao J, Yan J, Yi S, Tang X (2017) Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1077–1085
- Zhang X, Luo H, Fan X, Xiang W, Sun Y, Xiao Q, Sun J (2017) Alignedreid: Surpassing human-level performance in person re-identification. arXiv:1711.08184
- Xu J, Zhao R, Zhu F, Wang H, Ouyang W (2018) Attention-aware compositional network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2119–2128
- Su C, Li J, Zhang S, Xing J, Gao W, Tian Q (2017) Pose-driven deep convolutional model for person re-identification. In: Proceedings of the IEEE international conference on computer vision, pp 3960–3969
- Zhao L, Li X, Zhuang Y, Wang J (2017) Deeply-learned part-aligned representations for person re-identification. In: Proceedings of the IEEE international conference on computer vision, pp 3219–3228
- Zheng L, Huang Y, Lu H, Yang Y (2019) Pose-invariant embedding for deep person re-identification. IEEE Trans Image Process 28(9):4500–4509
- Ristani E, Solera F, Zou R, Cucchiara R, Tomasi C (2016) Performance measures and a data set for multi-target, multi-camera tracking. In: European conference on computer vision. Springer, Cham, pp 17–35
- Zheng Z, Zheng L, Yang Y (2017) Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: Proceedings of the IEEE International Conference on Computer Vision, pp 3754–3762
- Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7132–7141
- Li X, Wang W, Hu X, Yang J (2019) Selective kernel networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 510–519
- Woo S, Park J, Lee JY, Kweon IS (2018) Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV), pp 3–19
- Wei SE, Ramakrishna V, Kanade T, Sheikh Y (2016) Convolutional pose machines. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp 4724–4732
- Cao Z, Simon T, Wei SE, Sheikh Y (2017) Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7291–7299
- Sun K, Xiao B, Liu D, Wang J (2019) Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 5693–5703
- Wang G, Yang S, Liu H, Wang Z, Yang Y, Wang S, Yu G, Zhou E, Sun J (2020) High-Order Information Matters: Learning

- Relation and Topology for Occluded Person Re-Identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
29. Li Y, He J, Zhang T, Liu X, Zhang Y, Wu F (2021) Diverse Part Discovery: Occluded Person Re-Identification With Part-Aware Transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 2898–2907
  30. Yang Y, Yang J, Yan J, Liao S, Yi D, Li S (2014) Salient color names for person re-identification. In European conference on computer vision. Springer, Cham, pp 536–551
  31. Qian X, Fu Y, Xiang T, Jiang YG, Xue X (2019) Leader-based multi-scale attention deep architecture for person re-identification. *IEEE Trans Pattern Anal Mach Intell* 99(1):1–1
  32. Zheng WS, Gong S, Xiang T (2012) Reidentification by relative distance comparison. *IEEE Trans Pattern Anal Mach Intell* 35(3):653–668
  33. Koestinger M, Hirzer M, Wohlhart P, Roth PM, Bischof H (2012) Large scale metric learning from equivalence constraints. In 2012 IEEE conference on computer vision and pattern recognition. IEEE, 2288–2295
  34. Liao S, Li S (2015) Efficient psd constrained asymmetric metric learning for person re-identification. In: Proceedings of the IEEE international conference on computer vision, pp 3685–3693
  35. Liu M, Wang K, Ji R, Ge S, Chen J (2021) Person image generation with attention-based injection network. *Neurocomputing* 460:345–359
  36. Xu J, Zhao R, Zhu F, Wang H, Ouyang W (2018) Attention-aware compositional network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2119–2128
  37. Sarfraz MS, Schumann A, Eberle A, Stiefelhagen R (2018) A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 420–429
  38. Kalayeh MM, Basaran E, Gökmen M., Kamasak ME, Shah M (2018) Human semantic parsing for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1062–1071
  39. Li W, Zhu X, Gong S (2018) Harmonious attention network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2285–2294
  40. Zheng Z, Zheng L, Yang Y (2017) A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing. Commun Appl (TOMM)* 14(1):1–20
  41. Lin TY, Maire M, Belongie S, Hays J, Zitnick CL (2014) Microsoft COCO: Common Objects in Context. European Conference on Computer Vision. Springer International Publishing
  42. Ristani E, Tomasi C (2018) Features for multi-target multi-camera tracking and re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6036–6046
  43. Zhao L, Li X, Zhuang Y, Wang J (2017) Deeply-learned part-aligned representations for person re-identification. In: Proceedings of the IEEE international conference on computer vision, pp 3219–3228
  44. Wang C, Zhang Q, Huang C, Liu W, Wang X (2018) Mancs: a multi-task attentional network with curriculum sampling for person re-identification. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 365–381
  45. Zheng Z, Zheng L, Yang Y (2018) Pedestrian alignment network for large-scale person re-identification. *IEEE Trans Circ Syst Video Technol* 29(10):3037–3045
  46. Li Z, Lv J, Chen Y, Yuan J (2021) Person re-identification with part prediction alignment. *Comput Vis Image Underst* 205(103172)
  47. Si J, Zhang H, Li CG, Kuen J, Kong X, Kot AC, Wang G (2018) Dual attention matching network for context-aware feature sequence based person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 5363–5372
  48. Liu J, Ni B, Yan Y, Zhou P, Cheng S, Hu J (2018) Pose transferrable person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4099–4108
  49. Qian X, Fu Y, Xiang T, Wang W, Qiu J, Wu Y, Xue X (2018) Pose-normalized image generation for person re-identification. In: Proceedings of the European conference on computer vision (ECCV), pp 650–667
  50. Li D, Chen X, Zhang Z, Huang K (2017) Learning deep context-aware features over body and latent parts for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 384–393
  51. Suh Y, Wang J, Tang S, Mei T, Lee KM (2018) Part-aligned bilinear representations for person re-identification. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 402–419
  52. Shan L, Li H, Li CT, Kot AC (2018) Multi-task Mid-level Feature Alignment Network for Unsupervised Cross-Dataset Person Re-Identification *BMVC* 2018
  53. Deng W, Zheng L, Ye Q, Kang G, Yang Y, Jiao J (2018) Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 994–1003
  54. Zhong Z, Zheng L, Zheng Z, Li S, Yang Y (2018) Camstyle: a novel data augmentation method for person re-identification. *IEEE Trans Image Process* 28(3):1176–1190
  55. Wang J, Zhu X, Gong S, Li W (2018) Transferable joint attribute-identity deep learning for unsupervised person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2275–2284
  56. Liu J, Zha ZJ, Chen D, Hong R, Wang M (2019) Adaptive transfer network for cross-domain person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 7202–7211
  57. Zhong Z, Zheng L, Li S, Yang Y (2018) Generalizing a person retrieval model hetero-and homogeneously. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 172–188
  58. Zhong Z, Zheng L, Luo Z, Li S, Yang Y (2019) Invariance matters: Exemplar memory for domain adaptive person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, p p 598–607
  59. Bai X, Yang M, Huang T, Dou Z, Yu R, Xu Y (2020) Deep-person: Learning discriminative deep features for person re-identification. *Pattern Recogn* 98(107036)
  60. Fan X, Luo H, Zhang X, He L, Zhang C, Jiang W (2018) Scpnet: Spatial-channel parallelism network for joint holistic and partial person re-identification. In: Asian conference on computer vision. Springer, Cham, pp 19–34
  61. Dai Z, Chen M, Gu X, Zhu S, Tan P (2019) Batch dropout network for person re-identification and beyond. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 3691–3701
  62. Wang G, Yuan Y, Chen X, Li J, Zhou X (2018) Learning discriminative features with multiple granularities for person re-identification. In: Proceedings of the 26th ACM international conference on Multimedia, pp 274–282

63. Yang F, Yan K, Lu S, Jia H, Xie X, Gao W (2019) Attention driven person re-identification, vol 86
64. Chen T, Ding S, Xie J, Yuan Y, Chen W, Yang Y, Wang Z (2019) Abd-net: Attentive but diverse person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 8351–8361
65. Yu R, Dou Z, Bai S, Zhang Z, Xu Y, Bai X (2018) Hard-aware point-to-set deep metric for person re-identification. In: Proceedings of the European conference on computer vision (ECCV), pp 188–204
66. Suh Y, Wang J, Tang S, Mei T, Lee KM (2018) Part-aligned bilinear representations for person re-identification. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 402–419
67. Shen Y, Xiao T, Li H, Yi S, Wang X (2018) End-to-end deep kronecker-product matching for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6886–6895
68. Sun Y, Xu Q, Li Y, Zhang C, Li Y, Wang S, Sun J (2019) Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 393–402
69. Miao J, Wu Y, Liu P, Ding Y, Yang Y (2019) Pose-guided feature alignment for occluded person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 542–551
70. Yu T, Li D, Yang Y, Hospedales TM, Xiang T (2019) Robust person re-identification by modelling feature uncertainty. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 552–561
71. Liu Z, Wang J, Gong S, Lu H, Tao D (2019) Deep reinforcement active learning for human-in-the-loop person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 6122–6131
72. Yao H, Zhang S, Hong R, Zhang Y, Xu C, Tian Q (2019) Deep representation learning with part loss for person re-identification. *IEEE Trans Image Process* 28(6):2860–2871
73. Lin Y, Zheng L, Zheng Z, Wu Y, Hu Z, Yan C, Yang Y (2019) Improving person re-identification by attribute and identity learning. *Pattern Recogn* 95:151–161
74. Liu M, Yan X, Wang C, Wang K (2021) Segmentation mask-guided person image generation. *Appl Intell* 51(2):1161–1176
75. Li R, Zhang B, Teng Z, Fan J (2021) A divide-and-union deep network for person re-identification. *Appl Intell* 51(3):1479–1491
76. Serbetci A, Akgul YS (2020) End-to-end training of CNN ensembles for person re-identification. *Pattern Recogn* 104 (107319)
77. Wang C, Song L, Wang G, Zhang Q, Wang X (2020) Multi-scale multi-patch person re-identification with exclusivity regularized softmax. *Neurocomputing* 382:64–70
78. Liu C, Chang X, Shen YD (2020) Unity style transfer for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 6887–6896
79. Yin J, Fan Z, Chen S, Wang Y (2020) In-depth exploration of attribute information for person re-identification. *Appl Intell* 50(11):3607–3622
80. Chen Y, Zhu X, Gong S (2017) Person re-identification by deep learning multi-scale representations. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp 2590–2600
81. Zhong Z, Zheng L, Kang G, Li S, Yang Y (2020) Random erasing data augmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence 34(07):13001–13008
82. Chang X, Hospedales TM, Xiang T (2018) Multi-level factorisation net for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2109–2118
83. Liang Y, He F, Zeng X (2020) 3d mesh simplification with feature preservation based on whale optimization algorithm and differential evolution. *Integr Comput Aided Eng* 27(4):417–435
84. Pan Y, He F, Yu H (2020) Learning social representations with deep autoencoder for recommender system. *World Wide Web* 23(4):2259–2279
85. Quan Q, He F, Li H (2020) A multi-phase blending method with incremental intensity for training detection networks. *Vis Comput* 37(2):245–259
86. Zhang S, He F (2020) DRCDN: Learning deep residual convolutional dehazing networks. *Vis Comput* 36(9):1797–1808
87. Huang H, Yang W, Chen X, Zhao X, Huang K, Lin J, Huang G, Du D (2018) Eanet: Enhancing alignment for cross-domain person re-identification. [arXiv:1812.11369](https://arxiv.org/abs/1812.11369)

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