Learning Feature Recovery Transformer for Occluded Person Re-Identification

Boqiang Xu, Lingxiao He, *Member, IEEE*, Jian Liang, *Member, IEEE*, and Zhenan Sun, *Senior Member, IEEE*

***Abstract*— One major issue that challenges person re-identification (Re-ID) is the ubiquitous occlusion over the captured persons. There are two main challenges for the occluded person Re-ID problem, *i.e.*, the interference of noise during feature matching and the loss of pedestrian information brought by the occlusions. In this paper, we propose a new approach called Feature Recovery Transformer (FRT) to address the two challenges simultaneously, which mainly consists of visibility graph matching and feature recovery transformer. To reduce the interference of the noise during feature matching, we mainly focus on visible regions that appear in both images and develop a visibility graph to calculate the similarity. In terms of the second challenge, based on the developed graph similarity, for each query image, we propose a recovery transformer that exploits the feature sets of its k-nearest neighbors in the gallery to recover the complete features. Extensive experiments across different person Re-ID datasets, including occluded, partial and holistic datasets, demonstrate the effectiveness of FRT. Specifically, FRT significantly outperforms state-of-the-art results by at least 6.2% Rank-1 accuracy and 7.2% mAP scores on the challenging Occluded-Duke dataset.**

***Index Terms*— Occluded person re-identification, transformer, graph, occlusion recovery.**

I. Introduction

PERSON re-identification (Re-ID) [1]–[3] aims to retrieve the same person from overlapping cameras, which is widely used in security, video surveillance and smart city. Recently, considerable Re-ID methods have been proposed in this field [4]–[7]. However, most of them rely on a strong assumption that the entire body of the pedestrian is available, however, this is not always the case in practice. As shown in Fig. 1(a), in realistic Re-ID systems, people are always occluded by some obstacles especially in crowded places such as malls, railway stations and airports. Thus, it is necessary to study the occluded person Re-ID problem [8].

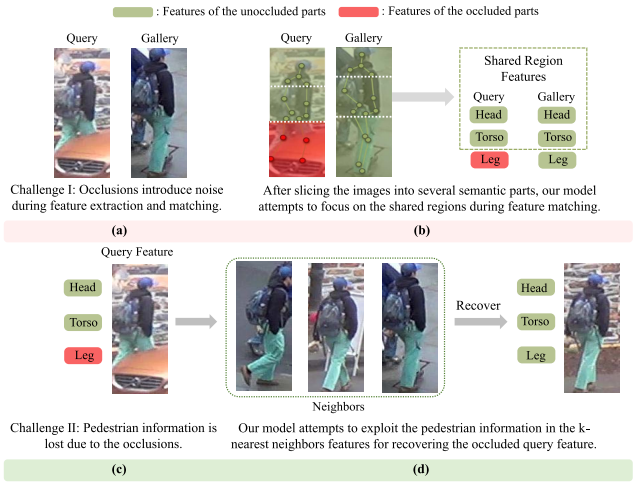


Fig. 1. Two challenges in the occluded Re-ID problem and our solutions. The figure is only for illustration, our method is processed in the feature level.

There are two main challenges for the occluded person Re-ID problem. Firstly, as shown in Fig. 1(a), occlusions always bring noise during feature extraction and feature matching. Secondly, as shown in Fig. 1(b), the pedestrian information in the occluded regions is always lost, making the extracted features not discriminative anymore. Recently, some occluded person Re-ID methods have been proposed [9]–[12] for the first challenge. They try to use key-points [9], [10] or probability maps [11], [12] for feature alignment and increase the robustness of the representations. A recent work [13] views key-points as nodes to construct a graph for learning the human-topology information and has proved the effectiveness of the graph for solving the occluded person Re-ID problem. For the second challenge, some researches [14]–[16] attempt to predict the occluded parts in the images with GANs. However, occluded regions generation is not so convincing, especially when the occlusion is serious. Therefore, the performance gains of these methods are very limited.

In this work, we propose a novel framework named Feature Recovery Transformer (FRT) to address the two challenges simultaneously. Specifically, the proposed framework mainly consists of three phases, semantic feature extraction, visibility graph matching, and occluded feature recovery. In the first phase, we employ the pose information to extract semantic

Manuscript received 12 January 2022; revised 21 May 2022; accepted 17 June 2022. Date of publication 4 July 2022; date of current version 12 July 2022. This work was supported in part by the National Natural Science Foundation of China under Grant U1836217 and in part by the Beijing Nova Program under Grant Z211100002121108. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Lucio Marcenaro. *(Corresponding author: Jian Liang.)*

Boqiang Xu, Jian Liang, and Zhenan Sun are with the Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (email: boqiang.xu@cripac.ia.ac.cn; liangjian92@gmail.com; znsun@nlpr.ia.ac.cn).

Lingxiao He is with the AI Research of JD, Beijing 100020, China (e-mail: helingxiao3@jd.com).

Digital Object Identifier 10.1109/TIP.2022.3186759

features (*i.e.*, global feature, head feature, torso feature and leg feature) and calculate the visibility score for the corresponding regions. In the second feature matching phase, we construct a directional graph with each node containing the semantic features from the same part within a pair of images. The weights of edges are determined by the visibility score of the starting node. As shown in Fig. 1(a), by promoting the flow of messages in the shared regions (which regions are visible in both images), the visibility graph pays more attention to these areas during feature matching, which is not sensitive to the occlusions. Based on the similarity computed by visibility graph matching, we readily obtain its k-nearest neighbors in the gallery for each query. In the third occluded feature recovery phase, as shown in Fig. 1(d), different from other methods [14]–[16] which use GANs to predict the occluded parts, we propose a Feature Recovery Transformer (FRT) to exploit the pedestrian information in its k-nearest neighbors features for occluded feature recovery. FRT considers the local information of each semantic feature in the k-nearest neighbors, including position, visibility score and similarity between the query. By this way, FRT is able to filter out noise in the k-nearest neighbors features and exploit valuable information to recover the occluded query feature. Finally, the recovered query feature is used for person Re-ID. Extensive experiments validate the consistent superiority of our FRT over prior state-of-the-art methods. Specifically, FRT outperforms state-of-the-art results by at least 6.2% Rank-1 accuracy and 7.2% mAP scores on the challenging Occluded-Duke dataset [10].

The main contributions of this paper are summarized as follows:

- We propose a Feature Recovery Transformer (FRT) to exploit the pedestrian information in the features of k-nearest neighbors for occluded feature recovery. Compared to other methods [14]–[16] which use GANs to predict the occluded parts, our approach is more convincing and could bring much more improvements to the occluded person Re-ID performance.
- We propose a novel visibility graph to learn the human-topology information among body parts, which is able to promote the information in the shared regions and suppress the noisy message in the occluded parts.
- Extensive experiments on occluded, partial and holistic Re-ID datasets validate the effectiveness of our method for solving occluded person Re-ID problem.

II. Related Work

A. Occluded and Partial Person Re-Identification

Occluded [10] and partial person re-identification [17] aim to find the same person, who is occluded or partially detected in the query image, from dis-joint cameras. These two problems are usually studied as the same issue in research. There are two main challenges for the occluded person Re-ID problem. Firstly, occlusions always bring noise during feature extraction and feature matching. Secondly, the pedestrian information in the occluded regions is always lost, making the extracted features not discriminative anymore. Recently, some methods have been proposed for the first challenge [10]–[13], [18]–[21]. Miao *et al.* [10] propose a feature alignment method based on the semantic key-points. In addition, they design a matching strategy to calculate the distance of representations in an unoccluded region. He *et al.* [18] propose a reconstruction method for soft feature alignment and further introduce foreground-background mask to avoid the influence of backgrounds in [11]. Sun *et al.* [12] propose a Visibility-aware Part Model (VPM) to learn to perceive the visibility of regions through self-supervision. Wang *et al.* [13] utilize GCN, which considers different key-points as nodes, to embed the high-order information between various body parts. Zheng *et al.* [19] propose a Guided Feature Learning with Knowledge Distillation (PGFL-KD) network to learn aligned representations of different body parts. Benefiting from the knowledge distillation and interaction-based learning, the pose estimator could be discarded in testing. Chen *et al.* [20] propose an Occlusion Aware Mask Network (OAMN), which incorporates an attention-guided mask module to extract features of body parts precisely regardless of the occlusion. Yang *et al.* [21] propose to discretize pose information to the visibility label of body parts for reducing the interference of noisy pose information in the occluded Re-ID problem. Zhang *et al.* [22] and Jia *et al.* [23] attempt to extract semantically aligned features and eliminate occlusion noises for solving the occluded Re-ID problem. Tan *et al.* [24] propose a Multi-Head Self-Attention Network (MHSA-Net) to prune noise and capture key local information from images for occluded Re-ID. Despite the promising results achieved in occlude Re-ID, all these methods ignore the lost pedestrian information in the occluded regions. To solve this problem, [14]–[16] attempt to predict the occluded parts in the images with GANs for the occlusion recovery. However, occluded regions generation is not so convincing, especially when the occlusion is serious. Therefore, the performance gains of these methods [14]–[16] are very limited. Although Hou *et al.* [25] propose a Spatio-Temporal Completion network (STCnet) for recovering the appearance of the occluded parts with spatial and temporal information for video person reid, temporal information is unavailable in image-based occluded Re-ID. Different from previous methods, our method attempts to filter out noise in the k-nearest neighbors and fuse the query feature with valuable information in the k-nearest neighbors for occluded feature recovery.

B. Transformer

Vaswani *et al.* [26] propose the Transformer to dispense the recurrence and convolutions involved in the encoding step entirely. Transformer only relies on attention mechanisms to capture the global relations between input and output for transduction problems such as machine translation and language modeling [27]–[29]. Some methods have tried to exploit Transformer in computer vision tasks such as image processing [30], object detection [31], semantic segmentation [32], feature matching [33], etc. For example, Chen *et al.* [30] propose Image Processing Transformer (IPT) for utilizing large-scale pre-training and achieves the state-of-the-art performance on

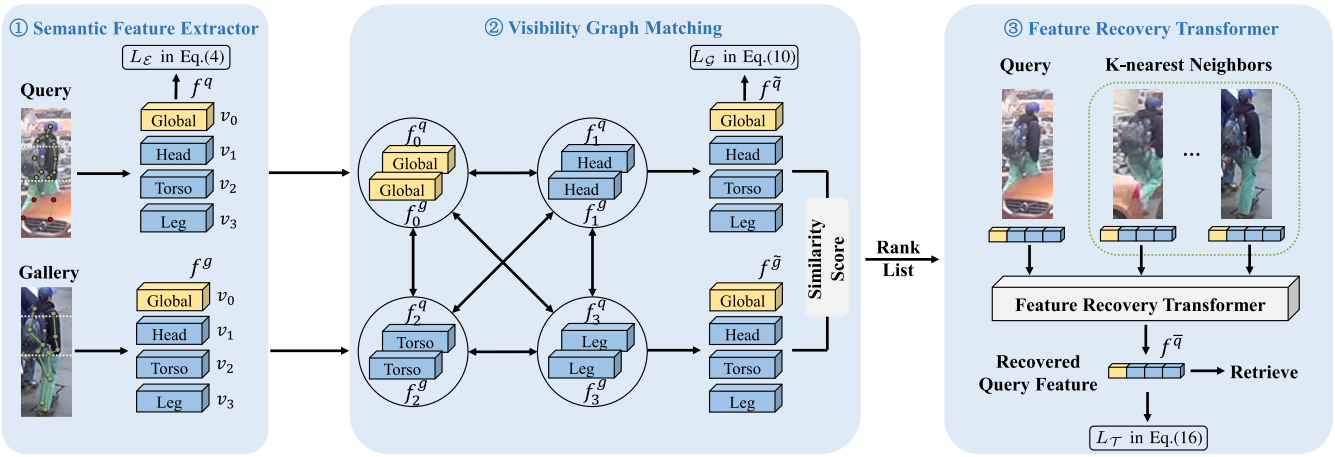


Fig. 2. Overview of the proposed framework. It consists of semantic feature extractor $\mathcal{E}$, visibility graph matching $\mathcal{G}$ and feature recovery transformer $\mathcal{T}$. In $\mathcal{E}$ we utilize key-points to extract semantic features and calculate the visibility score $\{v_i\}_{i=0}^{3}$ for each part. In $\mathcal{G}$, we input the features extracted by $\mathcal{E}$ and consider the same semantic features within a pair of images as nodes of a graph to calculate the similarity scores. According to the feature matching by $\mathcal{G}$, a rank list is produced for each query. In $\mathcal{T}$, we input the query feature and features of its k-nearest neighbors for recovering the occluded query feature. The recovered query feature is then utilized for retrieving.

several image processing tasks like denoising, de-raining and super-resolution. Sarlin *et al.* [33] incorporate Transformer to establish pointwise correspondences between a pair of images for feature matching. They utilize self- (intra-image) and cross- (inter-image) attention to simulate the procedure that humans look back-and-forth at two images when matching them. Recently, Li *et al.* [34] try to solve the occluded Re-ID problem by the transformer encoder-decoder architecture and propose a Part-Aware Transformer (PAT). PAT works on precisely extracting features of visible body parts by a pixel context based transformer encoder and a part prototype based transformer decoder. Different from [34], we utilize Transformer to exploit the pedestrian information in the k-nearest neighbors for recovering the occluded query features.

C. Graph Convolutional Network

Graph Convolutional Networks (GCN) is firstly proposed in [35] to build the relationship between graph nodes, and has been proved to be effective in many computer vision tasks [36]–[38]. Recently, Re-ID methods combined with GCN have also been explored [13], [39], [40]. Wang *et al.* [13] construct the graph based on the visibility of key-points intra image, and take advantage of the affinities between various key-points. Cheng *et al.* [39] formulate the structured distance into the graph Laplacian form to consider the relationships among training samples. Yan *et al.* [40] attempt to solve the person search by considering the context information with GCN.

III. Our Approach

This section introduces our proposed framework, including 1) Semantic feature extractor ($\mathcal{E}$) to extract the semantic features with pose assistance and calculate visibility scores for them; 2) Visibility graph matching ($\mathcal{G}$) to promote the information in the shared regions and learn the similarity; 3) Feature recovery transformer ($\mathcal{T}$) to recover the occluded query features with pedestrian information in the features of k-nearest neighbors. An overview of the proposed method is shown in Fig. 2.

A. Semantic Feature Extractor

The semantic feature extractor ($\mathcal{E}$) is demonstrated in Fig. 2. The module $\mathcal{E}$ is inspired by two cues. Firstly, part-based models have been proved to be effective for person re-identification task as they can employ both global and fine-grained local features [5]. Secondly, occlusions always cause spatial misalignment during feature matching. Therefore, accurate feature alignment is necessary for occluded Re-ID [11], [12], [18]. Following the ideas above, we use HR-Net [41] pre-trained on the COCO dataset [42] for pose estimation. The model predicts 12 key-points, including shoulders, elbows, wrists, hips, knees and ankles. We exploit the pose information to divide the person image into three parts: head part, torso part and leg part. Then, the local features of these three parts together with the global feature are extracted for alignment. Additionally, we calculate the visibility score for each part as follows:

$$v_i = \frac{\sum_{s_n \in \mathbb{R}_i} s_n}{|\mathbb{R}_i|}, \quad i = 0, 1, 2, 3, \tag{1}$$

where v_0, v_1, v_2, v_3 are the visibility scores for the global, head, torso and leg regions respectively. $\{\mathbb{R}_i\}_{i=0}^{3}$ is the set of key-points in the region i and $|\cdot|$ denotes the number of elements in the set. s_n is the confidence score calculated by the pose estimator for the n-th key-point. The visibility score is calculated by the average confidence scores for all the key-points in the corresponding region. For example, we firstly use pose estimator to detect 6 key-points in the torso region and calculate their confidence scores by the pose estimator. Then the visibility score of the torso region is calculated by the average confidence scores for these 6 key-points. Furthermore,

we set a threshold δ to determine whether the region $\mathbb{R}_i$ is completely occluded. If the visibility score v_i is smaller than the δ, we would regard the region $\mathbb{R}_i$ as fully occluded and set the feature of region $\mathbb{R}_i$ to zero.

Training Loss: To train the module $\mathcal{E}$, we use cross-entropy loss $\mathcal{L}^{\mathcal{E}}_{cross}$ and triplet $\mathcal{L}^{\mathcal{E}}_{tri}$ for all the semantic features (i.e. global, head, torso and leg feature) as follows:

$$\mathcal{L}^{\mathcal{E}}_{cross} = -\sum_{j=1}^{N} \log \frac{\exp(W^{\mathcal{E}}_{y_j} f_j + b^{\mathcal{E}}_{y_j})}{\sum_{k=1}^{C} \exp(W^{\mathcal{E}}_{k} f_j + b^{\mathcal{E}}_{k})}, \tag{2}$$

$$\mathcal{L}^{\mathcal{E}}_{tri} = |\theta_{\mathcal{E}} + d_{f^q_j, f^P_j} - d_{f^q_j, f^N_j}|_+, \tag{3}$$

$$\mathcal{L}_{\mathcal{E}} = \mathcal{L}^{\mathcal{E}}_{cross} + \mathcal{L}^{\mathcal{E}}_{tri}, \tag{4}$$

where N is the number of images in a mini-batch, C is the number of classes and y_j is the label for the feature f_j. W_k and b_k are the weights and bias of classifier for the k-th class, respectively. $d_{f^q_j, f^P_j}$ and $d_{f^q_j, f^N_j}$ are the distance between a positive pair (f^q_j, f^P_j) from the same identity and a negative pair (f^q_j, f^N_j) from different identities, respectively. $\theta_{\mathcal{E}}$ is a hyper-parameter to control the margin between the negative and positive pairs in feature space. Especially, we only utilize the $\mathcal{L}_{\mathcal{E}}$ to monitor the feature of the part that is not completely occluded. The reason for this is that monitoring the feature of the completely occluded regions, which is manually set to zero, would confuse the classifiers.

B. Visibility Graph Matching

Although we have obtained the aligned pedestrian representations, occluded Re-ID is still challenging due to the interference of the occlusions during feature matching. Thus, it is necessary to suppress the meaningless message of occluded parts and enhance the meaningful features of shared regions. We resort to the graph convolutional network (GCN) [43] which is effective in message propagation and aggregation. As shown in Fig. 2, given two images q and g, their feature maps $\{f^q_i\}^3_{i=0}$and $\{f^g_i\}^3_{i=0}$ along with their corresponding visibility scores $\{v^q_i\}^3_{i=0}$ and $\{v^g_i\}^3_{i=0}$ could be extracted and calculated by the module $\mathcal{E}$ above. We aim to construct a graph which could focus on the shared regions when matching features.

1) Visibility Graph Building: In particular, considering a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, which consists of 4 vertices $\mathcal{V}$ and a set of edges $\mathcal{E}$ as shown in Fig. 2, we assign corresponding features $\{f^q_i, f^g_i\}^3_{i=0}$ to each node. We use $\mathbf{A} \in \mathbb{R}^{4\times4}$ to denote the adjacent matrix. The adjacent matrix $\mathbf{A}$ is set as follows:

$$A_{i,j} = \begin{cases} 1, & i = j \\ \varphi_i + \mathbb{1}(\varphi_i - \Gamma)[1 - cosine(f^q_i, f^g_i)], & otherwise \end{cases} \tag{5}$$

$$\varphi_i = \min\{v^q_i, v^g_i\}, \tag{6}$$

where $A_{i,j}$ indicates the information propagated from node i to node j, φ_i denotes the shared visibility of i *th* part, Γ is a margin and $\mathbb{1}(\cdot)$ is the indicator function. A high valued φ_i denotes that part i is a shared region while a small valued φ_i means that at least one of i *th* parts of the image q and g is occluded. The first term φ_i in Eq. 5 indicates that the lower the shared visibility φ_i is, the less information of node i is spread. The second term $\mathbb{1}(\varphi_i - \Gamma)[1 - cosine(f^q_i, f^g_i)]$ indicates that when the shared visibility φ_i is larger than Γ, the greater the difference between f^q_i and f^g_i, the more their message will be propagated. This is designed to enhance the comparisons of the shared regions.

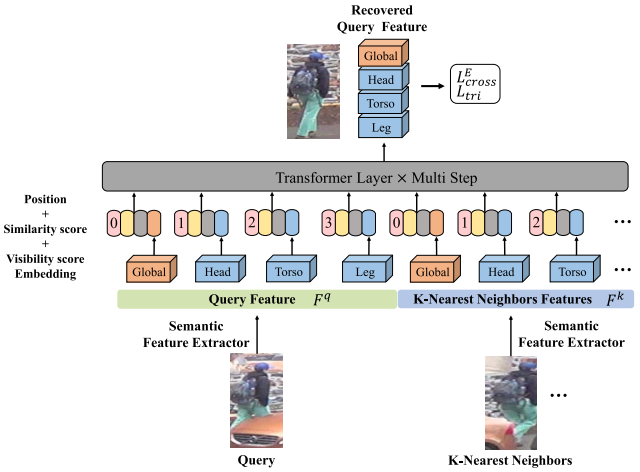


Fig. 3. Illustration of the proposed feature recovery transformer ($\mathcal{T}$). F^q, F^k are query feature and k-nearest neighbors features respectively.

2) Message Propagation: Following GCN [33], we denote $m^{(l)}_i$ the message aggregated from all nodes to the i *th* node at layer l, which can be defined as:

$$m^{(l)}_i = \sigma(\widehat{A} f^{(l)}_i W^{(l)}_m), \tag{7}$$

where $\widehat{A}$ is the normalized adjacency matrix, $W^{(l)}_m$ is a learnable parameter matrix, $f^{(l)}_i$ represents an element from $\{f^{q(l)}_i, f^{g(l)}_i\}$ and σ is the ReLU activation function. Furthermore, we then use the residual message passing to update all the nodes by:

$$f^{(l+1)}_i = f^{(l)}_i + W^{(l)}_r[f^{(l)}_i, m^{(l)}_i], \tag{8}$$

where $[\cdot,\cdot]$ denote concatenation and $W^{(l)}_r$ is a learnable parameter matrix.

3) Feature Matching: After message propagation, we obtain the updated features $f^{\tilde{q}}$ and $f^{\tilde{g}}$. Then, the cosine distance is used to calculate the similarity score as follows:

$$s_{q,g} = cosine(f^{\tilde{q}}, f^{\tilde{g}}). \tag{9}$$

According to the feature matching by $\mathcal{G}$, a rank list is produced for each query.

4) Training Loss: We use triplet and classification losses to monitor module $\mathcal{G}$ as in Eq. 10. The definition of $\mathcal{L}^{\mathcal{G}}_{cross}$ and $\mathcal{L}^{\mathcal{G}}_{tri}$ is the same as Eq. 2 and Eq. 3 respectively.

$$\mathcal{L}_{\mathcal{G}} = \mathcal{L}^{\mathcal{G}}_{cross} + \mathcal{L}^{\mathcal{G}}_{tri}. \tag{10}$$

C. Feature Recovery Transformer

Even so, learned features still suffer from loss of pedestrian information caused by the occlusions. To solve this problem,

we propose feature recovery transformer ($\mathcal{T}$) for recovering the occluded features. In the gallery, there possesses lots of pedestrian information, which hides the cues about complete features recovery. Inspired by the success of Transformer [26] and local feature matching [33], we want to employ the attention mechanism of Transformer to aggregate the pedestrian information in the k-nearest neighbors features for recovering the occluded query feature.

Although both the Feature Recovery Transformer and re-ranking strategy [44] utilize the k-nearest neighbors information, we emphasize that the motivation and implementation are different. Re-ranking [44] works on re-calculating the distance in the k-nearest neighbors to re-rank the retrieval results. Feature Recovery Transformer focuses on recovering the occluded query feature by filtering out noise in the k-nearest neighbors and fusing the query feature with valuable information in the k-nearest neighbors features.

1) Local Information Embedding: The feature recovery transformer is illustrated in Fig. 3. We input the concatenation of query feature and its k-nearest neighbors features to the $\mathcal{T}$. For each part feature f, we embed its position, similarity score between the query and visibility score into a high-dimensional vector with a Multilayer Perceptron (MLP) as:

$$f = f + MLP(p, cos, v), \tag{11}$$

where p is the position for f, $p = 0, 1, 2, 3$ stand for the global, head, torso and leg part respectively, cos is the cosine distance between the pedestrian feature and query feature and v is the visibility score for f. This embedding enables $\mathcal{T}$ to consider the local information of each part feature during the query feature recovery.

2) Transformer Layer: The transformer layer works on aggregating the information from k-nearest neighbors features to recover query feature. Inspired by Transformer [26], the key, query and value can be calculated by:

$$\begin{aligned} q^{(l)} &= W_1^{(l)} f^{q(l)} + b_1^{(l)}, \\ \begin{bmatrix} k \\ s \end{bmatrix}^{(l)} &= \begin{bmatrix} W_2 \\ W_3 \end{bmatrix}^{(l)} f^k + \begin{bmatrix} b_2 \\ b_3 \end{bmatrix}^{(l)}, \end{aligned} \tag{12}$$

where, q, k, s, f^q, f^k indicate query, key, value, query feature and k-nearest neighbors features respectively and l is the layer number. Each layer l has its own learnable projection matrix, shared for all part features f^q and f^k. Then, the message propagated to f^q is computed as:

$$m_{f^{t(l)} \to f^{q(l)}} = \sum_{f^k \in F^k} Softmax(q^{(l)\top} k^{(l)}) s^{(l)}. \tag{13}$$

The final projection is a linear projection:

$$f^{\overline{q}} = W_{final} f^{q(L)} + b, \tag{14}$$

where $f^{\overline{q}}$ is the recovered query representation for better retrieval. It is worth noting that, the k-nearest neighbors features are not updated during the whole process. Restricted by pages, we do not show more details of Transformer, please refer to the paper [26], [33]. In addition, we employ multi-step mechanism as follows:

$$f^{\overline{q}(s)} = \mathcal{T}_s \mathcal{T}_{s-1} \ldots \mathcal{T}_0 (f^{\overline{q}(0)}, f^{\overline{q}(0)}), \tag{15}$$

TABLE I

DATASETS DETAILS. WE EVALUATE OUR METHOD ON 6 PUBLIC DATASETS, INCLUDING 2 OCCLUDED, 2 PARTIAL AND 2 HOLISTIC ONES. OCCLUDED-REID, PARTIAL-REID AND PARTIAL-ILIDS DATASETS ADOPT CROSS-DOMAIN SETTING

Dataset	Nums (ID/Image)		
	Training	Query	Gallery
Occluded-Duke [10]	702/15,618	519/2210	1,110/17,661
Occluded-ReID [8]	-	200/1,000	200/1,000
Partial-REID [17]	-	60/300	60/300
Partial-iLIDS [18]	-	119/119	119/119
Market-1501 [47]	751/12,936	750/3,368	750/19,732
DukeMTMC-reID [45]	702/16,522	702/2,228	1,110/17,661

where s denotes conduct $\mathcal{T}$ for s times. The recovered query feature $f^{\overline{q}(s)}$ is then utilized for retrieval.

3) Training Loss: The $\mathcal{T}$ only recover the query features and would leave the k-nearest neighbors features unchanged. Therefore, we use the classifier $\mathcal{L}_{cross}^{\mathcal{E}}$ designed for $\mathcal{E}$ to train the $\mathcal{T}$, and freeze $\mathcal{L}_{cross}^{\mathcal{E}}$ during the whole training process of $\mathcal{T}$ to ensure that the gallery features and recovered query features are in the same feature space. The training loss of the module $\mathcal{T}$ can be defined as:

$$\mathcal{L}_{\mathcal{T}} = \mathcal{L}_{cross}^{\mathcal{E}} + \mathcal{L}_{tri}^{\mathcal{T}}, \tag{16}$$

where $\mathcal{L}_{cross}^{\mathcal{E}}$ is the classifier designed for the module $\mathcal{E}$ and the definition of $\mathcal{L}_{tri}^{\mathcal{T}}$ is the same as Eq. 3.

IV. EXPERIMENTS

A. Setup

1) Datasets: We evaluate our method on the following six datasets and compare it with the state-of-the-art methods, the details of the datasets are illustrated in Table I. 1) The Occluded-Duke dataset [10] is derived from DukeMTMC-reID [45] by filtering out some overlap pictures and leaving occluded images. It consists of 15, 618 training images, 2, 210 occluded query images and 17, 661 images in gallery. 2) The Occluded-ReID dataset [8] contains 1000 occluded query images and 1000 full-body gallery pictures. 3) The Partial-REID dataset [17] includes 600 images from 60 people in test set, with five full-body images and five partial images for each person. 4) Partial-iLIDS dataset [18] is selected from iLIDS [46]. It contains 238 occluded images from 119 people captured in the airport. Specifically, the Occluded-ReID [8], Partial-REID [17] and Partial-iLIDS dataset [18] only contain test sets, following [11], [13], the Market-1501 [47] is used for training. 5) The Market-1501 [47] dataset consists of 32,688 images of 1,501 subjects captured by six cameras, and only few of occluded or partial person images are included. 6) DukeMTMC-reID [45] is a holistic dataset which contains 1,404 identities, 16,522 training images, 2,228 queries, and 17,661 gallery images.

2) Training Details: Our baseline is built based on the open-source project "fastreid" [53]. We resize all the training images into 384×128. We set the number of feature channels c to 512 and batch size N to 64. Following the work of [54],

TABLE II

PERFORMANCE (%) COMPARISONS WITH THE STATE-OF-THE-ART METHODS ON THE TWO OCCLUDED DATASETS, *i.e.*, OCCLUDED-DUKE [10] AND OCCLUDED-REID [8]. OUR METHOD ACHIEVES THE BEST PERFORMANCE ON THE OCCLUDED-DUKE DATASET. THE BEST PERFORMANCE IS HIGHLIGHTED IN BOLD

Methods	Reference	Occluded-Duke		Occluded-REID		Average	
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
Part-Aligned [6]	ICCV 2017	28.8	20.2	-	-	-	-
PCB [5]	ECCV 2018	42.6	33.7	41.3	38.9	42.0	36.3
Part Bilinear [48]	ECCV 2018	36.9	-	-	-	-	-
FD-GAN [49]	NIPS 2018	40.8	-	-	-	-	-
DSR [18]	CVPR 2018	40.8	30.4	72.8	62.8	56.8	46.6
Ad-Occluded [50]	CVPR 2018	44.5	32.2	-	-	-	-
PGFA [10]	ICCV 2019	51.4	37.3	-	-	-	-
HOReID [13]	CVPR 2020	55.1	43.8	80.3	70.2	67.7	57.0
PVPM [9]	CVPR 2020	47.0	37.7	70.4	61.2	58.7	49.5
Pirt [51]	ACM MM 2021	60.0	50.9	-	-	-	-
PGFA-KD [19]	ACM MM 2021	63.0	54.1	80.7	70.3	71.9	62.2
Yang et al. [21]	ICCV 2021	62.2	46.3	81.0	71.0	71.6	58.7
OAMN [20]	ICCV 2021	62.6	46.1	-	-	-	-
PAT [34]	CVPR 2021	64.5	53.6	**81.6**	**72.1**	73.1	62.9
FRT (ours)		**70.7**	**61.3**	80.4	71.0	**75.6**	**66.2**

TABLE III

PERFORMANCE (%) COMPARISONS WITH THE STATE-OF-THE-ART METHODS ON THE TWO PARTIAL DATASETS, *i.e.*, PARTIAL-REID [17] AND PARTIAL-iLIDS [18]. OUR METHOD ACHIEVES THE BEST PERFORMANCE ON THE PARTIAL-REID [17]. THE BEST PERFORMANCE IS HIGHLIGHTED IN BOLD

Methods	Reference	Partial-REID		Partial-iLIDS		Average	
		Rank-1	Rank-3	Rank-1	Rank-3	Rank-1	Rank-3
DSR [18]	CVPR 2018	58.8	67.2	50.7	70.0	54.8	68.6
AFPB [8]	ICME 2018	78.5	-	-	-	-	-
FPR [11]	ICCV 2019	68.1	-	81.0	-	74.6	-
PGFA [10]	ICCV 2019	69.1	80.9	68.0	80.0	68.6	80.5
VPM [12]	CVPR 2019	65.5	74.8	67.7	81.9	66.6	78.4
STNReID [52]	TMM 2020	66.7	80.3	54.6	71.3	60.7	75.8
PVPM [9]	CVPR 2020	78.3	87.7	-	-	-	-
HOReID [13]	CVPR 2020	85.3	91.0	72.6	86.4	79.0	88.7
PGFA-KD [19]	ACM MM 2021	85.1	90.8	74.0	86.7	80.0	88.8
OAMN [20]	ICCV 2021	86.0	-	**77.3**	-	81.7	-
PAT [34]	CVPR 2021	88.0	92.3	76.5	**88.2**	**82.2**	**90.3**
FRT (ours)		**88.2**	**93.2**	73.0	87.0	80.6	90.1

the global average pooling (GAP) and fully connected layers are removed from the original ResNet-50 [55] architecture and the stride of the last convolution layer is set to 1. The parameter s in Eq. 15 equals 3 and we input the 5-nearest neighbors features to the $\mathcal{T}$. We exploit one GCN layer in our method.

3) Evaluation Metrics: We utilize mean average precision (mAP) and Cumulative Matching Characteristic (CMC) curves to evaluate the performance of various Re-ID models. All the experiments are conducted in a single query setting.

B. Comparison With State-of-the-Art Methods

1) Results on the Occluded Datasets: In Table II, we compare our method with the state-of-the-art Re-ID methods on the two occluded datasets, *i.e.*, Occluded-Duke [10] and Occluded-ReID [8]. Four kinds of methods are compared, they are holistic methods [5], [6], key-points based methods [48], [49], partial Re-ID methods [18] and occluded Re-ID methods [9], [10], [13], [19]–[21], [34], [50], [51]. The result shows that FRT outperforms other methods on Occluded-Duke dataset [10], which demonstrates the effectiveness of our FRT in dealing with the occluded Re-ID problem. Specifically, on the Occluded-Duke [10] dataset, FRT achieves the best result with Rank-1 accuracy of 70.7% and mAP of 61.3%, which is at least 6.2% and 7.7% higher than the corresponding metrics of other methods. On the Occluded-REID dataset [8], FRT achieves the competitive results to PAT [34] with 80.4% Rank-1 accuracy and 71.0% mAP.

2) Results on the Partial Datasets: In Table III, we compare our method with the state-of-the-art Re-ID methods on the two partial datasets, *i.e.*, Partial-REID [17] and

TABLE IV
PERFORMANCE (%) COMPARISONS WITH THE STATE-OF-THE-ART RE-ID METHODS ON HOLISTIC DATASETS *i.e.* MARKET-1501 [47] AND DUKEMTMC-REID [45]. OUR METHOD ACHIEVES BEST PERFORMANCE ON HOLISTIC RE-ID. THE BEST PERFORMANCE IS HIGHLIGHTED IN BOLD

	Methods	Reference	Market-1501		DukeMTMC-reID		Average	
			Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
Holistic Methods	PCB [5]	ECCV 2018	92.3	77.4	81.8	66.1	87.1	71.8
	BOT [56]	ICCV 2019	94.1	85.7	86.4	76.4	90.3	81.1
Key-points Based Methods	Part Bilinear [48]	ECCV 2018	90.2	76.0	82.1	64.2	86.2	70.1
	FD-GAN [49]	NIPS 2018	90.5	77.7	80.0	64.5	85.3	71.1
Partial Re-ID Methods	DSR [18]	CVPR 2018	83.5	64.2	-	-	-	-
Occluded Re-ID Methods	Ad-Occluded [50]	CVPR 2018	84.4	66.9	79.1	62.1	81.8	64.5
	FPR [11]	ICCV 2019	95.4	86.5	88.6	76.4	92.0	81.5
	PGFA [10]	ICCV 2019	91.2	76.8	82.6	65.5	86.9	71.2
	HOReID [13]	CVPR 2020	91.0	85.3	86.4	72.6	88.7	79.0
	Pirt [51]	ACM MM 2021	94.1	86.3	88.9	77.6	91.5	82.0
	PGFA-KD [19]	ACM MM 2021	95.3	87.2	89.6	79.5	92.5	83.4
	OAMN [20]	ICCV 2021	93.2	79.8	86.3	72.6	89.8	76.2
	PAT [34]	CVPR 2021	95.4	88.0	88.8	78.2	92.1	83.1
	FRT (ours)		**95.5**	**88.1**	**90.5**	**81.7**	**93.0**	**84.9**

Partial-iLIDS [18]. Accompanied by occluded images, partial ones often occur due to outliers of camera views, imperfect detection, and so on. As we can see, our method achieves the best results on the Partial-REID dataset [17], which outperforms other methods by at least 0.2% Rank-1 accuracy and 0.9% Rank-3 accuracy. FRT also achieves competitive results on the Partial-iLIDS dataset [18]. We think there are three reasons for the less performance gains on the three small-scale occluded and partial datasets, *i.e.*, Occluded-ReID [8], Partial-REID [17] and Partial-iLIDS [18] than on the Occluded-Duke [10]: 1) The three small-scale occluded and partial datasets *i.e.* Occluded-ReID [8], Partial-REID [17] and Partial-iLIDS [18] adopt cross-domain setting, which utilize Market-1501 as the training set and test on the other domains. Domain bias in the cross-domain evaluation would have a negative impact on the performance of our model. 2) The gallery of the Occluded-Duke [10], Occluded-ReID [8], Partial-REID [17] and Partial-iLIDS [18] contain 16,5,5,1 images for each identity respectively. Therefore, on Occluded-Duke [10], the feature recovery transformer is able to employ more pedestrian information in the gallery, resulting in better recovered query features and bigger Re-ID performance improvements on Occluded-Duke than other three small-scale occluded and partial datasets. 3) Occluded-Duke [10] is the largest dataset for studying the occluded Re-ID problem, and the results on the Occluded-Duke are more reliable for occluded Re-ID performance evaluation.

3) Results on Holistic Datasets: Although recent occluded/partial person Re-ID methods have made progress on occluded/partial datasets, their performances are always unsatisfying on the holistic datasets. In this part, we show that our method can also achieve comparable state-of-the-art performances on the holistic datasets Market-1501 [47] and DukeMTMC-reID [45]. The results are shown in Table IV. We compare the proposed FRT with two holistic Re-ID methods [5], [56], two key-points based methods [48], [49], one partial Re-ID method [18] and eight occluded Re-ID methods [10], [11], [13], [19], [20], [34], [50], [51]. The result shows that on the Market-1501 [47] our proposed FRT achieves the best results with 95.5% Rank-1 accuracy and 88.1% mAP and on the DukeMTMC-reID [45] FRT achieves the best results with 90.5% Rank-1 accuracy and 81.7% mAP, which outperforms other methods by at least 0.9% and 2.2% respectively.

C. Comparison With Post-Processing Techniques

As the visibility graph and feature recovery transformer work in the feature matching stage, we additionally compare FRT with other state-of-the-art post-processing techniques, *i.e.* re-ranking [44] and average query expansion (AQE) [57] on Occluded-Duke dataset. The result is shown in Table V. From the result we can see that FRT has the highest Rank-1 accuracy of 70.7%. We think the main reason for the higher Rank-1 accuracy of FRT is that $\mathcal{T}$ is able to filter out the noisy message in the k-nearest neighbors, and employ valuable information instead of weighted sum of all the features to recover the query features. However, re-ranking achieves the highest mAP of 63.7%, which is 2.4% higher than the FRT. We think the reason for this is that re-ranking attempts to calculate the k-nearest neighbors for all the candidates in the rank list and recalculate the Jaccard distance for re-ranking, which is better for the mean accuracy. In addition, the result in the last row indicates that FRT and re-ranking are not conflicting, they can be integrated for better performance. We visualize the comparison of FRT, average query expansion (AQE) [57] and re-ranking [44] in Fig. 4.

D. Further Analysis

1) Ablation Study of Proposed Modules: In this part, we analyze our proposed semantic feature extractor ($\mathcal{E}$), visibility graph matching ($\mathcal{G}$) and feature recovery transformer ($\mathcal{T}$)

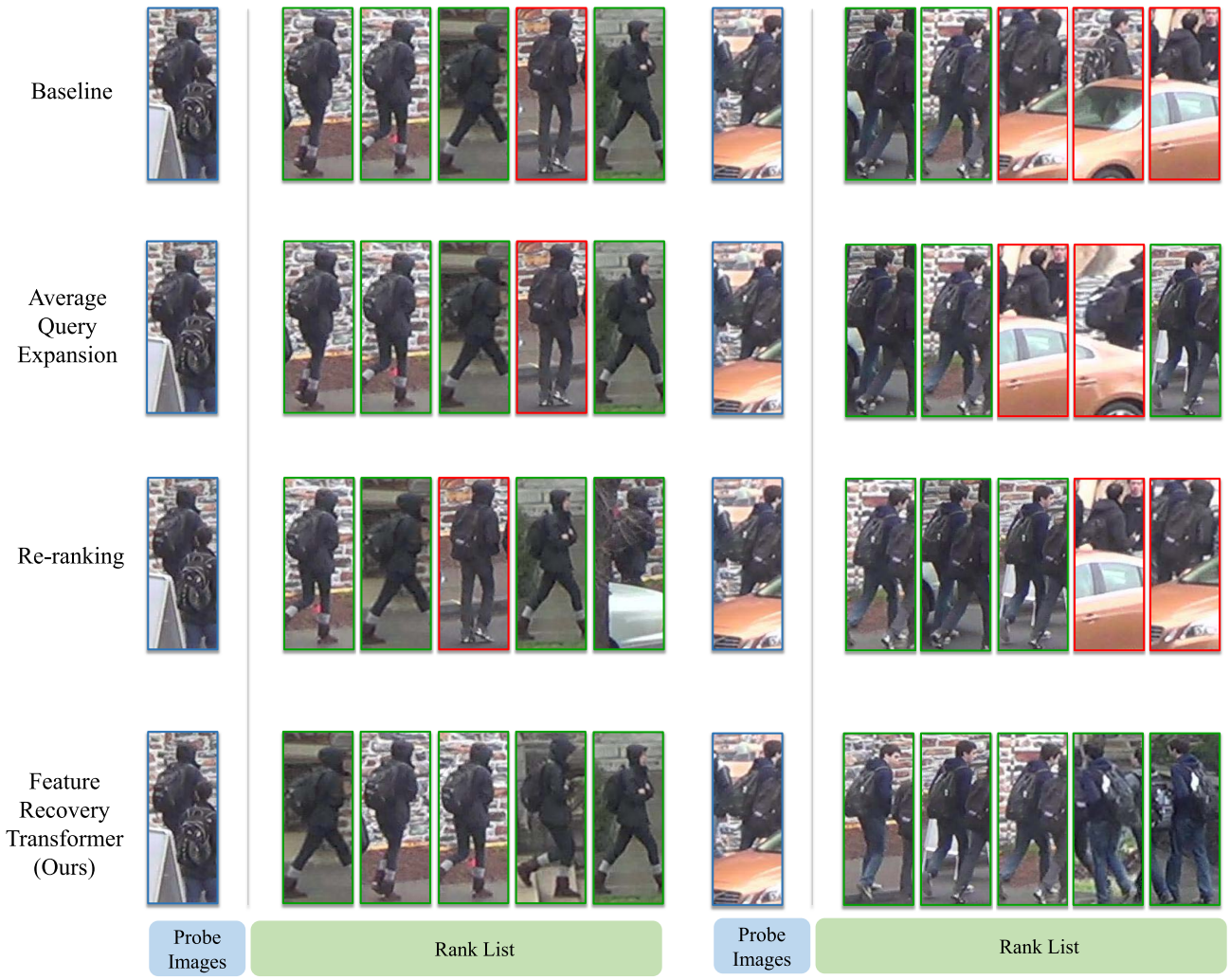


Fig. 4. Visualization of the comparison of our proposed FRT and other state-of-the-art post-processing techniques, *i.e.* average query expansion [57] and re-ranking [44]. Green and red rectangles indicate correct and error retrieval results,respectively.

on Occluded-Duke dataset. The results are shown in Table VI. Firstly, in index-1, we can see that thanks to the feature alignment by extracting semantic features with key-points, $\mathcal{E}$ is able to achieve 56.5% Rank-1 accuracy and 48.6% mAP on the Occluded-Duke dataset. Secondly, In index 2, affinities among different body parts are considered and the information in the shared regions is promoted. This gives 3.4% and 3.2% improvement to the Rank-1 accuracy and mAP respectively and demonstrates the effectiveness of $\mathcal{G}$. Thirdly, in index 1 and 3, we can see that $\mathcal{T}$ is able to give Rank-1 accuracy of 13% and mAP of 11.2% improvement to the $\mathcal{E}$. Finally, in index 3 and 4, $\mathcal{G}$ gives another 1.2% and 1.5% higher points to Rank-1 accuracy and mAP respectively to the $\mathcal{T}$.

2) Analysis of Parameters: We evaluate the effects of parameters s, k and δ in Fig. 5. From Fig. 5(a), we can see that multi-step mechanism is able to give about 1.5%, 1.9% higher points to Rank-1 accuracy and mAP respectively. The performance is at its best when s equals three. From Fig. 5(b) we can see that the parameter k has a great impact on the performance. When we input the 5-nearest neighbors features to the feature recovery transformer, FRT achieves the best result with Rank-1 accuracy of 70.7% and mAP of 61.3%.

TABLE V

PERFORMANCE (%) COMPARISONS WITH THE STATE-OF-THE-ART POST-PROCESSING TECHNIQUES ON OCCULUDED-DUKE DATASET. $\mathcal{E}$ INDICATES SEMANTIC FEATURE EXTRACTOR. OUR MODEL ACHIEVES THE BEST RESULTS AND COULD BE INTEGRATED WITH OTHER POST-PROCESSING TECHNIQUES FOR BETTER RESULTS

Methods	Rank-1	mAP
PGFA [10]	51.4	37.3
PGFA [10] + re-ranking	52.4	46.8
HOReID [13]	55.1	43.8
HOReID [13] + re-ranking	58.3	49.2
Pirt [51]	60.0	50.9
Pirt [51] + re-ranking	62.1	59.3
$\mathcal{E}$	56.5	48.6
$\mathcal{E}$ + AQE [57]	62.8	60.2
$\mathcal{E}$ + re-ranking [44]	64.6	63.7
FRT (*ours*)	70.7	61.3
FRT (*ours*) + re-ranking [44]	**70.8**	**65.0**

From Fig. 5(c) we can find that when δ equals 0.2, it is able to give about 1.5%, 1.7% higher points to Rank-1 accuracy and mAP respectively than δ equals 0. The reason is that the

TABLE VI

ABLATION STUDY OF THE PROPOSED MODULES ON OCCULUDED-DUKE DATASET. $\mathcal{E}$ IS THE SEMANTIC FEATURE EXTRACTOR, $\mathcal{G}$ IS THE VISIBILITY GRAPH MATCHING AND $\mathcal{T}$ IS THE FEATURE RECOVERY TRANSFORMER. THE RESULTS VALIDATE THE EFFECTIVENESS OF THE THREE PROPOSED MODULES

Index	$\mathcal{E}$	$\mathcal{G}$	$\mathcal{T}$	Rank-1	mAP
1	√	×	×	56.5	48.6
2	√	√	×	59.9	51.4
3	√	×	√	69.5	59.8
4	√	√	√	**70.7**	**61.3**

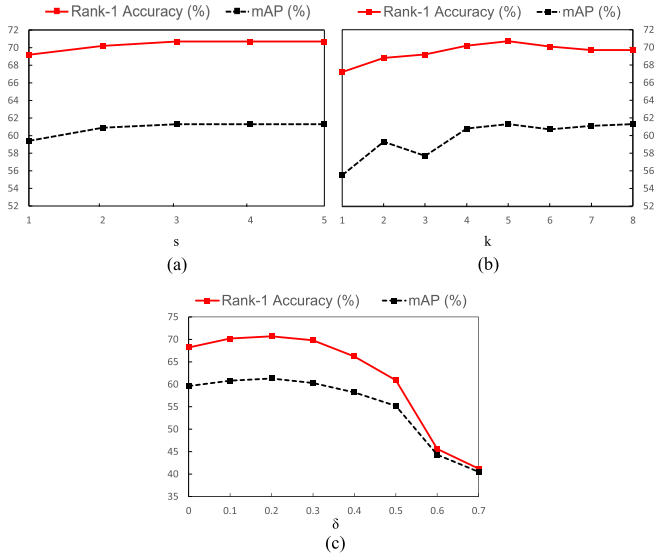


Fig. 5. Analysis of parameters s, k and δ on the Occluded-Duke dataset. (a) s indicates conduct $\mathcal{T}$ for s times; (b) k indicates input the k-nearest neighbors features to the $\mathcal{T}$; (c) δ is the threshold for the semantic feature extractor.

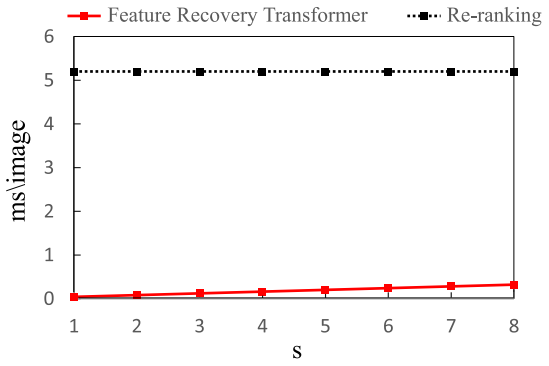


Fig. 6. Analysis of the time consumption of the feature recovery transformer. s indicates conduct $\mathcal{T}$ for s times. The results show that our feature recovery transformer has less time consumption.

threshold δ can eliminate the effectiveness of occlusion during the training.

3) Analysis of the Feature Recovery Transformer Time Consumption: We evaluate the time consumption of the feature recovery transformer in Fig. 6. Fig. 6 shows the time consumption on each image in inference. The result indicates that when s equals three, the time consumption is about 0.12 ms per image, which is nearly 1/40 of the Re-ranking costs.

TABLE VII

EVALUATION OF THE FEATURE RECOVERY TRANSFORMER ON OCCLUDED-DUKE DATASET. THE RESULT DEMONSTRATES THAT FEATURES ARE RECOVERED AFTER BEING PROCESSED BY OUR FEATURE RECOVERY TRANSFORMER

Feature	Before Feature Recovery		After Feature Recovery	
	Rank-1	mAP	Rank-1	mAP
Global	55.6	46.1	**69.3**	**59.1**
Head	57.3	40.7	**62.6**	**45.9**
Torso	49.2	35.5	**67.4**	**49.6**
Leg	26.8	18.9	**62.8**	**45.5**
Concat	56.5	48.6	**70.7**	**61.3**

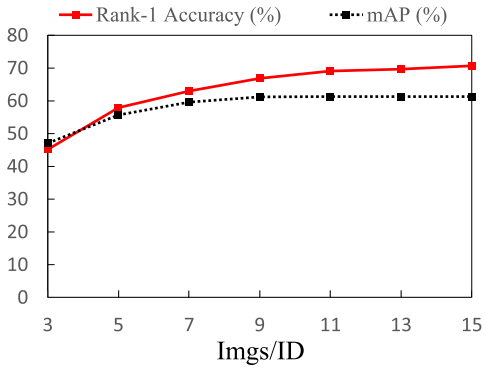


Fig. 7. Evaluation of the effects of the gallery size. The gallery size is changed from 3 images per ID to 15 images per ID.

The results demonstrate that our feature recovery transformer is able to achieve better performance with less time consumption.

4) Evaluation of the Feature Recovery Transformer: In Table VII, we evaluate the Re-ID performance of each part feature and the final representation before and after being processed by feature recovery transformer on Occluded-Duke dataset. From the result we can find that before the feature recovery transformer, the leg part feature has the worst performance with Rank-1 accuracy of 26.8% and mAP of 18.9%, indicating that the lower part of most of the images is occluded. After being processed by the feature recovery transformer, we can find that the Re-ID accuracy of the global feature, head feature, torso feature and leg feature are all improved. In particular, the Rank-1 accuracy and mAP of the leg feature are improved from 26.8% to 62.8% and 18.9% to 45.5% respectively, indicating that the occluded leg feature has been recovered by the feature recovery transformer. This experiment proves that the feature recovery transformer is able to recover occluded features and improve the Re-ID performance of both the global and local features.

5) Visualization of the Feature Recovery Transformer: We visualize the recovery process of the feature recovery transformer in Fig. 8. We illustrate the top-5 nearest neighbors. Green and red rectangles indicate correct and error retrieval results,respectively. The numbers above the pictures indicate the contribution of corresponding neighbors to the feature

Fig. 8. Visualization of the contributions of k-nearest neighbors when conducting the feature recovery transformer. We illustrate the top-5 nearest neighbors. Green and red rectangles indicate correct and error retrieval results,respectively. The numbers above the pictures indicate the contribution of corresponding neighbors to the feature recovery. The result shows that the feature recovery transformer is able to filter out the noise in the k-nearest neighbors and exploit valid information for feature recovery.

recovery. In the first example, the k-nearest neighbors contain an error retrieval with a similar appearance to the probe. The biggest difference between the correct and error retrieval is the shoes. During the process the feature recovery, we find that the $\mathcal{T}$ exploit little information from the error retrieval, and thus the recovered query feature is reliable for representing the probe identity. In the second example, the probe image is occluded by a car, which leads to the two error retrievals in the k-nearest neighbors. Fortunately, the $\mathcal{T}$ is able to distinguish the noise and exploit the message from the other 3 correct retrievals for the recovery of the occluded feature in the probe. The result shows that the feature recovery transformer is able to filter out the noisy information, even for some hard cases, in the k-nearest neighbors, and then exploit the valid message for feature recovery.

6) Evaluation of the Effects of the Gallery Size: Feature recovery transformer utilizes the pedestrian information in the gallery for features recovery. Therefor, we attempt to evaluate the effects of the gallery size on the feature recovery transformer. We change the gallery size from 3 images per identity to 15 images per identity, and the result is shown in Fig. 7. From the result we can see that the performance of the FRT improves with the increase of the gallery size. Specifically, the Rank-1 accuracy and mAP are improved from 45.2%, 47.1% to 70.7%, 61.3% respectively when the gallery size increases from 3 images per ID to 15 images per ID. The reason is that when the gallery size increases, the feature recovery transformer is able to employ more pedestrian information in the gallery for features recovery.

V. Conclusion

In this paper, we propose a novel framework called Feature Recovery Transformer (FRT) for the occluded person re-identification. Firstly, we employ key-points to extract semantic features for alignment and calculate the visibility scores for them. Then, we consider the semantic features from same part within a pair of images as nodes to construct a directional graph. We set the edge based on the visibility score of starting nodes for promoting the propagation of information in the shared regions. In terms of the occluded feature recovery, we propose a feature recovery transformer to exploit the pedestrian information in the features of its k-nearest neighbors. Finally, the recovered query feature is utilized for retrieving. Extensive experiments on occluded, partial and holistic datasets demonstrate that our proposed framework is able to recover the occluded features and achieve the best Re-ID performance.

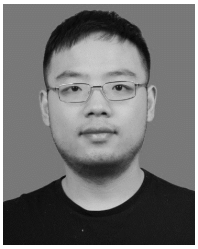
Acknowledgment

The authors would like to thank associate editor and anonymous reviewers for providing valuable suggestions to improve this paper.

References

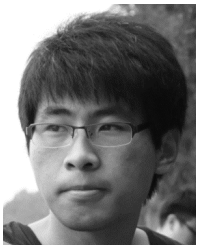
[1] W.-H. Li, F.-T. Hong, and W.-S. Zheng, "Learning to learn relation for important people detection in still images," in *Proc. CVPR*, Jun. 2019, pp. 5003–5011.
[2] E. Ristani and C. Tomasi, "Features for multi-target multi-camera tracking and re-identification," in *Proc. CVPR*, Jun. 2018, pp. 6036–6046.
[3] B. Xu, L. He, X. Liao, W. Liu, Z. Sun, and T. Mei, "Black re-ID: A head-shoulder descriptor for the challenging problem of person re-identification," in *Proc. ACM-MM*, Oct. 2020, pp. 673–681.
[4] X. Jin, C. Lan, W. Zeng, Z. Chen, and L. Zhang, "Style normalization and restitution for generalizable person re-identification," in *Proc. CVPR*, Jun. 2020, pp. 3143–3152.
[5] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. ECCV*, 2018, pp. 480–496.
[6] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *Proc. ICCV*, Oct. 2017, pp. 3219–3228.
[7] J. Zhou, B. Su, and Y. Wu, "Online joint multi-metric adaptation from frequent sharing-subset mining for person re-identification," in *Proc. CVPR*, Jun. 2020, pp. 2909–2918.
[8] J. Zhuo, Z. Chen, J. Lai, and G. Wang, "Occluded person re-identification," in *Proc. ICME*, Jul. 2018, pp. 1–6.
[9] S. Gao, J. Wang, H. Lu, and Z. Liu, "Pose-guided visible part matching for occluded person ReID," in *Proc. CVPR*, Jun. 2020, pp. 11744–11752.
[10] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, "Pose-guided feature alignment for occluded person re-identification," in *Proc. ICCV*, Oct. 2019, pp. 542–551.
[11] L. He, Y. Wang, W. Liu, H. Zhao, Z. Sun, and J. Feng, "Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8450–8459.
[12] Y. Sun *et al.*, "Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification," in *Proc. CVPR*, Jun. 2019, pp. 393–402.
[13] G. Wang *et al.*, "High-order information matters: Learning relation and topology for occluded person re-identification," in *Proc. CVPR*, Jun. 2020, pp. 6449–6458.
[14] S. Iodice and K. Mikolajczyk, "Partial person re-identification with alignment and hallucination," in *Proc. ACCV*, 2018, pp. 101–116.

[15] X. Jin, C. Lan, W. Zeng, G. Wei, and Z. Chen, "Semantics-aligned representation learning for person re-identification," in *Proc. AAAI*, 2020, pp. 11173–11180.
[16] Y. Zhouy and L. Shao, "Aware attentive multi-view inference for vehicle re-identification," in *Proc. CVPR*, Jun. 2018, pp. 6489–6498.
[17] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong, "Partial person re-identification," in *Proc. ICCV*, Dec. 2015, pp. 4678–4686.
[18] L. He, J. Liang, H. Li, and Z. Sun, "Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach," in *Proc. CVPR*, Jun. 2018, pp. 7073–7082.
[19] K. Zheng, C. Lan, W. Zeng, J. Liu, Z. Zhang, and Z.-J. Zha, "Pose-guided feature learning with knowledge distillation for occluded person re-identification," in *Proc. ACM-MM*, Oct. 2021, pp. 4537–4545.
[20] P. Chen *et al.*, "Occlude them all: Occlusion-aware attention network for occluded person re-ID," in *Proc. ICCV*, Oct. 2021, pp. 11833–11842.
[21] J. Yang *et al.*, "Learning to know where to see: A visibility-aware approach for occluded person re-identification," in *Proc. ICCV*, Oct. 2021, pp. 11885–11894.
[22] Z. Zhang, C. Lan, W. Zeng, and Z. Chen, "Densely semantically aligned person re-identification," in *Proc. CVPR*, Jun. 2019, pp. 667–676.
[23] M. Jia, X. Cheng, S. Lu, and J. Zhang, "Learning disentangled representation implicitly via transformer for occluded person re-identification," *IEEE Trans. Multimedia*, early access, Jan. 7, 2022, doi: 10.1109/TMM.2022.3141267.
[24] H. Tan, X. Liu, B. Yin, and X. Li, "MHSA-Net: Multihead self-attention network for occluded person re-identification," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 21, 2022, doi: 10.1109/TNNLS.2022.3144163.
[25] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "VRSTC: Occlusion-free video person re-identification," in *Proc. CVPR*, Jun. 2019, pp. 7183–7192.
[26] A. Vaswani *et al.*, "Attention is all you need," in *Proc. NeurIPS*, 2017, pp. 1–11.
[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[28] J. Lee *et al.*, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020.
[29] C.-C. Chiu *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *Proc. ICASSP*, Apr. 2018, pp. 4774–4778.
[30] H. Chen *et al.*, "Pre-trained image processing transformer," 2020, *arXiv:2012.00364*.
[31] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. ECCV*, 2020, pp. 213–229.
[32] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. CVPR*, Jun. 2019, pp. 3146–3154.
[33] P.-E. Sarlin, D. De Tone, T. Malisiewicz, and A. Rabinovich, "SuperGlue: Learning feature matching with graph neural networks," in *Proc. CVPR*, Jun. 2020, pp. 4938–4947.
[34] Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang, and F. Wu, "Diverse part discovery: Occluded person re-identification with part-aware transformer," in *Proc. CVPR*, Jun. 2021, pp. 2898–2907.
[35] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 61–80, Jan. 2009.
[36] J. Gao, T. Zhang, and C. Xu, "Graph convolutional tracking," in *Proc. CVPR*, Jun. 2019, pp. 4649–4659.
[37] W. Wang, X. Lu, J. Shen, D. Crandall, and L. Shao, "Zero-shot video object segmentation via attentive graph neural networks," in *Proc. ICCV*, Oct. 2019, pp. 9236–9245.
[38] S. Yun, M. Jeong, R. Kim, J. Kang, and H. J. Kim, "Graph transformer networks," in *Proc. NeurIPS*, 2019, pp. 1–11.
[39] D. Cheng, Y. Gong, X. Chang, W. Shi, A. G. Hauptmann, and N. Zheng, "Deep feature learning via structured graph Laplacian embedding for person re-identification," *Pattern Recognit.*, vol. 82, pp. 94–104, Oct. 2018.
[40] Y. Yan, Q. Zhang, B. Ni, W. Zhang, M. Xu, and X. Yang, "Learning context graph for person search," in *Proc. CVPR*, Jun. 2019, pp. 2158–2167.
[41] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. CVPR*, Jun. 2019, pp. 5693–5703.
[42] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. ECCV*, 2014, pp. 740–755.
[43] F. Wu, A. H. Souza, Jr., T. Zhang, C. Fifty, T. Yu, and K. Q. Weinberger, "Simplifying graph convolutional networks," in *Proc. ICML*, 2019, pp. 6861–6871.
[44] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proc. CVPR*, Jul. 2017, pp. 1318–1327.
[45] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. ECCV*, 2016, pp. 17–35.
[46] W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *Proc. CVPR*, Jun. 2011, pp. 649–656.
[47] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. ICCV*, Dec. 2015, pp. 1116–1124.
[48] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee, "Part-aligned bilinear representations for person re-identification," in *Proc. ECCV*, 2018, pp. 402–419.
[49] Y. Ge *et al.*, "FD-GAN: Pose-guided feature distilling GAN for robust person re-identification," in *Proc. NeurIPS*, 2018, pp. 1–12.
[50] H. Huang, D. Li, Z. Zhang, X. Chen, and K. Huang, "Adversarially occluded samples for person re-identification," in *Proc. CVPR*, Jun. 2018, pp. 5098–5107.
[51] Z. Ma, Y. Zhao, and J. Li, "Pose-guided inter- and intra-part relational transformer for occluded person re-identification," in *Proc. ACM-MM*, Oct. 2021, pp. 1487–1496.
[52] H. Luo, W. Jiang, X. Fan, and C. Zhang, "STNReID: Deep convolutional networks with pairwise spatial transformer networks for partial person re-identification," *IEEE Trans. Multimedia*, vol. 22, no. 11, pp. 2905–2913, Nov. 2020.
[53] L. He, X. Liao, W. Liu, X. Liu, P. Cheng, and T. Mei, "FastReID: A PyTorch toolbox for general instance re-identification," 2020, *arXiv:2006.02631*.
[54] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proc. ACM-MM*, Oct. 2018, pp. 274–282.
[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Jun. 2016, pp. 770–778.
[56] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–9.
[57] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total recall: Automatic query expansion with a generative feature model for object retrieval," in *Proc. ICCV*, Oct. 2007, pp. 1–8.



Boqiang Xu received the B.E. and M.S. degrees from China Agricultural University in 2016 and 2019, respectively. He is currently pursuing the Ph.D. degree with the University of Chinese Academy of Sciences and the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences. His research interests include person ReID and computer vision.

Lingxiao He (Member, IEEE) received the B.E. degree in information engineering from the Chengdu University of Technology, Chengdu, China, in 2014, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences (CASIA), in 2019. Since 2019, he has been with the JD AI Research, JD.com, as a Research Scientist. He has authored/coauthored over 20 technical articles. His research interests include biometrics, pattern recognition, and computer vision. He is a member of the IEEE Computer society and the IEEE Signal Processing Society.



Jian Liang (Member, IEEE) received the B.E. degree in electronic information and technology from Xi'an Jiaotong University in July 2013 and the Ph.D. degree in pattern recognition and intelligent systems from the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), in January 2019. From June 2019 to April 2021, he was a Research Fellow at the National University of Singapore. He is currently joins NLPR, as an Associate Professor. His research interests focus on transfer learning, pattern recognition, and computer vision.

Zhenan Sun (Senior Member, IEEE) received the B.S. degree in industrial automation from the Dalian University of Technology, Dalian, China, in 1999, the M.S. degree in system engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2002, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China, in 2006. Since 2006, he has been a Faculty Member with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, CASIA, where he is currently a Professor. He has authored or coauthored more than 200 technical articles. His current research interests include biometrics, pattern recognition, and computer vision. He is a fellow of the IAPR. He serves as the Chair for IAPR Technical Committee on Biometrics. He is an Associate Editor of the IEEE TRANSACTIONS ON BIOMETRICS, BEHAVIOR, AND IDENTITY SCIENCE.