CAAI Transactions on Intelligence Technology

WILEY

The Institution of Engineering and Technology

ORIGINAL RESEARCH OPEN ACCESS

A Keypoint-Guided Feature Partition Network for Occluded Person Re-Identification

Die Dai[1] | Xu Zhang[2,3,4] | Zhiguang Wu[2] | Hongying Meng[5] | Zuyu Zhang[2,3,4]

[1]School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, China | [2]School of Artificial Intelligence, Chongqing University of Posts and Telecommunications, Chongqing, China | [3]Key Laboratory of Tourism Multisource Data Perception and Decision, Ministry of Culture and Tourism, Chongqing, China | [4]Key Laboratory of Big Data Intelligent Computing, Chongqing University of Posts and Telecommunications, Chongqing, China | [5]Department of Electronic and Electrical Engineering, Brunel University London, London, United Kindom

Correspondence: Xu Zhang (zhangx@cqupt.edu.cn) | Hongying Meng (hongying.meng@brunel.ac.uk)

Received: 19 December 2023 | **Revised:** 10 December 2024 | **Accepted:** 19 December 2024

Handling Editor: Zan Gao

Funding: This work is supported in part by the National Natural Science Foundation of China (No. 62276038), and in part by the Scientific and Technological Research Program of Chongqing Municipal Education Commission (Grant KJZD-M202400603), and in part by the Project of Key Laboratory of Tourism Multisource Data Perception and Decision, Ministry of Culture and Tourism (No. H2023009).

Keywords: computer vision | human recognition | object recognition | occluded person re-identification

ABSTRACT

Existing occluded person re-identification methods employ hard or soft partition strategies to explore fine-grained information. However, the hard partition strategy which extracts region-level features may impair the semantic connectivity of correlated human body parts. A pose-guided soft partition establishes correlations among human keypoints, while the generated pixel-level embeddings may lose the surrounding semantic information. In this paper, we propose a keypoint-guided feature partition (KGFP) method that consists of a feature extractor, a hard partition branch, and a soft partition branch. Specifically, we adopt a vision transformer and a pose estimator to extract features and keypoint information. In the hard partition branch, we partition features into distinct groups and classify them into nonoccluded, semi-occluded, and occluded features to obtain region-level features and filter out occlusions. Furthermore, we design a dissimilarity loss to reduce the similarity between semi-occluded and occluded features. In the soft partition branch, we introduce a graph attention network and consider global and keypoint embeddings as nodes of a graph to discover interrelationships. Additionally, we formulate image alignment as a graph matching problem and propose a feature alignment-based graph to reduce position misalignment. Extensive experiments demonstrate that the proposed method achieves superior performance compared to state-of-the-art methods on Occluded-DukeMTMC, Markt1501, and DukeMTMC-reID.

1 | Introduction

Person re-identification (Re-ID) presents a challenging task in image-based retrieval, which aims to match pedestrian images of the same identity across multiple nonoverlapping cameras [1] deployed in different locations or at different times. Because of challenges posed by various body poses, different views of cameras and cluttered backgrounds, the Re-ID task has gained considerable attention. Most existing methods [2–4] assume that a pedestrian's entire body is completely visible. However, pedestrians are easily occluded by various obstacles (i.e., vehicles, nontarget pedestrians and warning signs). These occlusions, including differences in colour, size, shape and structural position, have a detrimental impact on the overall person re-

identification performance. Therefore, numerous advancements have introduced occluded person re-identification to mitigate the effect of occlusions.

In occluded person Re-ID tasks, the presence of obstacles within occluded regions often leads to mismatches. Therefore, learning discriminative features from nonoccluded regions is crucial. Several methods [2–9] in the holistic person re-identification task have proved that various feature partition strategies are effective in extracting fine-grained features, which is helpful in capturing discriminative features. Inspired by that, most methods [10–16] also introduce these strategies for occluded person re-identification tasks, which can extract discriminative features and further alleviate the negative impact of occlusions. Feature partition strategies can be divided into two categories: *hard partition strategy* and *soft partition strategy*. The hard partition strategy does not require part labelling, which can divide the feature maps into stripes or patches. However, the soft partition strategy leverages external cues obtained from a pose estimator to capture part features. Figure 1a–c employ a hard partition strategy that splits the image into horizontal stripes, vertical stripes and patches. These splittings offer region-level fine-grained information. However, they may separate highly semantically correlated regions, thus destroying the interconnections among body parts. In contrast, Figure 1d demonstrates the partitioning outcome obtained by the pose estimator, which generates keypoint embeddings and preserves the relationships among keypoints. However, the keypoint embeddings obtained are pixel-level feature representations, which lose the detailed information around keypoints.

The advantages and limitations of these two partition strategies are found to be complementary. Inspired by the above observations, we propose a keypoint-guided feature partition (KGFP) method, aiming to explore fine-grained region-level information and the interconnections among human body parts. Specifically, the proposed KGFP method comprises of a feature extractor, a hard partition branch and a soft partition branch. In the feature extractor, we employ the Vision Transformer (ViT) [17] to capture the global feature and patch tokens, alongside the pose estimator to generate the landmark and heatmap of each keypoint. In the hard partition branch, the patch tokens are divided into groups and then fed into a shared transformer layer to investigate region-level information. To extract nonoccluded features, we classify groups into three categories including nonoccluded, semi-occluded and occluded groups via coordinates and visibility of these keypoints. Additionally, we introduce a dissimilarity loss to push semi-occluded and occluded features away to focus on human body parts. In the soft partition branch, a graph attention network is introduced, then the global feature and keypoint embeddings are viewed as node features for the purpose of mining their interconnections inside. Furthermore, image alignment is regarded as a graph matching problem and the corresponding keypoint embeddings are compared to calculate the distance for query and gallery images during testing. Therefore, we further alleviate the impact of position misalignment and occlusion problem, and achieve an improved retrieval accuracy. The main contributions are summarised as belows:

1. We propose a novel keypoint-guided feature partition (KGFP) method, which integrates hard and soft partition strategies, enabling the extraction of discriminative region-level information while preserving interconnections among human body parts. Extensive experiments demonstrate its remarkable performance.
2. We propose a hard partition branch that splits patch tokens obtained from ViT into different groups, which have been classified into nonoccluded, semi-occluded, and occluded features based on coordinates and visibility of body keypoints. Additionally, a dissimilarity loss is introduced to effectively separate semi-occluded and occluded features and enhance focus on the human body.
3. We propose a soft partition branch to investigate the connectivity among global and keypoint embeddings, which aggregates and updates semantic information with a graph attention network.
4. We consider image alignment as a graph matching problem and propose a feature-alignment-based graph strategy to minimise the retrieval misalignment.

2 | Related Works

2.1 | Occluded Person Re-Identification

Occluded person Re-ID differs from holistic person Re-ID, as the latter assumes the whole human body is visible [18]. In contrast,

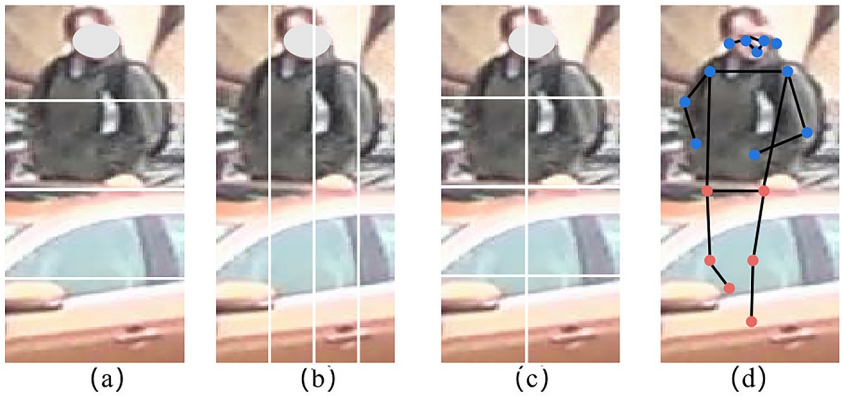


FIGURE 1 | Different strategies for splitting an image. (a) Horizontal stripes, (b) vertical stripes, (c) patches, and (d) pose estimator.

occluded person Re-ID concentrates on pedestrian images that are obstructed by diverse obstacles (e.g., cars, warning signs, trees, and nontarget pedestrians). With the recent progress in deep learning, numerous methodologies for occluded person Re-ID have been introduced. To learn discriminative features from nonoccluded regions, numerous methods split the feature map to concentrate on nonoccluded features. Generally, splitting strategies can be divided into hard partition strategy and soft partition strategy. The hard partition strategy, which does not require part labelling, primarily splits an image into stripes or patches to filter out occluded regions and background clutters. PCB [2] simply evenly partitions the image into horizontal stripes to mine fine-grained information. Zhang et al. [10] and Kim et al. [19] also split the image based on PCB and extracts centre parts from the vertical direction under the presumption that the target pedestrian is positioned in the middle of the bounding box. As local features make varying contributions, Wang et al. [11] propose to predict quality scores of each part and introduce an attention mechanism focusing on common nonoccluded regions. To further reduce the impact of occlusions, He et al. [9] introduce the ViT to split the image into patches. However, these methods may lose some highly semantically correlated information and interconnections among body parts. The soft partition strategy needs to utilise external cues to focus on the nonoccluded features. DRL-Net [15] and PAT [13] introduce a learnable semantic representation to target discriminative human body parts. Additionally, PMFB [12], Yang et al. [14], and DAReID [16] split feature maps into horizontal stripes and categorise them as either nonoccluded or occluded features by comparing the confidence of keypoints with the threshold. They determine that if a local region contains visible keypoints, it is considered as a nonoccluded region, whereas if not, it is classified as occluded. However, if a local region contains both visible and invisible keypoints simultaneously, it will be regarded as a nonoccluded region in the above methods, even though it still introduces the noise caused by invisible keypoints.

Unlike existing methods, our approach involves constructing a graph to establish the linkage of each human body part in the soft partition branch. Furthermore, we classify region-level features into nonoccluded, semi-occluded, and occluded features to more effectively filter out occlusions in the hard partition branch.

2.2 | Feature Alignment

Person Re-ID presents numerous challenges including occlusion, pose changes, viewpoint variation, and background clutter. These obstacles frequently lead to misalignment among pedestrians of the same identity, and subsequently, a reduction in matching accuracy. Sun et al. [2] have attempted to resolve these issues by splitting the feature map into horizontal stripes and matching corresponding stripes in a one-to-one manner. However, it ignores the mismatching between pedestrians and their backgrounds, as well as the significance of global features. Zhang et al. [20] propose to compute the shortest path between two sets of local features, while also retaining the global feature to compute the global similarity. Zhao et al. [21] and Li et al. [22] employ an attention mechanism that concentrates on human body parts, thereby successfully achieving feature alignment. To obtain discriminative features and robust alignment, Wang et al. [23] perceive person Re-ID as a graph matching problem, and propose a CGEA layer to achieve local feature alignment with topology information. We consider person Re-ID as a graph matching problem and design a graph-based new feature alignment strategy that relies on human skeleton to improve the retrieval precision.

2.3 | Graph Convolution Network

Graph convolution networks (GCNs) have been successful in computer vision applications including gait recognition [24], image recognition [25] and person re-identification. Specifically, Zhang et al. [26] and Pan et al. [27] split the feature map into stripes and use GCNs to model potential relations between local features of image patches for holistic person Re-ID. HOReID [23] constructs a graph based on human keypoints and proposes a cross-graph embedded alignment method to alleviate the impact of occlusions for occluded person Re-ID. Liang et al. [28] utilise the GCN based on the pose to alleviate the impact of modality discrepancy for visible-infrared person Re-ID. Chen et al. [29] propose pose-assisted GCN to mine temporal information for video person Re-ID.

Our method differs from previous studies in two main aspects. Firstly, we jointly combine GCN with human skeleton information to investigate the connectivity between human body parts. Secondly, whilst GCN treats all neighbouring nodes equally for a node, certain node features pertain to occluded features, we introduce the attention mechanism in GCN to minimise the interference of occlusions.

3 | The Proposed Method

As illustrated in Figure 2, the proposed KGFP mainly includes three parts: *feature extractor*, *hard partition branch,* and *soft partition branch*. Specifically, feature extractors extract patch tokens with a ViT architecture and keypoint information by pose estimator. For the hard partition branch, we split patch tokens into groups and feed them into a shared transformer layer to mine the region-level local information, and further classify them into nonoccluded, semi-occluded, and occluded features by the visibility label of keypoints to alleviate the impact of occlusions. Moreover, we develop a dissimilarity loss to push the semi-occluded and occluded features away to focus on human body parts. For soft partition branch, we view the global feature and keypoint embeddings as node features and introduce a graph attention network to explore the linkage of global and keypoint embeddings. Additionally, we consider image alignment as a graph matching problem, and propose a feature-alignment-based graph to reduce the impact of occlusions and position misalignment in testing.

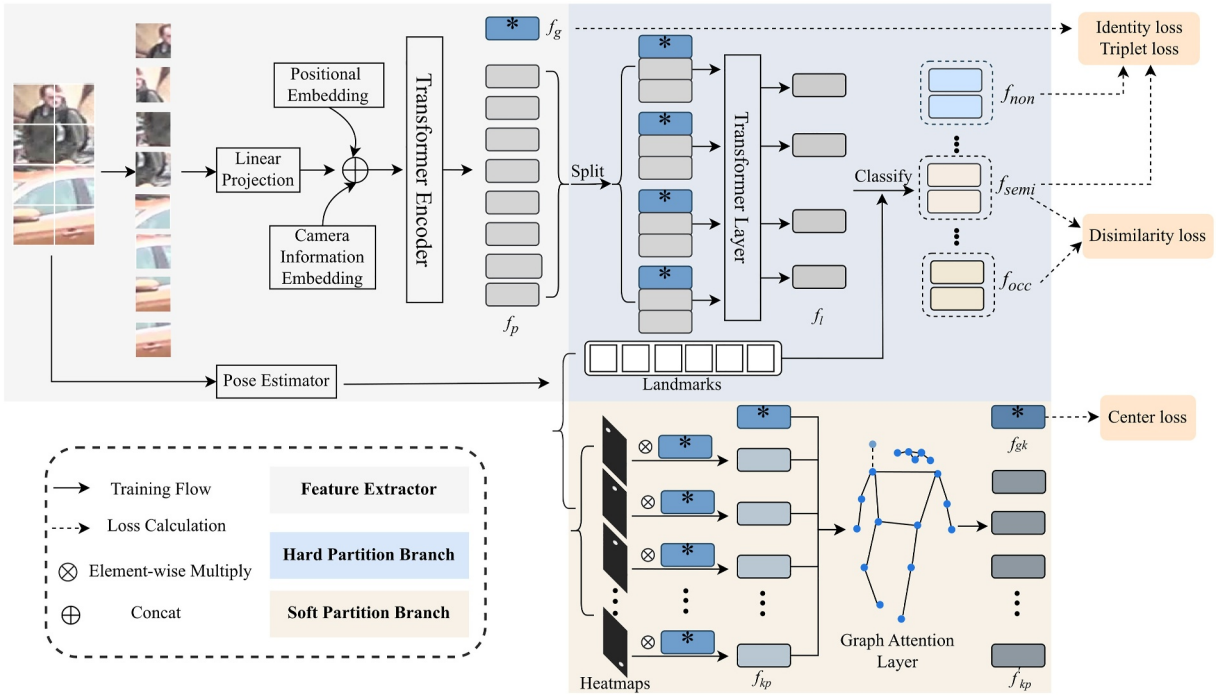


FIGURE 2 | The architecture of the proposed keypoint-guided feature partition (KGFP).

3.1 | Feature Extractor

3.1.1 | Transformer-Based Framework

Vision Transformer possesses a strong capability to extract the patch tokens. We build a feature extractor and use ViT as the basis. When given an image $x \in \mathbb{R}^{H\times W\times C}$, with H, W and C denoting the height, width, and channel dimensions of x respectively, we split the image into N nonoverlapping patches $\{x_p \mid p = 1, 2, ..., N\}$, and the number of patch tokens can be described as follows:

$$N = \left\lfloor \frac{H-P}{S} + 1 \right\rfloor \times \left\lfloor \frac{W-P}{S} + 1 \right\rfloor. \tag{1}$$

where S denotes the stride size, and P denotes the size of each image patch. And then we map the N fixed-size patches sequences to D dimensions via a trainable linear projection and obtain the patch embeddings $E \in \mathbb{R}^{N\times D}$. We introduce a learnable position embedding E_P to preserve position information. Additionally, we set a learnable parameter E_{cm} to alleviate the impact of camera perspective following the existing method [9]. Simultaneously, a learnable class token x_{cls} is defined to serve as global feature representation, which is presented to the patch embeddings. Finally, the definitive input can be described as follows:

$$E_{\text{input}} = \{x_{\text{cls}}; E\} + E_p + \lambda_{\text{cm}} E_{\text{cm}}. \tag{2}$$

where λ_{cm} is the ratio of camera embeddings. Then the ViT takes E_{input} as input. The output of the feature extractor is $f \in \mathbb{R}^{(N+1)\times D}$, which contains one global feature $f_g \in \mathbb{R}^{1\times D}$ and N patch tokens $f_p \in \mathbb{R}^{N\times D}$.

3.1.2 | Pose Estimator

Pedestrian images with occlusions suffer from performance degradation as they lack full-body information, and the occlusions and backgrounds could be similar. In order to address this issue, we utilise a human pose estimator to retrieve crucial keypoint information. Specifically, given an image x, the pose estimator extract M landmarks, with the coordinate and confidence score of each landmark $\{L_i = \{y_i, x_i, c_i\} \mid i = 1, 2, ..., M\}$. These landmarks are utilised to generate heatmaps $\text{HT} = \{\text{HT}_1, \text{HT}_2, ..., \text{HT}_M\}$. We set a threshold δ to distinguish visible and occluded landmarks in pedestrian images. If a landmark is occluded, the confidence score is below δ. Conversely, if a landmark is visible, the confidence score exceeds δ. Formally, the visible label of a landmark vl can be illustrated as follows:

$$vl_i = \begin{cases} 0 & c_i < \delta \\ 1 & c_i \geq \delta \end{cases} (i = 1, ..., M). \tag{3}$$

where c_i denotes the confidence score of i-th landmark.

3.2 | Hard Partition Branch

The soft partition strategy via the pose estimator provides the adjacency relationship of human body parts. However, the keypoint features extracted are at pixel level, which loses the highly correlated information around keypoints. Therefore, we propose to extract region-level features in the hard partition branch. The local features are obtained by regrouping patch tokens and feeding them into a shared transformer layer. Then, we refine and classify them by filtering occlusions based on the visibility of keypoints. Specifically, patch tokens f_p are firstly

rearranged into a new form according to their original position information. The size of the rearranged patch tokens is $\lfloor \frac{H-P}{S} + 1 \rfloor \times \lfloor \frac{W-P}{S} + 1 \rfloor$, and then it is divided into m groups from a horizontal view. The shared global token f_g is concatenated with each group. After that, m groups are inputted into a shared transformer layer to learn the relationship of global and local features, represented as $\left\{f_l^j \mid j = 1, 2, \cdots, m\right\}$ and f_l^j is the output of j-th group. For the occluded person Re-ID task, it is crucial to explore discriminative features from nonoccluded regions. We introduce the coordinate and visible label of human keypoints offered by the pose estimator to judge whether a local region is nonoccluded. Existing methods state that if a local region contains visible keypoints, it is deemed to be nonoccluded. However, when a local region has visible and invisible key points simultaneously, some nonoccluded regions may still contain occlusions. Therefore, we propose to classify local features into three categories: *nonoccluded features*, *semi-occluded features,* and *occluded features* for better filtering of occlusions. Formally, since the canvas size of coordinates is same with the size of the input image, we map the canvas size of coordinates to be consistent with the rearranged patch tokens, which facilitates the retrieving of the amount of keypoints in group features. Then, group features are divided into nonoccluded, semi-occluded, and occluded features according to the visibility of all key points in each group feature. And the division criterion is: if all human keypoints are visible within a certain local region, it is termed a nonoccluded region f_{non}; Conversely, if all human keypoints are invisible or nonexistent within a local region, it is considered as an occluded region f_{occ}; Otherwise, it is considered as a semi-occluded region f_{semi}.

3.3 | Soft Partition Branch

Although region-level features obtained by the hard partition strategy preserve the local region information, they ignore the correlated relationship of each human body part. Therefore, we propose to learn interconnections among human body parts in the soft partition branch. By introducing a pose estimator, we can directly extract keypoint embeddings. Furthermore, we introduce a graph attention network to identify interconnections between different human body parts and aggregate features from global and keypoint embeddings.

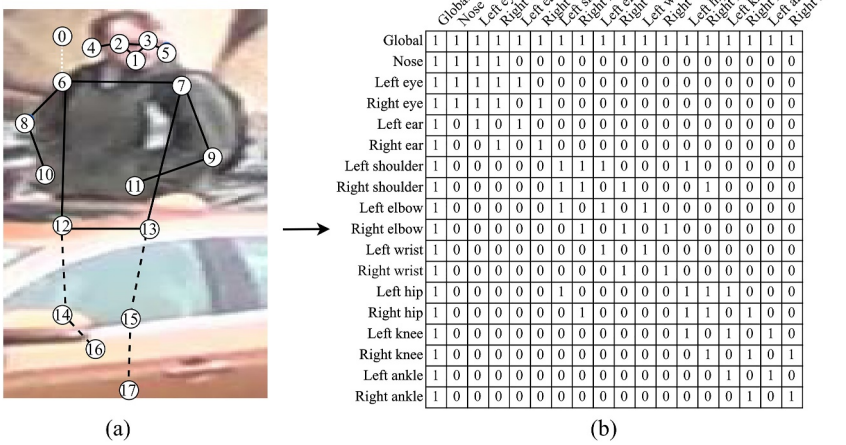
Firstly, we construct the graph, denoted by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, and generate the keypoint embeddings $\mathcal{V} = \left\{f_{kp}^1, f_{kp}^2, ..., f_{kp}^M\right\}$ via element-wise product of the global feature f_g and the heatmaps $\text{HT} = \{\text{HT}_1, \text{HT}_2, ..., \text{HT}_M\}$. As represented in Figure 3a, the obtained keypoint embeddings serve as nodes within the graph, with edges categorised into two distinct types: keypoint-keypoint edges $\mathcal{E}_{kk}$ and global-keypoint edges $\mathcal{E}_{gk}$. The former defines interconnections among the inherent joints in the human keypoints, and also includes the self-connection of nodes. In contrast, the latter defines fully relationships extending from global aspect to each keypoint (for simplicity, only one dotted line is used to illustrate global-keypoint edges). Besides, we define the adjacent matrix in Figure 3b, where the value is set to 1 when there is an adjacent relationship between keypoints, or 0 otherwise.

Then, we introduce a graph attention network to update the embedding of each node by aggregating features from its adjacency nodes. The graph attention network consists of eight graph attention layers, and the structure of each layer [30] is shown in Figure 4. Specially, as represented in Figure 4a, we initially compute the weight e_{ij} for each edge, representing the relevance of the neighbouring node j to the current node:

$$e_{ij} = a\left(\left[Wv_i \| Wv_j\right]\right); j \in \mathcal{N}_i. \tag{4}$$

where $W \in \mathbb{R}^{d \times d}$ and $a \in \mathbb{R}^{2d}$ are learnable parameters for transforming input node embeddings, the $[\,\|\,]$ denotes the operation of embedding concatenate and $\mathcal{N}_i$ is the set of adjacent nodes. Then we normalise the edge weights across all the neighbouring nodes for current node to calculate the attention coefficient α_{ij}:

$$\alpha_{ij} = \text{Softmax}\left(e_{ij}\right) = \frac{\exp\left(\text{LeakyReLU}\left(e_{ij}\right)\right)}{\sum_{k \in \mathcal{N}_i} \exp\left(\text{LeakyReLU}\left(e_{ik}\right)\right)}. \tag{5}$$



(a) (b)

FIGURE 3 | The designs in adjacent matrix. (a) The setting of nodes and edges in the defined graph. Solid line defines the relationships between keypoints, and the dotted line defines the relationships between global and keypoints. (b) The definition of adjacent matrix.

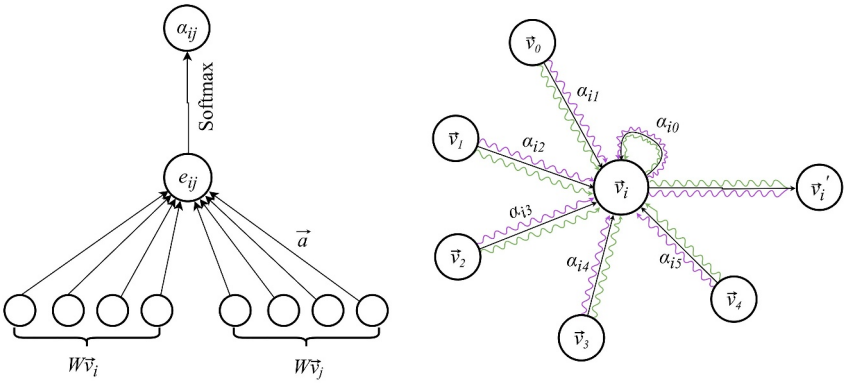


FIGURE 4 | The framework of graph attention network. (a) The process of computing the weight e_{ij} via the attention mechanism and normalising the edge weight to calculate the attention coefficient α_{ij}. (b) The aggregating operation of keypoint graph attention.

With normalised attention coefficient, the updated node embedding v_i is computed by all neighbouring nodes and summed by the average weights in Figure 4b:

$$v_i' = \text{LeakyReLU}\left(\sum_{j\in\mathcal{N}_i}\alpha_{ij}Wv_j\right). \tag{6}$$

Finally, we obtain global-keypoint embedding f_{gk} and keypoint embeddings f_{kp}' as outputs. Graph representation learning is introduced to establish relationships between the keypoint and global embeddings. It aggregates semantic features with specified edges, preserving the semantic information of human body parts.

The graph attention network evaluates the significance of each edge, and aggregates the node embedding by the attention coefficient among adjacent nodes, lessening the unfavourable impact of occlusions.

3.4 | Optimisation

The KGFP framework is trained in an end-to-end manner using the multi-task loss functions: *an identity loss*, *a triplet loss*, *a centre loss* and *a dissimilarity loss*.

3.4.1 | Identity Loss

The person re-identification can be regarded as a multi-class task for feature learning, where each person ID is a class. We adopt a cross-entropy loss as the identity loss for global features, nonoccluded features, and semi-occluded features. The loss is formulated as follows:

$$L_{\text{id}} = L_{\text{id}}(f_g) + L_{\text{id}}(f_{\text{non}}) + L_{\text{id}}(f_{\text{semi}}). \tag{7}$$

3.4.2 | Triplet Loss

The triplet loss involves an anchor sample, a positive sample, and a negative sample. By minimising the distance between positive sample pairs, while maximising the distance between negative pairs to make the model learn more discriminative features, we implement the triplet loss for global features, non-occluded features, and semi-occluded features, and the loss function is expressed as follows:

$$L_{\text{tri}} = L_{\text{tri}}(f_g) + L_{\text{tri}}(f_{\text{non}}) + L_{\text{tri}}(f_{\text{semi}}). \tag{8}$$

3.4.3 | Dissimilarity Loss

The semi-occluded feature contains partial human body parts, while the occluded feature contains nonhuman body parts. The semi-occluded feature and occluded feature should not have strong similarities. If the semi-occluded and occluded features are similar, the model is prone to mistakenly learn human-body parts as nonhuman body features. To mitigate this, we develop a dissimilarity loss to make the model focus on human body parts in the hard partition branch. The loss can be formulated as follows:

$$f'_{\text{semi}} = \text{avgpooling}(f_{\text{semi}}). \tag{9}$$

$$f'_{\text{occ}} = \text{avgpooling}(f_{\text{occ}}). \tag{10}$$

$$L_{\text{dis}} = D(f'_{\text{semi}}, f'_{\text{occ}}). \tag{11}$$

where $D(,)$ denotes the cosine similarity.

3.4.4 | Centre Loss

Because of the varying capture times of images and the continuous movement of pedestrians, there are pose variations among images of the same pedestrians. We introduce the centre loss from the face recognition problem, which can further mitigate the influence of pose variation. We employ the centre loss for global-keypoint embedding to constrain the feature distribution of each keypoint and the loss can be formulated as follows:

$$\mathcal{L}_{\text{ce}} = \frac{1}{2}\sum_{j=1}^{B}\|f_{gk_j} - \mathbf{c}_{y_j}\|_2^2. \tag{12}$$

where B is the number of batch size, c_{y_j} denotes the class centre feature labelled y_j and y_j denotes the label of j-th image in a mini-batch. The overall objective function can be illustrated as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{id}} + \lambda_2 \mathcal{L}_{\text{tri}} + \mathcal{L}_{\text{dis}} + \mathcal{L}_{\text{ce}}. \tag{13}$$

where λ_1 and λ_2 are set to 1, which denotes the balance factor between identity loss and triplet loss.

3.5 | Feature-Alignment-Based Graph

To improve the accuracy and performance in testing an image retrieval task Re-ID, most existing methods propose matching strategies to align human body parts. As shown in Figure 5, PCB [4] splits the image into stripes and matches the corresponding local features between the query and gallery images. However, occlusions and misalignment of positions are disregarded. PMFB [12] proposes to compare the commonly visible parts in the query and gallery images. However, this method is only suitable for images with strong occlusion. Otherwise, it still introduces occlusions. To reduce the impact of occlusions, when aligning two images, we take it as a graph matching problem and further propose a feature-alignment-based graph strategy to calculate the distance of graph representation between the query and gallery images. Furthermore, the confidence is considered as a weight to prevent matching occluded keypoint embeddings, allowing us to fully utilise non-occluded features and improve performance. The skeleton distance between the query and gallery images is defined as follows:

$$d_{sk} = \sum_{i=1}^{M} c_i^q c_i^g d\left(f'^q_{kp_i}, f'^g_{kp_i}\right). \tag{14}$$

where c_i^q and c_i^g are the confidence score of i-th keypoint of query and gallery image; $f'^q_{kp_i}$ and $f'^g_{kp_i}$ denote the keypoint embeddings of i-th keypoint of query and gallery image, and $d(,)$ denotes the cosine similarity.

Besides, we compute part-based distance by the following equation:

$$d_p = d\left(f_g^q, f_g^g\right) + d\left(f_{\text{non}}^q, f_{\text{non}}^g\right) + d\left(f_{gk}^q, f_{gk}^g\right). \tag{15}$$

where f_g and f_{non} and f_{gk} are the global feature, nonoccluded features, and global-keypoint feature of the query and gallery images, respectively.

The total distance is computed through simply adding the distance of skeleton graph and part-based features as follows:

$$\text{dist} = d_{sk} + d_p. \tag{16}$$

where dist denotes the final distance.

4 | Experiments

4.1 | Datasets and Evaluation Metrics

In order to demonstrate the effectiveness of the proposed method, we conduct extensive experiments on three large-scale datasets, including Occluded-DukeMTMC [31], Market1501 [32] and DukeMTMC-reID [33]. The Cumulative Matching Characteristic (CMC) and mean average precision (mAP) metrics are adopted to evaluate the performance.

4.2 | Implementation Details

Both the training and testing images are resized to 256 × 128. The training images are augmented through random horizontal flipping, padding, random cropping, and random erasing as outlined in ref. [34]. The initial weights of ViT are pre-trained on ImageNet21K and then finetuned on ImageNet-1K. In this paper, we perform all the experiments with one Nvidia A100 GPU using PyTorch. The batch size is set to 64 with 4 images per ID, and the model is trained for 120 epochs. SGD optimiser is employed with a momentum of 0.9 and a weight decay of 1e-4. The learning rate is initialised as 0.008 with cosine learning rate decay. The stride size is 16 × 16. To detect landmarks from images, we adopt HRNet [35] pre-trained on the COCO dataset. The threshold δ is set to 0.2. The number of groups m is set to 16.

4.3 | Comparison With State-of-The-Art Methods

4.3.1 | Results on Occluded Dataset

In experiments, we compare our proposed method with the prevailing occluded person Re-ID methods, and implement the standard ViT method as baseline. As shown in Table 1, we present the evaluation results on the Occluded-DukeMTMC. We compare the accuracy of Rank-1, Rank-5, and Rank-10, as well as the mAP metric of these methods. Existing methods are roughly divided

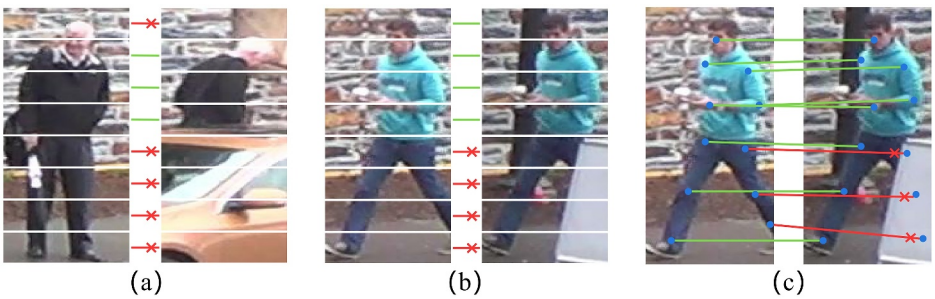


FIGURE 5 | Different matching strategies. (a) Matching corresponding local features. (b) Matching commonly visible local features. (c) The proposed matching strategy matches commonly visible keypoints.

TABLE 1 | Performance comparison with state-of-the-art methods on occluded dataset.

	Methods	Ref	Occluded-DukeMTMC				Market1501		DukeMTMC-reID	
			R-1	R-5	R-10	mAP	R-1	mAP	R-1	mAP
Hard partition	PCB [2]	ECCV 18	42.6	57.1	62.9	33.7	92.3	77.4	81.8	66.1
	TransReID [9]	ICCV 21	64.2	—	—	55.7	95.1	88.2	89.5	80.5
	PAGCN [26]	SPIC 22	57.1	73.3	80.2	44.0	94.4	87.3	86.7	78.0
	OCNet [19]	CIASSP 22	59.9	—	—	49.7	94.9	87.2	87.8	77.2
	QPM [11]	TMM 22	64.4	79.3	84.2	49.3	—	—	—	—
	FGMFN [10]	JCVIR 22	65.5	80.4	84.4	55.8	90.6	85.0	88.5	80.4
	RTGAT [36]	TIP 23	61.0	69.7	73.6	50.1	93.3	85.1	88.0	76.9
	Liu et al. [37]	CVIU 23	62.4	—	—	54.9	—	—	—	—
	SCAT [38]	TII 23	62.5	78.1	83.1	54.9	95.1	88.0	89.3	79.8
	SCSRL [39]	Neurocomputing 23	62.6	—	—	51.4	95.0	86.2	—	—
	AET-Net [40]	Appl. Sci 23	64.5	—	—	54.5	94.8	87.5	89.5	80.1
Soft partition	PGMANet [41]	IJCNN 21	51.3	66.5	73.4	40.9	—	—	—	—
	HOReID [23]	CVPR 22	55.1	—	—	43.8	94.2	84.9	86.9	75.6
	DSF [42]	NCA 23	56.8	—	—	45.9	94.6	86.2	88.2	76.3
	EGMN [43]	PAA 23	60.5	—	—	49.2	94.3	84.6	87.7	75.6
	MFEN [44]	ACCESS 23	60.8	—	—	47.6	95.3	85.7	88.3	77.3
	EcReID [45]	Symmetry 23	64.8	—	—	52.7	95.5	87.2	88.3	78.5
Hybrid partition	PGFA [31]	ICCV 19	51.4	68.6	74.9	37.3	91.2	76.8	82.6	65.5
	Pirt [46]	ACM MM 21	60.0	—	—	50.9	94.1	86.3	88.9	77.6
	Yang et al. [14]	ICCV 21	62.2	—	—	46.3	—	—	—	—
	DAReID [16]	K0BS 21	63.4	77.5	82.5	53.2	94.6	87.0	88.9	78.4
	PMFB [12]	TNNLS 22	56.3	72.4	78.0	43.5	92.7	81.3	86.2	72.6
	Pang et al. [47]	JVCIR 23	62.2	—	—	50.7	94.3	85.9	86.8	75.2
	FPTE [48]	MTA 23	57.8	—	—	48.3	94.3	87.0	88.7	77.2
	GPEOG [49]	ICME 23	64.1	78.7	83.5	51.2	94.8	84.5	87.5	75.5
	ViT (baseline)		60.2	76.2	81.8	53.2	94.3	86.9	89.1	79.5
	KGFP (ours)		**67.0**	**80.4**	**85.3**	**58.4**	**94.5**	**87.7**	**89.8**	**80.8**

Note: The best results are bolded and the runner-ups are underlined.

into three categories, including methods based on hard partition strategy, methods based on soft partition strategy, and methods based on hybrid partition strategies.

In Table 1, the proposed method KGFP exhibits impressive results, with a performance of 67.0% Rank-1 and 58.4% mAP accuracy, outperforming all existing methods. Notably, compared to the methods in the first group, the proposed KGFP addresses the limitations of the hard partition strategy by additionally exploring the interconnections among human body parts through a graph attention network. Compared to the methods in the second group, the proposed KGFP introduces an attention mechanism in the graph convolution network to reduce the impact of occlusions. This is in contrast to the graph convolution network that equally treats every node. In addition, the keypoint embedding is a pixel-level feature representation that loses the contextual information around the keypoint. To address these limitations, we add region-level features to enrich the feature representation in the hard partition strategy branch. Compared to the methods in the third group, which use the hybrid partition strategies for feature learning and achieve image matching by strict alignment, the proposed KGFP considers image alignment as a graph matching problem and has surpassed them by at least +2.6% Rank-1 and +7.1% mAP, which demonstrates the effectiveness of feature-alignment-based graph in the proposed method.

4.3.2 | Results on Holistic Datasets

We also conduct experiments on two holistic datasets, including Market1501 and DukeMTMC-reID, where the full body of the pedestrian is available. As shown in Table 1, these methods can also be divided into three categories: hard-partition based, soft-partition based, and hybrid-partition based. It is evident that while the results on Market1501 may not meet initial expectations, they still demonstrate comparable performance to some existing methods, suggesting that our method possesses generality and universality for the Re-ID task. In comparison to other methods, our method attains the best Rank-1 and mAP on the DukeMTMC-reID dataset with 89.8% and 80.8%. It is obvious that the proposed method can obtain an impressive performance on DukeMTM-reID, which contains complex scenes and occlusions in real environment. These results also indicate that the proposed method is robust and applicable to both occluded Re-ID and holistic Re-ID tasks.

4.4 | Ablation Studies

In this section, we conducted extensive ablation experiments on the proposed modules to validate their effectiveness. As depicted in Table 2, we present various settings for the proposed three modules and perform a detailed comparison in each module. The ablation experiments adopt the Occluded-DukeMTMC dataset and all research modes by default.

4.4.1 | Effectiveness of the Hard Partition Branch

As shown in Table 2, we first train the baseline method. In the hard partition branch, the $hard_1$ (Index-2) denotes that local features are divided into nonoccluded and occluded features under the guidance of the visibility of keypoints, and the $hard_2$ (Index-3) denotes that local features are divided into nonoccluded, semi-occluded, and occluded features. It can be noticed that splitting local features by the visibility of keypoints can contribute to performance improvements. The $hard_1$ divides local features into two categories, and as long as a local region has visible keypoints, it is a nonoccluded region. However, the region still contains noise when both visible and invisible keypoints exist simultaneously. In contrast, the $hard_2$ further classify the nonoccluded regions into a nonoccluded region when all keypoints are visible and a semi-occluded region when both visible and invisible keypoints exist simultaneously. The results exhibit that the proposed method improves with +0.7% Rank-1 and +0.5% mAP, which exhibits the effectiveness of the hard partition branch.

TABLE 2 | Ablation studies of the hard partition branch, the soft partition branch and the feature-alignment-based graph.

Index	Methods	R-1	R-5	R-10	mAP
1	Baseline (B)	60.2	76.5	82.0	53.2
2	B + $hard_1$	64.5	78.5	83.3	56.3
3	B + $hard_2$	65.2	78.8	82.9	56.8
4	B + $soft_1$	60.6	76.7	82.3	53.5
5	B + $soft_2$	62.5	78.9	84.1	54.5
6	B + $soft_3$	64.7	78.8	83.5	57.0
7	B + $hard_2$ + $soft_3$ + $match_1$	66.5	80.0	83.8	57.4
8	B + $hard_2$ +$soft_3$ + $match_2$	67.0	80.4	85.3	58.4

4.4.2 | Effectiveness of the Soft Partition Branch

In the soft branch, we set up three different experimental settings. The $soft_1$ (Index-4) involves extracting keypoint embeddings by keypoint heatmaps without incorporating graph learning. Subsequently, we introduce a graph learning to explore the relationship of keypoint-keypoint embeddings in the $soft_2$ (Index-5), and add global-keypoint edges to mine interconnections among global and keypoints embeddings in the $soft_3$ (Index-6). Compared with Index-1, adding keypoint information by pose estimator B+ $soft_1$ can bring a slight improvement in performance. By introducing the graph attention network to explore relationships among keypoints, B+ $soft_2$ achieves the performance 62.5% Rank-1 and 54.5% mAP, which shows the crucial significance of exploring keypoint relationships. Additionally, by adding the global keypoint edges in the graph attention network, the performance of the Index-6 further increases by +1.8% Rank-1 and 2.5% mAP. It demonstrates that we can extract discriminative local features under the guidance of keypoints and graph learning.

4.4.3 | Effectiveness of Feature-Alignment-Based Graph

The $match_1$ (Index-7) denotes computing the distance only by part-based features, and we add the graph matching in the $match_2$ (Index-8) to calculate the distance of keypoint features.

Existing image-matching strategies split images into stripes and compare the corresponding stripes, which cannot achieve impressive performance when occlusions occur or position misalignment happens. We view image alignment as a graph matching problem, which can solve the position misalignment problem and alleviate the influence of occlusions. As shown in Index-8, our method achieves the best performance at 67.0% Rank-1 and 58.4% mAP, which further increases by +0.5% Rank-1 and +1.0% mAP compared to Index-7, which further illustrates the effectiveness of graph matching.

4.5 | Parameters Analysis

4.5.1 | Analysis of Groups m

In the hard partition branch, we divide the feature map into m groups and the number of groups denotes the size of region. For the convenience of locating key points in that horizontal stripe, we map the coordinates into 16 × 8. We set the m as 4, 8, and 16 to analyse the impact of groups. And we consider that when all keypoints are visible in a region, it is a nonoccluded feature. Therefore, when m is smaller, a group contains more keypoints, and it is more easily defined as a semi-occluded feature. It will further negatively affect the discriminative feature learning of nonoccluded features. As shown in Table 3, it can be observed that when $m = 16$, we get the best results 67.0% Rank-1 and 58.4% mAP, which demonstrates the importance of region-level features.

TABLE 3 | Parameter analysis of groups m.

m	R-1	R-5	R-10	mAP
4	64.0	78.7	83.6	55.5
8	66.6	79.7	85.0	57.3
16	67.0	80.4	85.3	58.4

Note: Comparison in CMC and mAP with different settings of the group m in hard partition branch on Occluded-DukeMTMC.

4.5.2 | Analysis of Threshold δ

The threshold is defined to generate the visibility label of keypoints, which contributes to explicitly classifying local features. To find the most suitable threshold δ, we conduct extensive experiments by varying δ from 0 to 0.7. When the threshold is set smaller than 0.2, all keypoints may be viewed as visible. Specifically, when $\delta = 0$, it indicates that all local features are visible. Although it is inevitable to introduce noise, including background clutters and occlusions, we still achieve 65.0% Rank-1 and 56.7% mAP. When the threshold is set larger than 0.2, it is easy to incorrectly classify visible local features as invisible features, which leads to some certain visible body part regions being lost. As shown in Figure 6, when the threshold $\delta = 0.2$, it achieves the best performance, which demonstrates the robustness of the proposed method to filter out background clutters and occlusions.

4.6 | Effectiveness of Each Loss

4.6.1 | Identity Loss

$\mathcal{L}_{id}$ trains Re-ID models as a classification task by predicting the identity for every image. As shown in Table 4, when $\mathcal{L}_{id}$ is removed, the performance drops by +7.5% Rank-1 and +6.0%

TABLE 4 | Effectiveness of each component of the overall loss function over the Occluded-DukeMTMC.

$\mathcal{L}_{id}$	$\mathcal{L}_{tri}$	$\mathcal{L}_{ce}$	$\mathcal{L}_{dis}$	R-1	R-5	R-10	mAP
✓	✓	✓	✓	67.0	80.4	85.3	58.4
✗	✓	✓	✓	59.2	74.9	81.0	52.3
✓	✗	✓	✓	61.4	75.4	80.5	50.1
✓	✓	✗	✓	66.0	80.5	84.8	57.9
✓	✓	✓	✗	66.0	79.3	83.8	57.0

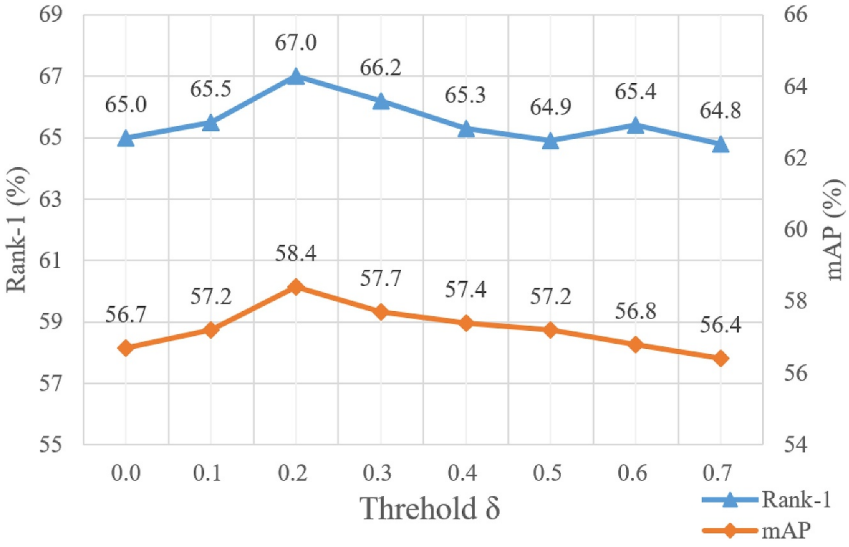


FIGURE 6 | Parameter analysis of threshold δ. Comparison in Rank-1 and mAP with different settings of threshold δ on Occluded-DukeMTMC.

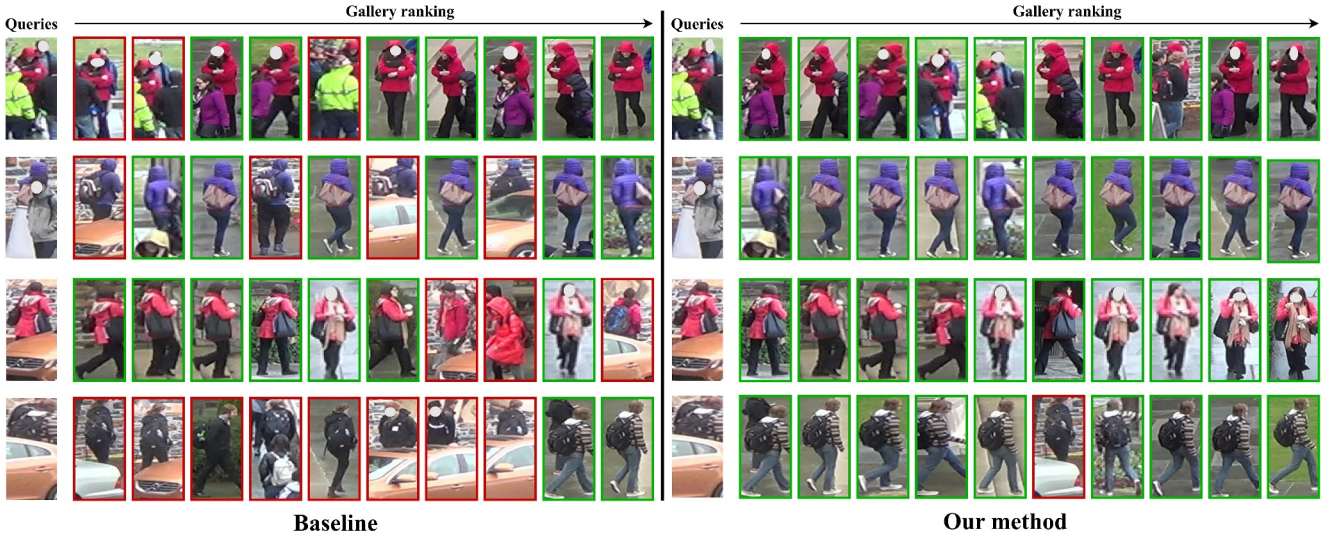


FIGURE 7 | Retrieval results of the baseline and the proposed KGFP method. Green and red rectangles indicate correct and error retrieval results.

mAP, which indicates that the $\mathcal{L}_{id}$ plays a fundamental role in discriminative feature learning.

4.6.2 | Triplet Loss

$\mathcal{L}_{tri}$ trains Re-ID models as a rank problem, ensuring that an image of a specific person is closer to intra-class images than inter-class images. As depicted in Table 4, by removing $\mathcal{L}_{tri}$, the performance drops +5.3% Rank-1 and +8.2% mAP, which shows that $\mathcal{L}_{tri}$ can enhance the discriminative capability of the Re-ID task.

4.6.3 | Centre Loss

$\mathcal{L}_{ce}$ aims to provide a centre feature for each class, minimising the distance between intra-class. To this end, it further alleviates the interference of pose variation. As shown in Table 4, when $\mathcal{L}_{ce}$ is removed, the performance drops +0.7% Rank-1 and +0.4% mAP, validating its effectiveness in mitigating the impacts of pose variation.

4.6.4 | Dissimilarity Loss

Dissimilarity loss $\mathcal{L}_{dis}$ aims to push the semi-occluded and occluded features away, making the model focus on human body parts. To this end, it further mitigates the influence of occlusions. As shown in Table 4, without the $\mathcal{L}_{dis}$, the performance drops +0.7% Rank-1 and +1.3% mAP, which validates the effectiveness in reducing the impacts of occlusions.

4.7 | Visualisation

We compare the retrieval results between the baseline and the proposed KGFP method on Occluded-DukeMTMC. The results are organised from left to right based on their similarity score. The green rectangles represent correct matches, while the red rectangles represent incorrect matches. As presented in Figure 7, our method can almost accurately match the same person, which proves the effectiveness of our proposed method.

5 | Conclusion

In this paper, we propose a new keypoint-guided feature partition (KGFP) method for occluded person Re-ID tasks. We develop innovative modules in the hard partition branch and soft partition branch under the guidance of keypoint information, respectively. In the hard partition branch, we firstly split patch tokens obtained by ViT into groups to explore the region-level features, which can repair information around keypoints in the soft partition branch. And then, we classify them into nonoccluded, semi-occluded, and occluded features by the visibility label of each keypoint to reduce the impact of occlusions. In the soft partition branch, in order to enhance interconnections among human parts, a graph attention network is introduced to further mine the human semantic graph structure. Moreover, person Re-ID is treated as a graph-matching problem, and a feature-alignment-based graph strategy is proposed to improve the retrieval accuracy during the reference stage. Besides, a dissimilarity loss is designed that pushes the semi-occluded and occluded features away to focus on human parts. Experiments on the Occluded-DukeMTMC, Market1501, and DukeMTMC-reID demonstrate the effectiveness of the proposed KGFP method.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China (Grant No. 62276038), and in part by the Scientific and Technological Research Program of Chongqing Municipal Education Commission (Grant No. KJZD-M202400603), and in part by the Project of Key Laboratory of Tourism Multisource Data Perception and Decision, Ministry of Culture and Tourism (Grant No. H2023009).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

[Occluded-DukeMTMC]

The Occluded-DukeMTMC dataset used in this study is derived from the DukeMTMC-reID dataset. The resources are available at https://github.com/lightas/Occluded-DukeMTMC-Dataset. Researchers wishing to use this dataset must first obtain the DukeMTMC-reID dataset independently and then apply the conversion script [31].

[Market1501]

The Market-1501 dataset that supports the findings of this study is openly available in the public domain. The dataset contains training, query, and gallery partitions for person re-identification research and can be downloaded from the original release site at https://www.kaggle.com/datasets/sachinsarkar/market1501, or from other repositories hosting the dataset. These data are freely accessible under the original licencing terms [32].

[DukeMTMC-reID]

The DukeMTMC-reID dataset used in this study is a subset of the original DukeMTMC multi-camera tracking dataset. The dataset is available at https://www.kaggle.com/datasets/igorkrashenyi/dukemtmc-reid. Researchers should ensure compliance with licencing and ethical use requirements when downloading and using this dataset [33].

References

1. L. Zheng, Y. Yang, and A. G. Hauptmann, "Person Re-Identification: Past, Present and Future," *arXiv preprint arXiv:1610.02984* (2016).

2. 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 *Proceedings of the European Conference on Computer Vision (ECCV)* (Springer, 2018), 480–496.

3. C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-Guided Contrastive Attention Model for Person Re-Identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE CVPR, 2018), 1179–1188.

4. Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, "Relation-Aware Global Attention for Person Re-Identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE CVPR, 2020), 3186–3195.

5. G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning Discriminative Features With Multiple Granularities for Person Re-Identification," in *Proceedings of the 26th ACM International Conference on Multimedia* (ACM MM, 2018), 274–282.

6. Y. Shen, H. Li, S. Yi, D. Chen, and X. Wang, "Person Re-Identification With Deep Similarity-Guided Graph Neural Network," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Springer, 2018), 486–504.

7. G. Chen, T. Gu, J. Lu, J. A. Bao, and J. Zhou, "Person Re-Identification via Attention Pyramid," *IEEE Transactions on Image Processing* 30 (2021): 7663–7676, https://doi.org/10.1109/TIP.2021.3107211.

8. X. Qian, Y. Fu, T. Xiang, Y. G. Jiang, and X. Xue, "Leader-Based Multi-Scale Attention Deep Architecture for Person Re-Identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, no. 2 (2019): 371–385, https://doi.org/10.1109/TPAMI.2019.2928294.

9. S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "TransReID: Transformer-Based Object Re-Identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (IEEE ICCV, 2021), 15013–15022.

10. G. Zhang, C. Chen, Y. Chen, H. Zhang, and Y. Zheng, "Fine-Grained-Based Multi-Feature Fusion for Occluded Person Re-Identification," *Journal of Visual Communication and Image Representation* 87 (2022): 103581, https://doi.org/10.1016/J.JVCIR.2022.103581.

11. P. Wang, C. Ding, Z. Shao, Z. Hong, S. Zhang, and D. Tao, "Quality-Aware Part Models for Occluded Person Re-Identification," *IEEE Transactions on Multimedia* 25 (2022): 3154–3165, https://doi.org/10.1109/TMM.2022.3156282.

12. J. Miao, Y. Wu, and Y. Yang, "Identifying Visible Parts via Pose Estimation for Occluded Person Re-Identification," *IEEE Transactions on Neural Networks and Learning Systems* 33, no. 9 (2021): 4624–4634, https://doi.org/10.1109/TNNLS.2021.3059515.

13. 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE CVPR, 2021), 2898–2907.

14. J. Yang, J. Zhang, F. Yu, et al., "Learning to Know Where to See: A Visibility-Aware Approach for Occluded Person Re-Identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (IEEE ICCV, 2021), 11885–11894.

15. M. Jia, X. Cheng, S. Lu, and J. Zhang, "Learning Disentangled Representation Implicitly via Transformer for Occluded Person Re-Identification," *IEEE Transactions on Multimedia* 25 (2022): 1294–1305, https://doi.org/10.1109/TMM.2022.3141267.

16. Y. Xu, L. Zhao, and F. Qin, "Dual Attention-Based Method for Occluded Person Re-Identification," *Knowledge-Based Systems* 212 (2021): 106554, https://doi.org/10.1016/J.KNOSYS.2020.106554.

17. A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale," *arXiv preprint arXiv:2010.11929* (2020).

18. K. Wang, H. Wang, M. Liu, X. Xing, and T. Han, "Survey on Person Re-Identification Based on Deep Learning," *CAAI Transactions on Intelligence Technology* 3, no. 4 (2018): 219–227, https://doi.org/10.1049/TRIT.2018.1001.

19. M. Kim, M. Cho, H. Lee, S. Cho, and S. Lee, "Occluded Person Re-Identification via Relational Adaptive Feature Correction Learning," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), 2719–2723.

20. X. Zhang, H. Luo, X. Fan, et al., "AlignedReID: Surpassing Human-Level Performance in Person Re-Identification," *arXiv preprint arXiv:1711.08184* (2017): abs/1711.08184.

21. L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-Learned Part-Aligned Representations for Person Re-Identification," in *Proceedings of the IEEE International Conference on Computer Vision* (IEEE ICCV, 2017), 3219–3228.

22. W. Li, X. Zhu, and S. Gong, "Harmonious Attention Network for Person Re-Identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE CVPR, 2018), 2285–2294.

23. G. Wang, S. Yang, H. Liu, et al., "High-Order Information Matters: Learning Relation and Topology for Occluded Person Re-Identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE CVPR, 2020), 6449–6458.

24. J. Zhang, G. Ye, Z. Tu, et al., "A Spatial Attentive and Temporal Dilated (SATD) GCN for Skeleton-Based Action Recognition," *CAAI Transactions on Intelligence Technology* 7, no. 1 (2022): 46–55, https://doi.org/10.1049/CIT2.12012.

25. Z. M. Chen, X. S. Wei, P. Wang, and Y. Guo, "Multi-Label Image Recognition With Graph Convolutional Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE CVPR, 2019), 5177–5186.

26. J. Zhang, J. P. Ainam, W. Song, Lh Zhao, X. Wang, and H. Li, "Learning Global and Local Features Using Graph Neural Networks for

Person Re-Identification," *Signal Processing: Image Communication* 107 (2022): 116744, https://doi.org/10.1016/J.IMAGE.2022.116744.

27. H. Pan, Y. Bai, Z. He, and C. Zhang, "AAGCN: Adjacency-Aware Graph Convolutional Network for Person Re-Identification," *Knowledge-Based Systems* 236 (2022): 107300, https://doi.org/10.1016/J.KNOSYS.2021.107300.

28. T. Liang, Y. Jin, W. Liu, S. Feng, T. Wang, and Y. Li, "Keypoint-Guided Modality-Invariant Discriminative Learning for Visible-Infrared Person Re-Identification," in *Proceedings of the 30th ACM International Conference on Multimedia* (ACM MM, 2022), 3965–3973.

29. D. Chen, A. Döring, S. Zhang, J. Yang, J. Gall, and B. Schiele, "Keypoint Message Passing for Video-Based Person Re-Identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36 (AAAI, 2022), 239–247.

30. P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph Attention Networks," *arXiv preprint arXiv:1710.10903* (2017).

31. J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, "Pose-Guided Feature Alignment for Occluded Person Re-Identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (IEEE ICCV, 2019), 542–551.

32. L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable Person Re-Identification: A Benchmark," in *Proceedings of the IEEE International Conference on Computer Vision* (IEEE ICCV, 2015), 1116–1124.

33. Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled Samples Generated by GAN Improve the Person Re-Identification Baseline in Vitro," in *Proceedings of the IEEE International Conference on Computer Vision* (IEEE ICCV, 2017), 3754–3762.

34. Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random Erasing Data Augmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34 (AAAI, 2020), 13001–13008.

35. K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep High-Resolution Representation Learning for Human Pose Estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE CVPR, 2019), 5693–5703.

36. M. Huang, C. Hou, Q. Yang, and Z. Wang, "Reasoning and Tuning: Graph Attention Network for Occluded Person Re-Identification," *IEEE Transactions on Image Processing* 32 (2023): 1568–1582, https://doi.org/10.1109/tip.2023.3247159.

37. Z. Liu, X. Mu, Y. Lu, T. Zhang, and Y. Tian, "Learning Transformer-Based Attention Region With Multiple Scales for Occluded Person Re-Identification," *Computer Vision and Image Understanding* 229 (2023): 103652, https://doi.org/10.1016/J.CVIU.2023.103652.

38. H. Fan, X. Wang, Q. Wang, S. Fu, and Y. Tang, "Skip Connection Aggregation Transformer for Occluded Person Reidentification," *IEEE Transactions on Industrial Informatics* 20, no. 1 (2023): 442–451, https://doi.org/10.1109/tii.2023.3266372.

39. S. Han, D. Liu, Z. Zhang, and D. Ming, "Spatial Complementary and Self-Repair Learning for Occluded Person Re-Identification," *Neurocomputing* 546 (2023): 126360, https://doi.org/10.1016/J.NEUCOM.2023.126360.

40. J. Wang, P. Li, R. Zhao, R. Zhou, and Y. Han, "CNN Attention Enhanced ViT Network for Occluded Person Re-Identification," *Applied Sciences* 13, no. 6 (2023): 3707, https://doi.org/10.3390/app13063707.

41. Y. Zhai, X. Han, W. Ma, X. Gou, and G. Xiao, "PGMANet: Pose-Guided Mixed Attention Network for Occluded Person Re-Identification," in *2021 International Joint Conference on Neural Networks (IJCNN)* (IEEE IJCNN, 2021), 1–8.

42. Y. Fan, X. Gong, and Y. He, "DSF-Net: Occluded Person Re-Identification Based on Dual Structure Features," *Neural Computing and Applications* 35, no. 4 (2023): 3537–3550, https://doi.org/10.1007/S00521-022-07927-6.

43. S. Zhou and M. Zhang, "Occluded Person Re-Identification Based on Embedded Graph Matching Network for Contrastive Feature Relation," *Pattern Analysis and Applications* 26, no. 2 (2023): 487–503, https://doi.org/10.1007/S10044-022-01123-X.

44. Z. Liu, Q. Wang, M. Wang, and Y. Zhao, "Occluded Person Re-Identification With Pose Estimation Correction and Feature Reconstruction," *IEEE Access* 11 (2023): 14906–14914, https://doi.org/10.1109/ACCESS.2023.3243113.

45. M. Zhu and H. Zhou, "EcReID: Enhancing Correlations From Skeleton for Occluded Person Re-Identification," *Symmetry* 15, no. 4 (2023): 906, https://doi.org/10.3390/SYM15040906.

46. Z. Ma, Y. Zhao, and J. Li, "Pose-Guided Inter-and Intra-Part Relational Transformer for Occluded Person Re-Identification," in *Proceedings of the 29th ACM International Conference on Multimedia* (ACM MM, 2021), 1487–1496.

47. Y. Pang, H. Zhang, L. Zhu, D. Liu, and L. Liu, "Feature Generation Based on Relation Learning and Image Partition for Occluded Person Re-Identification," *Journal of Visual Communication and Image Representation* 91 (2023): 103772, https://doi.org/10.1016/J.JVCIR.2023.103772.

48. S. Zhou and W. Zou, "Fusion Pose Guidance and Transformer Feature Enhancement for Person Re-Identification," *Multimedia Tools and Applications* 83, no. 7 (2023): 1–19, https://doi.org/10.1007/s11042-023-15303-2.

49. Z. Li, H. Zhang, L. Zhu, J. Sun, and L. Liu, "Effective Occlusion Suppression Network via Grouped Pose Estimation for Occluded Person Re-Identification," in *2023 IEEE International Conference on Multimedia and Expo* (IEEE ICME, 2023), 2645–2650.