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

Guan’an Wang1*, Shuo Yang3*, Huanyu Liu2, Zhicheng Wang2, Yang Yang1, Shuliang Wang3, Gang Yu2, Erjin Zhou2 and Jian Sun2

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

{wangguanan2015@ia.ac.cn, liuhuanyu,wangzhicheng,yugang,zej,sunjian}@megvii.com

{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 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 (S), high-order relation module (R) and high-order human-topology module (T). (1) In the 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 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 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
alignment of the partial Re-ID tasks. Besides, several efforts are put on the spatial align with the partial ones, then calculate the distance of the spatial transform module to transform the holistic image to fluence of backgrounds clutter in [10]. Sun et al. prove it by a foreground-background mask to avoid the information. He et al. use a global-to-local matching model to capture the spatial layout 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 $S$ to extract semantic features of human key-point regions, a high-order relation module $R$ to model the relation-information among different semantic local features, and a high-order human-topology module $T$ 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 ($g(\cdot)$),

\[
\text{Inputs} \xrightarrow{\text{CNN Backbone}} \text{Feature Map} \xrightarrow{\text{Pose Estimator}} \text{Pose Estimator} \xrightarrow{S: \text{One-Order Semantic Module}} \xrightarrow{R: \text{High-Order Relation Module}} \xrightarrow{T: \text{High-Order Human-Topology Module}} \text{Outputs}
\]

Figure 2. Illustration of our proposed framework. It consists of an one-order semantic module $S$, a high-order relation module $R$ and a high-order topology module $T$. The module $S$ learns semantic local features of key-point regions. In $R$, 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 $T$, 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.
we can get a group of semantic local features of key-points regions $V^S_i$ and a global feature $V^S_g$. 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 experiment section.

$$
\begin{align*}
V^S_i &= \{v^S_k\}_{k=1}^K = g(m_{cnn} \otimes m_{kp}) \\
V^S_g &= v^S_{K+1} = g(m_{cnn}) 
\end{align*}
$$

(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, $p_{v^S_k}$ is the probability of feature $v^S_k$ belonging to its ground truth identity predicted by a classifier, $\alpha$ is a margin, $d_{v^S_{ak},v^S_{pk}}$ is the distance between a positive pair $(v^S_{ak},v^S_{pk})$ from the same identity, $(v^S_{ak},v^S_{pk})$ is from different identities. The classifiers for different local features are not shared.

$$
\mathcal{L}_S = \frac{1}{K+1} \sum_{k=1}^{K+1} \beta_k [\mathcal{L}_{cls}(v^S_k) + \mathcal{L}_{tri}(v^S_k)]
$$

\begin{align*}
\mathcal{L}_S &= \frac{1}{K+1} \sum_{k=1}^{K+1} \beta_k [-\log p_{v^S_k} + \alpha + d_{v^S_{ak},v^S_{pk}} - d_{v^S_{ak},v^S_{nk}}] \\
\end{align*}

(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, 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 $V_i$, and a pre-defined graph (adjacent matrix is $A$). We use differences between local features $V_i$ 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 $V_i$ and $A^{adp}$. To stabilize training, we fuse the input local features $V_i$ 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.

$$
V^{out} = [f_1(A^{adp} \otimes V^R_{in}) + f_2(V^R_{in}), V^R_{g}]
$$

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 $V^S = \{v^S_k\}_{k=1}^{K+1}$ via Eq.(1). Then its relation features $V^R = \{v^R_k\}_{k=1}^{K+1}$ can be formulated as below:

$$
V^R = f_R(V^S)
$$

(4)

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

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

(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 \(V_2^R = \{v_{2k}^R\}_{k=1}^{K+1}\) via Eq.(4), and calculate their similarity with cosine distance as in Eq.(6).

\[
s^R_{x_1,x_2} = \frac{1}{K+1} \sum_{k=1}^{K+1} \sqrt{\beta_{1k} \beta_{2k}} \cosine(v_{1k}^R, v_{2k}^R) \tag{6}
\]

### 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 \(M \in R^{K \times K}\) 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:

\[
U^* = \arg\max_U U^T M U, \text{ s.t. } ||U|| = 1 \tag{7}
\]

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 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 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 \(V_{in}^1 \in R^{(K+1) \times C_{in}}\) and \(V_{in}^2 \in R^{(K+1) \times C_{in}}\), we embed them to a hidden space with a fully-connected layer and a ReLU layer, getting two groups of hidden features \(V^h_1 \in R^{(K+1) \times C_{out}}\) and \(V^h_2 \in R^{(K+1) \times C_{out}}\). Secondly, we perform graph matching between \(V^h_1\) and \(V^h_2\) via Eq.(7), and get an affinity matrix \(U^{k \times k}\) between \(V^h_1\) and \(V^h_2\). Here, \(U(i,j)\) means correspondence between \(v^h_1_i\) and \(v^h_2_j\). Finally, the output can be formulated in Eq.(8), where \([\cdot, \cdot]\) means concatenation operation along channel dimension, \(f\) is a fully-connected layer.

\[
\begin{align*}
V^1_{out} &= f([V^h_1, U \otimes V^h_2]) + V^h_1 \\
V^2_{out} &= f([V^h_2, U^T \otimes V^h_1]) + V^h_2 
\end{align*} \tag{8}
\]

We implement our high-order topology module \((T)\) 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 \((V^R_1, V^R_2)\) via Eq.(4), and then their topology features of \((V^T_1, V^T_2)\) via Eq.(9). After getting the topology features pair \((V^T_1, V^T_2)\), we can compute their similarity using Eq.(10), where \(|\cdot|\) is element-wise absolute operation, \(f_s\) is a fully-connected layer from \(C_T\) to 1, \(\sigma\) is sigmoid activation function.

\[
\begin{align*}
(V^T_1, V^T_2) &= F_T(V^R_1, V^R_2) \\
s^T_{x_1, x_2} &= \sigma(f_s(-|V^T_1 - V^T_2|)) 
\end{align*} \tag{9, 10}
\]

**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 \(y = 0\).

\[
\mathcal{L}_T = y \log s^T_{x_1, x_2} + (1 - y) \log (1 - s^T_{x_1, x_2}) \tag{11}
\]

### 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 $\mathcal{L}$.

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

(12)

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

$$s = \gamma s_{x_1,x_2}^R + (1 - \gamma) s_{x_1,x_2}^T$$

(13)

When inferring, given an query image $x_q$, we first compute its similarity $x_q^R$ 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 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-

\(^1\)https://github.com/wangguanana/HOReID

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

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

<table>
<thead>
<tr>
<th>Methods</th>
<th>Occluded-Duke Rank-1 mAP</th>
<th>Occluded-ReID Rank-1 mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part-Aligned</td>
<td>28.8</td>
<td>-</td>
</tr>
<tr>
<td>PCB</td>
<td>42.6</td>
<td>41.3</td>
</tr>
<tr>
<td>Part Bilinear</td>
<td>36.9</td>
<td>-</td>
</tr>
<tr>
<td>FD-GAN</td>
<td>40.8</td>
<td>-</td>
</tr>
<tr>
<td>AMC+SWM</td>
<td>-</td>
<td>31.2</td>
</tr>
<tr>
<td>DSR</td>
<td>40.8</td>
<td>72.8</td>
</tr>
<tr>
<td>SFR</td>
<td>42.3</td>
<td>68.0</td>
</tr>
<tr>
<td>Ad-Occluded</td>
<td>44.5</td>
<td>-</td>
</tr>
<tr>
<td>TCSDO</td>
<td>-</td>
<td>73.7</td>
</tr>
<tr>
<td>FPR</td>
<td>-</td>
<td>78.3</td>
</tr>
<tr>
<td>PGF</td>
<td>51.4</td>
<td>-</td>
</tr>
<tr>
<td>HOReID (Ours)</td>
<td>55.1</td>
<td>80.3</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Methods</th>
<th>Partial-REID Rank-1</th>
<th>Partial-REID Rank-3</th>
<th>Partial-ilLIDS Rank-1</th>
<th>Partial-ilLIDS Rank-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSR [8]</td>
<td>50.7</td>
<td>70.0</td>
<td>58.8</td>
<td>67.2</td>
</tr>
<tr>
<td>SFR [9]</td>
<td>56.9</td>
<td>78.5</td>
<td>63.9</td>
<td>74.8</td>
</tr>
<tr>
<td>VPM [34]</td>
<td>67.7</td>
<td>81.9</td>
<td>65.5</td>
<td>74.8</td>
</tr>
<tr>
<td>PGFA [26]</td>
<td>68.0</td>
<td>80.0</td>
<td>69.1</td>
<td>80.9</td>
</tr>
<tr>
<td>AFPB [48]</td>
<td>78.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FPR [10]</td>
<td>81.0</td>
<td>-</td>
<td>68.1</td>
<td>-</td>
</tr>
<tr>
<td>TCSDO [49]</td>
<td>82.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>HORelID(Ours)</strong></td>
<td><strong>85.3</strong></td>
<td><strong>91.0</strong></td>
<td><strong>72.6</strong></td>
<td><strong>86.4</strong></td>
</tr>
</tbody>
</table>

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

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

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-ilLIDS [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-ilLIDS 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 fail 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 DukeMTMC-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 as human-parsing and key-points information perform worse. For example, SPReID [14] uses human-parsing information and only achieves 92.5% Rank-1 score on Market-1501 dataset. PGFA [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 discriminative 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 (S), high-
order relation module ($R$) and high-order human-topology module ($T$). 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 $R$. Finally, in index-4, our full framework achieves the best accuracy at 55.1% Rank-1 score, showing the effectiveness of our module $T$.

**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. The experimental results show the effectiveness of our proposed three modules.

**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 optimal 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 parameters. 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 meaning-less 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.

**Acknowledgment**

This research was supported by National Key R&D Program of China (No. 2017YFA0700800).

---

<table>
<thead>
<tr>
<th>Index</th>
<th>$S$</th>
<th>$R$</th>
<th>$T$</th>
<th>Rank-1</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>49.9</td>
<td>39.5</td>
</tr>
<tr>
<td>2</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>52.4</td>
<td>42.8</td>
</tr>
<tr>
<td>3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>53.9</td>
<td>43.2</td>
</tr>
<tr>
<td>4</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>55.1</td>
<td>43.8</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>NORM</th>
<th>ADGC</th>
<th>CGEA</th>
<th>Rank-1</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>52.5</td>
<td>40.4</td>
</tr>
<tr>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>53.7</td>
<td>42.2</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>54.4</td>
<td>43.5</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>55.1</td>
<td>43.8</td>
</tr>
</tbody>
</table>

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

![Figure 5. Analysis of parameters $\gamma$ and $n$ in Eq.(13).](image)
References


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


[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.


