# High-Order Information Matters: Learning Relation and Topology for Occluded Person Re-Identification

Guan'an Wang[1*], Shuo Yang[3*], Huanyu Liu[2], Zhicheng Wang[2], Yang Yang[1], Shuliang Wang[3], Gang Yu[2], Erjin Zhou[2] and Jian Sun[2]

[1]Institute of Automation, CAS     [2]MEGVII Technology     [3]Beijing Institute of Technology

[1]wangguanan2015@ia.ac.cn [2]{liuhuanyu,wangzhicheng,yugang,zej,sunjian}@megvii.com
[3]{shuoyang,slwang2011}@bit.edu.cn [1]{yang.yang}@nlpr.ia.ac.cn

## Abstract

*Occluded person re-identification (ReID) aims to match occluded person images to holistic ones across dis-joint cameras. In this paper, we propose a novel framework by learning high-order relation and topology information for discriminative features and robust alignment. At first, we use a CNN backbone and a key-points estimation model to extract semantic local features. Even so, occluded images still suffer from occlusion and outliers. Then, we view the local features of an image as nodes of a graph and propose an adaptive direction graph convolutional (ADGC) layer to pass relation information between nodes. The proposed ADGC layer can automatically suppress the message passing of meaningless features by dynamically learning direction and degree of linkage. When aligning two groups of local features from two images, we view it as a graph matching problem and propose a cross-graph embedded-alignment (CGEA) layer to jointly learn and embed topology information to local features, and straightly predict similarity score. The proposed CGEA layer not only take full use of alignment learned by graph matching but also replace sensitive one-to-one matching with a robust soft one. Finally, extensive experiments on occluded, partial, and holistic ReID tasks show the effectiveness of our proposed method. Specifically, our framework significantly outperforms state-of-the-art by 6.5% mAP scores on Occluded-Duke dataset. Code is available at https://github. com/wangguanan/HOReID.*

## 1. Introduction

Person re-identification (ReID) [6, 43] aims to match images of a person across dis-joint cameras, which is widely

---

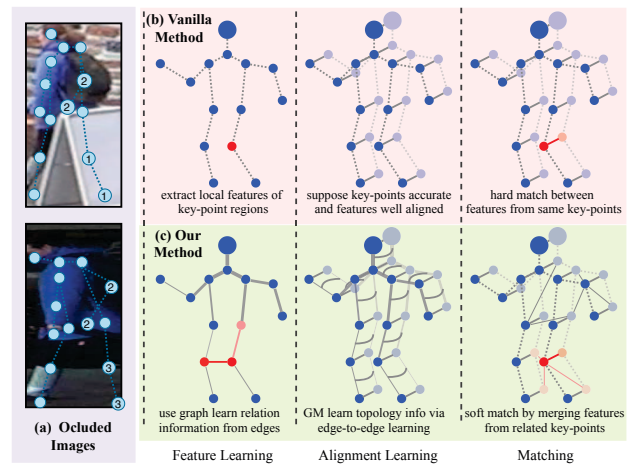*Equal contribution, works done as interns in Megvii Research.



Figure 1. Illustration of high-order relation and topology information. (a) In occluded ReID, key-points suffer from occlusions (①②) and outliers (③). (b) Vanilla method relies on one-order key-points information in all three stages, which is not robust. (c) Our method learn features via an graph to model relation information , and view alignment as a graph matching problem to model topology information by learning both node-to-node and edge-to-edge correspondence.

used in video surveillance, security and smart city. Recently, various of methods [25, 39, 18, 44, 16, 19, 43, 11, 35] have been proposed for person ReID. However, most of them focus on holistic images, while neglecting occluded ones, which may be more practical and challenging. As shown in Figure 1(a), persons can be easily occluded by some obstacles (*e.g.* baggage, counters, crowded public, cars, trees) or walk out of the camera fields, leading to occluded images. Thus, it is necessary to match persons with occluded observation, which is known as occluded person Re-ID problem [48, 26].

Compared with matching persons with holistic images,

occluded ReID is more challenging due to the following reasons [45, 48]: (1) With occluded regions, the image contains less discriminative information and is more likely to match wrong persons. (2) Part-based features have been proved to be efficient [35] via part-to-part matching. But they require strict person alignment in advance, thus cannot work very well in seriously occluded situations. Recently, many occluded/partial person ReID methods [48, 49, 26, 10, 8, 34, 23] are proposed, most of them only consider one-order information for feature learning and alignment. For example, the pre-defined regions [35], poses [26] or human parsing [10] are used to for feature learning and alignment. We argue that besides one-order information, high-order one should be imported and may work better for occluded ReID.

In Figure 1(a), we can see that key-points information suffers from occlusion (①②) and outliers (③). For example, key-points ① and ② are occluded, leading to meaningless features. Key-points ③ are outliers, leading to misalignment. A common solution is shown in Figure 1(b). It extracts local features of key-point regions, supposes all key-points are accurate and local features well aligned. In this solution, all three stages rely on the one-order key-points information, which is not very robust. In this paper, as shown in Figure 1(c), we propose a novel framework for both discriminative feature and robust alignment. In feature learning stage, we view local features of an image as nodes of a graph to learn relation information. By passing message in the graph, the meaningless features caused by occluded key-points can be improved by their neighbor meaningful features. In alignment stage, we use graph matching algorithm [40] to learn robust alignment. Besides aligning with node-to-node correspondence, it models extra edge-to-edge correspondence. We then embed the alignment information to features by constructing a cross-images graph, where node message of an image can be passed to nodes of the other images. Thus, the features of outlier key-points can be repaired by its corresponding features on the other image. Finally, instead of computing similarity with predefined distance, we use a network to learn similarity supervised by a verification loss.

Specifically, we propose a novel framework jointly modeling high-order relation and human-topology information for occluded person re-identification. As shown in Figure 2, our framework includes three modules, *i.e.* one-order semantic module ($\mathcal{S}$), high-order relation module ($\mathcal{R}$) and high-order human-topology module ($\mathcal{T}$). (1) In the $\mathcal{S}$, we utilize a CNN backbone to learn feature maps and a human key-points estimation model to learn key-points. Then we can extract semantic features of corresponding key-points. (2) In the $\mathcal{R}$, we view the learned semantic features of an image as nodes of a graph and propose an adaptive-direction graph convolutional (ADGC) layer to learn and pass messages of edge features. The ADGC layer can automatically decide the direction and degree of every edge. Thus it can promote the message passing of semantic features and suppress that of meaningless and noisy ones. At last, the learned nodes contain both semantic and related information. (3) In the $\mathcal{T}$, We propose a cross-graph embedded-alignment (CGEA) layer. It takes two graphs as inputs, learns correspondence of nodes across the two graphs using graph-matching strategy, and passes messages by viewing the learned correspondence as an adjacency matrix. Thus, the related features can be enhanced, and alignment information can be embedded in features. Finally, to avoid hard one-to-one alignment, we predict the similarity of the two graphs by mapping them to a logit and supervise with a verification loss.

The main contributions of this paper are summarized as follows: (1) A novel framework of jointly modeling high-order relation and human-topology information is proposed to learn well and robustly aligned features for occluded ReID. To our best of our knowledge, this is the first work that introduces such high-order information to occluded ReID. (2) An adaptive directed graph convolutional (ADGC) layer is proposed to dynamically learn the directed linkage of the graph, which can promote message passing of semantic regions and suppress that of meaningless regions such as occlusion or outliers. With it, we can better model the relation information for occluded ReID. (3) A cross-graph embedded-alignment (CGEA) layer conjugated with verification loss is proposed to learn feature alignment and predict similarity score. They can avoid sensitive hard one-to-one person matching and perform a robust soft one. (4) Extensive experimental results on occluded, partial, and holistic ReID datasets demonstrate that the proposed model performs favorably against state-of-the-art methods. Especially on the occluded-Duke dataset, our method significantly outperforms state-of-the-art by at least 3.7% and 6.5% in terms of Rank-1 and mAP scores.

## 2. Related Works

**Person Re-Identification.** Person re-identification addresses the problem of matching pedestrian images across disjoint cameras [6]. The key challenges lie in the large intra-class and small inter-class variation caused by different views, poses, illuminations, and occlusions. Existing methods can be grouped into hand-crafted descriptors [25, 39, 18], metric learning methods [44, 16, 19] and deep learning algorithms [43, 11, 35, 36, 37, 22]. All those ReID methods focus on matching holistic person images, but cannot perform well for the occluded images, which limits the applicability in practical surveillance scenarios.

**Occluded Person Re-identification.** Given occluded probe images, occluded person re-identification [48] aims to find the same person of full-body appearance in dis-joint
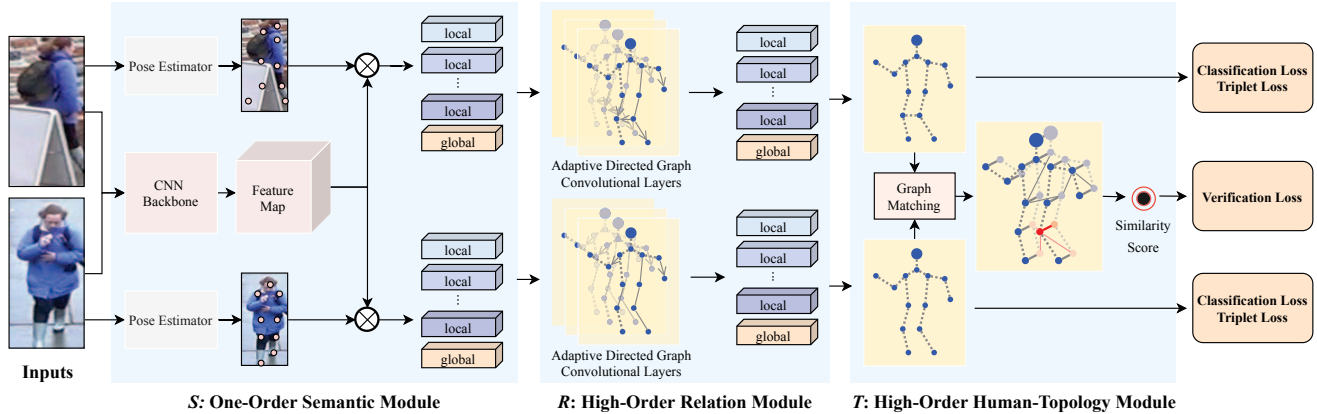
Figure 2. Illustration of our proposed framework. It consists of an one-order semantic module $\mathcal{S}$, a high-order relation module $\mathcal{R}$ and a high-order topology module $\mathcal{T}$. The module $\mathcal{S}$ learns semantic local features of key-point regions. In $\mathcal{R}$, we view the local features of an image as nodes of a graph and propose an adaptive direction graph convolutional (ADGC) layer to pass relation information between nodes. In $\mathcal{T}$, we view alignment problem as a graph matching problem and propose a cross-graph embedded-alignment (CGEA) layer to joint learn and embed topology information to local features, and straightly predict similarity scores.

cameras. This task is more challenging due to incomplete information and spatial misalignment. Zhuo *et al.*[48] use occluded/non-occluded binary classification(OBC) loss to distinguish the occluded images from holistic ones. In their following works, a saliency map is predicted to highlight the discriminative parts, and a teacher-student learning scheme further improves the learned features. Miao *et al.*[26] propose a pose guided feature alignment method to match the local patches of probe and gallery images based on the human semantic key-points. And they use a pre-defined threshold of key-points confidence to determine whether the part is occluded or not. Fan *et al.*[3] use a spatial-channel parallelism network (SCPNet) to encode part features to specific channels and fuse the holistic and part features to get discriminative features. Luo *et al.*[23] use a spatial transform module to transform the holistic image to align with the partial ones, then calculate the distance of the aligned pairs. Besides, several efforts are put on the spatial alignment of the partial Re-ID tasks.

**Partial Person Re-Identification.** Accompanied by occluded images, partial ones often occur due to imperfect detection and outliers of camera views. Like occluded person ReID, partial person ReID [45] aims to match partial probe images to gallery holistic images. Zheng *et al.*[45] propose a global-to-local matching model to capture the spatial layout information. He *et al.*[7] reconstruct the feature map of a partial query from the holistic pedestrian, and further improve it by a foreground-background mask to avoid the influence of backgrounds clutter in [10]. Sun *et al.*propose a Visibility-aware Part Model(VPM) in [34], which learns to perceive the visibility of regions through self-supervision.

Different from existing occluded and partial ReID methods which only use one-order information for feature learning and alignment, we use high-order relation and human-

topology information for feature learning and alignment, thus achieve better performance.

## 3. The Proposed Method

This section introduces our proposed framework, including a one-order semantic module ($\mathcal{S}$) to extract semantic features of human key-point regions, a high-order relation module ($\mathcal{R}$) to model the relation-information among different semantic local features, and a high-order human-topology module ($\mathcal{T}$) to learn robust alignment and predict similarities between two images. The three modules are jointly trained in an end-to-end way. An overview of the proposed method is shown in Figure 2.

**Semantic Features Extraction.** The goal of this module is to extract one-order semantic features of key-point regions, which is inspired by two cues. Firstly, part-based features have been shown to be efficient for person ReID [35]. Secondly, accurate alignment of local features is necessary in occluded/partial ReID [8, 34, 10]. Following the ideas above and inspired by recent developments on person ReID [43, 35, 24, 4] and human key-points prediction [2, 33], we utilize a CNN backbone to extract local features of different key-points. Please note that although the human key-points prediction have achieved high accuracy, they still suffer from unsatisfying performance under occluded/partial images [17]. Those factors lead to inaccurate key-points positions and their confidence. Thus, the following relation and human-topology information are needed and will be discussed in the next section.

Specifically, given a pedestrian image $x$, we can get its feature map $m_{cnn}$ and key-points heat map $m_{kp}$ through the CNN model and key-points model. Through an outer product ($\otimes$) and a global average pooling operations ($g(\cdot)$),

6450

we can get a group of semantic local features of key-points regions $V_l^S$ and a global feature $V_g^S$. The procedures can be formulated in Eq.(1), where $K$ is key-point number, $v_k \in R^c$ and $c$ is channel number. Note that $m_{kp}$ is obtained by normalizing original key-points heatmap with a softmax function for preventing from noise and outliers. This simple operation is shown to be effective in experimet section.

$$V_l^S = \{v_k^S\}_{k=1}^K = g(m_{cnn} \otimes m_{kp})$$
$$V_g^S = v_{K+1}^S = g(m_{cnn}) \qquad (1)$$

**Training Loss.** Following [43, 11], we utilize classification and triplet losses as our targets as in Eq.(2). Here, $\beta_k = max(m_{kp}[k]) \in [0,1]$ is the $k^{th}$ key-point confidence, and $\beta_{K+1} = 1$ for global features, $p_{v_k^S}$ is the probability of feature $v_k^s$ belonging to its ground truth identity predicted by a classifier, $\alpha$ is a margin, $d_{v_k^s,v_k^s}$ is the distance between a positive pair $(v_{ak}^S, v_{pk}^S)$ from the same identity, $(v_{ak}^S, v_{pk}^S)$ is from different identities. The classifiers for different local features are not shared.

$$\mathcal{L}_S = \frac{1}{K+1} \sum_{k=1}^{K+1} \beta_k [\mathcal{L}_{cls}(v_k^s) + \mathcal{L}_{tri}(v_k^s)]$$
$$= \frac{1}{K+1} \sum_{k=1}^{K+1} \beta_k [-log p_{v_k^s} + |\alpha + d_{v_{ak}^S,v_{pk}^S} - d_{v_{ak}^S,v_{nk}^S}|_+] \qquad (2)$$

## 3.1. High-Order Relation Learning

Although we have the one-order semantic information of different key-point regions, occluded ReID is more challenging due to incomplete pedestrian images. Thus, it is necessary to exploit more discriminative features. We turn to the graph convolutional network (GCN) methods [1] and try to model the high-order relation information. In the GCN, semantic features of different key-point regions are viewed as nodes. By passing messages among nodes, not only the one-order semantic information (node features) but also the high-order relation information (edge features) can be jointly considered.

However, there is still a challenge for occluded ReID. Features of occluded regions are often meaningless even noisy. When passing those features in a graph, it brings in more noise and has side effects on occluded ReID. Hence, we propose a novel adaptive-direction graph convolutional (ADGC) layer to learn the direction and degree of message passing dynamically. With it, we can automatically suppress the message passing of meaningless features and promote that of semantic features.

**Adaptive Directed Graph Convolutional Layer.** A simple graph convolutional layer [15] has two input, an adjacent matrix A of the graph and the features X of all node,
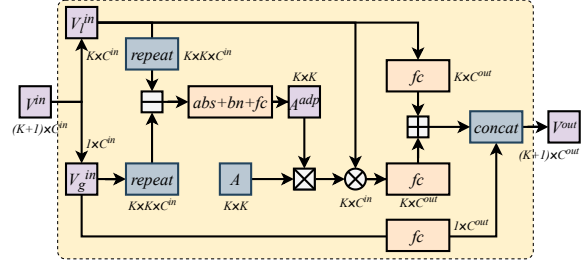


Figure 3. Illustration of the proposed adaptive directed graph convolutional (ADGC) layers. $A$ is a pre-defined adjacent matrix $\boxminus$, $\boxplus$, $\boxtimes$ are element-wise subtraction, add and multiplication. $abs$, $bn$ and $fc$ are absolution, batch normalization and fully connected layer, $trans$ is transpose. Please refer text for more details.

output can be calculated by:

$$O = \hat{A}XW$$

where $\hat{A}$ is normalized version of A and W refers to parameters.

We improve the simple graph convolutional layer by adaptively learning the adjacent matrix (the linkage of node) based on the input features. We assume that given two local features, the meaningful one is more similar to the global feature than that of meaningless one. Therefore, we propose an adaptive directed graph convolutional (ADGC) layer, whose inputs are a global feature $V_g$ and K local features $V_l$, and a pre-defined graph (adjacent matrix is A). We use differences between local features $V_l$ and global feature $V_g$ to dynamically update the edges' weights of all nodes in the graph, resulting $A^{adp}$. Then a simple graph convolutional can be formulated by multiplication between $V_l$ and $A^{adp}$. To stabilize training, we fuse the input local features $V_l$ to the output of our ADGC layer as in the ResNet [7]. Details are shown in Figure 3. Our adaptive directed graph convolutional (ADGC) layer can be formulated in Eq.(3), where $f_1$ and $f_2$ are two unshared fully-connected layers.

$$V^{out} = [f_1(A^{adp} \otimes V_l^{in}) + f_2(V_l^{in}), V_g^{in}] \qquad (3)$$

Finally, we implement our high-order relation module $f_R$ as cascade of ADGC layers. Thus, given an image $x$, we can get its semantic features $V^S = \{v_k^S\}_{k=1}^{K+1}$ via Eq.(1). Then its relation features $V^R = \{v_k^R\}_{k=1}^{K+1}$ can be formulated as below :

$$V^R = f_R(V^S) \qquad (4)$$

**Loss and Similarity.** We use classification and triplet losses as our targets as in Eq.(5), where the definition of $\mathcal{L}_{ce}(\cdot)$ and $\mathcal{L}_{tri}(\cdot)$ can be found in in Eq.(2). Note that $\beta_k$ is the $k^{th}$ key-point confidence.

$$\mathcal{L}^R = \frac{1}{K+1} \sum_{k=1}^{K+1} \beta_k [\mathcal{L}_{cls}(v_k^R) + \mathcal{L}_{tri}(v_k^R)] \qquad (5)$$

Given two images $x_1$ and $x_2$, we can get their relation features $V_1^R = \{v_{1k}^R\}_{k=1}^{K+1}$ and $V_1^R = \{v_{2k}^R\}_{k=1}^{K+1}$ via Eq.(4), and calculate their similarity with cosine distance as in Eq.(6).

$$s_{x_1,x_2}^R = \frac{1}{K+1} \sum_{k=1}^{K+1} \sqrt{\beta_{1k}\beta_{2k}} \, cosine(v_{1k}^R, v_{2k}^R) \quad (6)$$

## 3.2. High-Order Human-Topology Learning

Part-based features have been proved to be very efficient for person ReID [35, 34]. One simple alignment strategy is straightly matching features of the same key-points. However, this one-order alignment strategy cannot deal with some bad cases such as outliers, especially in heavily occluded cases [17]. Graph matching [40, 38] can naturally take the high-order human-topology information into consideration. But it can only learn one-to-one correspondence. This hard alignment is still sensitive to outliers and has a side effect on performance. In this module, we propose a novel cross-graph embedded-alignment layer, which can not only make full use of human-topology information learned by graph matching algorithm, but also avoid sensitive one-to-one alignment.

**Revision of Graph Matching.** Given two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ from image $x_1$ and $x_2$, the goal of graph matching is to learn a matching matrix $U \in [0,1]^{K \times K}$ between $V_1$ and $V_2$. Let $U \in [0,1]$ be an indicator vector such that $U_{ia}$ is the matching degree between $v_{1i}$ and $v_{2a}$. A square symmetric positive matrix $M \in R^{KK \times KK}$ is built such that $M_{ia;jb}$ measures how well every pair $(i,j) \in E_1$ matches with $(a,b) \in E_2$. For pairs that do not form edges, their corresponding entries in the matrix are set to 0. The diagonal entries contain node-to-node scores, whereas the off-diagonal entries contain edge-to-edge scores. Thus, the optimal matching $u^*$ can be formulated as below:

$$U^* = \underset{U}{argmax} \, U^T M U, \; s.t. \; ||U|| = 1 \quad (7)$$

Following [40], we parameter matrix $M$ in terms of unary and pair-wise point features. The optimization procedure is formulated by a power iteration and a bi-stochastic operations. Thus, we can optimize $U$ in our deep-learning framework with stochastic gradient descent. Restricted by pages, we don not show more details of graph matching, please refer to the paper [38, 40].

**Cross-Graph Embedded-Alignment Layer with Similarity Prediction.** We propose a novel cross-graph embedded-alignment layer (CGEA) that both considering the high-order human-topology information learned by GM and avoiding the sensitive one-to-one alignment. The proposed CGEA layer takes two sub-graphs from two images as inputs and outputs the embedded features, including both
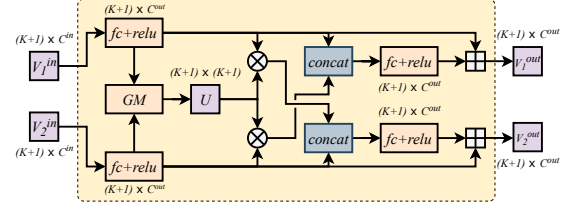


Figure 4. Illustration of the cross-graph embedded-alignment layer. Here, $\otimes$ is matrix multiplication, $fc + relu$ means fully-connected layer and Rectified Linear Unit, $GM$ means graph matching operation, $U$ is the learned affinity matrix. Please refer text for more details.

semantic features and the human-topology guided aligned features.

The structure of our proposed CGEA layer is shown in Figure 4. It takes two groups of features and outputs two groups of features. Firstly, with two groups of nodes $V_1^{in} \in R^{(K+1) \times C^{in}}$ and $V_2^{in} \in R^{(K+1) \times C^{in}}$, we embed them to a hidden space with a fully-connected layer and a ReLU layer, getting two groups of hidden features $V_1^h \in R^{(K+1) \times C^{out}}$ and $V_2^h \in R^{(K+1) \times C^{out}}$. Secondly, we perform graph matching between $V_1^h$ and $V_2^h$ via Eq.(7), and get an affinity matrix $U^{k \times k}$ between $V_1^h$ and $V_2^h$. Here, $U(i,j)$ means correspondence between $v_{1i}^h$ and $v_{2j}^h$. Finally, the output can be formulated in Eq.(8), where $[\cdot, \cdot]$ means concatenation operation along channel dimension, $f$ is a fully-connected layer.

$$\begin{aligned} V_1^{out} &= f([V_1^h, U \otimes V_2^h]) + V_1^h \\ V_2^{out} &= f([V_2^h, U^T \otimes V_1^h]) + V_2^h \end{aligned} \quad (8)$$

We implement our high-order topology module ($\mathcal{T}$) with a cascade of CGEA layers $f_T$ and a similarity prediction layer $f_P$. Given a pair of images $(x_1, x_2)$, we can get their relation features $(V_1^R, V_2^R)$ via Eq.(4), and then their topology features of $(V_1^T, V_2^T)$ via Eq.(9). After getting the topology features pair $(V_1^T, V_2^T)$, we can compute their similarity using Eq.(10), where $|\cdot|$ is element-wise absolution operation, $f_s$ is a fully-connected layer from $C_T$ to 1, $\sigma$ is sigmoid activation function.

$$(V_1^T, V_2^T) = F_T(V_1^R, V_2^R) \quad (9)$$

$$s_{x_1,x_2}^T = \sigma(f_s(-|V_1^T - V_2^T|)) \quad (10)$$

**Verification Loss.** The loss of our high-order human-topology module can be formulated in Eq.(11), where $y$ is their ground truth, $y = 1$ if $(x_1, x_2)$ from the same person, otherwise $y = 0$.

$$\mathcal{L}_T = y log s_{x_1,x_2}^T + (1-y) log(1 - s_{x_1,x_2}^T) \quad (11)$$

## 4. Train and Inference

During the training stage, the overall objective function of our framework is formulated in Eq.(12), where $\lambda_*$ are

weights of corresponding terms. We train our framework end-to-end by minimizing the $\mathcal{L}$.

$$\mathcal{L} = \mathcal{L}_S + \lambda_R \mathcal{L}_R + \lambda_T \mathcal{L}_T \tag{12}$$

For the similarity, given a pair of images $(x_1, x_2)$, we can get their relation information based similarity $s^R_{x_1,x_2}$ from Eq.(6) and topology information based similarity $s^T_{x_1,x_2}$ from Eq.(10). The final similarity can be calculated by combing the two kind of similarities.

$$s = \gamma s^R_{x_1,x_2} + (1 - \gamma)s^T_{x_1,x_2} \tag{13}$$

When inferring, given an query image $x_q$, we first compute its similarity $x^R$ with all gallery images and get its top $n$ nearest neighbors. Then we compute the final similarity $s$ in Eq.(13) to refine the top $n$.

# 5. Experiments

## 5.1. Implementation Details

**Model Architectures.** For CNN backbone, as in [43], we utilize ResNet50 [7] as our CNN backbone by removing its global average pooling (GAP) layer and fully connected layer. For classifiers, following [24], we use a batch normalization layer [13] and a fully connect layer followed by a softmax function. For the human key-points model, we use HR-Net [33] pre-trained on the COCO dataset [20], a state-of-the-art key-points model. The model predicts 17 key-points, and we fuse all key-points on head region and get final $K = 14$ key-points, including head, shoulders, elbows, wrists, hips, knees, and ankles.

**Training Details.** We implement our framework with Pytorch. The images are resized to $256 \times 128$ and augmented with random horizontal flipping, padding 10 pixels, random cropping, and random erasing [47]. When test on occluded/partial datasets, we use extra color jitter augmentation to avoid domain variance. The batch size is set to 64 with 4 images per person. During the training stage, all three modules are jointly trained in an end-to-end way for 120 epochs with the initialized learning rate 3.5e-4 and decaying to its 0.1 at 30 and 70 epochs. Please refer our code[1] for implementation details.

**Evaluation Metrics.** We use standard metrics as in most person ReID literatures, namely Cumulative Matching Characteristic (CMC) curves and mean average precision (mAP), to evaluate the quality of different person re-identification models. All the experiments are performed in single query setting.

## 5.2. Experimental Results

**Results on Occluded Datasets.** We evaluate our proposed framework on two occluded datasets, *i.e.* Occluded-Duke [26] and Occluded-ReID [48]. Occluded-Duke is

---

[1]https://github.com/wangguanan/HOReID

| Dataset | Train Nums (ID/Image) | Testing Nums (ID/Image) | |
|---|---|---|---|
| | | Gallery | Query |
| Market-1501 | 751/12,936 | 750/19,732 | 750/3,368 |
| DukeMTMC-reID | 702/16,522 | 1,110/17,661 | 702/2,228 |
| Occluded-Duke | 702/15,618 | 1,110/17,661 | 519/2,210 |
| Occluded-ReID | - | 200/1,000 | 200/1,000 |
| Partial-REID | - | 60/300 | 60/300 |
| Partial-iLIDS | - | 119/119 | 119/119 |

Table 1. Dataset details. We extensively evaluate our proposed method on 6 public datasets, including 2 holistic, 2 occluded and 2 partial ones.

| Methods | Occluded-Duke | | Occluded-REID | |
|---|---|---|---|---|
| | Rank-1 | mAP | Rank-1 | mAP |
| Part-Aligned [41] | 28.8 | 20.2 | - | - |
| PCB [35] | 42.6 | 33.7 | 41.3 | 38.9 |
| Part Bilinear [32] | 36.9 | - | - | - |
| FD-GAN [5] | 40.8 | - | - | - |
| AMC+SWM [45] | - | - | 31.2 | 27.3 |
| DSR [8] | 40.8 | 30.4 | 72.8 | 62.8 |
| SFR [9] | 42.3 | 32 | - | - |
| Ad-Occluded [12] | 44.5 | 32.2 | - | - |
| TCSDO [49] | - | - | 73.7 | 77.9 |
| FPR [10] | - | - | 78.3 | 68.0 |
| PGFA [26] | 51.4 | 37.3 | - | - |
| **HOReID** (*Ours*) | **55.1** | **43.8** | **80.3** | **70.2** |

Table 2. Comparison with state-of-the-arts on two occluded datasets, *i.e.* Occluded-Duke [26] and Occluded-REID [48].

selected from DukeMTMC-reID by leaving occluded images and filter out some overlap images. It contains 15,618 training images, 17,661 gallery images, and 2,210 occluded query images. Occluded-ReID is captured by the mobile camera, consist of 2000 images of 200 occluded persons. Each identity has five full-body person images and five occluded person images with different types of severe occlusions.

Four kinds of methods are compared, they are vanilla holistic ReID methods [41, 35], holistic ReID methods with key-points information [32, 5], partial ReID methods [45, 8, 9] and occluded ReID methods [12, 49, 10, 26]. The experimental results are shown in Table 2. As we can see, there is no significant gap between vanilla holistic ReID methods and holistic methods with key-points information. For example, PCB [34] and FD-GAN [5] both achieve approximately $40\%$ Rank-1 score on Occluded-Duke dataset, showing that simply using key-points information may not significantly benefit occluded ReID task. For partial ReID and occluded ReID methods, they both achieve an obvious improvement on occluded datasets. For example, DSR [8] get a $72.8\%$ and FPR [10] get a $78.3\%$ Rank-1 scores on Occluded-REID dataset. This shows that occluded and par-

6453

| Methods | Partial-REID | | Partial-iLIDS | |
|---|---|---|---|---|
| | Rank-1 | Rank-3 | Rank-1 | Rank-3 |
| DSR [8] | 50.7 | 70.0 | 58.8 | 67.2 |
| SFR [9] | 56.9 | 78.5 | 63.9 | 74.8 |
| VPM [34] | 67.7 | 81.9 | 65.5 | 74.8 |
| PGFA [26] | 68.0 | 80.0 | 69.1 | 80.9 |
| AFPB [48] | 78.5 | - | - | - |
| FPR [10] | 81.0 | - | 68.1 | - |
| TCSDO [49] | 82.7 | - | - | - |
| **HOReID**(*Ours*) | **85.3** | **91.0** | **72.6** | **86.4** |

Table 3. Comparison with state-of-the-arts on two partial datasets, *i.e.* Partial-REID [45] and Partial-iLIDS [8] datasets. Our method achieves best performance on the two partial datasets.

| Methods | Market-1501 | | DukeMTMC | |
|---|---|---|---|---|
| | Rank-1 | mAP | Rank-1 | mAP |
| PCB [35] | 92.3 | 77.4 | 81.8 | 66.1 |
| VPM [34] | 93.0 | 80.8 | 83.6 | 72.6 |
| BOT [24] | 94.1 | 85.7 | 86.4 | 76.4 |
| SPReID [14] | 92.5 | 81.3 | - | - |
| MGCAM [30] | 83.8 | 74.3 | 46.7 | 46.0 |
| MaskReID [27] | 90.0 | 75.3 | - | - |
| FPR [10] | 95.4 | 86.6 | 88.6 | 78.4 |
| PDC [31] | 84.2 | 63.4 | - | - |
| Pose-transfer [21] | 87.7 | 68.9 | 30.1 | 28.2 |
| PSE [29] | 87.7 | 69.0 | 27.3 | 30.2 |
| PGFA [26] | 91.2 | 76.8 | 82.6 | 65.5 |
| **HOReID**(*Ours*) | 94.2 | 84.9 | 86.9 | 75.6 |

Table 4. Comparison with state-of-the-arts on two holistic datasets, Market-1501 [42] and DukeMTMTc-reID [28, 46]. Our method achieves comparable performance on holistic ReID.

tial ReID task share similar difficulties, *i.e.* learning discriminative feature and feature alignment. Finally, our proposed framework achieves best performance on Occluded-Duke and Occlude-REID datasets at $55.1\%$ and $80.4\%$ in terms of Rank-1 score, showing the effectiveness.

**Results on Partial Datasets.** Accompanied by occluded images, partial ones often occur due to imperfect detection, outliers of camera views, and so on. To further evaluate our proposed framework, in Table 3 we also report the results on two partial datasets, Partial-REID [45] and Partial-iLIDS [8]. Partial-REID includes 600 images from 60 people, with five full-body images and five partial images per person, which is only used for the test. Partial-iLIDS is based on the iLIDS [8] dataset and contains a total of 238 images from 119 people captured by multiple non-overlapping cameras in the airport, and their occluded regions are manually cropped. Following [34, 10, 49], because the two partial datasets are too small, we use Market-1501 as training set and the two partial datasets as test set. As we can see, our proposed framework significantly outperforms the other methods by at least $2.6\%$ and $4.4\%$ in terms of Rank-1 score on the two datasets.

**Results on Holistic Datasets.** Although recent occluded/partial ReID methods have obtained improvements on occluded/partial datasets, they often fails to get a satisfying performance on holistic datasets. This is caused by the noise during feature learning and alignment. In this part, we show that our proposed framework can also achieve satisfying performance on holistic ReID datasets including Market-1501 and DuekMTMTC-reID. Market-1501 [42] contains 1,501 identities observed from 6 camera viewpoints, 19,732 gallery images and 12,936 training images, all the dataset contains few of occluded or partial person images. DukeMTMC-reID [28, 46] contains 1,404 identities, 16,522 training images, 2,228 queries, and 17,661 gallery images.

Specifically, we conduct experiments on two common holistic ReID datasets Market-1501 [42] and DukeMTMC-reID [28, 46], and compare with 3 vanilla ReID methods [35, 34, 24], 3 ReID methods with human-parsing information [14, 30, 27, 10] and 4 holistic ReID methods with key-points information [31, 21, 29, 26]. The experimental results are shown in Table 4. As we can see, the 3 vanilla holistic ReID methods obtain very competitive performance. For example, BOT [24] gets a $94.1\%$ and $86.4\%$ Rank-1 score on two datasets. However, for the holistic ReID methods using external cues such human-parsing and key-points information perform worse. For example, SPReID [14] uses human-parsing information and only achieves $92.5\%$ Rannk-1 score on Market-1501 dataset. PFGA [26] uses key-points information and only gets a $82.6\%$ Rank-1 score on DukeMTMC-reID dataset. This shows that simply using external cues such as human-parsing and key-points may not bring improvement on holistic ReID datasets. This is caused by that the most images holistic ReID datasets are well detected, vanilla holistic ReID methods is powerful enough to learn discrimintive features. Finally, we propose a adaptive direction graph convolutional (ADGC) layer which can suppress noisy features and a cross-graph embedded-alignment (CGEA) layer which can avoid hard one-to-one alignment. With the proposed ADGC and CGEA layers, our framework also achieves comparable performance on the two holistic ReID datasets. Specifically, we achieve about $94\%$ and $87\%$ Rank-1 scores on Market-1501 and DukeMTMC-reID datasets.

### 5.3. Model Analysis

**Analysis of Proposed Modules.** In this part, we analyze our proposed one-order semantic module ($\mathcal{S}$), high-

| Index | $\mathcal{S}$ | $\mathcal{R}$ | $\mathcal{T}$ | Rank-1 | mAP |
|---|---|---|---|---|---|
| 1 | × | × | × | 49.9 | 39.5 |
| 2 | ✓ | × | × | 52.4 | 42.8 |
| 3 | ✓ | ✓ | × | 53.9 | 43.2 |
| 4 | ✓ | ✓ | ✓ | **55.1** | **43.8** |

Table 5. Analysis of one-order semantic module ($\mathcal{S}$), high-order relation module ($\mathcal{R}$) and high-order human-topology module ($\mathcal{T}$). The experimental results show the effectiveness of our proposed three modules.

| NORM | ADGC | CGEA | Rank-1 | mAP |
|---|---|---|---|---|
| × | ✓ | ✓ | 52.5 | 40.4 |
| ✓ | × | ✓ | 53.7 | 42.2 |
| ✓ | ✓ | × | 54.4 | 43.5 |
| ✓ | ✓ | ✓ | **55.1** | **43.8** |

Table 6. Analysis of normalization of key-point confidences (NORM), adaptive direction graph convolutional (ADGC) layer and cross-grpah embedded-alignment (CGEA) layer. The experimental results show the effectiveness of our proposed layers.

order relation module ($\mathcal{R}$) and high-order human-topology module ($\mathcal{T}$). The experimental results are shown in Table 5. Firstly, in index-1, we remove all the three modules degrading our framework to an IDE model [43], where only a global feature $V_g$ is available. Its performance is unsatisfying and only achieves 49.9% Rank-1 score. Secondly, in index-2, when using one-order semantic information, the performance is improved by 2.5% and up to 52.4% Rank-1 score. This shows that the semantic information from keypoints is useful for learning and aligning features. Thirdly, in index-3, extra high-order relation information is added, and the performance is further improved by 1.5% achieving 53.9%. This demonstrates the effectiveness of our module $\mathcal{R}$. Finally, in index-4, our full framework achieves the best accuracy at 55.1% Rank-1 score, showing the the effectiveness of our module $\mathcal{T}$.

**Analysis of Proposed layers.** In this part, we further analyze normalization of key-point confidences (NORM), adaptive direction graph convolutional (ADGC) layer and cross-graph embedded-alignment (CGEA) layer, which are the key components of for semantic module ($\mathcal{S}$), relation module ($\mathcal{R}$) and topology module ($\mathcal{T}$). Specifically, when removing NORM, straightly use the original confidence score. When removing ADGC, in Eq.(3), we replace $A^{adj}$ with a fixed adjacency matrix linked like a human-topology. Thus, the relation module ($\mathcal{S}$) degrades to a vanilla GCN, which cannot suppress noise information. When removing CGEA, in Eq.(8), we replace $U_1$ and $U_2$ with a fully-connected matrix. That is, every node of graph 1 is connected to all nodes of graph 2. Then, the topology module ($\mathcal{T}$) contains no high-order human-topology information for feature alignment and degrades to a vanilla verification module. The experimental results are shown in Table 6. As we can see, when removing NORM, ADGC or CGEA, the performance significantly drop by 2.6%, 1.4% and 0.7% rank-1 scores. The experimental results show the effectiveness of our proposed NORM, ADGC and CGEA components.

**Analysis of Parameters.** We evaluate the effects of parameters in Eq.(13), *i.e.* $\gamma$ and $n$. The results are shown Figure 5, and the optimal setting is $\gamma = 0.5$ and $n = 8$. When analyzing one parameter, the other is fixed at the op-
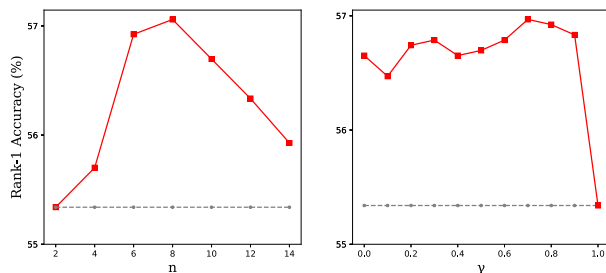


Figure 5. Analysis of parameters $\gamma$ and $n$ in Eq.(13). The optimal values are $\gamma = 0.5$ and $n = 8$. When analyze one of them, the other one is fixed as its optimal value. The The experimental results shows that our model is robust to different parameters.

timal value. It is clear that, when using different $\gamma$ and $n$, our model stably outperforms the baseline model. The experimental results show the our proposed framework is robust to different weights. Please note that the performances here are different from Table 2, where the former achieves 57% while and the latter 55%. This is because the latter is computed using average of 10 times for fair comparison.

## 6. Conclusion

In this paper, we propose a novel framework to learn high-order relation information for discriminative features and topology information for robust alignment. For learning relation information, we formulate local features of an image as nodes of a graph and propose an adaptive-direction graph convolutional (ADGC) layer to promote the message passing of semantic features and suppress that of meaningless and noisy ones. For learning topology information, we propose a cross-graph embedded-alignment (CGEA) layer conjugated with a verification loss, which can avoid sensitive hard one-to-one alignment and perform a robust soft alignment. Finally, extensive experiments on occluded, partial and holistic datasets demonstrate the effectiveness of our proposed framework.

## Acknowledge

# References

[1] Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vincius Flores Zambaldi, and Mateusz Malinowski. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

[2] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*, 2018.

[3] 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.

[4] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. Horizontal pyramid matching for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8295–8302, 2019.

[5] 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 *Advances in Neural Information Processing Systems*, pages 1222–1233, 2018.

[6] Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy. *Person Re-Identification*. 2014.

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

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

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

[10] Lingxiao He, Yinggang Wang, Wu Liu, Xingyu Liao, He Zhao, Zhenan Sun, and Jiashi Feng. Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification. *arXiv: Computer Vision and Pattern Recognition*, 2019.

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

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

[13] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *international conference on machine learning*, pages 448–456, 2015.

[14] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1062–1071, 2018.

[15] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[16] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2288–2295. IEEE, 2012.

[17] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10863–10872, 2018.

[18] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2197–2206, 2015.

[19] Shengcai Liao and Stan Z Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3685–3693, 2015.

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

[21] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4099–4108, 2018.

[22] 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, 2020.

[23] Hao Luo, Xing Fan, Chi Zhang, and Wei Jiang. Stnreid : Deep convolutional networks with pairwise spatial transformer networks for partial person re-identification. *arXiv preprint arXiv:1903.07072*, 2019.

[24] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[25] Bingpeng Ma, Yu Su, and Frederic Jurie. Covariance descriptor based on bio-inspired features for person re-identification and face verification. *Image and Vision Computing*, 32(6-7):379–390, 2014.

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

[27] Lei Qi, Jing Huo, Lei Wang, Yinghuan Shi, and Yang Gao. Maskreid: A mask based deep ranking neural network for person re-identification. *arXiv preprint arXiv:1804.03864*, 2018.

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

multi-target, multi-camera tracking. In *European Conference on Computer Vision*, pages 17–35. Springer, 2016.

[29] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 420–429, 2018.

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

[31] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3980–3989, 2017.

[32] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. *european conference on computer vision*, pages 418–437, 2018.

[33] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019.

[34] 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. pages 393–402, 2019.

[35] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 480–496, 2018.

[36] 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 *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[37] Guan-An Wang, Tianzhu Zhang, Yang Yang, Jian Cheng, Jianlong Chang, Xu Liang, and Zengguang Hou. Cross-modality paired-images generation for rgb-infrared person re-identification. In *AAAI-20 AAAI Conference on Artificial Intelligence*, 2020.

[38] Runzhong Wang, Junchi Yan, and Xiaokang Yang. Learning combinatorial embedding networks for deep graph matching. *arXiv preprint arXiv:1904.00597*, 2019.

[39] Yang Yang, Jimei Yang, Junjie Yan, Shengcai Liao, Dong Yi, and Stan Z Li. Salient color names for person re-identification. In *European conference on computer vision*, pages 536–551. Springer, 2014.

[40] Andrei Zanfir and Cristian Sminchisescu. Deep learning of graph matching. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2684–2693, 2018.

[41] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3239–3248, 2017.

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

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

[44] 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, 2013.

[45] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shaogang Gong. Partial person re-identification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4678–4686, 2015.

[46] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *arXiv preprint arXiv:1701.07717*, 2017.

[47] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017.

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

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