symmetry

MDPI

Article

EcReID: Enhancing Correlations from Skeleton for Occluded Person Re-Identification

Minling Zhu * and Huimin Zhou

Computer School, Beijing Information Science and Technology University, Beijing 100101, China
* Correspondence: zhuminling@bistu.edu.cn

Abstract: Person re-identification is a challenging task due to the lack of person image information in occluded scenarios. The current methods for person re-identification only take into account global information, neglect local information, and are not responsive to changes in input. Additionally, these methods do not address the issue of inaccurate joint detection caused by occlusion. In this paper, we propose an occluded person re-identification method based on a graph model and deformable method, which is able to simultaneously focus on global and local information and can flexibly adapt to local information and changes in the input, efficiently resolving issues such as occluded or incorrect joint information. Our method consists of three modules: the mutual help denoising module, inter-node aggregation and update module, and graph matching module. The mutual help denoising module acquires global features and person skeleton node features using a CNN backbone network and a pose estimation model, respectively. It uses symmetric deformable graph attention to obtain the local and global features of the joint points in different views, correcting the information of incorrect nodes and extracting favorable human features. The inter-node aggregation and update module employs deformable graph convolution operations to enhance the relations between the nodes in the same view, resulting in higher-order information. The graph matching module uses graph matching methods based on the human topology to obtain a more accurate similarity calculation for masked images. Experimental results on the Occluded-Duck and Occluded-REID datasets show that our proposed method achieves Rank-1 accuracies of 64.8% and 84.5%, respectively, outperforming current mainstream methods such as HOReID. Our method also achieves good results on the MARKET-1501 and DukeMTMC-ReID datasets. These results demonstrate that our proposed method can extract person features well and effectively improve the accuracy of person re-identification tasks.

Keywords: person re-identification; occluded; deformable; attention; symmetry

Citation: Zhu, M.; Zhou, H. EcReID: Enhancing Correlations from Skeleton for Occluded Person Re-Identification. *Symmetry* **2023**, *15*, 906. https://doi.org/10.3390/sym15040906

Academic Editor: Zhixun Su

Received: 23 February 2023
Revised: 29 March 2023
Accepted: 7 April 2023
Published: 13 April 2023

1. Introduction

Person re-identification (Person ReID), a technology that uses computer vision to extract features from images of people across cameras for person matching, is widely used in intelligent security, smart retail, and other fields. Person ReID is also an essential tool in combining artificial intelligence technology with industry [1]. However, factors such as the camera's viewpoint, the person's pose, and lighting changes could lead to problems such as occluded and blurred person images, which significantly impact the effect of person re-identification. Therefore, the occluded person re-identification technique is a research hotspot in this field, aiming to enable the person re-identification model to obtain a relatively good recognition effect on occluded images.

Compared with re-identification techniques for clear person images, occluded person re-identification has more challenges. (1) Clear person images contain complete personal information, while occluded person images may cause the loss of crucial person features, and if feature extraction is performed without being able to discern the covered areas, it would cause interference in feature recognition, which could lead to a decrease in the accuracy of person recognition [2]. (2) the existing methods for person re-identification are

typically based on traditional attention or convolution methods, where the sampling grid is usually fixed and cannot adaptively adjust to the different shapes and sizes of the objects of interest, i.e., they cannot handle non-uniform sampling grids. Additionally, in traditional methods, sampling points typically only consider local information and cannot fully utilize global information, resulting in reduced model performance. Therefore, these methods have limited feature extraction capabilities on occluded images [3]. (3) In the traditional structure of local feature matching, the image features are extracted and then matched strictly on a one-to-one basis. If too many occluded areas exist, the matching performance would be reduced [4].

Based on the above analysis, to effectively and comprehensively extract the target information in occluded images and to be able to recognize the targets accurately, we propose a method of occluded person re-identification based on a graph model and deformable method, which use graph models to model human structures and then mine for more discriminative person features to improve model recognition. The global features of the image are first extracted through the backbone network, and the pre-trained human pose estimation model is used to obtain the human joints in the person image, and then the joint information is fused with the global features to obtain the person feature information representation. However, due to occlusions caused by body parts, the accuracy of human body node information may be affected, leading to reduced model recognition rates. To address this issue, symmetrical deformable graph attention is employed to aggregate and update the features of human body joint points, thereby correcting information for nodes that are poorly represented under different viewing conditions. In addition, based on the information of the human joints in each image already corrected in the previous module, the features are aggregated and updated between the different joints of the person in the same view using deformable graph convolution, thus further increasing the robustness of the model recognition effect. Then, the alignment phase uses a graph matching algorithm to measure the similarity of the two sets of person images.

The main contributions of this paper are as follows. (1) Deformable methods are used to flexibly adapt to input feature information and changes, which enables more comprehensive feature extraction. This enhances the model's flexibility and adaptability. (2) To tackle the issue of inaccurate node features in complex scenes, we suggest employing symmetry deformable graph attention to aggregate and update features. Specifically, by utilizing symmetry deformable graph attention, the node information from various person images can cooperate and rectify any erroneous node information, thereby enhancing the model's resilience against complex scene node features. (3) We employ a deformable graph convolution technique to aggregate and update features among nodes within a single view. This results in more precise person-specific features and consequently enhances the accuracy of the model's recognition. (4) The proposed framework was evaluated on two occluded datasets, Occluded-Duck and Occluded-ReID, and the publicly available datasets Market-1501 and DukeMTMC-reID, respectively, and it achieved high accuracy rates on all of them.

2. Related Work

2.1. Occluded Person Re-Identification

The objective of occluded person re-identification is to find persons with the same features in different cameras as in the occluded image. However, occluded person re-identification is made more challenging by incomplete information about the persons in the blocked image and spatial misalignment due to different angles [5]. Zhuo et al. [6] trained the network to adapt to varying types of occlusion by simulating an occluded scene, using either occluded or non-occluded binary classification loss to distinguish the occluded part from the overall image. They used multi-task training to classify both personal identification and whether it is occluded or not. Miao et al. [7] proposed a method to detect non-occluded regions based on the pose estimation of human semantic vital points and used a predefined confidence threshold for the joints to determine whether the part is

occluded. Fan et al. [8] used a Spatial Channel Alignment Network (SCPNet) to map local features into channels of global elements and fuse the overall and partial features to obtain the desired discriminant features, thus reducing the effect of occluded noise on global parts. Luo et al. [9] used a spatial transformation module to transform the overall image to align with the local vision and then calculate the distance to its correct alignment.

2.2. Graph Model-Based Person Re-Identification

Recently, graph models have been widely used in tasks such as human action pose recognition [10–13], multi-label image recognition [14–17], the identification of vehicles [18], and video classification [19–22], so related methods based on graph models have also been introduced into the field of person re-identification. Barman et al. [23] proposed an unsupervised algorithm that maps the process of person recognition ranking to the graph model theory problem. Ye et al. [24] focused on the issue of cross-camera label estimation in unsupervised learning. They proposed constructing a graph for each sample in each camera and then proposed dynamic graph matching methods for cross-camera label association. However, these methods are ineffective in recognizing persons with different and occluded clothing. In addition, there are two graph model-based methods similar to the above. Cheng et al. [25] proposed a structured Laplacian embedding algorithm, which transforms the structured distance relationship formula into a graph Laplacian matrix, thus optimizing the feature learning of the model for training samples and obtaining more robust and discriminative depth features. However, this method cannot be directly generalized to new nodes and is computationally complex, making it challenging to apply to more complex graph structures. Wu et al. [26] proposed a Graph Attention Network (GAT) and combined this network with a model feature extraction network to extract important discriminative features from the spatial and spatio-temporal domains, identify some regions with a relatively high contribution in the spatial field, and enable the network to focus on different parts. Then, the relationship between the spatio-temporal graph discovery frames and other areas is changed to learn the correlation weights within the feature graph continuously. However, this network only obtains first-order semantic information. It does not go deeper to obtain higher-order information. Li et al. [22] introduced a graph neural network that uses the relationship between human joint alignment and feature affinity to construct an adaptive structure-aware adjacency graph to model the intrinsic connections between graph nodes. However, extracting information such as human joints requires the inclusion of additional computations. It is not integrated into the overall model structure to achieve end-to-end training, so this method does not achieve the best results.

In this paper, the human topology is modeled and symmetry deformable attention is proposed to fuse the node information in the different views and fix the error node information. Meanwhile, the inter-nodes from the same view are aggregated and updated using deformable graph convolution to further enhance the robustness of the model's recognition results.

3. Network Structure Design

The network structure of this model is designed to address the issue of low accuracy in person recognition caused by occlusion problems in person re-identification. The overall structure, as shown in Figure 1, consists of three stages. The first stage is the mutual help denoising module. This module utilizes RestNet50 as the backbone network to extract global information, while also extracting local information about joints. By fusing these two sources of information, the module generates an output that is passed to a deformable graph attention (Deformable GAT) layer with a symmetric structure. In the second stage, the inter-node aggregation and update module utilizes deformable graph convolution (Deformable GConv) to establish connections between local and global information pertaining to the human body. This facilitates the aggregation and updating of information pertaining to different joints in the same view, resulting in the acquisition of higher-order semantic information and improved matching robustness. Additionally, the model employs multi-

head attention to concentrate on various parts of multiple inputs concurrently. This enables the model to comprehensively comprehend the input content, thereby enhancing the accuracy and efficiency of the recognition task. Finally, the graph matching module uses the graph matching algorithm (GM) to align images, resulting in a more accurate similarity calculation for occluded images. This three-stage approach improves the accuracy of person recognition in the presence of occlusion challenges.

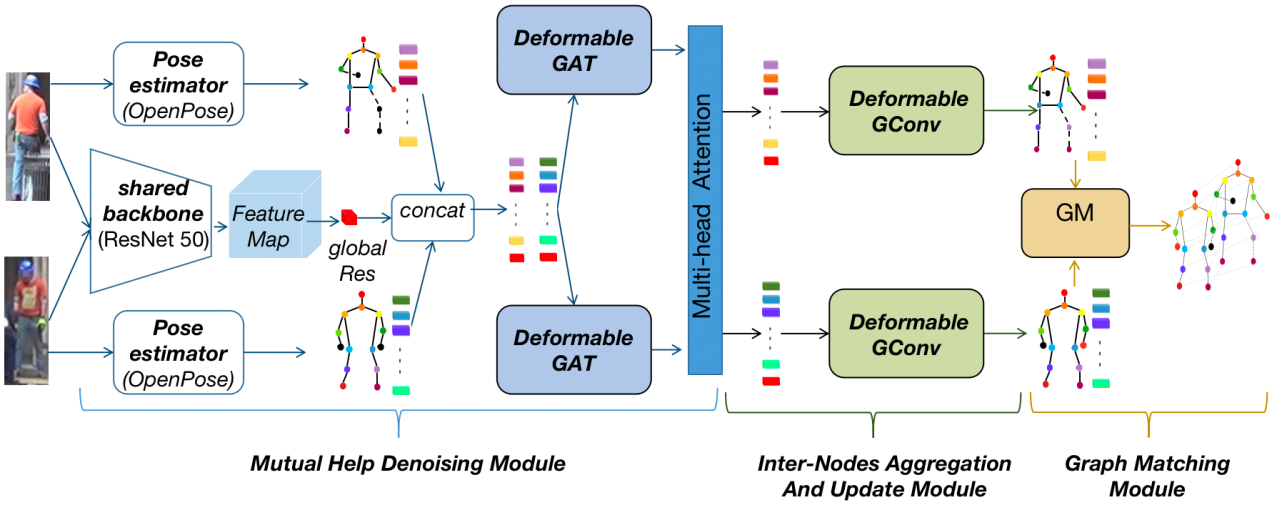


Figure 1. Overview of the proposed EcReID framework. Deformable GAT denotes deformable graph attention, Deformable GConv denotes deformable graph convolution, and GM denotes graph matching algorithm. concat denotes the concatenate operation.

3.1. Mutual Help Denoising Module

The mutual help denoising module is designed to extract person features and correct incorrect joint information. It has already been demonstrated that dividing the human body into partial regions can improve the effectiveness of the person re-identification task [27], and the person re-identification algorithm in occluded images requires the strict alignment of the local features of the image [25]. In this paper, we use a combination of human joint features and global features to obtain better person information, and then update and aggregate the human joint features in different views via symmetric deformable graph attention.

Specifically, the first step is the low-order feature extraction, as shown in Figure 2. This module first detects the original input image through the pre-trained OpenPose human pose estimation model, so as to obtain 13 pieces of human joint information and the corresponding confidence levels of the joints. If the confidence level is more significant, it means that the image is a less occluded area, and we then use the Gaussian function to obtain the heat map related to the 13 joints of the human body. Meanwhile, the fully connected layer and pooling layer of ResNet50 will be removed from the backbone network of this model, and then the original image is inputted to obtain the global features. Secondly, the global feature map m_g and the joint feature map m_{np} obtained above are passed through an outer product ($\otimes$) and global average pooling (g($\cdot$)) operations, and then the local features V_l^F and the global features V_g^F for a set of joints are obtained.

$$V_l^F = \{v_{nd}^F\}_{nd=1}^N = g(m_g \otimes m_{np}) \tag{1}$$

$$V_g^F = v_{N+1}^F = g(m_g) \tag{2}$$

where N indicates the number of critical points in the human body, $N = 13$.

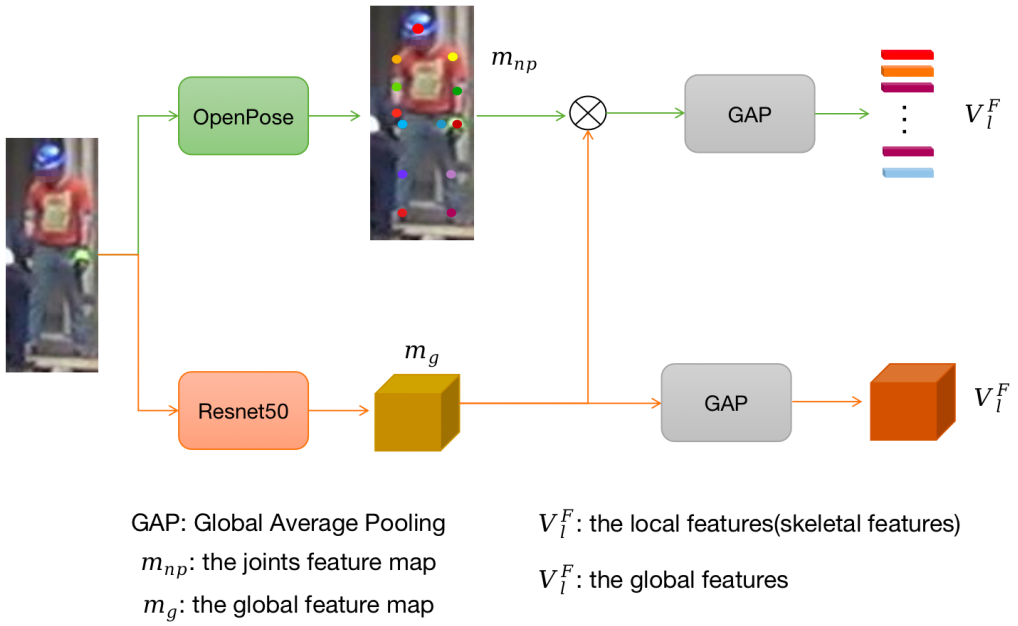


Figure 2. The flowchart of global and skeletal features. ⊗ denotes an outer product.

Traditional attention plays an important role in deep learning, which can give different weights to different parts of the input and help the model to understand the input. However, it has some problems, such as only considering the global information, ignoring the local information, and low sensitivity to input changes. When dealing with long sequences, it is easily affected by long-distance dependence, which reduces the performance. However, deformable graph attention is a technique that employs a deformable network to establish the correlation between the content and spatial location in the input feature map. By generating a deformable mesh, this technique can dynamically apply attention weights to the deformable version of the input feature map. Deforming the grid allows for the precise tuning of each spatial location, which alters the receptive field of the attention mechanism and improves the neural network's ability to consider the relationship between the input feature map's content and the spatial location. Consequently, deformable graph attention can adapt flexibly to local information and changes in the input, assigning different weights to different parts of the input. We leverage deformable graph attention to perform feature aggregation and updates on human joint features in various views, correcting incorrect joint point information. Figure 3 illustrates this process in detail.

All the joint point features and global features in the two images to be matched are used as the input of the symmetric deformable graph attention layer. The process of deformable graph attention is shown in Figure 3. Deformable graph attention divides the input sequence into two parts: one is the feature of the input nodes, and the other is the offset of each node in different directions. These offsets can be learned through convolutional layers and are usually represented as a vector, where each element represents the magnitude of the offset in different directions. Then, for each node, its position is shifted according to its corresponding offset.

Upon completion of the positional shift of nodes, the deformable graph attention mechanism computes the similarity between nodes based on their adjusted positions and assigns a weight to each node. This weight can be utilized to adjust the weighting of information from different nodes, thereby amplifying or dampening the characteristics of these nodes.

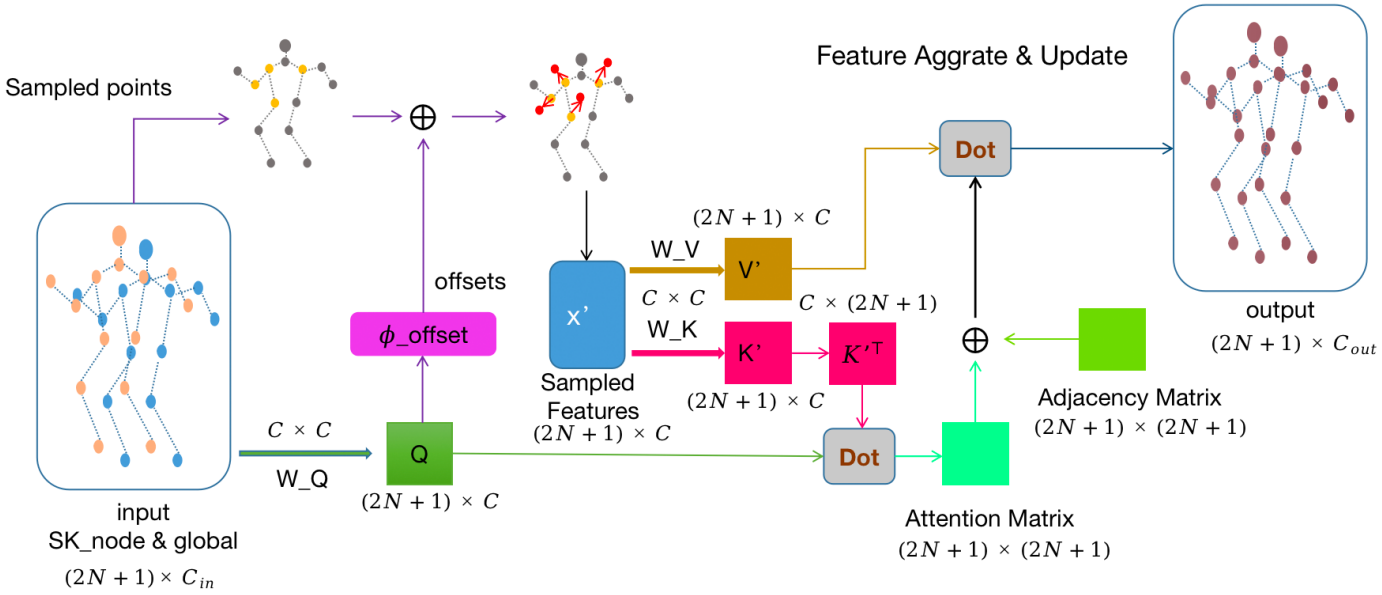


Figure 3. An illustration of the deformable graph attention architecture. SK_{node} denotes the skeleton node; C_{in} and C_{in} denote the number of channels in the input and output, respectively; N denotes the number of joints, and $\phi_{offset}(\cdot)$ is a sampling function. Only four sampling points are shown in the figure.

Overall, deformable graph attention utilizes position offset information and node similarity calculation to more comprehensively and accurately capture the relationships and spatial positions between different nodes in the input sequence, thereby improving the performance of the model.

In order to analyze joint attentional relationships with other joints, we utilize attention mechanisms applied to graphs. This involves taking a flattened feature map represented as $x \in R^{N \times C}$ as input. We then obtain three matrices, namely Q, K, and V, by applying different linear transformations to the input feature map x. The vectors q, k, and v are the constituents of Q, K, and V, respectively. Let v_i and v_j represent any two distinct human joints, and let $att_{i,j}$ denote the updated attention relationship between v_j and v_i, as shown in Equation (4).

$$Q = xW_Q, K = xW_K, V = xW_V \tag{3}$$

$$att_i = \sum_{j \in N} att_{i,j} = \sum_{j \in N(i)} softmax(\frac{q_i \cdot k_j^{\mathsf{T}}}{\sqrt{dim}})v_j \tag{4}$$

where $W_Q, W_K, W_V \in R^{C \times C}$ are the transform matrix; $q \in Q, k \in K, v \in V$; dim denotes the dimension of Q, K, V; and att_i denotes the attention relation between v_i and other joints.

As shown in Figure 3, the deformable graph attention inputs query tokens Q into a weight network $\theta_{offset}(\cdot)$ to generate offset $\Delta pe = \theta_{offset}(Q)$, and takes the node after the deformation operation as the key and value, and then the transformation matrix of Formula (3) becomes

$$Q = xW_Q, K^{'} = x^{'}W_K, V^{'} = x^{'}W_V \tag{5}$$

$$where\ \Delta_{pe} = \theta_{offset}(Q), x' = \phi(x; pe + \Delta pe). \tag{6}$$

where $K^{'}$ and $V^{'}$ denote the vector of the deformation, respectively. In addition, we define the sampling function, denoted by $\phi(\cdot;\cdot)$, to perform bilinear interpolation.

$$\phi(N; (pe_x, pe_y)) = \sum_{(n_x, n_y)} m(pe_x, n_x) m(pe_y, n_y) N[n_y, n_x, :] \tag{7}$$

where (n_x, n_y) denotes the index of all joints $N \in R^{N \times C}$, and $m(pe_x, n_x) = max(0, 1 - |pe_x - n_x|); m(pe_y, n_y|) = max(0, 1 - |pe_y - n_y|)$. Then, the attention output after the position shift is used as shown in Equation (8):

$$att_i = \sum_{j \in N} att_{i,j} = \sum_{j \in N(i)} softmax(\frac{q_i \cdot k_j^{'\mathsf{T}}}{\sqrt{dim}}) v_j^{'} \tag{8}$$

where $q \in Q, k^{'} \in K^{'}, v^{'} \in V^{'}$.

To improve the relevance of local joint information in the human body, we have introduced natural joint connections into an attention layer. This is achieved through the design of a symmetric adjacency matrix, denoted by $A \in R^{N \times N}$, where each element a_{ij} represents the relationship between joint v_i and joint v_j. By incorporating this adjacency matrix, we can express the attention relationship between joint v_i and other joints using Formula (9).

$$att_i = \sum_{j \in N} att_{i,j} = \sum_{j \in N(i)} softmax(\frac{q_i \cdot k_j^{'\mathsf{T}}}{\sqrt{dim}} + a_{ij}) v_j^{'} \tag{9}$$

Then, the attention representation of all joints is given by Equation (10).

$$att = softmax(\frac{Q \cdot K^{'\mathsf{T}}}{\sqrt{dim}} + A) V^{'} \tag{10}$$

To capture more comprehensive feature representations and improve model performance, we designed a symmetric deformable graph attention layer that calculates attention from two different perspectives on the input human joint information. Here, we used a multi-head attention approach to calculate the expression of the symmetric deformable graph attention layer. As shown in Figure 3, we set two attention heads, and then use the $concat(\cdot)$ operation to combine the attention of each head. The overall expression of the symmetric deformable graph attention is shown in Formula (11).

$$MHatt(Q, K, V) = Concat(att^1, att^2) \tag{11}$$

The loss function L_a of this module includes the triplet loss L_t and the classification loss L_c, as shown in Equation (12).

$$\begin{aligned} L_a &= \frac{1}{N+1} \sum_{j=1}^{N+1} \beta_j [L_t(V_j^F) + L_c(V_j^F)] \\ &= \frac{1}{N+1} \sum_{j=1}^{N+1} \beta_j [\max(\alpha + d_{V_{aj}^F, V_{pj}^F} - d_{V_{aj}^F, V_{nj}^F}, 0) - \log p_{V_j^F}] \end{aligned} \tag{12}$$

where N denotes the number of human joints, $N = 13$, β_j denotes the confidence of the jth joint, a and p are the same person, a and n are different persons, $d_{V_{aj}^F, V_{pj}^F}$ denotes the similarity between a and p, $d_{V_{aj}^F, V_{nj}^F}$ denotes the similarity between a and n, α denotes the minimum interval between $d_{V_{aj}^F, V_{pj}^F}$ and $d_{V_{aj}^F, V_{nj}^F}$, and $p_{V_j^F}$ denotes the probability that the feature V_j^F is correctly classified.

3.2. Inter-Node Aggregation and Update

The occluded person image is incomplete, and to utilize more person-related semantic information, a deformable graph convolution algorithm is used to aggregate and update features on the nodes in the same view. The general structure of deformable graph convo-

lution is shown in Figure 4, which adaptively changes the receptive field and aggregates the neighborhood information of the nodes on the image.

In a traditional Graph Convolutional Network (GCN), the relationship between nodes depends on the degree of the nodes, and GCN has a single transformation for non-2D convolutional relationships. In contrast, deformable graph convolution is more flexible, redefining the relationships between the nodes and the convolution kernel function. Firstly, the joints are embedded in a potential position space, and the position of the joints in the space is used to determine the coordinates of the joints; secondly, the relationship vector between the joints is calculated from the coordinates of the joints.

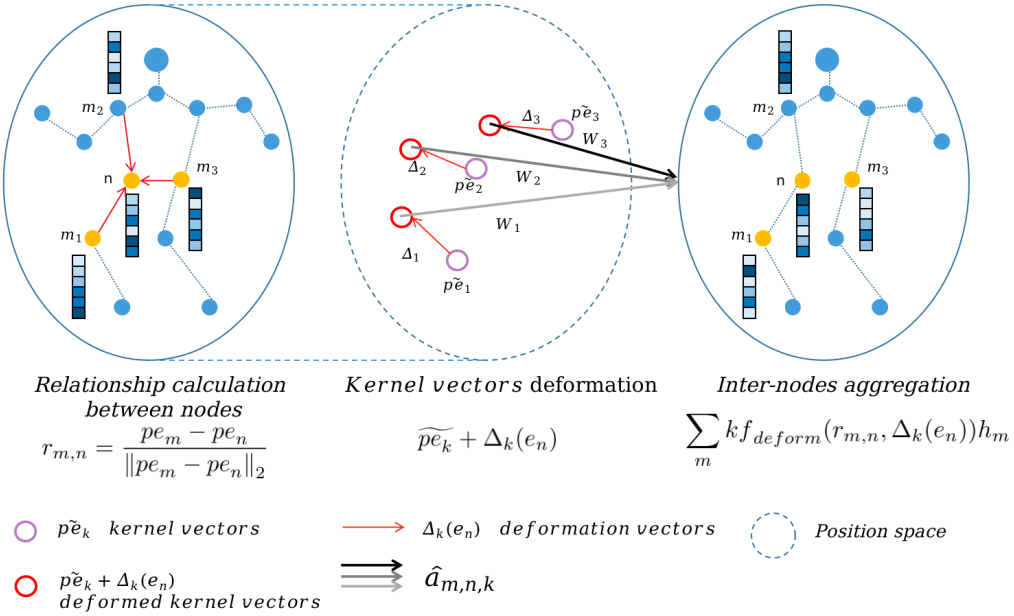


Figure 4. An illustration of deformable graph convolution. Only three joints are taken as examples in the figure. Deformable GConv, through kernel vector deformation $\Delta k(e_n)$, allows the convolution kernel $kf_{deform}(\cdot,\cdot)$ to adaptively deform for each central node n. This process facilitates more flexible graph convolutions and promotes feature aggregation and updates between nodes within the same view.

Specifically, each joint point in this module is embedded in a potential position space for localization, so that the coordinates and relationships of the joint points are obtained. The inputs to each node in the model are the node position coordinates and the node feature representation. The process of obtaining the node position coordinates is simple. Given an input image x, n is an joint in the original input feature map x. Then, the location embedding $pe_n^{(h)}$ of the joint n is obtained after smoothing h times the projection on the original input feature map x, as shown in Equation (13):

$$\begin{aligned} pe_n^{(h)} &= W_{pe}^{(h)} e_n^{(h)}, where \\ e_n^{(0)} = x_n, e_n^{(h)} &= \frac{1}{\widetilde{deg}(n)} \sum_{m \epsilon \widetilde{N}(n)} e_m^{(h-1)} \end{aligned} \tag{13}$$

$W_{pe}^{(h)}$ is a learnable weight matrix, $\widetilde{deg}(n)$ denotes the degree matrix of joint n, e_n^h is the feature obtained by smoothing h projections of joint n in the input image, and $\widetilde{N}(n)$ denotes the neighborhood of joint n.

The relationship between joints is represented by a relationship vector, i.e., $r_{m,n}$ represents the relationship between node n and its neighbor joints m. When joint n and neighbor joint m have different positional embeddings, the relative positions of the nodes in space are used to represent the relationship, and when joint n and neighbor joint m have the

same positional embedding, a normalized relationship vector is defined to encode the positional embedding. Then, the relationship between the joint n and the neighboring joint m is expressed as shown in Equation (14).

$$r_{m,n} = \begin{cases} r_{m,n}'||0] \in R^{d_{pe}+1} & \text{if } pe_m \neq pe_n \\ [0,0,...,1] \in R^{d_{pe}+1} & \text{otherwise} \end{cases} \tag{14}$$

where pe_m and pe_n denote the positional embedding of neighboring joints m and n, respectively. $r'_{m,n} = \frac{pe_m - pe_n}{\|pe_m - pe_n\|_2}$, where $[||]$ is a concatenation operator.

Then, the kernel function $kf(\dots)$ is defined as shown in Equation (15):

$$\begin{aligned} kf(r_{m,n}) &= \sum_{s}^{S} a_{m,n,s} W_s \\ where \quad a_{m,n,s} &= \frac{exp(r_{m,n}^T \widetilde{pe_k})}{Z} \end{aligned} \tag{15}$$

where $\widetilde{pe_k} \in R^{d_{pe}+1}$ denotes the kernel vector, $W_s \in R^{d_y \times d_h}$ is the weight matrix, and $Z \in R$ is a normalization factor.

In order to improve the flexibility of the model and to obtain more information between joints, the kernel function is then further optimized so that the kernel function considers not only the relationship between the central joint n and the neighboring node m, but also the relationships between neighboring joints. The kernel vector $\widetilde{pe_k}$ is dynamically translated by the deformable vector $\Delta_k(e_n) \in R^{d_{pe}+1}$ according to the feature $e_n \in R^{d_x}$ after the projection of the central node n. The deformable graph convolution method (DGCN) is shown in Equation (16).

$$y_n = \sum_{m \in \widetilde{N}(n)} kf_{deform}(r_{m,n}, \Delta_k(e_n)) h_m \tag{16}$$

where $kf_{deform}(r_{m,n}, \Delta_k(e_n)) = \sum_{s=1}^{S} \hat{a}_{m,n,k} W_s$ and $\hat{a}_{m,n,k} = \frac{exp(r_{m,n}^T(\widetilde{pe_k} + \Delta_k(e_n)))}{\sum_{m'} exp(r_{m',n}^T(\widetilde{pe_k} + \Delta_k(e_n)))}$; the deformation vector $\Delta_k(e_n) \in R^{d_{pe}+1}$ is generated by a simple MLP network.

A better person feature V^* is obtained by using a deformable graph convolution layer in the inter-node aggregation and update module, as shown in Equation (17).

$$V^* = DGCN(V^F) \tag{17}$$

The loss function of the graph convolution feature fusion module contains the triplet loss L_t and the classification loss L_c, and the loss function L_b is calculated as shown in Equation (18).

$$L_b = \frac{1}{nd+1} \sum_{j=1}^{N+1} \beta_j \left[L_t(V_j^F) + L_t(V_c^F) \right] \tag{18}$$

where β_j denotes the confidence level of the jth joint, and the meanings of triplet loss and classification loss are shown in the feature extraction module's loss function L_a.

Finally, for the two sets of input images x_1 and x_2, their higher-order relational features $V_1^* = \{V_{1,nd}^*\}_{nd=1}^N$ and $V_2^* = \{V_{2,nd}^*\}_{nd=1}^N$ can be obtained by Equation (17), where the cosine similarity of images x_1 and x_2 can be obtained, as shown in Equation (19).

$$s_{(x_1,x_2)}^* = \frac{1}{nd+1} \sum_{j=1}^{N+1} \sqrt{\beta_{1,j}\beta_{2,j}} \cos\left(V_{1,nd}^*, V_{2,nd}^*\right) \tag{19}$$

$\beta_{1,j}$ denotes the confidence of the jth joint of picture x_1, and $\beta_{2,j}$ denotes the confidence of the jth joint of picture x_2.

3.3. Graph Matching Module

Traditional person re-identification measures the similarity of local features with one-to-one feature matching. It cannot effectively solve the problem of matching the content of occluded person images. In order to more accurately measure the similarity between occluded persons, this paper uses an improved fusion graph matching method for feature matching, as shown in Figure 5. Firstly, the higher-order features corresponding to two different person images are obtained from the graph convolution information fusion module as the input of the person feature matching module. Then, a similarity matrix is obtained using the fusion graph matching method to measure the similarity of the person images. Finally, the crossover process of similarity prediction is used afterward to obtain the respective matched feature results separately.

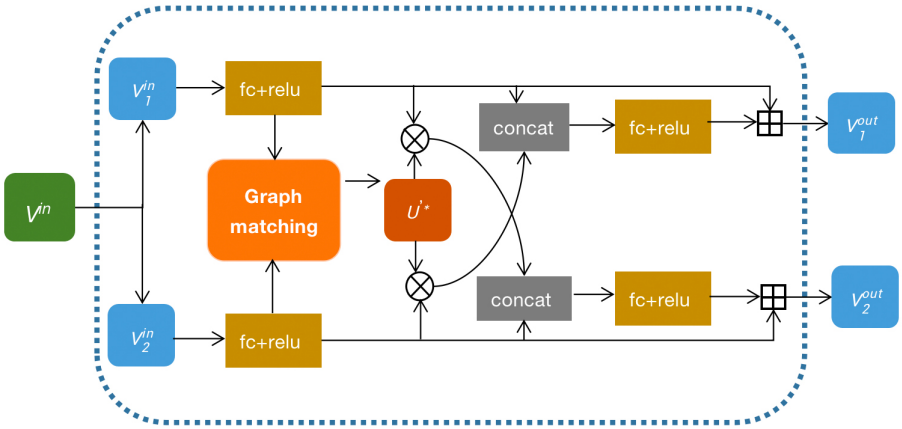


Figure 5. Figure matching operation process. ⊗ denotes the outer product operation and ⊞ denotes the matrix summation operation.

Specifically, for the matching matrix U updating process, firstly, the graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ are obtained based on the given image pairs x_1, x_2 combined with the human body topology information, where V_1, V_2 denote the set of nodes corresponding to the joints of the human body, and E_1, E_2 denote the set of edges between the joints. For each pair of joints between edges in E_1, the matching degree between two points (i, j) within this edge and two nodes (a, b) of the corresponding edge in E_2 is computed in turn to obtain the similarity matrix $M \in [0, 1]^{KK \times KK}$. For example, $M_{ia;jb}$ denotes the matching degree between (i, j) in G_1 and (a, b) in G_2. For the absence edge of two joint pairs, its corresponding element in the similarity matrix M is set to 0. Therefore, the elements on the diagonal of the similarity matrix denote the matching degree between nodes in G_1 and G_2, and the elements on the non-diagonal denote the edge-to-edge confidence degree in the two graphs. Then, let $U \in [0, 1]^{nd \times nd}$ be a matching strategy between G_1 and G_2, where u_{ia} denotes the matching degree between nodes $v_i \in V_1$ and $v_a \in V_2$. Initialize U as a $nd \times nd$ unit matrix U_0 and expand the elements in U_0 in order of rank to obtain the indicator vector $U_0^{'}$, and then the power iteration algorithm [28] is used to iteratively update $U_0^{'}$ based on the similarity matrix M. The iterative updating process is as follows.

$$U_{i+1}^{'} = \frac{MU_i^{'}}{\left\| MU_i^{'} \right\|_2} \tag{20}$$

Then, the optimal matching vector $U^{''}$ is obtained, as shown in Equation (21).

$$U^{''} = \underset{u^{'}}{\operatorname{argmax}}\, U^{'T} M U^{'}, s.t. \left\| U^{'} \right\|_2 = 1 \tag{21}$$

The vector U'' is matrixed to obtain the optimal matching result. Then, U^* is normalized and the softmax activation function is performed to finally obtain the matching matrix U'^*.

The whole process of feature matching is shown in Figure 5. Two input sets of images V_1^{in} and V_2^{in} are passed through the fully connected layer (fc) and the relu activation function to obtain V_1^m and V_2^m, respectively. In addition, the optimal matching matrix U'^* can be obtained by Equations (20) and (21). The final output V_1^{out} and V_2^{out} can be derived from the following Equations (22) and (23), respectively.

$$V_1^{out} = fc(concat(V_1^m, U'^* \otimes V_2^m)) + V_1^m \tag{22}$$

$$V_2^{out} = fc(concat(V_2^m, U'^* \otimes V_1^m)) + V_2^m \tag{23}$$

Finally, the similarity $S^Q_{x_1,x_2}$ between image x_1 and image x_2 is obtained by combining the U'^* obtained by Equation (21) with the similarity $s^*_{x_1,x_2}$ of the feature cosine, as shown in Equation (24).

$$S^Q_{x_1,x_2} = \frac{1}{nd+1}(s^*_{x_1,x_2}U'^* + \cos(V^g_{1,nd+1}, V^g_{2,nd+1})) \tag{24}$$

V_1^* and V_2^* denote the features of higher order of images x_1 and x_2, $s^*_{x_1,x_2}$ is the cosine similarity of V_1^* and V_2^*, and $V^g_{1,nd+1}$ and $V^g_{2,nd+1}$ denote the global features of images x_1 and x_2, respectively. The loss function L_c of this module includes the verification loss L_v and the matching loss L_m.

$$L_m = -U''^T M U'' \tag{25}$$

$$L_v = y \log S^Q_{x_1,x_2} + (1-y)\log(1 - S^Q_{x_1,x_2}) \tag{26}$$

$$L_c = \mu_m L_m + \mu_v L_v \tag{27}$$

$S^Q_{x_1,x_2}$ denotes the similarity between person image x_1 and image x_2; y denotes the person image verification result, i.e., the value of y is 1 if x_1 and x_2 are images of the same person, otherwise the value of y is 0; and μ_m and μ_v denote the weight values of L_m and L_v, respectively. For two sets of images x_1 and x_2, the cosine similarity $S^*_{x_1,x_2}$ of images x_1 and x_2 can be obtained according to Equation (19), and the topological relationship similarity $S^Q_{x_1,x_2}$ of images x_1 and x_2 can be obtained from Equation (24); then, $S^*_{x_1,x_2}$ and $S^Q_{x_1,x_2}$ constitute the final similarity rule of this model, as shown in Equation (28).

$$s = \gamma S^Q_{x_1,x_2} + (1-\gamma)S^*_{x_1,x_2} \tag{28}$$

The similarity s between the query image and the image to be queried is calculated sequentially according to Equation (28). Then, the top n most similar images are retrieved according to the size of s.

4. Experiments and Analysis

In this paper, the effectiveness of the model is evaluated using two occluded-type datasets, Occluded-Duck [7] and Occluded-ReID [6], and two larger person re-identification datasets, Market-1501 [29] and DukeMTMC-ReID [30]. The Occluded-Duck dataset is composed of occluded and non-repeated person images selected from the DukeMTMC-ReID dataset, which contains 15,618 training images, 17,661 test images, and 2210 occluded query images. The Occluded-ReID dataset contains a total of 2000 images of 200 occluded persons taken with moving cameras. Each person ID consists of 5 full-body images and 5 occluded images, where the person images are all resized to 128 × 64. The Market-1501 dataset includes 1501 persons captured by 6 cameras, with a total of 32,668 person images.

The training set contains 751 persons with 12,936 person images, the test set contains 750 persons with 19,732 images, and the query image set contains 3368 person images. The DukeMTMC-ReID dataset contains a total of 1404 persons with 36,411 pictures. The training set contains 16,522 images, and the test set contains 17,661 images and 2228 person images as the query set.

In this paper, the Cumulative Match Characteristic (CMC) and mean Average Precision (mAP) are used to evaluate the performance of the model. CMC seeks to rank the target images in the query set with the images in the test set according to the magnitude of image similarity. Rank-1 is used as the evaluation criterion of CMC in this paper, where Rank-1 indicates the average accuracy of the first image retrieval. mAP indicates the actual ranking result, i.e., the mean value of the average precision among all retrievals is calculated. mAP can show the stability of the recognition accuracy of the model.

4.1. Comparison with Other Methods

4.1.1. Comparison on Occluded Datasets

In order to verify the effectiveness of the methods in this paper, a comparison is made with some occluded person ReID-based methods on two datasets, Occluded-Duck and Occluded-ReID, including four categories of methods, namely overall-based person ReID methods—Part-Aligned [31] and PCB [27]; overall-based combined with joint information ReID methods—Part Bilinear [32] and FD-GAN [33]; local-based ReID methods—AMC + SWM [2], DSR [34], and SFR [35]; and occluded ReID methods—Ad-Occluded [36], TCSDO [37], FPR [38], PGFA [4], HOReID [39], and MoS [40], and other methods.

As shown in Table 1, both PCB [27] and FD-GAN [33] obtained approximately 40% Rank-1 scores on the Occluded-Duke dataset, i.e., there is no significant difference between the standard holistic-based ReID method and the holistic-based ReID method combined with joint information, which indicates that the use of joint information alone cannot significantly solve the occluded person re-identification problem. The local-based ReID method and the occluded ReID method improve significantly on the occluded dataset; for example, DSR [34] and FPR [38] have Rank-1 scores of 72.8% and 98.3% on the Occluded-ReID dataset, respectively. The above results show that both the local-based ReID task and the occluded-ReID task have difficulty extracting discriminative features and feature alignment. Our method achieves the best performance on the Occluded-Duke and Occluded-ReID datasets, with Rank-1 scores of 64.8% and 84.5%, respectively, demonstrating our method's effectiveness.

Table 1. Comparison of different methods on two occluded datasets.

Methods	Occluded-Duke		Occluded-ReID	
	Rank-1 (%)	mAP (%)	Rank-1 (%)	mAP (%)
Part-Aligned [31]	28.8	20.2	-	-
PCB [27]	42.6	33.7	41.3	38.9
Part Bilinear [32]	36.9	-	-	-
FD-GAN [33]	40.8	-	-	-
AMC + SWM [2]	-	-	31.2	27.3
DSR [34]	40.8	30.4	72.8	62.8
SFR [35]	42.3	32	-	-
Ad-Occluded [36]	44.5	32.2	-	-
TCSDO [37]	-	-	73.7	77.9
FPR [38]	-	-	78.3	68.0
PGFA [4]	51.4	37.3	-	-
HOReID [39]	55.1	43.8	80.3	70.2
MoS [40]	67.0	49.2	-	-
EcReID (ours)	64.8	52.7	84.5	75.1

4.1.2. Comparison on Market-1501 and DukeMTMC-ReID

In order to verify the effectiveness of our method, this paper compares it with other mainstream person re-identification methods on two larger public datasets of person re-identification, Market-1501 and DukeMTMC-ReID. These include overall-based person re-identification methods—PCB [27], VPM [41], and BOT [42]; local-based ReID methods—MGCAM [43] and FPR [38]; and human joint-based ReID methods—Pose-transfer [44], PSE [45], PGFA [7], HOReID [39], and MoS [40]. The experimental results are shown in Table 2.

Table 2. Comparison with other methods on Market-1501 and DukeMTMC datasets.

Methods	Market-1501		DukeMTMC-ReID	
	Rank-1(%)	mAP(%)	Rank-1(%)	mAP(%)
PCB [27]	92.3	77.4	81.8	66.1
VPM [41]	93.0	80.8	83.6	72.6
BOT [42]	94.1	85.7	86.4	76.4
MGCAM [43]	83.8	74.3	46.7	46.0
FPR [38]	95.4	86.6	88.6	78.4
Pose-transfer [44]	87.7	68.9	30.1	28.2
PSE [45]	97.7	69.0	27.3	30.2
PGFA [7]	91.2	76.8	82.6	65.5
HOReID [39]	94.2	84.9	86.9	75.6
MoS [40]	94.7	86.8	88.7	77.0
EcReID (ours)	95.5	87.2	88.3	78.5

As can be seen from Table 2, the three person re-identification methods based only on the whole PCB [27], VPM [41], and BOT [42] achieved good results in terms of recognition accuracy; for example, BOT [42] achieved 94.1% and 85.7% for Rank-1 and mAP on the Market-1501 dataset, respectively. However, the person re-identification method based only on the whole could not better identify the human body part information as well as joint information accurately, and it could not achieve the desired effect when solving the occluded problem. In addition, FPR [38] and MGCAM [43] utilize partial human information, but the recognition accuracy is not high; for example, the Rank-1 and mAP of MGCAM [43] only reach 83.8% and 74.3%, respectively, on the Market-1501 dataset, and only 46.7% and 46.0%, respectively, on the DukeMTMC-ReID dataset. The three methods of Pose-transfer [44], PSE [45], and PGFA [7] use human keypoint information but ignore the overall associations of person features, resulting in poor recognition accuracy; for example, PSE achieves only 27.3% and 30.2% of Rank-1 and mAP on the DukeMTMC-ReID dataset. The HOReID [39] method uses human joint information and fuses the joints using adaptive graph convolution, but it does not extract person information very well and its Rank-1 on the DukeMTMC-ReID dataset is only 86.9%. In contrast, we use symmetry deformable graph attention to correct the incorrect node information and fuse the global and local features of the image. In addition, we use a deformable graph convolution feature fusion and update of joints in the same view. Finally, the model recognition results are obtained using the graph matching algorithm. The Rank-1 and mAP of this model on the Market-1501 dataset reached 95.5% and 87.2%, respectively; the Rank-1 and mAP on the DukeMTMC-ReID dataset were 88.3% and 78.5%, respectively, and the experiments showed that this model achieved better results on both the Market-1501 dataset and the DukeMTMC-ReID dataset, proving the effectiveness of the method in this paper.

4.2. Ablation Experiments

To verify the effectiveness of our method, several different sets of ablation experiments were conducted on the Occluded-Duke dataset for different model ablations and different network layers.

4.2.1. Ablation Analysis of Different Models

The first is the ablation of different modules. They are divided into four cases. The first (Index-1) is the base person re-identification model after removing the methods in this paper, i.e., person features are extracted using ResNet50, and then the target image is retrieved based on the similarity. The second (Index-2) is a model with added symmetry of deformable graph attention to correct incorrect joint information. The third (Index-3) is a feature update and fusion of human joints in the same view using deformable graph convolution. The fourth (Index-4) is the model that uses the graph matching algorithm to match person features, which is the model proposed in this paper. The experimental results are shown in Table 3.

Table 3. Ablation experiments on different models.

Index	F	G	M	Rank-1 (%)	mAP (%)
1	×	×	×	49.9	39.5
2	√	×	×	62.2	45.9
3	√	√	×	62.7	48.4
4	√	√	√	64.8	52.7

As can be seen from Table 3, the most basic person re-identification model, Index-1, only uses the global features of images to recognize persons, and its Rank-1 and mAP only reach 49.9% and 39.5%, respectively. On this basis, Index-2 extracts the local and global features of human joints and uses symmetry deformable graph attention to obtain more discriminative person features. Its mAP is improved by 6.4 %, and the Rank-1 is improved by 12.3%, which shows that the use of the local features of human joints and symmetry deformable graph attention can effectively improve the accuracy rate of person recognition. Index-3 is based on Index-2, but we add the deformable graph convolutional network to enhance the relations between the joints in the same view, and its mAP is improved by 2.5%, which shows that the deformable graph convolutional network fuses local features and global features, enhances the correlations of joints, and effectively suppresses the occluded area's impact on person re-identification. Index-4 shows the whole network architecture of this paper, whose Rank-1 and mAP reach 64.8% and 52.7%, respectively. It can be seen that each module working together can improve the model effect and make the extracted person features more discriminative and robust.

4.2.2. Ablation Analysis of Different Network Layers

In this section, the symmetry deformable graph attentional layer (symDGAT), the deformable graph convolutional feature fusion layer (DGCN), and the graph matching layer (GM) are analyzed, respectively. The experimental results are shown in Table 4.

Table 4. Ablation experiments on different network layers.

Index	symDGAT	DGCN	GM	Rank-1 (%)	mAP (%)
1	×	√	√	59.8	47.3
2	√	×	√	61.2	48.9
3	√	√	×	63.3	50.8
4	√	√	√	64.8	52.7

The experiments are divided into the following four cases. The first one (Index-1) involves removing the symDGAT layer and using only the ordinary feature extraction network, whose Rank-1 and mAP scores reach only 59.8% and 47.3%. The second one (Index-2) involves removing the adaptive adjacency matrix Z^{adp} obtained from Equation (6) in the GCIF layer and using only the simple human keypoint adjacency matrix, whose Rank-1 is 61.2% and mAP is 48.9%. The third one (Index-3) involves removing the graph matching layer and using only the simple alignment matching with a Rank-1 of 63.3% and

mAP of 50.8%. The fourth one (Index-4) involves using the symDGAT layer, GCIF layer, and GM layer, i.e., the complete network architecture layer. Its Rank-1 is 64.8%, which is 5.0%, 3.6%, and 1.5% higher than those of Index-1, Index-2, and Index-3, respectively. Its mAP is 52.7% higher than those of Index-1, Index-2, and Index-3, which are 5.4%, 3.8%, and 1.9%. This result demonstrates the effectiveness of the symDGAT layer, GCIF layer, and GM layer for solving the occluded person re-identification model.

4.3. Parameter Analysis

To prove the rationality of the hyperparameter selection, we test the parameters γ and n in Equation (28) on the Occluded-Duke dataset, where n denotes the top n most similar images retrieved by the model. The control variable method is used during the experiment, i.e., one of the parameters is controlled not to change and the other parameter is changed to observe the effect on the experimental results. The experiments prove that the best results are achieved with $\gamma = 0.6$ and $n = 5$.

Firstly, $n = 5$ was fixed and the effect on the model results was observed by adjusting the value of γ. The results of Rank-1 and mAP are shown in Figure 6.

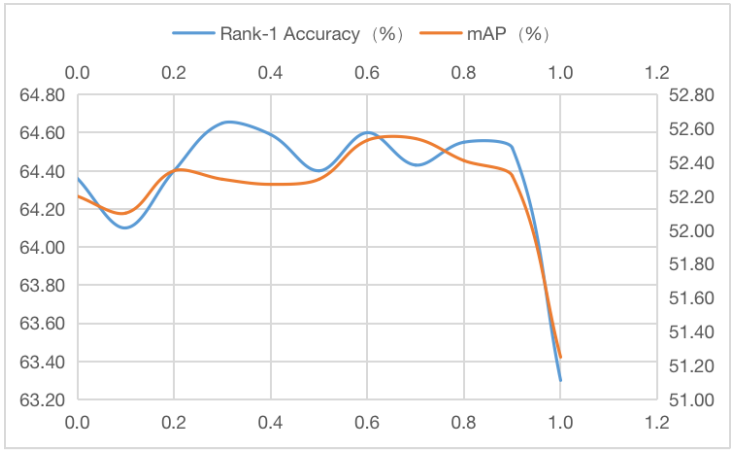


Figure 6. Effects of different values of γ on Rank-1 and mAP.

As can be seen from Figure 6, the best results of Rank-1 and mAP were achieved at $\gamma = 0.6$.

Secondly, we fixed $\gamma = 0.6$ and observed the effect on the model results by adjusting the value of n. The results of Rank-1 and mAP are shown in Figure 7.

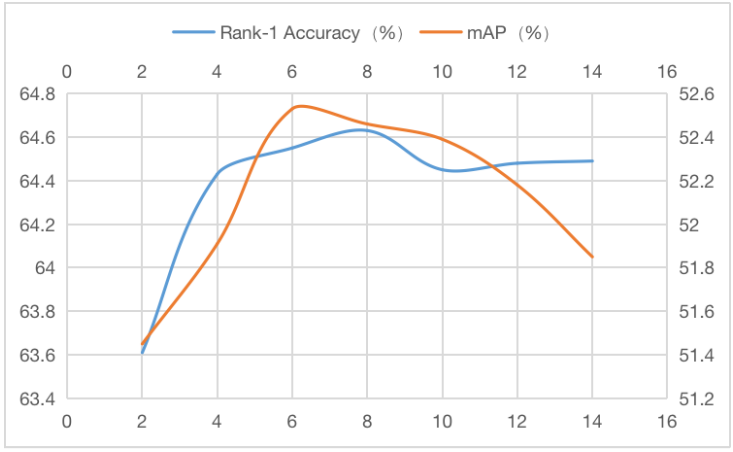


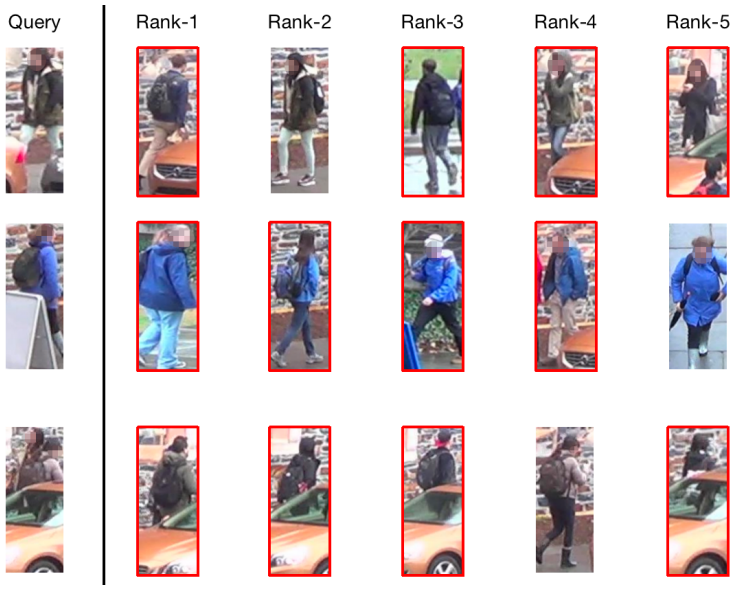
Figure 7. Effects of different values of n on Rank-1 and mAP.

As shown in Figure 7, the best values of Rank-1 and mAP are achieved when $n = 5$, i.e., when the first five matched images are taken.

As seen from the above graphs, the retrieval accuracy fluctuates within a small range for different hyperparameters, and the experiments prove that the model is robust to different hyperparameters.

4.4. Visualization of Experimental Results

In this paper, the visual analysis of the person re-identification results is shown in Figure 8. Three sets of person image recognition results are visualized. In Figure 8a, the recognition effect of the model is not satisfactory and the recognition rate is meager because the person images are affected by occlusion, a low resolution, image similarity, and other factors, while, in Figure 8b, the detection accuracy rate is greatly improved. The first five retrieval results are matched with the target image. The red box in the figure indicates a retrieval error, while no red box mark indicates a correct retrieval. Therefore, the results prove that the proposed method can improve the recognition accuracy of occluded person re-identification.



(**a**) Base model

(**b**) Our model

Figure 8. Comparison of the identification effect between the method in this paper and the benchmark model method.

Figure 9 displays a comparison of the thermal images between our method and the baseline method. It is evident that the recognition process can be easily disrupted in the occluded areas. Overall, our model demonstrates more comprehensive recognition and higher accuracy when compared to the baseline method.

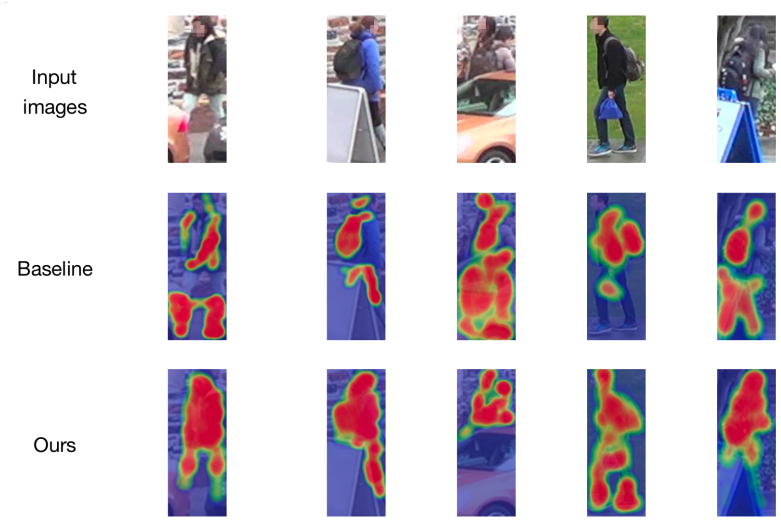


Figure 9. Heat map visualization of our method compared with the baseline method.

5. Conclusions

We propose a model based on the graph and deformable method, which extracts more discriminative features and enhanced correlations from the skeleton. It is demonstrated experimentally that the mutual help denoising block can effectively enhance the feature extraction effect by using the mutual help of images in different views to correct the incorrect joints. In addition, deformable graph convolution is used to aggregate and update the joints in the same view to further increase the robustness of model recognition and maximize the use of effective person feature information, and we also use the human topological map matching algorithm to enhance the person image similarity calculation. Finally, experiments on the Market-1501 dataset, DukeMTMC-ReID dataset, and two occlusion datasets (Occluded-Duke and Occluded-ReID) prove the effectiveness of the proposed method.

Author Contributions: Conceptualization: M.Z. and H.Z.; methodology: M.Z. and H.Z.; formal analysis: M.Z. and H.Z.; investigation: M.Z. and H.Z.; data curation: M.Z. and H.Z.; writing—original draft preparation: M.Z. and H.Z.; writing—review and editing: M.Z. and H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work is partially supported by the National Key Research and Development Plan, Ministry of Science and Technology of the People's Republic of China, under Grants 2018AAA0101000, 2017YFF0205306, and WQ20141100198. This work is also partially supported by the National Natural Science Foundation of China under Grant 91648117, the Promotion of the Classified Development of Universities—Computer Science and Technology First-Level Discipline Construction, 5112211036, and the Computer Science College—Research Construction Project, 5029923412.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; Hoi, S.C. Deep learning for person re-identification: A survey and outlook. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 2872–2893. [CrossRef] [PubMed]
2. Zheng, W.S.; Li, X.; Xiang, T.; Liao, S.; Lai, J.; Gong, S. Partial person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Seville, Spain, 17–19 March 2015; pp. 4678–4686.

3. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 764–773.
4. Zheng, K.; Lan, C.; Zeng, W.; Liu, J.; Zhang, Z.; Zha, Z.J. Pose-guided feature learning with knowledge distillation for occluded person re-identification. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual, 20–24 October 2021; pp. 4537–4545.
5. Li, Y.; He, J.; Zhang, T.; Liu, X.; Zhang, Y.; Wu, F. Diverse part discovery: Occluded person re-identification with part-aware transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2898–2907.
6. Zhuo, J.; Chen, Z.; Lai, J.; Wang, G. Occluded person re-identification. In Proceedings of the 2018 IEEE International Conference on Multimedia and Expo (ICME) IEEE, San Diego, CA, USA, 23–27 July 2018; pp. 1–6.
7. Miao, J.; Wu, Y.; Liu, P.; Ding, Y.; Yang, Y. Pose-guided feature alignment for occluded person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 542–551.
8. Fan, X.; Luo, H.; Zhang, X.; He, L.; Zhang, C.; Jiang, W. Scpnet: Spatial-channel parallelism network for joint holistic and partial person re-identification. In Proceedings of the Asian Conference on Computer Vision, Perth, Australia, 2–6 December 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 19–34.
9. Luo, H.; Jiang, W.; Fan, X.; Zhang, C. Stnreid: Deep convolutional networks with pairwise spatial transformer networks for partial person re-identification. *IEEE Trans. Multimed.* **2020**, *22*, 2905–2913. [CrossRef]
10. Li, Y.; He, Z.; Ye, X.; He, Z.; Han, K. Spatial temporal graph convolutional networks for skeleton-based dynamic hand gesture recognition. *EURASIP J. Image Video Process.* **2019**, *2019*, 1–7. [CrossRef]
11. Xing, Y.; Zhu, J. *Deep Learning-Based Action Recognition with 3D Skeleton: A Survey*; Wiley: Hoboken, NJ, USA, 2021.
12. Zhou, J.; Lin, K.Y.; Li, H.; Zheng, W.S. Graph-based high-order relation modeling for long-term action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8984–8993.
13. Zhang, J.; Ye, G.; Tu, Z.; Qin, Y.; Qin, Q.; Zhang, J.; Liu, J. A spatial attentive and temporal dilated (SATD) GCN for skeleton-based action recognition. *CAAI Trans. Intell. Technol.* **2022**, *7*, 46–55. [CrossRef]
14. Chen, Z.M.; Wei, X.S.; Wang, P.; Guo, Y. Multi-label image recognition with graph convolutional networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5177–5186.
15. Zheng, K.; Liu, W.; He, L.; Mei, T.; Luo, J.; Zha, Z.J. Group-aware label transfer for domain adaptive person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 5310–5319.
16. Zhao, J.; Yan, K.; Zhao, Y.; Guo, X.; Huang, F.; Li, J. Transformer-based dual relation graph for multi-label image recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 163–172.
17. Chen, Z.; Wei, X.S.; Wang, P.; Guo, Y. Learning graph convolutional networks for multi-label recognition and applications. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**. [CrossRef] [PubMed]
18. Huang, Z.; Qiao, S.; Han, N.; Yuan, C.A.; Song, X.; Xiao, Y. Survey on vehicle map matching techniques. *CAAI Trans. Intell. Technol.* **2021**, *6*, 55–71. [CrossRef]
19. Wang, X.; Gupta, A. Videos as space-time region graphs. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 399–417.
20. Pan, B.; Cai, H.; Huang, D.A.; Lee, K.H.; Gaidon, A.; Adeli, E.; Niebles, J.C. Spatio-temporal graph for video captioning with knowledge distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10870–10879.
21. Yang, J.; Zheng, W.S.; Yang, Q.; Chen, Y.C.; Tian, Q. Spatial-temporal graph convolutional network for video-based person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3289–3299.
22. Wu, Y.; Bourahla, O.E.F.; Li, X.; Wu, F.; Tian, Q.; Zhou, X. Adaptive graph representation learning for video person re-identification. *IEEE Trans. Image Process.* **2020**, *29*, 8821–8830. [CrossRef] [PubMed]
23. Barman, A.; Shah, S.K. Shape: A novel graph theoretic algorithm for making consensus-based decisions in person re-identification systems. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1115–1124.
24. Ye, M.; Ma, A.J.; Zheng, L.; Li, J.; Yuen, P.C. Dynamic label graph matching for unsupervised video re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5142–5150.
25. Cheng, D.; Gong, Y.; Chang, X.; Shi, W.; Hauptmann, A.; Zheng, N. Deep feature learning via structured graph Laplacian embedding for person re-identification. *Pattern Recognit.* **2018**, *82*, 94–104. [CrossRef]
26. Wu, X.; An, W.; Yu, S.; Guo, W.; García, E.B. Spatial-temporal graph attention network for video-based gait recognition. In *Proceedings of the Asian Conference on Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 274–286.
27. Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; Wang, S. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 480–496.

28. Wang, R.; Yan, J.; Yang, X. Learning combinatorial embedding networks for deep graph matching. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3056–3065.
29. Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; Tian, Q. Scalable person re-identification: A benchmark. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1116–1124.
30. Zheng, Z.; Zheng, L.; Yang, Y. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3754–3762.
31. Zhao, L.; Li, X.; Zhuang, Y.; Wang, J. Deeply-learned part-aligned representations for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3219–3228.
32. Suh, Y.; Wang, J.; Tang, S.; Mei, T.; Lee, K.M. Part-aligned bilinear representations for person re-identification. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 402–419.
33. Ge, Y.; Li, Z.; Zhao, H.; Yin, G.; Yi, S.; Wang, X.; Li, H. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. *Adv. Neural Inf. Process. Syst.* **2018**, 31. [CrossRef]
34. He, L.; Liang, J.; Li, H.; Sun, Z. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7073–7082.
35. He, L.; Sun, Z.; Zhu, Y.; Wang, Y. Recognizing partial biometric patterns. *arXiv* **2018**, arXiv:1810.07399.
36. Huang, H.; Li, D.; Zhang, Z.; Chen, X.; Huang, K. Adversarially occluded samples for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5098–5107.
37. Zhuo, J.; Lai, J.; Chen, P. A Novel Teacher-Student Learning Framework For Occluded Person Re-Identification. *arXiv* **2019**, arXiv:1907.03253.
38. He, L.; Wang, Y.; Liu, W.; Zhao, H.; Sun, Z.; Feng, J. Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8450–8459.
39. Wang, G.; Yang, S.; Liu, H.; Wang, Z.; Yang, Y.; Wang, S.; Yu, G.; Zhou, E.; Sun, J. 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, Seattle, WA, USA, 13–19 June 2020; pp. 6449–6458.
40. Jia, M.; Cheng, X.; Zhai, Y.; Lu, S.; Ma, S.; Tian, Y.; Zhang, J. Matching on sets: Conquer occluded person re-identification without alignment. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 1673–1681.
41. Sun, Y.; Xu, Q.; Li, Y.; Zhang, C.; Li, Y.; Wang, S.; Sun, J. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 10–15 June 2019; pp. 393–402.
42. Luo, H.; Gu, Y.; Liao, X.; Lai, S.; Jiang, W. Bag of tricks and a strong baseline for deep person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019; pp. 1487–1495.
43. Song, C.; Huang, Y.; Ouyang, W.; Wang, L. Mask-guided contrastive attention model for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1179–1188.
44. Liu, J.; Ni, B.; Yan, Y.; Zhou, P.; Cheng, S.; Hu, J. Pose transferrable person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4099–4108.
45. Sarfraz, M.S.; Schumann, A.; Eberle, A.; Stiefelhagen, R. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 420–429.