Diverse Part Discovery: Occluded Person Re-identification with Part-Aware Transformer

Yulin Li[1*], Jianfeng He[1*], Tianzhu Zhang[1†], Xiang Liu[2], Yongdong Zhang[1], Feng Wu[1]
[1] University of Science and Technology of China [2] Dongguan University of Technology
{liyulin, hejf}@mail.ustc.edu.cn {tzzhang, fengwu, zhyd73}@ustc.edu.cn
succeedpkmba2011@163.com

Abstract

Occluded person re-identification (Re-ID) is a challenging task as persons are frequently occluded by various obstacles or other persons, especially in the crowd scenario. To address these issues, we propose a novel end-to-end Part-Aware Transformer (PAT) for occluded person Re-ID through diverse part discovery via a transformer encoder-decoder architecture, including a pixel context based transformer encoder and a part prototype based transformer decoder. The proposed PAT model enjoys several merits. First, to the best of our knowledge, this is the first work to exploit the transformer encoder-decoder architecture for occluded person Re-ID in a unified deep model. Second, to learn part prototypes well with only identity labels, we design two effective mechanisms including part diversity and part discriminability. Consequently, we can achieve diverse part discovery for occluded person Re-ID in a weakly supervised manner. Extensive experimental results on six challenging benchmarks for three tasks (occluded, partial and holistic Re-ID) demonstrate that our proposed PAT performs favorably against stat-of-the-art methods.

1. Introduction

Person re-identification (Re-ID) aims to match images of a person captured from non-overlapping camera views [7, 44, 54]. It is one of the most important research topics in the computer vision field with various applications, such as video surveillance, autonomous driving, and activity analysis [47, 23, 56, 49, 48]. Recently, person Re-ID has drawn a growing amount of interest from academia and industry, and various methods have been proposed [20, 24, 14, 38, 18, 27, 42, 43]. Most of these approaches assume that the entire body of the pedestrian is available for the Re-ID model designing. However, when conducting person Re-ID in real-world scenarios, e.g., airports, railway stations, hospitals, and malls, it is difficult to satisfy this assumption due to the inevitable occlusions. For example, a person may be occluded by some obstacles (e.g., cars, trees, walls, and other persons), and the camera fails to capture the holistic person. Therefore, it is essential to design an effective model to solve this occluded person Re-ID problem [61, 30].

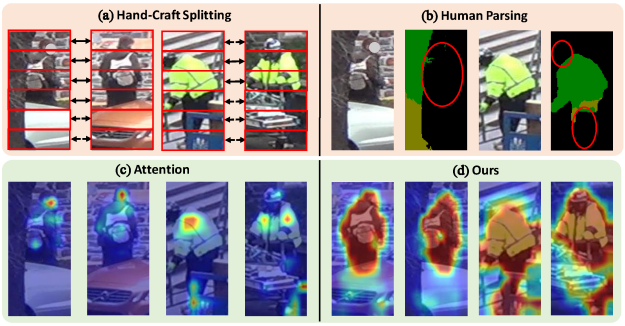

Figure 1. Examples of part-based methods for occluded person Re-ID. (a) The hand-crafted splitting based methods require strict person alignment. (b) The extra semantic based methods exclude the personal belongings and are error-prone when a person is seriously occluded. (c) The attention-based methods tend to mainly focus on the most discriminative region. (d) The attention maps produced by our proposed PAT by fusing all part-aware masks. The discriminative parts are highlighted.

In the occluded person Re-ID task, occluded regions usually contain some noise that results in mismatching, and the key issue is how to learn discriminative features from unoccluded regions. Recently, leveraging local features extracted from human body parts to improve representations of the pedestrian has been the mainstream for robust feature learning of the occluded Re-ID task. Generally, these part-based occluded Re-ID methods can be divided into three main categories. (1) The hand-crafted splitting based methods divide the image or feature map into small patches [10, 57, 12] or rigid stripes [38, 4] and then extract part features from the local patches or stripes. However, hand-crafted splitting is too coarse to align the human parts well and introduces lots of background noise. (2) The

*Equal contribution
†Corresponding author

extra semantic based methods [13, 30, 40, 5, 11] directly utilize human parsing [13, 11] or pose estimation models [30, 40, 5] as part localization modules to achieve more accurate human part localization. However, their success heavily relies on the accuracy of the off-the-shelf human parsing or pose estimation models. Since there exist differences between training datasets of human parsing/pose estimation and person Re-ID, the off-the-shelf human parsing/pose estimation models are error-prone when pedestrians are seriously occluded. (3) The attention based methods [37, 62] exploit attention mechanisms to localize discriminative human parts. Typically, the predicted attention maps distribute most of the attention weights on human parts, which can help decrease the negative effect of cluttered background. To sum up, most existing occluded Re-ID methods focus on locating discriminative human parts and leveraging local part features to develop powerful representations of the pedestrian.

Based on the above discussions, the part-based representations have been proven to be effective for the occluded Re-ID problem. To capture accurate human parts, an intuitive idea is to detect non-occluded body parts using body part detectors and then match the corresponding body parts. However, there are no extra annotations for the body detector learning. Thus, we propose to localize discriminative human parts only with identity labels. To achieve this goal, there are two main challenges as follows. On the one hand, background with diverse characteristics, such as colors, sizes, shapes, and positions, increase the difficulty of getting robust features for the target person. Intuitively, the appearance of pixels of the same human part region is similar, while quite different from the background pixels. Therefore, it is necessary to model the correlation between pixels for robust feature representation. On the other hand, as shown in Figure 1, the occluded parts vary between different pedestrian images. As there are no groundtruth annotations for human parts, it is difficult to cope with diverse appearance of pedestrians and adaptively locate all unoccluded parts only with the identity labels. As a result, as shown in Figure 1 (c), most of the attention based methods tend to put the main focus on the most discriminative region. They always ignore other human parts including personal belongings, e.g., backpack and reticule, which also provide important clues for person Re-ID.

To deal with the above issues, we propose a novel Part-Aware Transformer (PAT) for occluded person Re-ID through diverse part discovery via a transformer encoder-decoder architecture [39, 2], including a pixel context based transformer encoder and a part prototype based transformer decoder. In the **pixel context based transformer encoder**, we adopt a self-attention mechanism to capture the full image context information. Specifically, we model the correlation of pixels of the feature map and aggregate pixels with similar appearances. In this way, we can obtain the pixel context aware feature map, which is more robust to background clutters. In the **part prototype based transformer decoder**, we introduce a set of learnable part prototypes to generate part-aware masks focusing on discriminative human parts. In specific, given the feature map of a pedestrian, we take the learnable part prototypes as queries and pixels of the feature map as keys and values of the transformer decoder. We can obtain part-aware masks by calculating the similarity between all pixels in the feature map and part prototypes. Each part-aware mask is expected to denote the spatial distribution of one specific human part, e.g., head or body part. With part-aware masks, human part features can be further obtained from the values by a weighted pooling. However, without the assistance of part annotations, it is challenging to constraint these part prototypes to capture accurate human parts. Thus, to guide part prototype learning, we propose two mechanisms including part diversity and part discriminability. Intuitively, different part features of the same pedestrian should focus on different human parts. Therefore, the part diversity mechanism is adopted to encourage lower correlation between part features and make part prototypes focus on different discriminative foreground regions. The part discriminability mechanism is to make part features maintain identity discriminative via part classification and a triplet loss. By optimizing the transformer encoder and decoder jointly, part prototypes can be learned through the whole dataset. Consequently, we can achieve robust human part discovery for occluded person Re-ID in a weakly supervised manner.

The contributions of our method could be summarized into three-fold: (1) We propose a novel end-to-end Part-Aware Transformer for occluded person Re-ID through diverse part discovery via a transformer encoder-decoder architecture, including a pixel context based transformer encoder and a part prototype based transformer decoder. To the best of our knowledge, our PAT is the first work by exploiting the transformer encoder-decoder architecture for occluded person Re-ID in a unified deep model. (2) To learn part prototypes only with identity labels well, we design two effective mechanisms, including part diversity and part discriminability. Consequently, we can achieve robust human part discovery for occluded person Re-ID in a weakly supervised manner. (3) To demonstrate the effectiveness of our method, we perform experiments on three tasks, including occluded Re-ID, partial Re-ID and holistic Re-ID on six standard Re-ID datasets. Extensive experimental results demonstrate that the proposed method performs favorably against state-of-the-art methods.

2. Related Work

In this section, we briefly overview methods that are related to holistic person Re-ID, partial Re-ID and occluded

person Re-ID respectively.

Holistic Person Re-Identification. Person re-identification (Re-ID) aims to match images of a person captured from non-overlapping camera views [7, 44, 54]. Existing Re-ID methods can be summarized to hand-crafted descriptors [47, 23], metric learning methods [56, 20, 24] and deep learning methods[38, 25, 33, 35, 45, 52, 19, 21, 34, 26, 22]. Recent works utilizing part-based features have achieved state-of-the-art performance for the holistic person Re-ID task. Kalayeh *et al*. [19] extract several region parts with human parsing methods and assemble final discriminative representations with part-level features. Sun *et al*. [38] uniformly partition the feature map and learn part-level features by multiple classifiers. Zhao *et al*. [51] and Liu *et al*. [26] extract part-level features by attention-based methods. But all these Re-ID methods focus on matching holistic person images with the assumption that the entire body of the pedestrian is available. Different from these methods, our model can adaptively capture discriminative human part features via a transformer encoder-decoder architecture for the occluded person Re-ID task.

Partial Person Re-Identification. Partial person Re-ID aims to match partial probe images to holistic gallery images. Zheng *et al*. [57] propose a local-level matching model called Ambiguity-sensitive Matching Classifier (AMC) based on the dictionary learning and introduce a local-to-global matching model called Sliding Window Matching to provide complementary spatial layout information. He *et al*. [10] propose an alignment-free approach namely Deep Spatial feature Reconsruction (DSR) that exploits the reconstruction error based on sparse coding. Luo et al.*et al*. [29] proposed STNReID that combines a spatial transformer network (STN) and a Re-ID network for partial Re-ID. Sun *et al*. [37] introduce a Visibility-aware Part Model (VPM) to perceive the visibility of part regions through self-supervision. However, all these methods need a manual crop of the occluded target person in the probe image and then use the non-occluded parts as the new query. The manual cropping is not efficient in practice and might introduce human bias to the cropped results.

Occluded Person Re-Identification. Given occluded probe images, occluded person Re-ID aims to find the same person with holistic or occluded appearance in disjoint cameras. This task is more challenging due to incomplete information and spatial misalignment. Zhuo *et al*. [61] combine the occluded/unoccluded classification task and person ID classification task to extract key information from images. He *et al*. [13] reconstruct the feature map of unoccluded regions and propose a spatial background-foreground classifier to avoid the influence of background clutters. Besides, the Pose-Guided Feature Aligment (FGFA) [30] utilizes pose landmarks to mine discriminative parts to address the occlusion noise. Gao *et al*. [5] propose a Pose-guided Visible Part Matching (PVPM) model to learn discriminative part features with pose-guided attentions. Wang *et al*. [40] exploit graph convolutional layers to learn high-order human part relations for robust alignment. Although the above methods can solve the occlusion problem to some extent, most of them heavily rely on off-the-shelf human parsing models or pose estimators. Different from them, our model can exploit diverse parts with only identity labels in a weakly supervised manner via a transformer encoder-decoder architecture.

3. Part-Aware Transformer

In this section, we introduce the proposed Part-Aware Transformer (PAT) in detail. As shown in Figure 2, the proposed PAT mainly consists of two modules, including the pixel context based transformer encoder and the part prototype based transformer decoder. Here we give a brief introduction to the full process. First, we obtain the feature map of each pedestrian image through a CNN backbone. Then we flatten the feature map and carry out the self-attention operation to obtain the pixel context aware feature map with the transformer encoder. After obtaining the pixel context aware feature map, we calculate the similarity between the feature map and a set of learnable part prototypes to obtain part-aware masks. Part features can be further obtained by a weighted pooling where part-aware masks are treated as different spatial attention maps. Finally, we introduce the part diversity mechanism and part discriminability mechanism to learn part prototypes well with only identity labels.

3.1. Pixel Context based Transformer Encoder

Background regions with diverse characteristics increase the difficulty of getting robust features for the target person. Therefore, we adopt a self-attention mechanism to capture the full image context information. In this way, we can obtain the pixel context aware feature map, which is more robust to background clutters. Following [38], our method uses ResNet-50 [9] without the average pooling layer and fully connected layer as the backbone to extract global feature maps from given images. We also set the stride of conv4_1 to 1 to increase the feature resolution as in [38]. As a result, an input image with a size of $H \times W$ will get the feature map with the spatial dimension of $H/16 \times W/16$, which is larger than that of the original ResNet-50. A larger feature map has been proved to be effective in person Re-ID. Formally, the feature map extracted from the backbone is denoted as $\mathbf{Z} \in \mathbb{R}^{h \times w \times c}$, where h, w, c are the height, width and channel of the global feature map, respectively.

We first utilize a 1×1 convolution to reduce the channel dimension of the feature map $\mathbf{Z}$ to a smaller dimension d, creating a new feature map $\mathbf{F} \in \mathbb{R}^{h \times w \times d}$. The transformer encoder requires a 1D sequence as input. To handle 2D feature maps, we flatten the spatial dimensions of $\mathbf{F}$ into one dimension, resulting in a $hw \times d$ feature. In

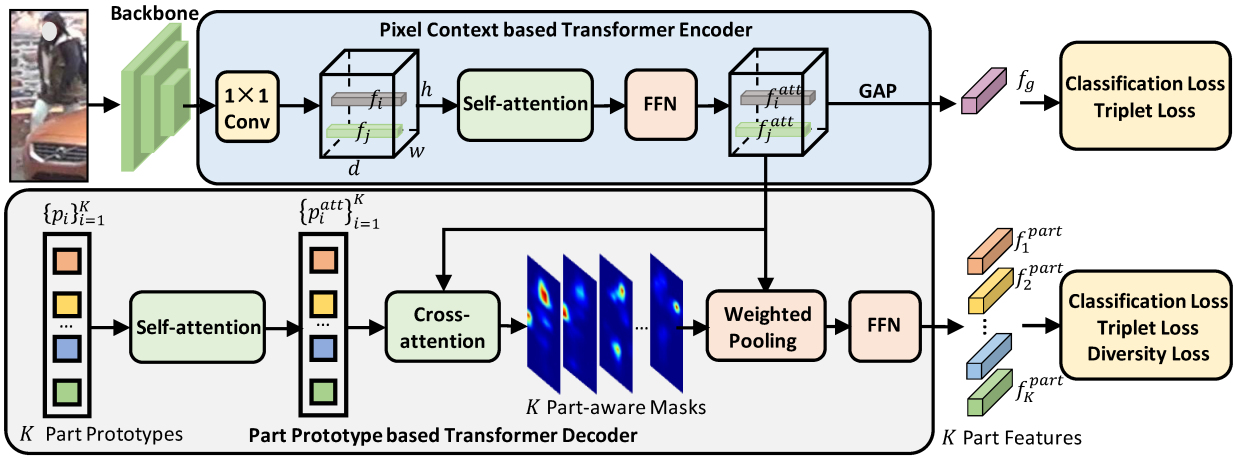

Figure 2. The pipeline of the proposed PAT consists of a pixel context based transformer encoder and a part prototype based transformer decoder. Here, "self-attention" denotes the self-attention layer, "cross-attention" denotes the cross-attention layer, and "FFN" denotes the feed forward layer. For more details, please refer to the text and please see supplemental materials for detailed architecture.

the self-attention mechanism, given the feature map $\mathbf{F} = [f_1; f_2; \ldots; f_{hw}]$ ($f_i \in \mathbb{R}^{1\times d}$ indicates the feature of the i^{th} spatial position), both keys, queries and values arise from pixels of the feature map. Formally,

$$\mathbf{Q}_i = f_i\mathbf{W}^Q, \quad \mathbf{K}_j = f_j\mathbf{W}^K, \quad \mathbf{V}_j = f_j\mathbf{W}^V, \tag{1}$$

where $i, j \in 1, 2, \ldots, hw$ and $\mathbf{W}^Q \in \mathbb{R}^{d\times d_k}, \mathbf{W}^K \in \mathbb{R}^{d\times d_k}, \mathbf{W}^V \in \mathbb{R}^{d\times d_v}$ are linear projections. For the i^{th} query $\mathbf{Q}_i$, the attention weights are calculated based on the dot-product similarity between each query and the keys:

$$s_{i,j} = \frac{\exp(\beta_{i,j})}{\sum_{j=1}^{hw}\exp(\beta_{i,j})}, \beta_{i,j} = \frac{\mathbf{Q}_i\mathbf{K}_j^T}{\sqrt{d_k}}, \tag{2}$$

where $\sqrt{d_k}$ is a scaling factor. The output of the self-attention mechanism is defined as weighted sum over all values according to the attention weights:

$$\hat{f}_i^{att} = \mathrm{Att}(\mathbf{Q}_i, \mathbf{K}, \mathbf{V}) = \sum_{j=1}^{hw} s_{i,j}\mathbf{V}_j, \tag{3}$$

The normalized attention weight $s_{i,j}$ models the interdependency between different spatial pixels f_i and f_j, and the weight sum of the values can aggregate these semantically related spatial pixels to update f_i. Since pixels of the same human part have high similarities while are different from background pixels, the feature map capturing the full-image context information would be more robust to background clutters. We implement Eq.(3) with the multi-head attention mechanism and can get $\hat{f}_i^{att} \in \mathbb{R}^{1\times d}$. The updated feature map can be obtained by aggregating pixel context information of all positions:

$$\hat{\mathbf{F}}^{att} = \left[\hat{f}_1^{att}; \hat{f}_2^{att}; \ldots; \hat{f}_{hw}^{att}\right] \in \mathbb{R}^{hw\times d}, \tag{4}$$

where $\hat{\mathbf{F}}^{att}$ represents the updated feature map with the self-attention mechanism. Following the standard transformer architecture, we use the feed-forward network to produce the final pixel context aware feature map as defined in:

$$\mathbf{F}^{att} = \mathrm{FFN}(\hat{\mathbf{F}}^{att}), \tag{5}$$

where $\mathrm{FFN}(\cdot)$ is a simple neural network using two fully connected layers [39]. The residual connections followed by the layer normalization [1] are also applied. Please see supplemental materials for more details about the architecture. Through the self-attention operation, the pixel context aware feature map $\mathbf{F}^{att} = [f_1^{att}; f_2^{att}; \ldots; f_{hw}^{att}] \in \mathbb{R}^{hw\times d}$ can be obtained, which is more robust to background clutter.

Encoder Training Loss. To make the pixel context aware feature focus on ID-related discriminative information and train our encoder, we use an identity classification loss and a triplet loss as the objective function. First of all, we utilize a global average pooling operation $f^g = GAP(\mathbf{F}^{att})$ and constrain the global feature $f^g \in \mathbb{R}^{1\times d}$ to satisfy the objective function. The objective function is formulated as:

$$\begin{aligned}\mathcal{L}_{En} &= \lambda_{cls}\mathcal{L}_{cls}(f^g) + \lambda_{tri}\mathcal{L}_{tri}(f^g) \\ &= -\lambda_{cls}\log p_g + \lambda_{tri}\left[\alpha + d_{f^g,f_p^g} - d_{f^g,f_n^g}\right]_+ .\end{aligned} \tag{6}$$

Where p_g is the probability predicted by a classifier, α is the margin, d_{f^g,f_p^g} is the distance between a positive pair (f^g, f_p^g) from the same identity, and (f^g, f_n^g) is the negative pair from different identities.

3.2. Part Prototype based Transformer Decoder

In the part prototype based transformer decoder, to localize discriminative human parts only with identity labels, we introduce a set of learnable part prototypes focusing on discriminative human parts and propose two mechanisms, including part diversity and part discriminability to guide part

prototype learning only with identity labes. In this way, we can achieve robust human part discovery in a weakly supervised manner. First of all, we introduce a set of part prototypes $\mathcal{P}_K = \{p_i\}_{i=1}^K, p_i \in \mathbb{R}^{1\times d}$ represents a part classifier that determines whether pixels of the feature map $\mathbf{F}^{att}$ belong to the part i. These part prototypes are set as learnable parameters.

Self-attention Layer. Following the standard architecture of the transformer, we first use a self-attention mechanism to further incorporate the local context of human parts to part prototypes. This process allows the local context information propagation between prototypes during part prototype learning. The implementation is the same as in Section 3.1, and both keys, queries and values arise from part prototypes. We can obtain the updated part prototype set $\{p_i^{att}\}_{i=1}^K$. The weights of self-attention encode the relations between part prototypes p_i and p_j. The updated part prototypes incorporate the local context of different parts.

Cross-attention Layer. The cross-attention layer aims to extract foreground part features from the feature map $\mathbf{F}^{att}$ with the learnable part prototypes. As shown in Figure 2, in the cross-attention layer, given the feature map $\mathbf{F}^{att} = [f_1^{att}; f_2^{att}; \dots; f_{hw}^{att}]$, queries arise from part prototypes $\{p_i^{att}\}_{i=1}^K$, keys and values arise from pixels of the feature map. Formally,

$$\mathbf{Q}_i = p_i^{att}\mathbf{W}^Q,\ \mathbf{K}_j = f_j^{att}\mathbf{W}^K,\ \mathbf{V}_j = f_j^{att}\mathbf{W}^V, \quad (7)$$

where $i \in 1, 2, \dots, K$, $j \in 1, 2, \dots, hw$, and $\mathbf{W}^Q \in \mathbb{R}^{d\times d_k}, \mathbf{W}^K \in \mathbb{R}^{d\times d_k}, \mathbf{W}^V \in \mathbb{R}^{d\times d_v}$ are linear projections. Note that they are different from Eq.(3). For each part prototype p_i^{att}, we illustrate how to compute the part-aware mask and the corresponding part feature. Formally,

$$m_{i,j} = \frac{\exp(\beta_{i,j})}{\sum_{j=1}^{hw}\exp(\beta_{i,j})}, \beta_{i,j} = \frac{\mathbf{Q}_i\mathbf{K}_j^T}{\sqrt{d_k}}, \quad (8)$$

where $\sqrt{d_k}$ is a scaling factor. The attention weight $m_{i,j}$ indicates the probability of the spatial feature f_j^{att} belonging to the foreground part i. The attention weights of all hw positions make up a part-aware mask $\mathbf{M}_i = [m_{i,1}; m_{i,2}; \dots; m_{i,hw}]$, which has high response values at pixels belonging to the part i. We can further obtain i^{th} part feature by a weighted pooling, which is defined as the weighted sum over all values:

$$\hat{f}_i^{part} = \mathrm{Att}(\mathbf{Q}_i, \mathbf{K}, \mathbf{V}) = \sum_{j=1}^{hw} m_{i,j}\mathbf{V}_j, \quad (9)$$

By computing over all part prototypes, we can obtain K part-aware masks (each mask is a $h \times w$ attention map) and further obtain K part features, as shown in Figure 2. We implement Eq.(9) with the multi-head attention mechanism and can get $\hat{f}_i^{part} \in \mathbb{R}^{1\times d}$. Then, two fully-connected layers are adopted, which is the same as the standard transformer architecture. The final part feature is formulated as:

$$f_i^{part} = \mathrm{FFN}(\hat{f}_i^{part}), \quad (10)$$

where $i \in 1, 2, \dots, K$ and $\mathrm{FFN}(\cdot)$ denotes the feed-forward network as in Eq.(5).

Since there are no human part annotations, part prototype learning tends to focus on the same discriminative part (e.g., the body), which may result in a suboptimal solution. Thus, to learn part prototypes only with identity labes, we propose two mechanisms including part diversity and part discriminability. (1) The part diversity mechanism is to make part prototypes focus on different discriminative foreground parts. A diversity loss is imposed to expand the discrepancy among different part features $\{f_i^{part}\}_{i=1}^K$:

$$\mathcal{L}_{div} = \frac{1}{K(K-1)}\sum_{i=1,j=1}^{K}\sum_{i\neq j}^{K}\frac{\langle f_i^{part}, f_j^{part}\rangle}{\|f_i^{part}\|_2\|f_j^{part}\|_2}, \quad (11)$$

The intuition behind this loss is obvious. If the i^{th} and the j^{th} prototypes give a high attention weight to the same foreground part, the $\mathcal{L}_{div}$ will be large and prompt these prototypes to adjust themselves adaptively. (2) The part discriminability mechanism is to make part features maintain identity discriminative. The part classification and triplet loss are employed to guide part feature representation learning as in Eq.(12), where the definitions of $\mathcal{L}_{cls}(\cdot)$ and $\mathcal{L}_{tri}(\cdot)$ can be found in Eq.(6).

$$\mathcal{L}_{dis} = \lambda_{cls}\sum_{i=1}^{K}\mathcal{L}_{cls}\left(f_i^{part}\right) + \lambda_{tri}\sum_{i=1}^{K}\mathcal{L}_{tri}\left(f_i^{part}\right). \quad (12)$$

In the triplet loss, part features f_i^{part} from different identities form negative pairs, and those from the same identity form positive pairs. As a result, the features obtained from the same prototype with different identities are pushed away and the identity discriminative part features can be obtained.

3.3. Training and Inference

For the occluded person Re-ID task, our proposed PAT is trained by minimizing the overall objective with identity labels as defined in Eq.(13).

$$\mathcal{L}_{PAT} = \mathcal{L}_{En} + \mathcal{L}_{div} + \mathcal{L}_{dis}, \quad (13)$$

During the testing stage, for each image of an unseen identity, we concatenate the global feature f^g and part features $\{f_i^{part}\}_{i=1}^K$ as its representation:

$$v = \left[f^g, f_1^{part}, \cdots, f_K^{part}\right]. \quad (14)$$

where $[\cdot]$ denotes a concatenation operation.

4. Experiments

In this section, we first verify the effectiveness of our proposed model for occluded person Re-ID, partial Re-ID, and holistic Re-ID. Then, we report a set of ablation studies to validate the effectiveness of each component. Finally, we provide more visualization results.

4.1. Datastes and Evaluation Metrics

To demonstrate the effectiveness of our method, we conduct extensive experiments on two occluded datasets: Occluded-Duke [30] and Occluded REID [30], two partial Re-ID datasets: Partial-REID [57] and Partial-iLIDS [55], and two holistic Re-ID datasets: Market-1501 [53] and DukeMTMC-reID [32, 58]. The details are as follows.

Occluded-Duke [30] contains 15,618 training images, 17,661 gallery images, and 2,210 occulded query images. It is selected from DukeMTMC-reID by leaving occluded images and filtering out some overlap images.

Occluded-REID [61] is an occluded person dataset captured by mobile cameras, including 2,000 images belonging to 200 identities. Each identity has five full-body person images and five occluded person images with different viewpoints and different types of severe occlusions.

Partial-REID [57] is a specially designed partial person Re-ID benchmark that includes 600 images from 60 people, with five full-body images in gallery set and five partial images in query set per person.

Partial-iLIDS [10] is a partial person Re-ID dataset based on the iLIDS dataset [55], and contains a total of 238 images from 119 people captured by multiple cameras in the airport, and their occluded regions are manually cropped.

Market-1501 [53] consists of 1,501 identities captured by 6 cameras. The training set consists of 12,936 images of 751 identities, the query set consists of 3,368 images, and the gallery set consists of 19,732 images.

DukeMTMC-reID [32, 58] contains 36,411 images of 1,404 identities captured by 8 cameras. The training set contains 16,522 images, the query set consists of 2,228 images and the gallery set consists of 17,661 images.

Evaluation Metrics. We adpot standard metrics as in most person Re-ID literature, namely Cumulative Matching Characteristic (CMC) curves and mean average precision (mAP), to evaluate the quality of different Re-ID models.

4.2. Implementation Details

We adopt ResNet-50 [9] pretrained on ImageNet as our backbone by removing the global average pooling (GAP) layer and fully connected layer. For classifiers, as in [28] we use a batch normalization layer [17] and a fully connected layer followed by a softmax function. The number of part prototypes K is set to 6 on Market-1501, and set to 14 on all other datasets. The images are resized to 256×128 and augmented with random horizontal flipping, padding 10 pixels, random cropping, and random erasing [59]. Extra color jitter is adopted on occluded-REID and partial datasets to avoid domain variance. The batch size is set to 64 with 4 images per person. During the training stage, all the modules are jointly trained for 120 epochs. The learning rate is initialized to 3.5×10^{-4} and decayed to its 0.1 and 0.01 at 40 and 70 epochs.

Table 1. Performance comparison with state-of-the-arts on Occluded-Duke and Occluded-REID. Our method achieves the best performance on the two occluded datasets.

Methods	Occluded-Duke		Occluded-REID	
	Rank-1	mAP	Rank-1	mAP
Part-Aligned [51]	28.8	20.2	-	-
PCB [38]	42.6	33.7	41.3	38.9
Part Bilinear [36]	36.9	-	-	-
FD-GAN [6]	40.8	-	-	-
AMC+SWM [57]	-	-	31.2	27.3
DSR [10]	40.8	30.4	72.8	62.8
SFR [12]	42.3	32	-	-
Ad-Occluded [16]	44.5	32.2	-	-
FPR [13]	-	-	78.3	68.0
PVPM [5]	47	37.7	70.4	61.2
PGFA [30]	51.4	37.3	-	-
GASM [11]	-		74.5	65.6
HOReID [40]	55.1	43.8	80.3	70.2
ISP [60]	62.8	52.3	-	-
PAT(*Ours*)	**64.5**	**53.6**	**81.6**	**72.1**

4.3. Comparison with State-of-the-art Methods

Results on Occluded Re-ID Datasets. Table 1 shows the performance of our model and previous methods on two occluded datasets. Four kinds of methods are compared, which are hand-crafted splitting based Re-ID methods [51, 38], holistic Re-ID methds with key-point information [36, 6], partial ReID methods [57, 10, 12] and occluded ReID methods [13, 5, 30, 11, 40, 60]. The Rank-1/mAP of our method achieves 64.5%/53.6% and 81.6%/72.1% on Occluded-Duke and Occluded-REID datasets, which set a new SOTA performance. Compared to the hand-crafted splitting based method PCB [38], our PAT surpasses it by $+21.9\%$ Rank-1 accuracy and $+19.9\%$ mAP on the Occluded-Duke dataset. This is because our PAT explicitly learns part-aware masks to depress the noisy information from the occluded regions. It can be seen that hand-crafted splitting based Re-ID methods and holistic methods with key-points information have similar performance on two occluded datasets. For example, PCB [38] and FD-GAN [6] both achieve about 40% Rank-1 score on the Occluded-Duke dataset, indicating that key-points information may not significantly benefit the occluded Re-ID task. Compared with PVPM and HOReID, which are SOTA occluded ReID methods with key-points information, our method achieves much better performance, surpassing them by at least $+9.4\%$ Rank-1 accuracy and $+9.8\%$ mAP on the

Table 2. Performance comparison with state-of-the-arts on Partial-REID and Partial-iLIDS datasets. Our method achieves the best.

Methods	Partial-REID		Partial-iLIDS	
	Rank-1	Rank-3	Rank-1	Rank-3
AMC+SWM [57]	37.3	46.0	21.0	32.8
DSR [10]	50.7	70.0	58.8	67.2
SFR [12]	56.9	78.5	63.9	74.8
STNReID [29]	66.7	80.3	54.6	71.3
VPM [37]	67.7	81.9	65.5	74.8
PGFA [30]	68.0	80.0	69.1	80.9
AFPB [61]	78.5	-	-	-
PVPM [5]	78.3	87.7	-	-
FPR [13]	81.0	-	68.1	-
HOReID [40]	85.3	91.0	72.6	86.4
PAT(*Ours*)	**88.0**	**92.3**	**76.5**	**88.2**

Occluded-Duke dataset. This is because their performance heavily relies on the accuracy of the off-the-shelf pose estimation models, while our method can capture more accurate human part information in a unified deep model. Furthermore, our PAT also outperforms the methods with the mask learning strategy, including GASM and ISP, which shows the effectiveness of our transformer encoder-decoder architecture and two learning mechanisms.

Results on Partial Datasets. To further evaluate our method, we compare the results on Partial-REID and Partial-iLIDS datasets with existing state-of-the-art methods. Like some previous methods [37, 13, 5, 40], since the two partial datasets are too small, we train our model on the Market-1501 training set and use two partial datasets as test sets. Therefore, it is also a cross-domain setting. As shown in Table 2, the Rank-1/Rank-3 of our method achieves 88.0%/92.3% and 76.5%/88.2% on Partial-REID and Partial-iLIDS datasets, respectively, which outperforms all the previous partial person Re-ID models. This suggests that the proposed PAT can be solid to address the occlusion problem. Compared to the most competing method HOReID [40], our PAT significantly surpasses it by $+2.7\%$ Rank-1 accuracy on Partial-REID, while surpasses it by $+3.9\%$ Rank-1 accuracy on Partial-iLIDS, which demonstrates the effectiveness of our proposed model.

Results on Holistic Re-ID Datasets. We also experiment on holistic person Re-ID datasets including Market-1501 and DukeMTMC-reID. We compare our method with state-of-the-art approaches of three categories, and the results are shown in Table 3. The methods in the first group are hand-crafted splitting based models. The methods in the second group are attention based approaches. The methods in the third group are extra semantic based methods. From the results, we can see that the proposed PAT achieves competitive performances with state-of-the-art on both datasets. Specifically, the Rank-1/mAP of our method achieves 95.4%/88.0% and 88.8%/78.2% on Market-1501 and DukeMTMC-reID datasets, respectively. Our PAT performs better than the hand-crafted splitting based model

Table 3. Performance comparison with state-of-the-art methods on Market-1501 and DukeMTMC-reID datasets.

Methods	Market-1501		DukeMTMC-reID	
	Rank-1	mAP	Rank-1	mAP
PCB [38]	92.3	77.4	81.8	66.1
BOT [28]	94.1	85.7	86.4	76.4
MGN [41]	95.7	86.9	88.7	78.4
VPM [37]	93.0	80.8	83.6	72.6
IANet [15]	94.4	83.1	87.1	73.4
CASN+PCB [31]	94.4	82.8	87.7	73.7
CAMA [46]	94.7	84.5	85.8	72.9
MHN-6 [3]	95.1	85.0	89.1	77.2
SPReID [19]	92.5	81.3	84.4	71.0
DSA-reID [50]	95.7	87.6	86.2	74.3
P^2 Net [8]	95.2	85.6	86.5	73.1
PGFA [30]	91.2	76.8	82.6	65.5
HOReID [40]	94.2	84.9	86.9	75.6
FPR [13]	95.4	86.6	88.6	78.4
PAT(*Ours*)	95.4	88.0	88.8	78.2

PCB, because the hand-crafted splitting is too coarse to align the human parts well. Furthermore, the proposed PAT is superior to those approaches with external cues. Specifically, compared to the pose-guided occluded Re-ID method HOReID [40], our PAT significantly surpasses it by $+3.1\%$ mAP on Market-1501, while surpasses it by $+2.6\%$ mAP on DukeMTMC-reID, which shows the effectiveness of the proposed part prototype learning mechanism. The extra semantic based approaches heavily rely on the external cues for person alignment, but they cannot always infer the accurate external cues in the case of severe occlusion. The above results also prove that the learnable part prototypes are robust to different views, poses, and occlusions.

4.4. Ablation Studies

In this section, we perform ablation studies on the Occluded-Duke dataset to analyze each component of our PAT, including the pixel context based transformer encoder ($\mathcal{P}$), the self-attention layer ($\mathcal{S}$) and the cross-attention layer ($\mathcal{C}$) of the part prototype based transformer decoder and the part diversity mechanism ($\mathcal{D}$). Note that the part discriminability mechanism is to make part features maintain identity discriminative, and it is the basis of our model. We reomve all the modules and set the ResNet-50 with the average pooling as our baseline, where only a global feature is available. The results are shown in Table 4.

Effectiveness of the Transformer Encoder. As shown in index-2, compared with the baseline model, when only the encoder is adopted and only the global feature f^g is used, the performance is improved by $+7.1\%$ mAP. This is because the self-attention mechanism of the encoder can capture the pixel context information well. From index-3 and index-5, we can also see that with the encoder, the performance is improved by $+0.8\%$ mAP since the pixel context aware feature is more robust to background clutters.

Table 4. Performance comparison with different components.

Index	$\mathcal{P}$	$\mathcal{S}$	$\mathcal{C}$	$\mathcal{D}$	R-1	R-5	R-10	mAP
1					46.0	65.7	71.9	38.8
2	✓				58.2	75.1	80.4	45.9
3		✓	✓		61.9	76.9	81.8	51.7
4			✓		59.9	75.6	81.2	50.8
5	✓	✓	✓		63.1	77.5	82.1	52.5
6	✓	✓	✓	✓	**64.5**	**78.3**	**83.4**	**53.6**

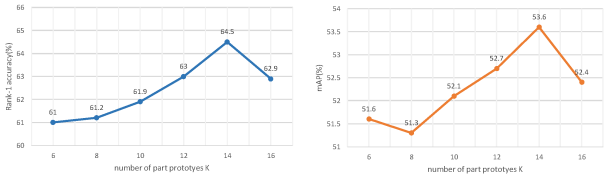

Figure 3. Comparison in Rank-1 accuracy and mAP with different settings of the part prototype number K on Occluded-Duke.

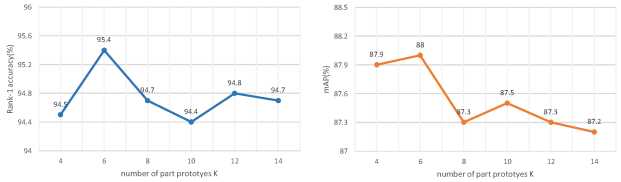

Figure 4. Comparison in Rank-1 accuracy and mAP with different settings of the part prototype number K on Market-1501.

Effectiveness of the Transformer Decoder. From index-1 and index-3, when the part prototype based decoder is added, the performance is greatly improved by $+12.9\%$ and up to 51.7% mAP. This shows that the part-aware masks obtained from part prototypes are useful for reducing the influence of background and aligning part features. From index-3 and index-4, when the self-attention layer in the decoder is added, the performance is further improved by $+0.9\%$ mAP. This demonstrates the effectiveness of the local context information propagation among all prototypes.

Effectiveness of the Part Diversity Mechanism. From index-5 and index-6, we can see that our full model achieves the best performance, which demonstrates the effectiveness of the proposed diversity loss. By adding the diversity loss, the learnable part prototypes are guided to discover diverse discriminative human parts for the occluded Re-ID task.

Analysis of the Number of Part Prototypes. The number of part prototypes K determines the granularity of the discovered parts. We perform quantitative experiments to clearly find the most suitable K on Occluded-Duke and Market-1501 datasets. As shown in Figure 3, with K increases, the performance keeps improving before K arrives 14 on Occluded-Duke, while the best performance is achieved when K is set to 6 on Market-1501, as shown in Figure 4. We conclude that this is because the scenarios in Occluded-Duke are more complex, and more fine-grained part features would be more useful.

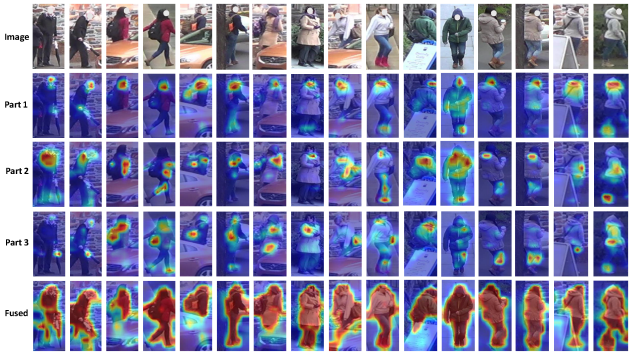

Figure 5. Visualization of the learned part-aware masks. The final mask is obtained by fusing all the part-aware masks. As we can see, these part-aware masks mainly focus on different discriminative human parts, including personal belongings.

4.5. Visualization of Discovered Parts

We visualize the part-aware masks generated from different part prototypes in Figure 5. From the results, we can observe that different part-aware masks can successfully capture diverse discriminative human parts for the same input image. For example, the part-aware mask obtained by the 1^{th} prototype mainly focuses on the head region, and the part-aware mask obtained by the 2^{th} prototype mainly focuses on the upper body. This also shows the effectiveness of our proposed part diversity mechanism. The final mask in Figure 5 is obtained by fusing all part-aware masks together. We can see that the fused masks almost span over the whole person rather than overfit in some local regions. In this way, these part-aware masks can reduce the background interference and occlusion, making the network more focus on discriminative human parts for the occluded Re-ID task.

5. Conclusion

In this work, we propose a novel Part-Aware Transformer to discover diverse discriminative human parts with a set of learnable part prototypes for occluded person Re-ID. To learn part prototypes only with identity labels well, we design two effective mechanisms, including part diversity and part discriminability, to discovery human parts in a weakly supervised manner. Extensive experimental results for three tasks on six standard Re-ID datasets demonstrate the effectiveness of the proposed method.

6. Acknowledgment

This work was partially supported by the National Key Research and Development Program under Grant No. 2018YFB0804204, Strategic Priority Research Program of Chinese Academy of Sciences (No.XDC02050500), National Defense Basic Scientific Research Program (JCKY2020903B002), National Nature Science Foundation of China (Grant 62022078, 62021001, 62071122), Open Project Program of the National Laboratory of Pattern Recognition (NLPR) under Grant 202000019, and Youth Innovation Promotion Association CAS 2018166.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020.

[3] Binghui Chen, Weihong Deng, and Jiani Hu. Mixed high-order attention network for person re-identification. In *ICCV*, pages 371–381, 2019.

[4] Xing Fan, Hao Luo, Xuan Zhang, Lingxiao He, Chi Zhang, and Wei Jiang. Scpnet: Spatial-channel parallelism network for joint holistic and partial person re-identification. In *Asian Conference on Computer Vision*, pages 19–34. Springer, 2018.

[5] Shang Gao, Jingya Wang, Huchuan Lu, and Zimo Liu. Pose-guided visible part matching for occluded person reid. In *CVPR*, pages 11744–11752, 2020.

[6] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, et al. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *NeurIPS*, pages 1222–1233, 2018.

[7] Shaogang Gong and Tao Xiang. Person re-identification. In *Visual Analysis of Behaviour*, pages 301–313. Springer, 2011.

[8] Jianyuan Guo, Yuhui Yuan, Lang Huang, Chao Zhang, Jin-Ge Yao, and Kai Han. Beyond human parts: Dual part-aligned representations for person re-identification. In *ICCV*, pages 3642–3651, 2019.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[10] Lingxiao He, Jian Liang, Haiqing Li, and Zhenan Sun. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *CVPR*, pages 7073–7082, 2018.

[11] Lingxiao He and Wu Liu. Guided saliency feature learning for person re-identification in crowded scenes. In *ECCV*, pages 357–373, 2020.

[12] Lingxiao He, Zhenan Sun, Yuhao Zhu, and Yunbo Wang. Recognizing partial biometric patterns. *arXiv preprint arXiv:1810.07399*, 2018.

[13] Lingxiao He, Yinggang Wang, Wu Liu, He Zhao, Zhenan Sun, and Jiashi Feng. Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification. In *ICCV*, pages 8450–8459, 2019.

[14] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[15] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Interaction-and-aggregation network for person re-identification. In *CVPR*, pages 9317–9326, 2019.

[16] Houjing Huang, Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. Adversarially occluded samples for person re-identification. In *CVPR*, pages 5098–5107, 2018.

[17] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[18] Kongzhu Jiang, Tianzhu Zhang, Yongdong Zhang, Feng Wu, and Yong Rui. Self-supervised agent learning for unsupervised cross-domain person re-identification. *IEEE Transactions on Image Processing*, 29:8549–8560, 2020.

[19] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *CVPR*, pages 1062–1071, 2018.

[20] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, pages 2288–2295, 2012.

[21] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, pages 2285–2294, 2018.

[22] Yaoyu Li, Tianzhu Zhang, Lingyu Duan, and Changsheng Xu. A unified generative adversarial framework for image generation and person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 163–172, 2018.

[23] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, pages 2197–2206, 2015.

[24] Shengcai Liao and Stan Z Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *ICCV*, pages 3685–3693, 2015.

[25] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *CVPR*, pages 4099–4108, 2018.

[26] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *ICCV*, pages 350–359, 2017.

[27] Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu. Cross-modality person re-identification with shared-specific feature transfer. In *CVPR*, pages 13379–13389, 2020.

[28] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPR Workshops*, 2019.

[29] Hao Luo, Wei Jiang, Xing Fan, and Chi Zhang. Stnreid: Deep convolutional networks with pairwise spatial transformer networks for partial person re-identification. *IEEE Transactions on Multimedia*, 2020.

[30] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *ICCV*, pages 542–551, 2019.

[31] Xuelin Qian, Yanwei Fu, Tao Xiang, Yu-Gang Jiang, and Xiangyang Xue. Leader-based multi-scale attention deep architecture for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):371–385, 2019.

[32] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for

multi-target, multi-camera tracking. In *ECCV*, pages 17–35, 2016.

[33] M Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *CVPR*, pages 420–429, 2018.

[34] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *CVPR*, pages 1179–1188, 2018.

[35] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *ICCV*, pages 3960–3969, 2017.

[36] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *ECCV*, pages 402–419, 2018.

[37] Yifan Sun, Qin Xu, Yali Li, Chi Zhang, Yikang Li, Shengjin Wang, and Jian Sun. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In *CVPR*, pages 393–402, 2019.

[38] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, pages 480–496, 2018.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[40] Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. High-order information matters: Learning relation and topology for occluded person re-identification. In *CVPR*, pages 6449–6458, 2020.

[41] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 274–282, 2018.

[42] Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In *ICCV*, pages 3623–3632, 2019.

[43] Guan'an Wang, Tianzhu Zhang, Yang Yang, Jian Cheng, Jianlong Chang, Xu Liang, and Zeng-Guang Hou. Cross-modality paired-images generation for rgb-infrared person re-identification. In *AAAI*, pages 12144–12151, 2020.

[44] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*, pages 1–16, 2014.

[45] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. In *CVPR*, pages 2119–2128, 2018.

[46] Wenjie Yang, Houjing Huang, Zhang Zhang, Xiaotang Chen, Kaiqi Huang, and Shu Zhang. Towards rich feature discovery with class activation maps augmentation for person re-identification. In *CVPR*, pages 1389–1398, 2019.

[47] Yang Yang, Jimei Yang, Junjie Yan, Shengcai Liao, Dong Yi, and Stan Z Li. Salient color names for person re-identification. In *ECCV*, pages 536–551, 2014.

[48] Tianzhu Zhang, Changsheng Xu, and Ming-Hsuan Yang. Learning multi-task correlation particle filters for visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):365–378, 2019.

[49] Tianzhu Zhang, Changsheng Xu, and Ming-Hsuan Yang. Robust structural sparse tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):473–486, 2019.

[50] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Densely semantically aligned person re-identification. In *CVPR*, pages 667–676, 2019.

[51] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, pages 3219–3228, 2017.

[52] Liang Zheng, Yujia Huang, Huchuan Lu, and Yi Yang. Pose-invariant embedding for deep person re-identification. *IEEE Transactions on Image Processing*, 28(9):4500–4509, 2019.

[53] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015.

[54] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.

[55] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by probabilistic relative distance comparison. In *ICCV*, pages 649–656, 2011.

[56] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Reidentification by relative distance comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):653–668, 2012.

[57] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shaogang Gong. Partial person re-identification. In *ICCV*, pages 4678–4686, 2015.

[58] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, pages 3754–3762, 2017.

[59] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, pages 13001–13008, 2020.

[60] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. Identity-guided human semantic parsing for person re-identification. In *ECCV*, 2020.

[61] Jiaxuan Zhuo, Zeyu Chen, Jianhuang Lai, and Guangcong Wang. Occluded person re-identification. In *ICME*, 2018.

[62] Jiaxuan Zhuo, Jianhuang Lai, and Peijia Chen. A novel teacher-student learning framework for occluded person re-identification. *arXiv preprint arXiv:1907.03253*, 2019.