Pose-guided Inter- and Intra-part Relational Transformer for Occluded Person Re-Identification

Zhongxing Ma[1], Yifan Zhao[1], Jia Li[1,2*]
[1]State Key Laboratory of Virtual Reality Technology and Systems, SCSE, Beihang University, Beijing, China
[2]Peng Cheng Laboratory, Shenzhen, China

ABSTRACT

Person Re-Identification (Re-Id) in occlusion scenarios is a challenging problem because a pedestrian can be partially occluded. The use of local information for feature extraction and matching is still necessary. Therefore, we propose a Pose-guided inter- and intra-part relational transformer (Pirt) for occluded person Re-Id, which builds part-aware long-term correlations by introducing transformer. In our framework, we firstly develop a pose-guided feature extraction module with regional grouping and mask construction for robust feature representations. The positions of a pedestrian in the image under surveillance scenarios are relatively fixed, hence we propose intra-part and inter-part relational transformer. The intra-part module creates local relations with mask-guided features, while the inter-part relationship builds correlations with transformers, to develop cross relationships between part nodes. With the collaborative learning inter- and intra-part relationships, experiments reveal that our proposed Pirt model achieves a new state of the art on the public occluded dataset, and further extensions on standard non-occluded person Re-Id datasets also reveal our comparable performances.

CCS CONCEPTS

• **Computing methodologies** → **Object identification**; **Object recognition**.

KEYWORDS

person re-identification, pose-guided, attention

ACM Reference Format:
Zhongxing Ma[1], Yifan Zhao[1], Jia Li[1,2*]. 2021. Pose-guided Inter- and Intra-part Relational Transformer for Occluded Person Re-Identification. In *Proceedings of the 29th ACM International Conference on Multimedia (MM'21), October 20–24, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3474085.3475283

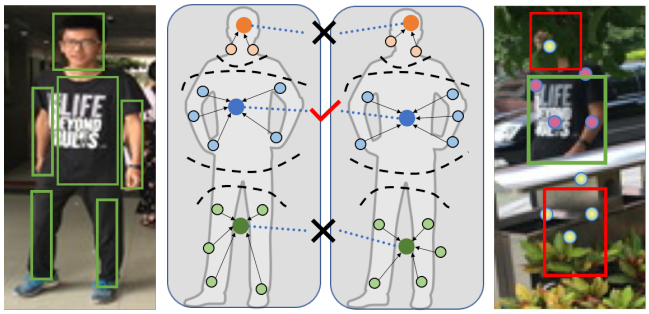

Figure 1: The motivation of our proposed method. We estimate the keypoints of the pedestrian in image and extract local information with limited region. Then we aggregate information in more abstract part region according to the position of each keypoint. We capture the most representative information and match it with its corresponding semantic information to get the final retrieval results.

1 INTRODUCTION

Person Re-Identification, which intends to retrieve the most similar person image to the query person, has made great progress due to the wide application of deep learning [39]. However, handling cross-camera identification in complex occlusion scenes still faces great challenges. In the conventional person Re-Id task, the aligned person image usually contains the information of the holistic body [37, 47]. Previous works [4, 32, 33, 41] tend to build strong backbone features or add part correlations, thereby the robust global feature expressions of the person can be learned. By contrast, in the occlusion scene, there usually exists a large number of occluding objects in aligned person images [52], which makes it difficult to produce reliable and discernible features.

By taking a deep investigation into the occluded person Re-Id task, the redundant occlusion does not only lead to the loss of target information but also introduces additional disturbing information. These occlusions of different characters such as color, size, shape, and structural position have affected the overall re-identification of the person. Most of the existing methods adopt the pose estimation [9, 21, 35], the mask guidance [15], or the class activations [13, 16, 18, 29, 31, 44] as auxiliary guidance to amplify the discriminative regions with higher confidence. Mask-guided information [16] helps discover the foreground region but loses the internal relation of keypoints which is provided by the pose estimation like [35].

To solve this problem, pose-guided person Re-Id models [9, 21, 35] adopt an auxiliary keypoint detector to detect visible body parts, such as heads, hands, and legs, which build the strong semantic representation of the object. In the occluded scenarios, the overlap of other agnostic objects usually leads to the disturbance of the right location of body parts. Hence directly matching these regions

*Jia Li is the corresponding author (E-mail: jiali@buaa.edu.cn).
Website: https://cvteam.net/.

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475283

would produce severe errors in the retrieval phase. To this end, most of the existing methods [9, 21, 35] adopt local feature matching, such as feature alignment and graph matching.

Wang *et al.* [35] propose a Graph Convolutional Network (GCN) based method to generate part features using pose guided keypoints and they apply graph matching to align these part features. However, some keypoints extracted from the image cover the background and occlusion or they have limited area. Miao *et al.* [21] propose a partial and pose-guided part feature alignment to select the useful information from the heavily occluded person. But the semantic information of different partial features is not always the same, and it neglects the relationship between these partial features.

Despite their performance differences, current occluded person Re-Id methods still show limitations in three aspects: 1) the pose estimation network generates unreliable localization, further feature matching with these parts is easily lead to misalignment; 2) the generation of local part regions usually neglects essential contextual information, leads to overfitting issues on visual patterns; 3) the structural relationship of keypoints structures is not deeply investigated, which makes it difficult to recognize some unreasonable matching results.

To solve the aforementioned problems, we propose a pose-guided inter- and intra-part relational transformer for the occluded person Re-Id task. For constructing reliable part features, we firstly expand the keypoint regions into larger masks and then merge the keypoint parts into several groups (3 groups for illustration in Fig. 1). Thus further feature representation learning is conducted within each group to form a stable feature. Moreover, we aggregate the keypoint heatmaps to form a holistic object to enhance the foreground semantic regions for identification.

After exerting pose-based information from an external pretrained model, we propose a joint part compositional model of three types of parts, encoding the contextual regions while strengthening the local features. We adopt the striped slice, patched grid, and pose-keypoint region as part representations of the image as shown in Fig. 2. After introducing contextual information and building strong part representations, we propose our relational transformer to construct structural understanding. We adopt the successfully practiced transformer architecture in the field of natural language processing. With part parsing, each part feature is regarded as one graph node to adaptively learn the robust representation, which handles the missing semantics with the occlusion. In Fig. 1, within each part group, every regularized keypoint region is aggregated into the holistic representation for this group. During the retrieval process, we search the nearest neighbor of the query image using global features and part-based local features. With the collaborative constraints of local features and global features, our proposed Pirt model achieves a new state of the art on the public occluded person Re-Id dataset, Occluded-Duke dataset [21] with the supervised setting. We also conduct detailed ablations to verify the effectiveness of the proposed pose guidance strategies and the inter- and intra-part relational architecture. Further extensions on standard non-occluded person Re-Id datasets also reveal the comparable generalization capability of our model.

Our contribution can be summarized as three-fold:

(1) We propose a pose-guided inter- and intra-part relational transformer for occluded person re-identification, which builds long-term correlations by introducing transformer architectures. Experimental results also verify our proposed method reaches a new state of the art on public benchmarks.
(2) We make insightful improvements for pose-guided feature extractions, in which detected keypoints are expanded to construct the holistic object while forming part groups.
(3) We propose to learn the intra-part relationship with self-correlations and construct a multi-source inter-part relationship learning with relational transformers, providing a structural understanding of different part components.

2 RELATED WORK

Person re-identification. Great improvements have been made in supervised Person Re-Id [39]. One solution to handle the challenge is the part-based model, which was proposed according to the fixed position of one person in the image [32, 43]. Sun *et al.* [32] proposed a uniform partitioning strategy to output the visual descriptor consisting of several horizontal part-level features, similar ideas can be founded in other fine-grained tasks [11, 46]. On the other hand, to enhance the person feature representation capability of Convolutional Neural Network (CNN), several methods [4, 17, 24, 50] tried to extend modules to get better performance. Hou *et al.* [17] proposed a spatial interaction-and-aggregation module to capture relationships between spatial features. Attention mechanism was used for designing CNNs to capture person information in images [2, 33, 38, 41, 45]. Zhang *et al.* [41] utilized the clustering-like information among spatial positions in the feature map and proposed two relation-aware global attention modules. In some mask-guided models [3, 27], external knowledge helps to capture the information of the foreground. In some pose-guided models [10, 28], the positional and semantic information of the keypoints were used to produce local features and explore the connection between them.

Occluded person re-identification. Recognizing an occluded person is much difficult because of the confusing information and spatial feature misalignment [35]. Most of the papers adopt the local feature matching method [8, 13, 15, 16, 18, 21, 29, 31, 35, 44, 51]. He *et al.* [13] proposed a method of spatial feature reconstruction that got robust local feature maps, then the author used the least square algorithm to solve the coefficient matrix to align corresponding spatial features. He *et al.* [16] proposed a method that was based on the inspiration of [13]. The paper [16] proposed a method aiming to extract probability scores from the foreground probability network. The scores are used for the spatial reconstruction by assigning the body parts with confidence scores. Wang *et al.* [35] proposed a keypoint semantic feature extraction method with pose guidance, then feature maps were passing messages using GNN. In addition, the method in the paper[21] used the basic idea of the paper [32] with pose estimation methods to generate robust features, while [15] used pose estimation and salience object detection simultaneously, which generated the semantic mask constrained by two keypoint detectors of the human body.

Image retrieval. The local feature matching used in the occluded person Re-Id is similar to the local feature matching used in image retrieval [1, 6, 19, 26, 36, 40]. Keypoints across images

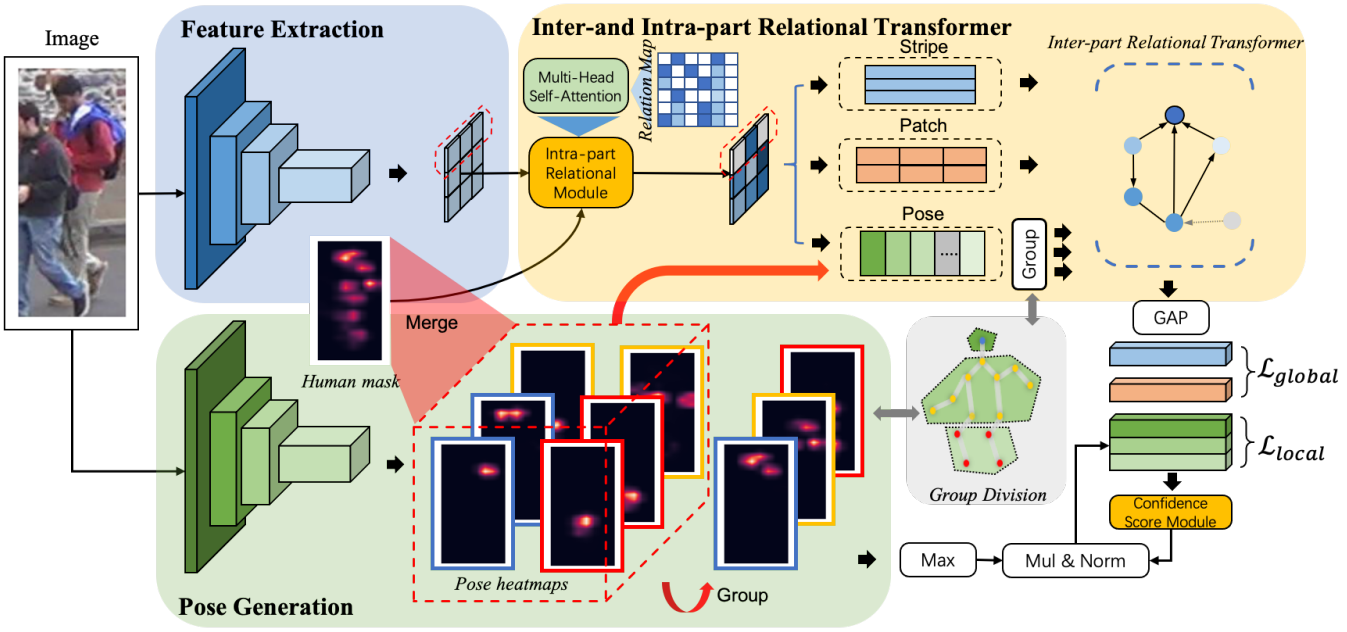

Figure 2: The overall architecture of our proposed model, which consists of three essential parts: feature extraction from the image, keypoints generation from the image, inter- and intra-part relational transformer to aggregate the information between different visible or invisible regions.

must comply some certain physical constraints. For example, one keypoint is associated with a certain keypoint in another image and some keypoints are mismatched due to occlusion. Sarlin *et al.* [26] proposed an attention-based graph neural network to jointly find corresponding keypoints and reject irrelevant keypoints. The paper [26] combined with traditional local feature detector and descriptor which extract sparse keypoints. Cao *et al.* [1] introduced an auto-encoder-based dimensionality reduction technique for local features and combined the global feature with local features to get better performances. The local features were generated from CNN in the paper [1], which were a sort of high-level information extracted from the original image. Our method also draws some inspiration from the paper [1].

Transformer. The transformer was firstly proposed by [34] for machine translation. Similar and improved methods reached the state of the art in many natural language processing tasks. One of the representative works [7] is a solid example to explain the strength of self-attention. With the development of the transformer, a lot of works [23, 42] were looking at how to design the transformer for image recognition. [42] explored several self-attention modules which were invariant to permutation, then they proposed a patch-wise self-attention module with the capacity to uniquely identify specific locations.

3 METHOD

3.1 Overview

The difficulties of occluded person Re-Id lie in the incomplete information and the unstable location of the person in the image. In most cases, a pedestrian in the surveillance scenario has a standing or walking posture with his head, body, and legs well aligned from the top to bottom. An intricate occlusion environment would cause more negative effects, especially on locality. We use a pose estimation network $\mathcal{P}$ to extract keypoints of the person in the image $\mathcal{I} \in [0, 255]^{H \times W \times 3}$. The keypoints cover the corresponding positions of human body joints in the image and keypoints are represented by heatmaps $\mathbf{M} \in (0, 1)^{H_m \times W_m \times P}$, where P is the number of keypoints. Besides, we use a Global Max Pooling (GMP) layer to expand the original limited coverage of keypoints to seize more essential contextual information. We extract the visible part information of the pedestrian through these heatmaps $\mathbf{M}$.

After the essential local part information is obtained, we merge all the heatmaps $\mathbf{M}$ into one human mask $\mathbf{C} \in (0, 1)^{H_f \times W_f}$. We combine human mask $\mathbf{C}$ together with the corresponding spatial feature map $\mathbf{F} \in \mathbb{R}^{H_f \times W_f \times C}$ generated by the backbone. $\mathbf{F}$ is passed into the Intra-part Relational Module (IRM) which is one of the components of our model. The feature maps $\mathbf{F}_{irm}$ generated by IRM are partitioned by different strategies into three-part feature sets as shown in Fig. 2 and Fig. 4. For each feature in each set, we use a Global Average Pooling (GAP) layer to mix the information of the partitioned area. We denote $\mathbf{F}_{local} \in \mathbb{R}^{N \times C}$ as our local features and N is the number of parts. In addition, all the keypoint parts are divided into three groups. We totally have three groups of keypoints that are represented by heatmaps: one group of head keypoints, one group of upper body keypoints, and the last one of lower body keypoints.

Finally, each type of local features $\mathbf{F}_{local}$ is sent into the Inter-part Relational Transform (IRT) to generate well representative features

as the final embedding features. Embedding features partitioned by the **M** are received by Confidence Score Module (CSM). The outputs are incorporated with max score **S** from **M** to produce final embedding features.

3.2 Pose-guided Feature Extraction

Person Re-Id is a branch of fine-grained task, therefore severe occlusion and multiple pedestrians in the image seriously impair the information of the target person. It is difficult for general models to learn the knowledge from the datasets and to reliably positioning the person in the image. To overcome the problems described above, pose information is critical to find keypoints that represent visible parts of the pedestrian.

We firstly use $\mathcal{P}$ to generate keypoints and their original heatmaps. For each heatmap, the value in each grid ranged from 0 to 1 represents the corresponding confidence score. Each initial pose heatmap only covers a limited region. Because the domain gap between the dataset trained for $\mathcal{P}$ and the dataset trained for Re-ID task is large. We use a GMP layer to expand the area and integrate more contextual information:

$$\mathbf{M} = GMP(\hat{\mathbf{M}}), \tag{1}$$

where $\hat{\mathbf{M}}$ is initial heatmaps generated by $\mathcal{P}$. After gathering all heatmaps **M** containing extra peripheral information, we choose the maximum confidence score **S** for all heatmaps on the identical grid to construct the human mask **C**:

$$\mathbf{C}_g = max(\mathbf{M}_g), \tag{2}$$

where g represents the gth grid in original heatmaps. We use local features as complementary information to get better performances. For each local feature generated by each keypoint, we set a threshold τ to convert the original heatmaps **M** into 0-1 masks. Then we apply a GAP layer to generate features:

$$\mathbf{F}_{pose} = GAP((\mathbf{M} > \tau) \cdot \mathbf{F}_{irm}), \tag{3}$$

where $\mathbf{F}_{pose}$ is one type of $\mathbf{F}_{local}$ in Sect. 3.4. We select the max confidence score in each grouped heatmaps **M** for subsequent fusion.

3.3 Intra-part Relational Module

With the guidance of pose information, learning methods for local part features from an image still have a large development space. Based on the mentioned traits of the image, we assume that each horizontal area probably contains part information of the pedestrian. After we obtain the spatial features **F** generated by the backbone, we design a bottleneck-like style module called intra-part relational module. For each horizontal part of **F**, our goal is to find the relationship between features in it. Our overall architecture of IRM is shown in Fig. 3. We split our architecture into three stages. At the first stage, we use a convolutional layer with 1×1 filters to encode the channel information. InstanceNorm layer, BatchNorm (BN) layer and the activation function ReLU to construct a block $\phi(\cdot)$, and for features **F**:

$$\mathbf{F}_\phi = \phi(\mathbf{F}), \tag{4}$$

where $\mathbf{F}_\phi \in \mathbb{R}^{H_f \times W_f \times d}$ indicates the intermediate features generated by $\phi(\cdot)$, and d is their dimension. To capture visible pedestrian information of each part, an attention mechanism called Multi-Head

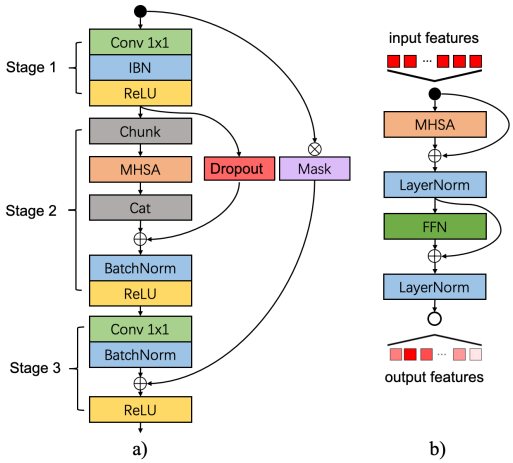

Figure 3: Brief illustration of our intra-part relational module and transform unit. a) is our proposed intra-part relational module and b) is the general component of our inter-part relational transformer.

Self-Attention (MHSA) [34] is used to find the relation between features in the horizontal part. This module is used in a standard transformer block as well. We then horizontally divide $\mathbf{F}_\phi$ into striped features $\widetilde{\mathbf{F}}_\phi \in \mathbb{R}^{1 \times W_f \times d}$ as shown in Fig. 3. The striped feature $\widetilde{\mathbf{F}}_\phi$ contains multiple individual features with dimension d. We use MHSA to build relationships between these features, then the relational guidance is built to construct different attention scores for each feature $\widetilde{\mathbf{F}}_\phi$. We place all features into original position. To avoid the over-fitting problem that might be raised by MHSA, we adopt a Dropout layer to be a stochastic identical mapping:

$$\mathbf{F}^*_\phi = concat(MHSA(chunk(\mathbf{F}_\phi))) + Dropout(\mathbf{F}_\phi). \tag{5}$$

We apply BN and ReLU to generate final features $\mathbf{F}^*_\phi$ of stage 2. MHSA aggregates and captures important messages between features. The input consists of query features and key features of dimension d_k, and value features of dimension d_v. For self-attention, query features, key features, and value features are identical. We compute the relation matrix as same as [34].

$$attn_i(\mathbf{X}) = softmax(\frac{\mathbf{X}\mathbf{W}_i{}^Q(\mathbf{X}\mathbf{W}_i{}^K)^T}{\sqrt{d_k}})\mathbf{X}\mathbf{W}_i{}^V, \tag{6}$$

$$\overline{\mathbf{X}} = concat(attn_i(\mathbf{X}))\mathbf{W}^H \quad i \in 1, \ldots, h, \tag{7}$$

where **X** indicates abstract input features, and $\overline{\mathbf{X}}$ indicates output features, and $\mathbf{W}^{Q,K,V}$ represent learnable weights of query, key, value features, and their projections. At last, we use $\mathbf{W}^H$ to enhance the features of one head. Applying multiple heads can get more relationships because of the diverse projections, and richer expressiveness of features through MHSA is then obtained. The next procedure of the final generated features $\mathbf{F}_{irm}$ is similar to function $\phi(\cdot)$. $\theta(\cdot)$ contains a convolutional layer and BN layer to restore original dimension as input features **F**. Instead of applying

the input features $\mathbf{F}$ as residual connection, we combine these features with human mask $\mathbf{C}$ as an auxiliary attention complement. Finally, $\mathbf{F}_{irm}$ is produced by the following equation:

$$\mathbf{F}_{irm} = ReLU(\theta(\mathbf{F}^*_\phi) + \mathbf{C} \cdot \mathbf{F}). \tag{8}$$

3.4 Inter-part Relational Transformer

In complex and extreme occlusion scenes, *e.g.*, few regions of the pedestrian are visible because of the huge size of occlusion or another person in the image, the pose estimation network $\mathcal{P}$ would generate heatmaps $\mathbf{M}$ that contain error locations with strong confidence scores. The different orientations of the body also influence the matching results. When optimizing local features generated by our model through the dataset, the above problems often lead to some fatal errors and an unstable training procedure.

Therefore, we intend to aggregate local part features extracted by pose estimation network $\mathcal{P}$ with the global feature of the whole image to train the model. The main challenge of generating global features is obliging the model to focus on visible parts of the pedestrian as much as possible, while the information of the occlusions occupies a small proportion in features. In the IRM model, for any horizontal region, we have paid attention to visible parts of the pedestrian, but a solitary part hardly represents the whole pedestrian. The relation between visible parts becomes one essential factor for representing the pedestrian. By constructing the relationship between the different parts, we can pass messages to each part which is abstracted as a graph node. The relationship provides a structural understanding between these nodes, and useless features like occlusion features are discarded.

For tackling these problems, we introduce the inter-part relational transformer [34]. In the field of natural language processing, the transformer takes sentences or words in some scenes as the input. We could build the internal relationship of various words from the global perspective. But in computer vision, the formation of image data is not constructed in a sequence style. In our proposed inter-part relational transformer, we incorporate transformer encoder architecture into our model. In each partitioning strategy, spatial features $\mathbf{F}_{irm}$ are used to generated different part features. Different partitioned abstract features $\mathbf{X}$ in each group are sent into a weight-sharing transformer unit as shown in Fig. 3 to capture the relationship between them. For the representation of a query part q, some value of the feature $\mathbf{X}$ is related to the key part k:

$$\alpha_{q,k} = softmax(\frac{\mathbf{X}\mathbf{W}_i{}^Q(\mathbf{X}\mathbf{W}_i{}^K)^T}{\sqrt{c}}), \tag{9}$$

where the attention score $\alpha_{q,k}$ is based on similarities over the query and key part features. c denotes the dimension of the abstract feature $\mathbf{X}$. We can obtain the attention-based features by using a learnable weight $\mathbf{W}^R$ and use multiple projections to generate final features:

$$\overline{\mathbf{X}} = FFN(\alpha_{q,k}\mathbf{X}\mathbf{W}^R), \tag{10}$$

where FFN denotes the linear projection on the $\mathbf{X}$ called Feed Forward Network. This module consists of two linear projections with the ReLU activation function and the Dropout function:

$$FFN(\mathbf{X}) = Dropout(ReLU(\mathbf{X}\mathbf{W}_1 + \mathbf{b}_1))\mathbf{W}_2 + \mathbf{b}_2. \tag{11}$$

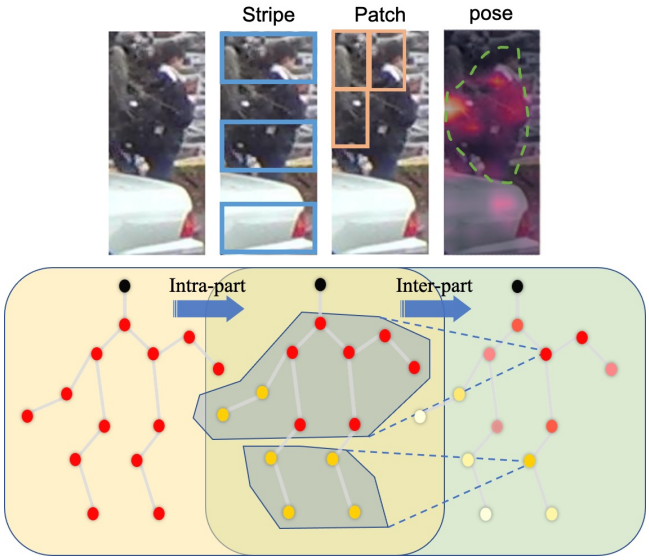

Figure 4: Different partitioning strategies and the procedure of relation building. Part features learn the self-correlation through intra-part relational module and understand the structural information in each group through inter-part relational transformer.

In this module, partitioned features $\mathbf{X}$ are firstly calculating relation through MHSA module, and output features are then combined with residual features and the layer normalization. The subsequent FFN module would process these features as the same.

Partitioning strategy. Based on the intra-part relation module, there are three strategies for partitioning as shown in Fig. 4. We firstly apply the stripe pooling which is closely related to the regional pooling to split features $\mathbf{F}_{irm}$ horizontally. We assume features $\mathbf{F}_{irm}$ of each part have already focused on the visible location of the pedestrian, and information of each part still well remains. We mark the striped features as $\mathbf{F}_{stripe} \in \mathbb{R}^{H_f \times 1 \times C}$.

The second partitioning strategy divides the spatial features $\mathbf{F}_{irm}$ into patched girds. We assume that the solitary utilization of intra-part relation module and horizontal partitioning strategy loses some information that is incorrectly ignored, resulting in the inability to underlie the thorough relation between their potential information. In the divided part, we aim to find the hidden spatial features and their relation, as a supplement for the global feature. We adopt a GMP layer to select features from $\mathbf{F}_{irm}$ without influencing their potential capability on representation. We mark patched features as $\mathbf{F}_{patch} \in \mathbb{R}^{\frac{H_f}{4} \times \frac{W_f}{2} \times C}$.

For the last partitioning strategy, we incorporate pose heatmaps $\mathbf{M}$ with features $\mathbf{F}_{irm}$ to select the corresponding part feature as shown in Fig. 4. In addition, we divide all features generated by this strategy into three disjoint groups. We denote the pose-guided features in one of the groups as $\mathbf{F}_{pose} \in \mathbb{R}^{M \times C}$, where M indicates the number of manually divided features.

We apply our inter-part relation transformer to build long-term relationships between partitioned areas. To build relations crossing

the part features, we individually send features into our transformer. For each set of local features $\mathbf{F}_{local} \in \{\mathbf{F}_{stripe}, \mathbf{F}_{patch}, \mathbf{F}_{pose}\}$ as mentioned in Sect. 3.2, we average the first two local features and regard these features as final embedding features $\mathbf{f}_{local}$:

$$\mathbf{f}_{local} = GAP(IRT(\mathbf{F}_{local})). \tag{12}$$

Because the pose estimation network $\mathcal{P}$ is unstable for confidence score generation due to the domain gap, we propose a confidence score module to generate self-scores $\mathbf{S}_{self}$ and combine with the scores S mentioned in Sect. 3.2. The CSM is consistent with a Multi-Layer Perceptron layer, and the $\mathbf{S}_{self}$ is generated as the following:

$$\mathbf{S}_{self} = ReLU(\hat{\mathbf{F}}_{pose}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2, \tag{13}$$

where $\hat{\mathbf{F}}_{pose}$ denotes three combined pose-guided features $\mathbf{f}_{pose}$. We use the ReLU activation function to build a non-linear mapping function. $\mathbf{S}_{self}$ is multiplied by S and a normalization function is applied to smooth the value. The $\widetilde{\mathbf{F}}_{pose}$ is combined with the score at last:

$$\widetilde{\mathbf{F}}_{pose} = \hat{\mathbf{F}}_{pose} \cdot norm(\mathbf{S}_{self} \cdot \mathbf{S}). \tag{14}$$

Matching strategy. Image retrieval task usually adopts global features and local features to do matching like [19]. Due to the huge differences between categories, robust local features could well complete traditional retrieval tasks. Different from the image retrieval task, many person re-identification methods use one single global feature with dimension d to represent the person in the image. The main reason is that the discrepancy between local features is not huge, and contributing to insufficient fine-grained details of traditional local features. Therefore the results of matching only based on local features are relatively poor.

On the other hand, occluded person Re-Id often requires local features for precise feature alignment to get better results. Therefore, several methods are proposed to handle this problem. [16] and [13] used spatial feature reconstruction and [35] directly used keypoints features. We choose pose-guided local features which contain the fine-grained information of the person. We combine global features and local features to produce results. An initial coarse rank list is established using the global feature, and local features are used to perform more precisely matching on the rank list. With the first ranking list obtained by measuring the global feature, we only conduct local ranking on the top-N images (*e.g.*, N=100) that appear in the list. In nearly all the cases, top-N-selected images cover all the correct results of a query image.

3.5 Loss

In the training process, we use cross entropy loss and triplet loss as the supervised classification signals. In every training step, we randomly sample K samples of P pedestrians from the training dataset. For triplet loss, we use hard triplet loss [16]. The final learning objective of pose-guided local features $\widetilde{\mathbf{F}}_{pose}$ is formulated as the following:

$$\mathcal{L}_{local} = \frac{1}{p}\sum_{i=1}^{p}[\mathcal{L}_{cls}(\mathbf{f}_{pose}^{i}) + \mathcal{L}_{tri}(\mathbf{f}_{pose}^{i})], \tag{15}$$

where p is the number of pose-guided groups. Learning objective of features $\mathbf{f}_{stripe}$ and $\mathbf{f}_{patch}$ which contain information in the inter- and intra-part manner can be formulated as:

$$\mathcal{L}_{global} = \mathcal{L}_{cls}(\mathbf{f}_{stripe}) + \mathcal{L}_{cls}(\mathbf{f}_{patch}), \tag{16}$$

The joint loss function is described below:

$$\mathcal{L} = \mathcal{L}_{local} + \mathcal{L}_{global}. \tag{17}$$

With the loss function above, our model can better understand the structural relationship of different part components.

4 EXPERIMENTS

4.1 Datasets

Occluded-DukeMTMC dataset [21]. There are 15,618 images for training, 17,661 gallery images and 2,210 occluded query images for testing. Training images contain about 9% occluded images and the portion of occluded images in the gallery is around 10%. There always exists one occluded image when computing the distance.

Market-1501 dataset [47]. The dataset includes 32,668 images of 1,501 identities. These pedestrians are captured by six cameras from different scenes and viewpoints. There are 12,936 training images of 751 identities. The testing set contains the remaining images. There is one label to represent the background.

DukeMTMC-reID dataset [25, 48]. The dataset includes 36,411 images of 1,404 identities. There are eight different cameras to capture the person. This dataset selects 702 identities for training. The remaining images are for testing. In the testing phase, there is only one query image of each person in each camera, and remaining images are reserved for gallery.

4.2 Implementation details

Model architectures. For our feature extraction backbone, we use ResNet50-ibn [12, 22] pretrained on ImageNet [5]. We drop its final GAP layer and fully connected layer to extend our module. For pose estimation network, we use HR-Net [30] pretrained on COCO dataset [20] which follows the identical setting as [35]. Pose estimation network predicts a total of 17 keypoints, including head, joints of arms, and joints of legs. The parameters of the pose estimation network are set to unchanged during training. We use three stacked transformer units with 512 hidden dimensions in FFN, 0.1 dropout ratio, and the ReLU activation layer. We set a threshold τ as 0.001 to filter the low confidence score. We use a BN layer to normalize our final embedding features. We use the modified early version of the platform [14] to implement our method.

Traning details. The input images are resized into 384×128. In the training stage, batch size is set to 64 by selecting 16 different identities and 4 samples for each identity. We choose some common data augmentation strategies including padding 10 pixels, random cropping, horizontal flipping, and random erasing [49] with a probability of 0.5. During training, Adam optimizer is adopted. We set weight decay 5×10^{-4} and we train our model for 60 epochs with initialized learning rate 3.5×10^{-4}. The learning rate is warmed up for 10 epochs first and decayed by cosine method from the 30th epoch to 1×10^{-6}. The model is implemented with the pytorch framework and trained with two NVIDIA RTX 2080 Ti.

Evaluation metrics. For evaluation, cumulative matching characteristic curve and mean average precision (mAP) are adopted.

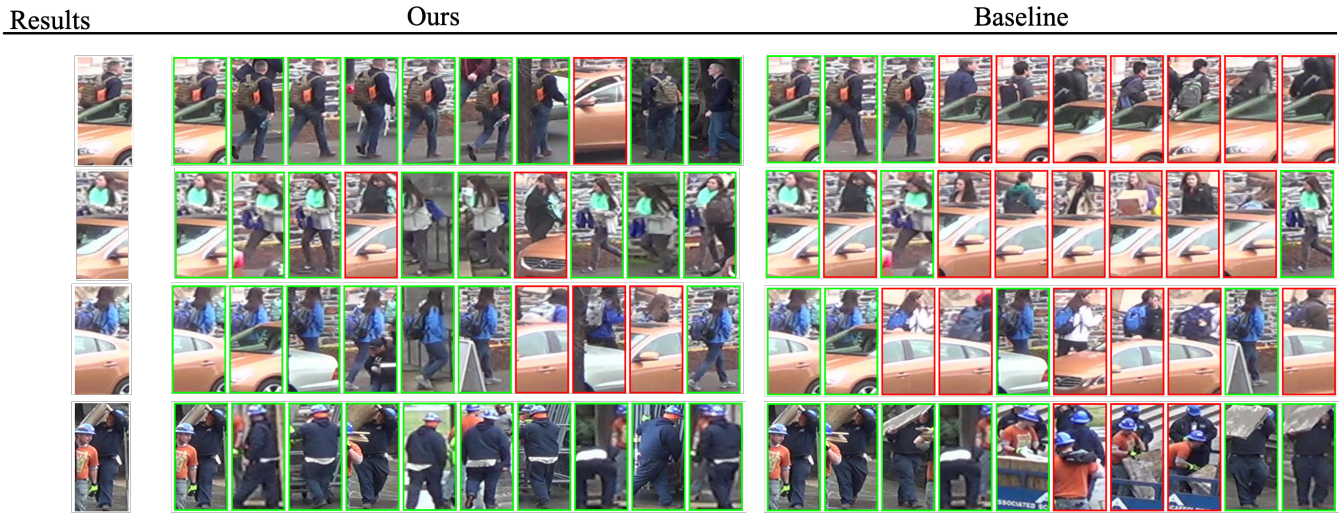

Figure 5: Comparison of retrieval results. The leftmost denotes query image which is followed by the ten closest matching results (including the same camera). The green color indicates correct retrieval results and the red color indicates error results.

Table 1: Performance (%) comparisons to the state-of-the-art occluded Re-Id results on the Occluded-DukeMTMC dataset.

Methods	Occluded-DukeMTMC	
	mAP	Rank-1
Part-Aligned [44]	20.2	28.8
PCB [32]	33.7	42.6
Adver Occluded [18]	32.2	44.5
FD-GAN [9]	-	40.8
Part Bilinear [29]	36.9	-
PGFA [21]	37.3	51.4
HONet [35]	43.8	55.1
DSR [13]	30.4	40.8
Baseline (ours)	43.4	51.8
Pirt (ours)	**50.9**	**60.0**

Table 2: The comparisons over the Market-1501 dataset and the DukeMTMC-reID dataset.

Methods	Market-1501		DukeMTMC-reID	
	mAP	Rank-1	mAP	Rank-1
PCB [32]	77.4	92.3	66.1	81.8
DSR [13]	64.3	83.6	-	-
FPR [16]	86.6	95.4	78.4	88.6
VPM [31]	80.8	93.0	72.6	83.6
PGFA [21]	76.8	91.2	65.5	82.6
HONet [35]	84.9	94.2	75.6	86.9
Pirt (ours)	86.3	94.1	77.6	88.9

4.3 Experimental Results

Results on the occluded dataset. Our experiments are conducted on the occluded Re-Id benchmark dataset with supervised setting, Occluded-DukeMTMC dataset [21]. Tab. 1 shows the comparisons over the dataset Occluded-DukeMTMC. Part-Aligned [44], PCB [32] and Adver Occluded [18] are designed for holistic ReID task. FD-GAN [9], Part Bilinear [29], PGFA [21], and HONet [35] use the extra pose information. DSR [13] does not use extra keypoints detector.

Results on holistic datasets. If one solution performs well in the occluded dataset, it should also perform well on holistic datasets. To verify whether our method works on holistic datasets, we compare our method with other methods for occluded Re-Id on holistic datasets Market-1501 [47] and DukeMTMC-reID [25, 48]. Tab. 2 shows that some methods perform well on the Occluded-DukeMTMC dataset, but have different performances on holistic datasets. The method in [16] can outperform two other methods proposed by [21, 35], which indicates that simply using external information like keypoints might not achieve the best performance on holistic datasets. We think it is mainly because holistic datasets contain few occlusions, and the ability of these methods on extracting and representing global information from the image to generate pedestrian features is relatively low. Occluded images in the dataset contain massive noise caused by occlusions, and directly using the global feature leads to serious confusion. Semantic information inside the global feature cannot make a good alignment to distinguish corresponding regions in images. In this case, we put forward different partition strategies and attention modules. On the basis of ensuring the robustness of feature representation ability, our method combines local features and global features for more detailed alignment to achieve better results.

In Fig. 5 we make a rank list to compare our baseline with our proposed method. In the first row, our method finds the holistic pedestrian according to visible parts in the query image. In the second and the third row, visible parts are fewer than occlusions thus baseline only capture obvious regions without considering the relation between the person and the occlusion in the image.

Table 3: Analysis of three different components: P represent pose generation, Intra represents intra-part relational module and Inter represents inter-part relational transformer.

P	Intra	Inter	mAP	Rank-1
×	×	×	43.40	51.81
✓	×	×	42.23	52.17
✓	✓	×	47.20	55.88
✓	✓	✓	**50.90**	**60.00**

Table 4: The performances on different number of our transformer unit N.

N	mAP	Rank-1
1	48.93	57.51
2	49.93	59.05
3	**50.90**	**60.00**
4	50.59	59.41

The last row shows that our method more robustly removes the noise generated by the information of another person. Our method shows the effectiveness to find occluded people.

4.4 Performance Analysis

Ablations studies. We evaluate the effectiveness of our proposed method by reconstructing our model to verify the influences of each module. As shown in Tab. 3, our proposed module improves the performance on the Occluded-DukeMTMC dataset. Baseline method uses only an individual feature vector extracted from the single backbone without using triplet loss function. Then we use HR-Net [30] to extract keypoints from the image and apply averaged local features to present each keypoint. We adopt a GMP and a GAP layer to select representative features as local features in Tab. 3. For each local feature, we use all the loss functions as described above to constraint the local part feature. But extracting features according to the keypoints from the backbone features hardly leads to a better result. DSR [13] shows another type of local features which consists of multiple max poolings with different scales. The final amount of the local features [13] is large and the used least square algorithm between two different local features set leads to inefficiency. Our method uses only three robust local features, and the computational complexity is similar to the baseline.

Effects of inter- and intra-part relational transformer. As shown in Tab. 3, we found that different dividing strategies applied on the spatial features into parts and building the relationship between grids inside each part can give a positive effect to recognize the person in the image as shown in Tab. 3. Different from [32], we do not restrict each striped part using independent cross entropy loss and regard the divided parts as separate local features. With the support of the attention mechanism, merging local features into global features well represents the pedestrian while eliminating lots of parameters. We visualize our model using its weight and activation function as shown in Fig. 6.

Effects of the number of our transformer units. Tab. 4 shows the influence of the different numbers of transformer units on the accuracy. One layer of the transformer unit is insufficient for extracting features. However, the person re-identification task heavily relies on pretrained backbone weights, more encoder layers lead to a tough training procedure while keeping a similar performance. As a result, we set the number of transformer units to 3 to balance the accuracy and the efficiency.

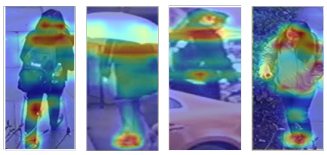

Figure 6: Visualization on features generated by our proposed intra-part relational module.

Table 5: The effects of different combinations of confidence scores on the matching results. Q and G represent confidence scores of query images and gallery images respectively.

Combination Methods	mAP	Rank-1
Q & G	**50.90**	**60.00**
Q only	49.91	59.86
G only	49.89	59.86
No confidence score	48.19	57.78

Effects of confidence score combination method. Corresponding local features are measured by cosine distance and multiplied with confidence scores generated by confidence score module. There are four situations as shown in Tab. 5. The best performance is conducted in the first situation by simultaneously uniting two confidence scores. The next two situations indicate that using only one of the scores impairs the performances. The reason behind this phenomenon mainly occurred in an occluded scene, *e.g.*, different degrees of occlusion on the corresponding part can get wrong matching results. Ignoring potential confidence score influenced by occlusions may lead to a weak performances.

5 CONCLUSIONS

We propose a novel framework namely Pose-guided inter- and intra-part relational transformers for occluded person Re-Id. In the pose generation stage, our detected keypoints are naturally fused into our module to well represent the holistic object with part groups. We combine these local part features with an attention mechanism to construct long-term correlations and provide a structural understanding of different part components by introducing transformer. With these two complementary relationships constructed from different perspectives, our method reaches the new state of the art on occluded person re-identification task and our experiments also show the effectiveness of our model.

6 ACKNOWLEDGEMENT

This work was supported by grants from National Natural Science Foundation of China (No. 61922006) and CAAI-Huawei MindSpore Open Fund.

REFERENCES

[1] Bingyi Cao, André Araujo, and Jack Sim. 2020. Unifying deep local and global features for image search. In *European Conference on Computer Vision*. Springer, 726–743.

[2] Binghui Chen, Weihong Deng, and Jiani Hu. 2019. Mixed high-order attention network for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 371–381.

[3] Xuesong Chen, Canmiao Fu, Yong Zhao, Feng Zheng, Jingkuan Song, Rongrong Ji, and Yi Yang. 2020. Salience-guided cascaded suppression network for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3300–3310.

[4] Zuozhuo Dai, Mingqiang Chen, Xiaodong Gu, Siyu Zhu, and Ping Tan. 2019. Batch dropblock network for person re-identification and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3691–3701.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 248–255.

[6] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2018. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 224–236.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.. In *NAACL-HLT (1)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. http://dblp.uni-trier.de/db/conf/naacl/naacl2019-1.html#DevlinCLT19

[8] Shang Gao, Jingya Wang, Huchuan Lu, and Zimo Liu. 2020. Pose-guided visible part matching for occluded person ReID. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11744–11752.

[9] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, and Hongsheng Li. 2018. FD-GAN: Pose-guided Feature Distilling GAN for Robust Person Re-identification. In *Advances in Neural Information Processing Systems*. 1229–1240.

[10] Jianyuan Guo, Yuhui Yuan, Lang Huang, Chao Zhang, Jin-Ge Yao, and Kai Han. 2019. Beyond human parts: Dual part-aligned representations for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3642–3651.

[11] Bing He, Jia Li, Yifan Zhao, and Yonghong Tian. 2019. Part-regularized near-duplicate vehicle re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3997–4005.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.

[13] Lingxiao He, Jian Liang, Haiqing Li, and Zhenan Sun. 2018. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7073–7082.

[14] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. 2020. FastReID: A Pytorch Toolbox for General Instance Re-identification. *arXiv preprint arXiv:2006.02631* (2020).

[15] Lingxiao He and Wu Liu. 2020. Guided Saliency Feature Learning for Person Re-identification in Crowded Scenes. In *European Conference on Computer Vision*. Springer, 357–373.

[16] Lingxiao He, Yinggang Wang, Wu Liu, He Zhao, Zhenan Sun, and Jiashi Feng. 2019. Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8450–8459.

[17] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. 2019. Interaction-and-aggregation network for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9317–9326.

[18] Houjing Huang, Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. 2018. Adversarially occluded samples for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5098–5107.

[19] Zhenzhong Lan, Yi Zhu, Alexander G Hauptmann, and Shawn Newsam. 2017. Deep local video feature for action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 1–7.

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[21] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. 2019. Pose-guided feature alignment for occluded person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 542–551.

[22] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. 2018. Two at Once: Enhancing Learning and Generalization Capacities via IBN-Net. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[23] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. 2018. Image transformer. In *International Conference on Machine Learning*. PMLR, 4055–4064.

[24] Ruijie Quan, Xuanyi Dong, Yu Wu, Linchao Zhu, and Yi Yang. 2019. Auto-reid: Searching for a part-aware convnet for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3750–3759.

[25] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. 2016. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*. Springer, 17–35.

[26] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2020. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4938–4947.

[27] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. 2018. Mask-guided contrastive attention model for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1179–1188.

[28] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. 2017. Pose-driven deep convolutional model for person re-identification. In *Proceedings of the IEEE international conference on computer vision*. 3960–3969.

[29] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. 2018. Part-aligned bilinear representations for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 402–419.

[30] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5693–5703.

[31] Yifan Sun, Qin Xu, Yali Li, Chi Zhang, Yikang Li, Shengjin Wang, and Jian Sun. 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*. 393–402.

[32] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 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)*. 480–496.

[33] Chiat-Pin Tay, Sharmili Roy, and Kim-Hui Yap. 2019. Aanet: Attribute attention network for person re-identifications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7134–7143.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[35] Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. 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*. 6449–6458.

[36] Runzhong Wang, Junchi Yan, and Xiaokang Yang. 2019. Learning combinatorial embedding networks for deep graph matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3056–3065.

[37] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. 2018. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 79–88.

[38] Bryan Ning Xia, Yuan Gong, Yizhe Zhang, and Christian Poellabauer. 2019. Second-order non-local attention networks for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3760–3769.

[39] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi. 2021. Deep Learning for Person Re-identification: A Survey and Outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021), 1–1. https://doi.org/10.1109/TPAMI.2021.3054775

[40] Andrei Zanfir and Cristian Sminchisescu. 2018. Deep learning of graph matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2684–2693.

[41] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. 2020. Relation-aware global attention for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3186–3195.

[42] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. 2020. Exploring self-attention for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10076–10085.

[43] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. 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*. 1077–1085.

[44] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. 2017. Deeply-learned part-aligned representations for person re-identification. In *Proceedings of the IEEE international conference on computer vision*. 3219–3228.

[45] Shizhen Zhao, Changxin Gao, Jun Zhang, Hao Cheng, Chuchu Han, Xinyang Jiang, Xiaowei Guo, Wei-Shi Zheng, Nong Sang, and Xing Sun. 2020. Do Not

Disturb Me: Person Re-identification Under the Interference of Other Pedestrians. In *European Conference on Computer Vision*. Springer, 647–663.

[46] Yifan Zhao, Ke Yan, Feiyue Huang, and Jia Li. 2021. Graph-Based High-Order Relation Discovery for Fine-Grained Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15079–15088.

[47] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*. 1116–1124.

[48] Zhedong Zheng, Liang Zheng, and Yi Yang. 2017. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision*. 3754–3762.

[49] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. 2020. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 13001–13008.

[50] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. 2019. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3702–3712.

[51] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. 2020. Identity-Guided Human Semantic Parsing for Person Re-Identification. *European Conference on Computer Vision* (2020).

[52] Jiaxuan Zhuo, Zeyu Chen, Jianhuang Lai, and Guangcong Wang. 2018. Occluded person re-identification. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.