applied sciences

MDPI

Article

A Multi-Scale Graph Attention-Based Transformer for Occluded Person Re-Identification

Ming Ma [1], Jianming Wang [2] and Bohan Zhao [3,*]

[1] School of Life Sciences, Tiangong University, Tianjin 300387, China; maming@tiangong.edu.cn
[2] Tianjin Key Laboratory of Autonomous Intelligence Technology and Systems, Tiangong University, Tianjin 300387, China; wangjianming@tiangong.edu.cn
[3] School of Computer Science and Technology, Tiangong University, Tianjin 300387, China
* Correspondence: bohanzhao610@163.com

Abstract: The objective of person re-identification (ReID) tasks is to match a specific individual across different times, locations, or camera viewpoints. The prevalent issue of occlusion in real-world scenarios affects image information, rendering the affected features unreliable. The difficulty and core challenge lie in how to effectively discern and extract visual features from human images under various complex conditions, including cluttered backgrounds, diverse postures, and the presence of occlusions. Some works have employed pose estimation or human key point detection to construct graph-structured information to counteract the effects of occlusions. However, this approach introduces new noise due to issues such as the invisibility of key points. Our proposed module, in contrast, does not require the use of additional feature extractors. Our module employs multi-scale graph attention for the reweighting of feature importance. This allows features to concentrate on areas genuinely pertinent to the re-identification task, thereby enhancing the model's robustness against occlusions. To address these problems, a model that employs multi-scale graph attention to reweight the importance of features is proposed in this study, significantly enhancing the model's robustness against occlusions. Our experimental results demonstrate that, compared to baseline models, the method proposed herein achieves a notable improvement in mAP on occluded datasets, with increases of 0.5%, 31.5%, and 12.3% in mAP scores.

Keywords: person re-identification; graph attention; occluded; GCN

check for updates

Citation: Ma, M.; Wang, J.; Zhao, B. A Multi-Scale Graph Attention-Based Transformer for Occluded Person Re-Identification. *Appl. Sci.* **2024**, *14*, 8279. https://doi.org/10.3390/app14188279

Academic Editors: Douglas O'Shaughnessy and Pedro Couto

Received: 26 July 2024
Revised: 6 September 2024
Accepted: 12 September 2024
Published: 13 September 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The primary goal of person re-identification (ReID) tasks is to identify and match the same individual across different camera viewpoints [1,2]. Consequently, achieving efficient and precise person re-identification (ReID), which entails identifying and associating the same individual across distinct camera viewpoints, has emerged as a paramount challenge that urgently needs to be addressed [3,4]. The task of occluded ReID is even more challenging than traditional ReID due to two primary factors: occlusions not only lead to the loss of pedestrian information but also introduce irrelevant features, which are inadvertently captured as noise during feature extraction by standard neural networks. This process results in ReID methods learning fewer discriminative features from pedestrian images, leading to incorrect retrievals. Of late, partial ReID approaches have been proposed to tackle such complexities. However, the process of ReID still encounters a multitude of challenges, including pose variations, viewpoint changes, occlusion noise, and missing information, further magnifying its complexity.

Recently, a series of models [5–10] have emerged that apply the transformer architecture to the domain of person re-identification. Under occlusion scenarios, pertinent human features can be partially obscured by obtrusive objects. The introduction of occlusion noise compromises the efficacy of global attention, as our analysis suggests that the self-attention mechanism of ViT may be influenced by occluding objects. This results in attention being

misdirected towards irrelevant areas. Most graph-based works [11–19] have used human key points as graph nodes to integrate key point information. Therefore, corresponding modules must be used to extract human body node information or perform pose estimation. However, the proportion of human content in features is unpredictable. In the case of occluded datasets in particular, considering the influence of occluded objects on human features, missing key point information may weaken the robustness of the model.

To address the above limitation, we introduce a novel module that employs multi-scale graph attention for feature weight redistribution. Our module does not need to extract human body node information, automatically encodes the basic features extracted with Vision Transformer (ViT) into the graph structure, and carries out multi-scale graph attention [20,21] aggregation to solve the problem of the imbalance of the human body image in the occlusion scene. The results of visualization experiments show that our module can refocus the ViT-extracted features on more important parts. Building upon this innovation, we propose a person re-identification model specifically designed to excel in occlusion scenarios, aiming to achieve enhanced robustness and accuracy under challenging conditions where occlusions are prevalent.

In this study, a feature extraction network is introduced that combines transformers and multi-scale graph convolutional feature fusion and is specifically designed for occluded person re-identification (ReID) datasets. Our model employs Vision Transformer [22] as the backbone for image feature extraction, where images are divided into multiple patches and input into ViT. The final layer's transformer outputs are treated as nodes [5,23], with the class token serving as the central node. Given that the class token aggregates information from all tokens, edges are established between the class token and other nodes; in comparison, additional edges are formed between other nodes based on their feature similarity, creating a graph of nodes. This graph, along with node features, is then fed into a multi-scale graph attention module featuring three branches, each utilizing a graph attention network of a different scale. Moreover, we use fully connected layers to map the features to a common dimension before concatenation. Within our model, this module is stacked multiple times, allowing the ViT-extracted features to incorporate inter-node relationships after passing through. Additionally, the class token from ViT is processed through a network, such as SENet [24], to derive a channel attention weight. Ultimately, the class token, after graph convolution computations, is combined with the channel weight through a weighted operation, and the resultant feature is fed into a ReID classification head.

The main contributions of our study can be summarized as follows:

1. A method is proposed to construct graphical data centered around the transformer's class token, where each output token from the transformer is treated as a node, with the class token serving as the core. This approach offers good versatility and can be flexibly integrated into other transformer-based models.
2. A multi-scale graph attention-based person re-identification model is proposed, which greatly improves the recognition accuracy through integrating features from different image patches using a multi-scale graph attention module.
3. Extensive experiments conducted on the occluded person ReID databases validate our proposed method, MSGA-ReID, as being effective for occluded ReID tasks.

The remainder of this paper is organized as follows. Section 2 provides a brief overview of related work on person re-identification under the perspective of occlusion. Section 3 presents the proposed method in detail, including the overall architecture, construction of graph structural information, the MSGA (multi-scale graph attention) module, and global channel attention. Section 4 reports the experimental results on partial and occluded Re-ID datasets along with comparisons to other methods. Finally, Section 5 discusses the experimental result and future work.

2. Related Works

2.1. Occluded Person Re-Identification

In the context of ReID, feature extraction is crucial for accurate pedestrian identification, with early models employing Convolutional Neural Network (CNN)-based feature extraction modules. However, CNN-based feature extraction predominantly focuses on local regions [25], overlooking the significance of distant information. A series of models have recently emerged that apply the transformer architecture to the domain of person re-identification. Transformer-based visual models, endowed with self-attention mechanisms, excel at capturing the global context from input images, facilitating comprehensive understanding of the overall appearance features. This capability has led to significant advancements in CNN-based models in recent years.

However, under occlusion scenarios, pertinent human features can be partially obscured by obtrusive objects. The introduction of occlusion noise compromises the efficacy of global attention, as analysis results suggest that the self-attention mechanism of ViT can be influenced by occluding objects. This process results in attention being misdirected toward irrelevant areas, consequently undermining the performance of ViT models in occluded settings.

The objective of occluded person re-identification is to locate a person (or persons) with the same characteristics as in an occluded image across different camera angles. However, this task becomes notably more challenging in the presence of incomplete information about the people in occluded images and spatial misalignment due to varying viewpoints. Specifically, there are two main influences of occlusion on person ReID: occlusion noise and image scale change caused by occlusion.

2.2. Transformer-Based ReID

Researchers [5–10] have recently employed transformer-based models for feature extraction, introducing multi-head attention modules which make transformer-based models well-suited to address challenges in this area. Research on transformer-based ReID has seen substantial progress in recent years. Compared to Convolutional Neural Networks (CNNs), Vision Transformers exhibit superior global information processing capabilities. Through leveraging self-attention mechanisms, ViTs excel at capturing long-range dependencies and global contextual information, thereby enabling the model to effectively extract inter-regional feature correlations when dealing with variations in pedestrian poses, occlusions, and images from different viewpoints. TransReID [6] was the first model to apply ViT in ReID, in which ViT offers a more flexible feature representation; through dividing images into multiple patches and directly applying self-attention to these patches, the model learns features without the constraints imposed by local receptive fields, providing a more adaptable expression for subtle differences in pedestrian identities. DC-Former [5] employs multiple class tokens to represent diverse embedding spaces. This approach incorporates SDC in the output of the final transformer block, thereby encouraging class tokens to be distanced from one another and embedded into distinct representational spaces. Part-Aware Transformer (PAT) [9] introduces a mechanism that adaptively identifies and highlights multiple key body parts of persons during the training process. Furthermore, through incorporating occlusion-aware loss functions and training strategies, the model is able to better comprehend and adapt to the effects of occlusions, thereby enhancing its robustness in recognizing individuals under such challenging conditions. PVT [10] uses a pose estimator to detect key points in the human body and uses these points to locate intermediate features. These intermediate features of key points are input into the pose-based transformer branch to learn point-level features.

However, the attention mechanism of ViT may be influenced by occluded objects, resulting in attention being dispersed to the occluded area. Features of these parts are ineffective for ReID tasks, leading to a decrease in the performance of ViT-based models when considering occluded scenes.

2.3. Graph-Based ReID

Graph Convolutional Networks (GCNs) [26], with their ability to model abstract node relationships, have recently been applied in the ReID domain to capture temporal and spatial information in video sequences. In pedestrian images, body parts inherently possess structural relationships, and GCNs capitalize on this prior knowledge by encoding spatial or semantic similarities between pedestrian image features through edge weights, fortifying the model's ability to learn pedestrian characteristics.

Graph-based person re-identification leverages the highly structured human skeleton to extract semantic information at the image pose level, thereby suppressing noise interference through guiding or fusion mechanisms. PFD [17] divides images into overlapping fixed-size patches, followed by the employment of a transformer encoder to capture contextual relationships among these patches. Subsequently, pose-guided feature aggregation and a feature-matching mechanism are utilized to explicitly emphasize visible body parts. Finally, pose heatmaps and a decoder are leveraged as keys and values to enhance the discriminability of individual body parts. HOReID [11] introduces a learnable relation matrix, treating human key points obtained from pose estimation as nodes in a graph and, ultimately, forming a topological graph that mitigates noise disturbances. The PMFB [18] utilizes pose estimation to acquire confidence scores and coordinates of human key points. Subsequently, thresholds are set to filter out occluded regions, and visible parts are employed to constrain feature responses at the channel level, addressing the challenge of occlusion. PGMANet [19] employs human heatmaps to generate attention masks, synergistically eliminating noise interference through both element-wise multiplication with feature maps and guidance from higher-order relations. EcReID [12] consists of three modules: a mutual denoising module, an inter-node aggregation and update module, and a graph matching module. Among them, the graph matching module uses a graph matching method based on the human body's topology to obtain a more accurate calculation of the mask image similarity.

The authors of some studies have partitioned each image's features horizontally [3], treating each segment as a node. However, human-related features may constitute a relatively small proportion of horizontal features in an image, and relying solely on horizontally segmented features can lead to interference from background characteristics in experimental outcomes. The authors of other works have integrated key point information, using human key points as graph nodes. However, he ratio of human content in features is unpredictable; this is particularly true in the case of occluded datasets, considering the fact that the impact of occluders on human features might result in missing key point information, undermining the model's robustness.

3. Methods

To better capture the features of irregularly shaped persons and associate local information with global information, thus extracting truly effective information, we propose a Multi-Scale Graph Attention-based Pedestrian Re-identification model, called MSGA-ReID. This approach comprises five main components: data augmentation, graph construction, feature extraction, multi-scale graph attention feature aggregation, and global channel attention. The network architecture is depicted in Figure 1. We employed two data augmentation strategies to fully leverage the information in the training dataset and utilized Vision Transformer for fundamental feature extraction.

To enhance the dependency among features, we constructed a graph-structured dataset based on the features extracted by ViT. Features output by the transformer serve as nodes, with images synthesized based on feature similarity. The class token, through multi-head attention mechanisms, engages in information exchange with all image patches, "summarizing" this local information to form an understanding of the entire image's content. This process enables the class token to carry high-level semantic features regarding the image's overall classification, thereby establishing connections between the central node and all other nodes. Subsequently, node features together with graph information are

fed into our proposed multi-scale graph attention module for feature aggregation. At this stage, we posit that our module is capable of performing relationship extraction among features and reallocating feature weights, thereby focusing the features on aspects that genuinely contribute to pedestrian re-identification.

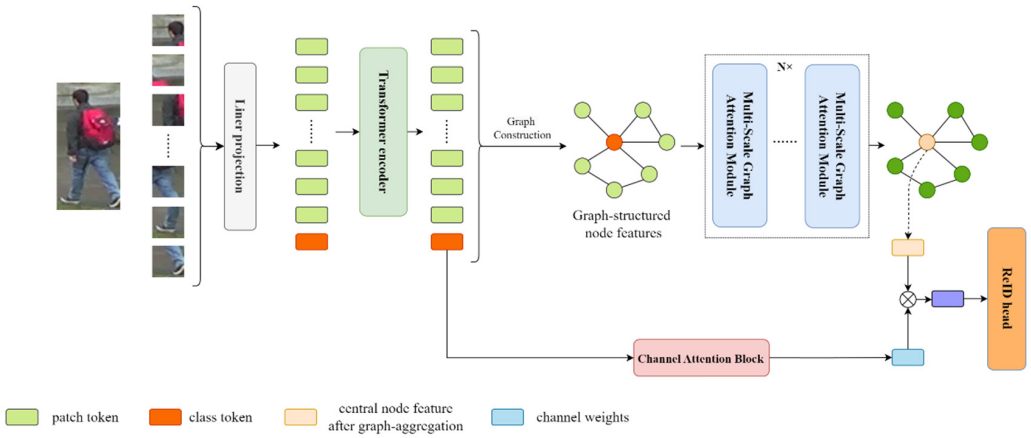

Figure 1. The framework of our MSGA-ReID approach is structured as follows: each token output by the transformer is regarded as a node, and a graph structure is established through computing the similarity between the features of these nodes and linking the central node to all other nodes. Subsequently, information from all nodes is aggregated into the central node via multiple instances of our proposed multi-scale graph attention (MSGA) modules. A global channel-wise attention mechanism is employed for weight allocation, integrating class label information with the features of the centralized node to facilitate effective re-identification.

3.1. Feature Extraction

In our model architecture, we opted for the Vision Transformer as the backbone network, which is tasked with efficiently extracting high-level semantic features from input images. The introduction of ViT has revolutionized conventional feature extraction methodologies based on Convolutional Neural Networks (CNNs), with it particularly excelling in handling global information and long-range dependencies. The images were segmented into uniformly sized patches and, through setting the stride of the ViT network to slightly less than the patch size, we ensured a degree of overlap between adjacent patches. This process increases information redundancy, enabling the model to capture continuity and detail within local regions to some extent, thereby enhancing its ability to learn fine-grained features while maintaining sensitivity to spatial structures.

Each patch is regarded as a visual word (visual token). Through linear projection, these patches are mapped into a fixed-dimensional vector space, ensuring encoding of the image's local features. Consequently, each patch is transformed into a D-dimensional vector after projection. A special class token is prepended to all patch vectors, serving as the sequence's start and guiding the model to learn a global feature representation that is conducive to classification. The serialized patch vectors, together with the class token, are then fed into a multi-layer transformer encoder. The self-attention mechanism within the transformer allows for global information exchange among different patches, capturing long-range dependencies. The hierarchical structure, in comparison, bolsters feature expression through deep learning.

The input image $x \in R^{H \times W \times C}$ is divided into N patches of size (P, P). We configured the ViT model with a patch size of P and a stride of S; consequently, the number of patches into which an image is divided is calculated as follows:

$$N = (H - P)(W - P)/S^2, \quad (1)$$

where (H, W) is the resolution of the input image and C is the number of channels. Subsequently, trainable linear projections are employed to map these patches onto D-dimensional

vectors, with the resultant output recorded as patch embeddings. An additional learnable embedding is incorporated into the input sequence of the transformer, serving as a class token xclass to learn contextual information from other embeddings. Next, a learnable position embedding is added to the input sequence, and the sequence is fed into multiple transformer layers. We express the feature $Z = \{z_{class}, z_1, z_2, \ldots, z_N\} \in R^{(N+1)\times d}$ (where d is the dimension of each feature vector) output by ViT as follows:

$$Z = ViT(I),\ I \in \left[I_{original}, I_{erasing}, I_{cropping}\right]. \tag{2}$$

3.2. Graph Construction

In our proposed method, we leverage the features output by ViT to construct graph-structured information, thereby enhancing the model's comprehension and expression of the pedestrian feature space. In our overall architecture, features are passed through this module multiple times to enhance their representations. The expression of the graph-structured information takes the form of

$$G = (F, A), \tag{3}$$

$$similarity(f_i,\ f_j) = \frac{f_i \cdot f_j}{\|f_i\| \cdot \|f_j\|}, \tag{4}$$

$$a_{i,j} = \begin{cases} 1, & i = 0 \ \ or\ j = 0 \\ 1, & \text{similarity}(f_i, f_j) > m \\ 0, & \text{similarity}(f_i, f_j) \le m \end{cases}, \tag{5}$$

where F denotes the features of nodes and A is the adjacency matrix. Specifically, we assess the degree of association between different nodes by computing the cosine similarity between their respective feature vectors. When the cosine similarity between the features of two nodes surpasses a predefined threshold, we infer a strong correlation between these nodes in the feature space, prompting us to establish an edge connection between them. Consequently, the corresponding entries in the constructed adjacency matrix are assigned a value of 1, visually representing the established connections between nodes. In alignment with the principles underlying the transformer architecture, we posit that the class token's features encapsulate information regarding global relationships among relevant features. Consequently, we establish connections between the class token and all other nodes. Then, all remaining connections are set to 0. For specific algorithmic pseudocode, please refer to Algorithm 1.

Algorithm 1: Calculate AdjacencyMatrix by CosineSimilarity

Input: n; Features; AdjMatrix
Output: AdjMatrix

```
var
  i, j: integer;
  dotProduct, magnitudeI, magnitudeJ, cosineSimilarity: real;
begin
  for i:= 1 to n do
    for j:= 1 to n do
      AdjMatrix[i, j]:= 0;
  for i:= 1 to n do
    for j:= 1 to n do
    begin
      dotProduct:= 0.0;
      for k := 1 to Length(Features[i]) do
        dotProduct:= dotProduct + Features[i][k] * Features[j][k];
      magnitudeI:= 0.0;
      magnitudeJ:= 0.0;
      for k:= 1 to Length(Features[i]) do
```

Algorithm 1: *Cont.*

```
17.            begin
18.              magnitudeI:= magnitudeI + Features[i][k] * Features[i][k];
19.              magnitudeJ:= magnitudeJ + Features[j][k] * Features[j][k];
20.            end;
21.            magnitudeI:= sqrt(magnitudeI);
22.            magnitudeJ:= sqrt(magnitudeJ);
23.            if (magnitudeI > 0.0) and (magnitudeJ > 0.0) then
24.              cosineSimilarity:= dotProduct / (magnitudeI * magnitudeJ)
25.            else
26.            cosineSimilarity:= 0.0;
27.            if cosineSimilarity > threshold then
28.              AdjMatrix[i, j]:= 1;
29.          end;
30.        for i := 1 to n do
31.          AdjMatrix[i, 0]:= 1;
32.          AdjMatrix[0, i]:= 1;
33.    end;
```

3.3. Multi-Scale Graph Attention Feature Aggregation

After acquiring the graph information, we feed the graph structural details, comprising the features of nodes and the adjacency matrix, into our proposed multi-scale graph attention feature fusion module. Details of this module's architecture are depicted in Figure 2. Each branch employs a dedicated graph attention module to extract features, with each branch configured to operate at a different feature scale. The feature calculation formula for the MSGA module is as follows:

$$e_{i,k,s}^{(l)} = W_{k,s}^{(l)} h_{i,k,s}^{(l)}, \tag{6}$$

$$\alpha_{ij,k,s}^{(l)} = softmax_j\left(LeakyReLU\left(a_k^{(l)^T}\left[e_{i,k,s}^{(l)}\right]\left[e_{i,k,s}^{(l)}\right]\right)\right), \tag{7}$$

$$f_{i,k,s}^{(l)} = MLP\left(ReLU\left(\sum_{j\in N}\alpha_{ij,k,s}^{(l)} e_{j,k,s}^{(l)}\right)\right) + h_{i,k,s}^{(l)},\ k = 1, 2, \ldots,\ K \tag{8}$$

$$F^{(l+1)} = F_{s1}^{l} + F_{s2}^{l} + F_{s3}^{l}. \tag{9}$$

In the formula, h represents the input features, α represents the multi head attention score in the graph attention module, f represents the features of a single node aggregated by the module, and F represents the overall feature map fused from multiple scales. For each attention head k, the graph attention computation is initiated by applying a weighting matrix to the feature vectors of each node, effectively transforming them into a new feature space. We control the scales of features with varying dimensionality by setting $W_{k,s}^{(l)}$, where k denotes the attention head and s is the scale of each branch. Subsequently, attention coefficients are computed for each pair of nodes within each head, following which, the features of neighboring nodes are aggregated through a weighted sum according to their respective attention coefficients. Then, the outputs from all heads are concatenated to obtain the feature representation after the nodes have undergone graph attention extraction. Subsequently, a Multi-layer Perceptron (MLP) layer is employed to map features from different scales into a common dimensionality. Finally, the features extracted from the three branches at different scales are aggregated through summation to achieve feature fusion. To mitigate the potential loss of original feature information in deeper layers, residual connections are introduced to reinforce the representation of the original features.

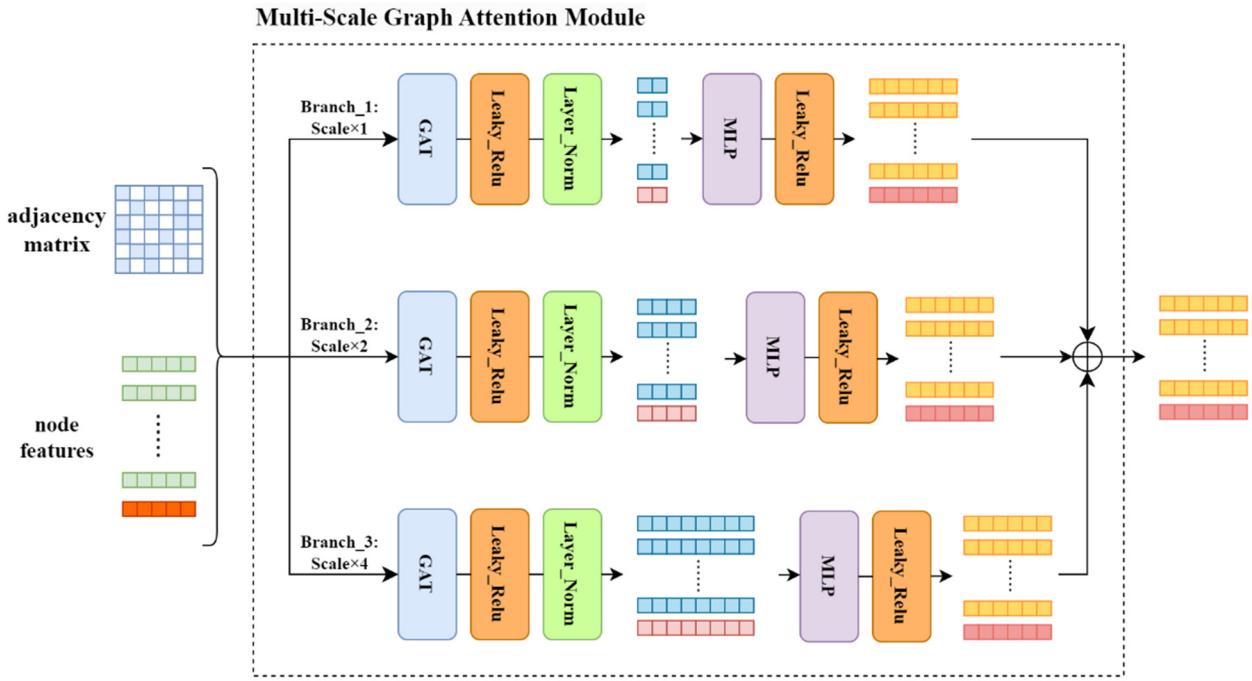

Figure 2. The structure of our proposed multi-scale graph attention (MSGA) module is as follows: features at different scales are extracted through three distinct branches. These features are projected to a common dimension via linear layers and subsequently fused. In our model, features are fed through the MSGA module multiple times in order to extract higher-order characteristics of occluded persons.

3.4. Global Channel Attention

Ultimately, our model employs a channel attention operation on the aggregated features. The raw class tokens output by the ViT model are passed through an MLP to derive weights for the channel features, which are then used as channel attention to perform a weighted multiplication with the features. The above process is illustrated by the following equation:

$$\hat{F} = softmax\left(MLP\left(Z_{class_token}\right)\right) \otimes F. \tag{10}$$

Integrating the global perspective of the transformer architecture with the channel attention mechanism, the class token—once transformed via the MLP—carries comprehensive information about the entire image sequence. Utilizing this token to generate channel-level attention weights ensures that the model not only considers local features but also leverages the global context effectively. This methodology enables the model to more accurately discern the correlations and significance among different feature channels, leading to richer and more discriminative feature representations.

4. Experiments and Results

4.1. Datasets

Experiments were conducted across three occlusion scenario datasets in order to assess the performance of our proposed methodology. Our primary objective was to address the challenge of occluded person re-identification; hence, the experiments utilized the Occluded-Duke [27], OccludedREID [28], and Partial-ReID [29] datasets. The Mean Average Precision (mAP) and Rank-1 accuracy were employed as standard evaluation metrics, consistent with prevailing practices in ReID research. Details about the dataset we used can be found in Table 1.

Table 1. Datasets.

Dataset	#ID	#Train	#Gallery	#Query	#Cam
Occ_Duke	1401	15,618	17,661	2210	8
Occ_ReID	200	-	1000	1000	-
Partial_ReID	60	-	300	300	-

The Occluded-Duke dataset was constructed based on the original DukeMTMC-reID dataset, yielding Occluded-DukeMTMC, which encompasses 15,618 training images, 17,661 gallery images, and 2210 occluded query images, making it the largest occluded ReID dataset to date. DukeMTMC-reID serves as a widely adopted benchmark for pedestrian re-identification, featuring a substantial collection of pedestrian images captured by surveillance cameras, ideal for assessing the performance of cross-camera tracking algorithms. The Occluded-Duke dataset was meticulously derived from DukeMTMC-reID by selecting instances with occlusions, aiming to escalate the difficulty of re-identification tasks, particularly when parts of pedestrians' bodies are obscured by objects such as pillars, backpacks, or other pedestrians.

The Occluded-REID dataset is specifically designed for occluded person re-identification, consisting of 2000 images of 200 distinct pedestrians captured by a moving camera. Each identity is represented by five full-body gallery images and five query images with different viewpoints and varying degrees of severe occlusion. All images have been uniformly resized to a dimension of 128 $\times$ 64, facilitating studies on the impact of occlusion on pedestrian recognition.

The Partial-ReID dataset comprises 600 images of 60 pedestrians, with each pedestrian represented by 5 partial images and 5 full-body images. Using the visible portions, the images are manually cropped to create new partial images. The images in the dataset were collected across a university campus under various viewpoints, backgrounds, and degrees of occlusion.

Occluded-Duke is a subdataset of DukeMTMC-reID, retaining the occluded images in the source dataset. We randomly divided the training dataset into training part and validation part at a ratio of 9:1. In addition, the Occluded-ReID and Partial-ReID datasets only provide the testing set, so we trained our model on Market-1501 and reused these two datasets for testing.

4.2. Data Augmentation

Almost all commonly utilized data augmentation methods involve the use of random erasing, cropping, flipping, filling, and adding noise operations to the training dataset during the data pre-processing stage. However, the random enhancement operation cannot be reproduced; that is, the method of each enhancement is not the same. In order to mitigate the above problems, we adopted a fixed enhancement method to make full use of the training data and strengthen the robustness of our model. We performed two enhancement operations by erasing and cropping on the input image x to obtain three groups of images ($I_{original}$, $I_{erasing}$, and $I_{cropping}$). Among them, $I_{original}$ is the original input image, $I_{erasing}$ adds obstacles to the image, and $I_{cropping}$ irregularly crops the image. Based on the original occluded data, we obtained two images with enhanced degrees of occlusion based on each original image. As shown in Figure 3, several examples of the data augmentation methods we used are provided.



Figure 3. Samples of images acquired using the data augmentation techniques employed in our approach: random cropping and random erasing.

4.3. Implementation

All experiments were conducted on four NVIDIA V100 GPUs. The fundamental architecture of our approach was based on VIT-B/16, comprising 12 transformer layers. The model's initial weights were pre-trained on ImageNet. In our experiments, we also incorporated methodologies from TransReID, employing both overlapping patch embedding and Scale-Invariant Embedding (SIE). When constructing the adjacency matrix, the threshold m was set to the mean value of the similarity matrix. The batch size was configured to 128, with each containing four images per identity. The initial learning rate for the experiments was set at 0.00196, and the learning rate was decayed following a cosine annealing schedule.

We used the Cumulative Matching Curve and Mean Average Precision (mAP) to evaluate different ReID methods.

The Rank-1 accuracy of CMC is a metric that evaluates whether the correct match can be found at the top of the retrieval list. Specifically, if the image ranked first in the retrieval results corresponding to a query image indeed belongs to the same individual, it is counted as a successful match. The Rank-1 accuracy is, thus, the ratio of successful matches to the total number of queries.

The Mean Average Precision (mAP) is a more comprehensive evaluation metric that takes into account all correct match positions in the retrieval results. mAP is the mean of the average precision, where precision is defined as the ratio of the number of correct matches to the total number of retrieved items up to a certain position in the retrieval list; recall, in contrast, is the ratio of the number of correct matches to the actual total number of matches. This metric provides a balanced view between precision and recall, offering a more holistic assessment of retrieval performance.

4.4. Results of Dataset Verification

We compared our approach with various state-of-the-art methods on the occluded person re-identification benchmark, as listed in Table 2. The comparison encompassed three

categories of methods: multi-scale convolutional approaches, methods utilizing spatial key point information, and those based on the transformer architecture. The results of our experiments demonstrate that our method outperforms other methods in occluded scenarios, excelling in extracting more effective features compared to its counterparts. Figure 4 shows the mAP improvement curve of our proposed method and the decrease in loss during the training process. Figure 5 shows the ROC curves of our model on different datasets. ROC provides a visual representation of the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) at various threshold settings.

Table 2. Comparison with state-of-the-art methods on occluded ReID datasets. The label ours∗ denotes configurations where the number of multi-head attention heads is set to four, while that for ours was set to eight (the best results are shown in bold, and the second-best result is underlined).

		Occ_Duke		Occ_ReID		Partial_ReID	
	Method	**mAP**	**Rank-1**	**mAP**	**Rank-1**	**mAP**	**Rank-1**
Multi-scale CNN	DSR [30]	30.4	40.8	62.8	72.8	-	50.7
	FPR [31]	-	-	68.0	78.3	-	68.1
Transformer	TransReID [6]	55.7	64.2	67.3	81.6	-	68.6
	DC-former [5]	56.6	63.3	45.7	49.0	-	73.0
	PAT [9]	53.6	64.5	72.1	70.2	-	-
	PVT [10]	**57.6**	<u>65.5</u>	74.0	79.1	-	81.0
Spatial key point graph	PVPM [13]	-	-	61.2	70.4	71.4	78.3
	OAMN [14]	46.1	62.6	-	-	-	**86.0**
	PAFM [15]	42.3	55.1	68.0	76.4	-	82.5
	HOReID [11]	43.8	55.1	70.2	80.3	-	85.3
	RFCNet [16]	54.5	63.9	-	-	-	-
	EcReID [12]	52.7	64.8	75.1	**84.5**	-	81.0
Others	QPM [32]	49.7	64.4	-	-	-	81.7
	MoS [33]	55.1	66.6	-	-	-	-
	BMM [34]	55.6	63.4	-	-	-	73.7
	ours*	56.9	<u>65.5</u>	**79.3**	<u>83.0</u>	**81.5**	84.9
	ours	<u>57.1</u>	**66.8**	<u>77.2</u>	81.3	<u>80.8</u>	<u>85.3</u>

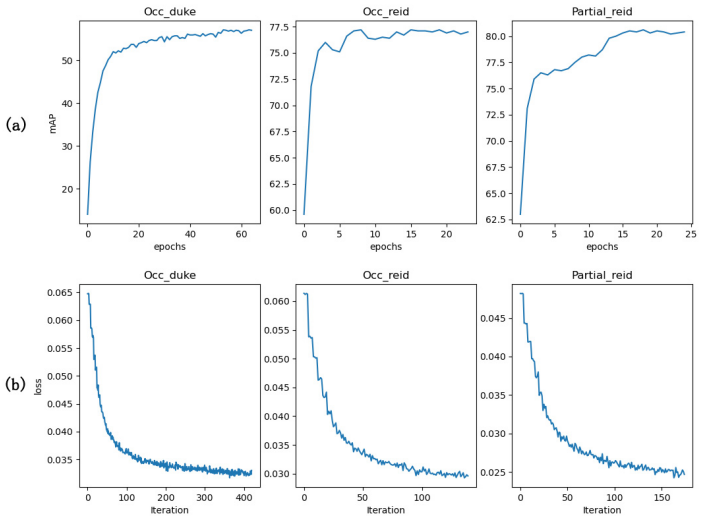

Figure 4. mAP curve and loss curve during training on different datasets. From left to right are the situations on the occ_duke, occ_reid, and partial_reid datasets. The first row of three images labeled as' (**a**) display changes in mAP metrics on the test set. The second row of three images labeled as' (**b**) display changes in loss on the train set.

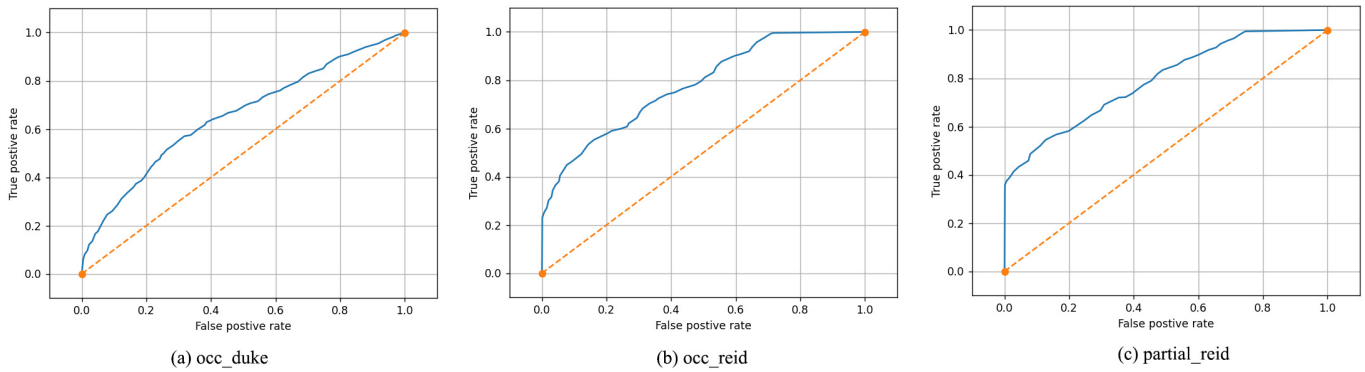

Figure 5. Our model's ROC curves on the test sets of various datasets, from left to right, are occ_cuke, occ_deid, and part_deid, and their AUC values are 0.655, 0.758, and 0.782, respectively. ROC provides a visual representation of the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) at various threshold settings. TPR, also known as sensitivity or recall, measures the proportion of actual positive instances that are correctly identified by the model. FPR represents the proportion of actual negative instances that are incorrectly identified as positive. The Area Under the Curve (AUC) is a single scalar value that quantifies the overall performance of the model across all possible thresholds.

4.5. Comparison of Parameter Tuning

In Table 3, we present the results of our evaluation of the effectiveness of data augmentation on the occluded datasets. To fully leverage the training information and mitigate the issue of data imbalance in the test set, three augmentation techniques were applied during the data pre-processing stage. The experimental results show that data augmentation significantly enhanced the performance of our model. For OCC_Duke, the augmentation strategies notably improved the method's efficacy, with an increase of 5.6% in mAP and 1.8% in Rank-1 accuracy. For OCC_ReID, the improvements in mAP and Rank-1 accuracy were even more pronounced, at 7.4% and 2.4%, respectively. Likewise, for Partial-ReID, the enhancements yielded an increase of 2.3% in Rank-1 accuracy. It is evident that the incorporation of data augmentation was efficacious across all three occlusion datasets.

Table 3. Comparison of data augmentation.

	Occ_Duke		**Occ_ReID**		**Partial_ReID**
	mAP	**Rank-1**	**mAP**	**Rank-1**	**Rank-1**
ours (no_aug)	51.5	65.0	69.8	78.9	83.0
ours (aug)	57.1	66.8	77.2	81.3	85.3

In Table 4, we present our evaluation of the impact of the number of layers in our proposed module on the results. The MSGA module was configured with one, two, or three layers, and the results of our experiments show that the best performance was attained when the number of layers was set to three. This result proves that deeper MSGA modules can learn more intricate feature representations, enabling the module to capture abstract ReID features.

Table 4. Parameter tuning of MSGA layers.

Num of MSGA Layers	**Occ_Duke**		**Occ_ReID**		**Partial_ReID**
	mAP	**Rank-1**	**mAP**	**Rank-1**	**Rank-1**
1	56.1	66.3	76.6	80.8	85.1
2	56.5	65.7	76.8	80.9	84.8
3	57.1	66.8	77.2	81.3	85.3

In Table 5, we present our analysis of the impact of the number of attention heads. We configured the number of heads to two, four, six, or eight. Moreover, we conducted experiments on occluded datasets. The use of multi-head attention enables GAT to concurrently capture diverse features from the input data. Each head is initialized with distinct weights. An increase in the number of heads allows for the learning of a broader range of distinct features; however, beyond a certain threshold, some heads may begin to learn redundant features, duplicating information already captured by others which, in turn, leads to a decline in performance.

Table 5. Parameter tuning of attention heads.

Num of Attention Heads	Occ_Duke		Occ_ReID		Partial_ReID
	mAP	Rank-1	mAP	Rank-1	Rank-1
2	53.3	63.3	76.4	81.1	83.5
4	56.9	65.5	79.3	83.0	84.9
6	56.5	66.4	77.3	82.5	85.6
8	57.1	66.8	77.2	81.3	85.3

4.6. Visual Experiment

4.6.1. Examples of Inference on the Occluded Dataset

In the following section, we show two examples of inference on the Occ_Duke dataset. As illustrated in Figure 6, both of the query images show occluded persons. It can be seen that, compared to the other two transformer-based baseline approaches, our approach was more clearly aware of persons being occluded. The other two methods presented more cases of using shelter features for matching, resulting in errors.

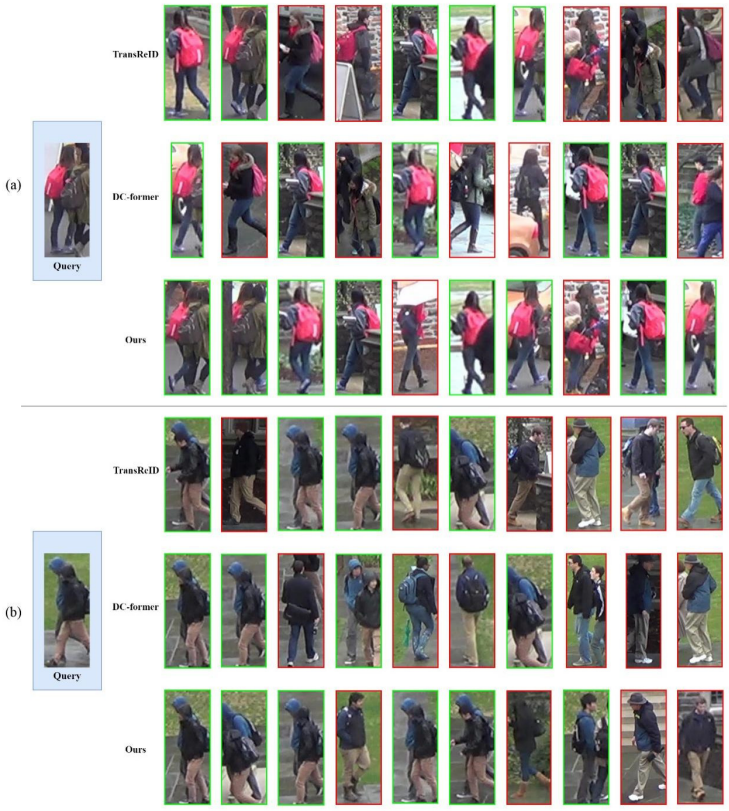

Figure 6. Two illustrative examples of inference on the occlusion dataset Occ_Duke. Panels (**a**) and (**b**) depict the top 10 matches (Rank-10) for person re-identification corresponding to a query image. The results from top to bottom were obtained with TransReID, DC-Former, and our proposed method. Green and red borders around images denote correct and incorrect match results.

4.6.2. Visual Comparison of Feature Attention

As shown in Figure 7, gradient-based visualization was employed to analyze the attention heatmap of the features ultimately fed into the ReID head. It is evident that, in comparison with the two baseline approaches, our model was notably effective in directing focus to salient portions of the person, rather than being distracted by occluding objects, particularly under conditions of occlusion.

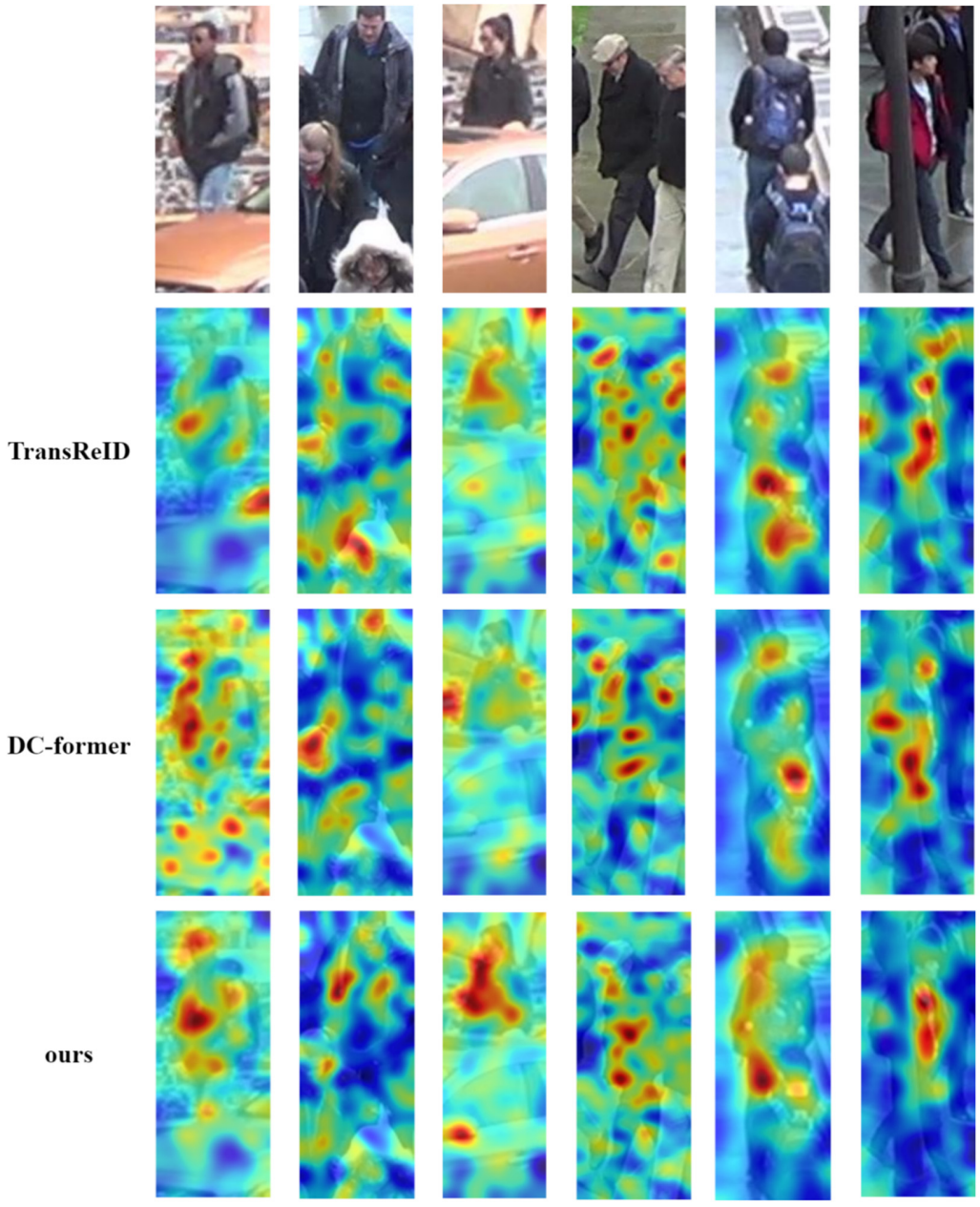

Figure 7. Heatmaps of feature maps after processing with three distinct transformer-based models. In the heatmaps, red regions signify areas of high attention focus by the model, whereas blue areas indicate low attention focus.

4.7. Conclusions of Experiments

In the experimental section, we compared our method with various listed reference methods. The comparison included three methods: the multi-scale convolution method, key point feature-based method, and transformer architecture-based method. Our experimental results indicate that our method outperforms other methods in occluded scenes. Because

ViT can capture long-range dependencies in images through its self-attention mechanism, the model is more robust to occlusion and viewpoint changes. The multi-scale graph attention mechanism can further enhance the sensitivity of the model to local changes, thereby improving pedestrian re-identification performance in different occlusion scenarios. Our model outperforms most of the three types of models mentioned above on occlusion datasets. Additionally, the heatmap of the attention visualization experiment also shows that our MSGA module can pull the attention for feature extraction towards parts more relevant to the task.

5. Discussion

In this study, we proposed a novel module that harnesses the graph attention mechanism to aggregate human feature information. After analyzing other transformer-based models, we believe that the self-attention mechanism in ViT might be vulnerable to interference caused by occluding objects, which can result in attention being scattered across irrelevant areas. To mitigate the impact of occlusions, some approaches have utilized pose estimation or human key point detection to create graph-structured data. However, these methods introduce additional noise, partly due to challenges like the invisibility of key points.

Our method addresses the reliance of graph-based models on human key point detection through reassembling the features extracted with ViT to concentrate on effective information segments under occlusion scenarios. Our overarching framework employs ViT as the backbone, treating features extracted by ViT as graph nodes, which are then processed with our multi-scale attention module for information aggregation. By adopting graph structures and attention-driven feature learning, the network is encouraged to focus on features that persist stably, even under occluded conditions, while integrating pedestrian features across various scales. This approach sustains robust recognition performance in the presence of occlusions. Through assigning attention weights to feature maps at different scales, the model effectively integrates both global and local information.

Our MSGA-ReID surpasses most existing models based on graph convolutional approaches—albeit not outperforming the most recent ones—which remains an area for future investigation and improvement. Currently, our method for constructing graph-structured information relies solely on feature similarity, which may not fully align with the feature distribution of person re-identification tasks, thus potentially degrading model performance.

We found that existing ReID models based on graph structures, include ours, have some limitations when dealing with changes in occlusion patterns. Specifically, the current dataset size is not sufficient to cover all population and environmental changes, resulting in limited generalization ability of the model on new samples. In addition, in terms of occlusion processing, the model may have difficulty effectively dealing with situations where the occlusion area is large or the occlusion position is uncertain. To overcome these limitations, future directions should include expanding the dataset size, developing more robust occlusion processing strategies, and improving the model's cross-domain adaptability. We plan to enhance the robustness of the model by designing new occlusion data augmentation methods and utilizing multi-scale feature representation and attention mechanisms. We will also explore efficient graph processing algorithms and techniques to reduce computational costs and enable better deployment of models in practical applications. We plan to explore these issues in future work.

Author Contributions: Methodology, J.W.; formal analysis, B.Z.; data curation, M.M.; writing original draft preparation, B.Z.; Conceptualization, J.W., M.M. and B.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/lightas/Occluded-DukeMTMC-Dataset; https://github.com/tinajia2012/ICME2018_Occluded-Person-Reidentification_datasets; https://hyper.ai/datasets/16904 (accessed on 11 September 2024).

Acknowledgments: The authors acknowledge the anonymous reviewers for their helpful comments on the manuscript.

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

References

1. Yan, C.; Pang, G.; Jiao, J.; Bai, X.; Feng, X.; Shen, C. Occluded Person Re-Identification with Single-Scale Global Representations. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 11875–11884.
2. Chen, Y.C.; Zhu, X.; Zheng, W.S.; Lai, J.H. Person re-identification by camera correlation aware feature augmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 392–408. [CrossRef] [PubMed]
3. Peng, Y.; Wu, J.; Xu, B.; Cao, C.; Liu, X.; Sun, Z.; He, Z. Deep Learning Based Occluded Person Re-Identification: A Survey. *ACM Trans. Multimed. Comput. Commun. Appl.* **2023**, *20*, 1–27. [CrossRef]
4. Ning, E.; Wang, C.; Zhang, H.; Ning, X.; Tiwari, P. Occluded person re-identification with deep learning: A survey and perspectives. *Expert Syst. Appl.* **2023**, *239*, 122419. [CrossRef]
5. Li, W.; Zou, C.; Wang, M.; Xu, F.; Zhao, J.; Zheng, R.; Cheng, Y.; Chu, W. Dc-Former: Diverse and Compact Transformer for Person Re-Identification. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 1415–1423.
6. He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; Jiang, W. Transreid: Transformer-based Object Re-Identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 15013–15022.
7. Xu, B.; He, L.; Liang, J.; Sun, Z. Learning feature recovery transformer for occluded person re-identification. *IEEE Trans. Image Process.* **2022**, *31*, 4651–4662. [CrossRef] [PubMed]
8. Zhu, K.; Guo, H.; Zhang, S.; Wang, Y.; Liu, J.; Wang, J.; Tang, M. Aaformer: Auto-Aligned Transformer for Person Re-Identification. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**. [CrossRef] [PubMed]
9. 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, 19–25 June 2021; pp. 2898–2907.
10. Gao, H.; Hu, C.; Han, G.; Mao, J.; Huang, W.; Guan, Q. Point-level feature learning based on vision transformer for occluded person re-identification. *Image Vis. Comput.* **2024**, *143*, 104929. [CrossRef]
11. Wang, P.; Zhao, Z.; Su, F.; Zu, X.; Boulgouris, N.V. Horeid: Deep high-order mapping enhances pose alignment for person re-identification. *IEEE Trans. Image Process.* **2021**, *30*, 2908–2922. [CrossRef] [PubMed]
12. Zhu, M.; Zhou, H. EcReID: Enhancing Correlations from Skeleton for Occluded Person Re-Identification. *Symmetry* **2023**, *15*, 906. [CrossRef]
13. Gao, S.; Wang, J.; Lu, H.; Liu, Z. Pose-Guided Visible Part Matching for Occluded Person Reid. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11744–11752.
14. Chen, P.; Liu, W.; Dai, P.; Liu, J.; Ye, Q.; Xu, M.; Chen, Q.; Ji, R. Occlude them All: Occlusion-Aware Attention Network for Occluded Person Re-Id. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 11833–11842.
15. Yang, J.; Zhang, C.; Tang, Y.; Li, Z. PAFM: Pose-drive attention fusion mechanism for occluded person re-identification. *Neural Comput. Appl.* **2022**, *34*, 8241–8252. [CrossRef]
16. Hou, R.; Ma, B.; Chang, H.; Gu, X.; Shan, S.; Chen, X. Feature completion for occluded person re-identification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 4894–4912. [CrossRef] [PubMed]
17. Wang, T.; Liu, H.; Song, P.; Guo, T.; Shi, W. Pose-Guided Feature Disentangling for Occluded Person Re-Identification based on Transformer. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; Volume 36, pp. 2540–2549.
18. Miao, J.; Wu, Y.; Yang, Y. Identifying visible parts via pose estimation for occluded person re-identification. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 4624–4634. [CrossRef] [PubMed]
19. Zhai, Y.; Han, X.; Ma, W.; Gou, X.; Xiao, G. Pgmanet: Pose-Guided Mixed Attention Network for Occluded Person Re-Identification. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Virtual, 18–22 July 2021.
20. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph attention networks. *arXiv* **2017**, arXiv:1710.10903.
21. Wang, Z.; Chen, J.; Chen, H. EGAT: Edge-Featured Graph Attention Network. In *Artificial Neural Networks and Machine Learning–ICANN 2021, Proceedings of the 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, 14–17 September 2021*; Proceedings, Part I 30; Springer International Publishing: Berlin/Heidelberg, Germany, 2021; pp. 253–264.
22. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
23. Han, K.; Wang, Y.; Guo, J.; Tang, Y.; Wu, E. Vision GNN: An image is worth graph of nodes. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 8291–8303.

24. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
25. 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.
26. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**, arXiv:1609.02907.
27. 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.
28. 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), San Diego, CA, USA, 23–27 July 2018; pp. 1–6.
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. 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, 18–22 June 2018; pp. 7073–7082.
31. 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.
32. Wang, P.; Ding, C.; Shao, Z.; Hong, Z.; Zhang, S.; Tao, D. Quality-aware part models for occluded person re-identification. *IEEE Trans. Multimed.* **2022**, *25*, 3154–3165. [CrossRef]
33. 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.
34. Wang, Y.; Wang, L.; Zhou, Y. Bi-level deep mutual learning assisted multi-task network for occluded person re-identification. *IET Image Process.* **2023**, *17*, 979–987. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.