symmetry

Article

Constructing Adaptive Multi-Scale Feature via Transformer-Aware Patch for Occluded Person Re-Identification

Zhi Liu *,†, **Xingyu Mu** †, **Shidu Dong**, **Yunhua Lu** and **Mingzi Jiang**

School of Artificial Intelligence, Chongqing University of Technology, Chongqing 401120, China; muxy@2020.cqut.edu.cn (X.M.); dongsd@cqut.edu.cn (S.D.); yhlu@cqut.edu.cn (Y.L.); mingziJ@stu.cqut.edu.cn (M.J.)

* Correspondence: liuzhi@cqut.edu.cn; Tel.: +86-139-8398-2460

† These authors contributed equally to this work.

Abstract: Person re-identification (Re-ID) aims to retrieve a specific pedestrian across a multi-disjoint camera in a surveillance system. Most of the research is based on a strong assumption that images should contain a full human torso. However, it cannot be guaranteed that all the people have a clear foreground because they are out of constraint. In the real world, a variety of occluded situations frequently appear in video monitoring, which impedes the recognition process. To settle the occluded person Re-ID issue, a new Dual-Transformer symmetric architecture is proposed in this work, which can reduce the occluded impact and build a multi-scale feature. There are two contributions to our proposed model. (i) A Transformer-Aware Patch Searching (TAPS) module is devised to learn visible human region distribution using a multiheaded self-attention mechanism and construct a branch of distributed information attention scale. (ii) An Adaptive Visible-Part Cropping (AVPC) Strategy, with two steps of cropping and weakly-supervised learning, is used to generate a fine-scale visible image for another branch. Only ID labels are utilized to restrain TAPS and AVPC without any extra visible-part annotation. Extensive experiments are conducted on two occluded person Re-ID benchmarks, confirming that our approach performs a SOTA or comparable effect.

Keywords: person re-identification; multiheaded self-attention; weakly-supervised learning; multi-scale feature

check for updates

Citation: Liu, Z.; Mu, X.; Dong, S.; Lu, Y.; Jiang, M. Constructing Adaptive Multi-Scale Feature via Transformer-Aware Patch for Occluded Person Re-Identification. *Symmetry* **2022**, *14*, 1454. https://doi.org/10.3390/sym14071454

Academic Editor: Mihai Postolache

Received: 21 June 2022
Accepted: 13 July 2022
Published: 15 July 2022

1. Introduction

With the rapid development of artificial intelligence [1–4], deep neural networks have enjoyed extensive attention in computer vision. Person Re-ID, as one of its important applications, has made remarkable achievements in recent years [5]. It has been widely applied in video surveillance analysis, criminal investigation, and solving crimes, which greatly improves the analysis efficiency and significantly reduces labor costs. Also, it provides sufficient technical support for public safety. However, most of the best approaches are designed based on holistic person Re-ID. This is a relatively ideal situation in the closed world; that is, the pedestrian video frame images used for retrieval all contain the complete human torso [6]. For the increasingly complex video surveillance system, the captured scene is often complex and changeable. In the public environment, it is common for the target pedestrian to be occluded by various objects, such as pillars, plants, railings, and even other unrelated pedestrians. This would pose a significant obstacle to the retrieval process, as the occluded parts of the person may contain decisive clues. In addition, if we still adopt the approach designed for the conventional, holistic person Re-ID, the noise of obstacles also interferes with the correct feature extraction process. Therefore, there is an urgent and pressing need to tackle such an occluded situation.

At first, to reduce the influence of occlusion, the researchers manually cropped the video frame images to exclude the occluded parts and then adopted the cropped images for retrieval in the gallery composed of holistic pedestrian images, which is the partial re-ID [7] method. However, cropping will not only increase a lot of labor costs, contrary to the original intention of convenience but also introduce additional artificial cropping bias, bringing new challenges to the retrieval process [8].

Recently, different from partial Re-ID, researchers have directly matched the unprocessed query images with the library of occluded images, which is the ubiquitous occluded Re-ID [9] in reality. This kind of Re-ID not only can avoid the shortcomings of the partial re-ID method, but the idea of not advanced cropping is also more consistent with the practical operation logic. Thereby, several occlusion-robust approaches are derived in this way, which can be summarized in chronological order as follows. Initially, researchers dealt with occlusion mainly based on the extension of the holistic Re-ID method. One mainstream strategy was image partitioning, which directly compared the part-level features in the raw image to improve the occlusion robustness. However, the positioning accuracy of this method is not high and the performance improvement is limited. To make further use of the idea of partitioning, a large number of methods introduce existing pose-estimation models to assist in locating pedestrians and find segmentation containing visible areas of pedestrians to match [8,10,11], thus improving the effective utilization of part features. In addition, HOReID [12] used landmarks generated by the pose estimation model combined with graph convolution to improve performance further. However, the granularity of part-level division is too coarse, which has limitations on division and discrimination of complex occlusion. Recently, there have been a lot of attention-based efforts [13,14] to solve the occluded problem. The segmentation accuracy of the attention mechanism is much higher than that of the hard partitioning strategy. This is because it can more accurately perceive the area of the pedestrian rather than other interferences, leading to purer extracted features than before and thus higher accuracy.

Inspired by [15], accurate positioning of distinguishing object regions is carried out to construct fine-scale feature representation and then combine original scale features to form multi-scale representation so that a more complete feature representation can be obtained. In this paper, we propose a novel Dual-Transformer structure to accurately capture multi-scale features of occluded pedestrians. In detail, because Transformer [16] has enjoyed great success in computer vision in recent years, the two branches are made up of the same Vision Transformer (ViT) [17] for extracting pedestrian features at different scales. To obtain the visible region cropping of the pedestrians, we first design Transformer-Aware Patch Searching (TAPS) by using the multiheaded self-attention mechanism of the Transformer to learn the general distribution of the visible region of the pedestrians and generate the distributed information attention branch. Then, the Adaptive Visible Part Cropping (AVPC) strategy is used to generate cropping images of the visible pedestrians' part as the input of the fine-scale branch, and only the weakly-supervised signals generated by ID labels are used to constrain the formation of distribution and the cropping quality. In the inference stage, the three symmetric branches, including global scale, cropped fine scale, and distributed information attention scale, are directly concatenated to represent pedestrian features at multiple scales. Our contribution can be summarized as the following three points:

- We propose an end-to-end Dual-Transformer symmetric multi-scale occluded person Re-ID structure, which contains three scales from coarse to fine into original scale, cropped fine scale, and distributed information attention scale.
- TAPS module was designed for occluded images to search for the initial distribution of pedestrians and to construct the branch of distributed information attention scale.
- The AVPC strategy only with weakly-supervised signals generated by ID labels is proposed to constrain the cropping of visible areas for pedestrians and simultaneously constraints TAPS to obtain accurate distribution information.

2. Related Work

With the development of person Re-ID, Holistic Person Re-ID has made a good achievement. However, in real scenes, a person's body information is often incomplete in the picture, so Person Re-ID in complex situations has attracted much more attention and more research is still needed on Partial Re-ID and Occluded Re-ID. Next, we will introduce three cases of Holistic Person Re-ID, Partial Person Re-ID, and Occluded Person Re-ID, respectively.

2.1. Conventional Person Re-ID

The Holistic Person Re-ID task is divided into two types: the traditional method based on the CNN backbone and the recently emerging method based on the Transformer backbone.

There are roughly three methods based on CNN. (1) Extracting fine-grained features of a person by local features, such as PCB [18], DSR [19], and VPM [20]. (2) Methods based on global features, mainly including MVP [21], SFT [22], and IANet [23]. (3) Through the method of additional information, mainly including P^2Net [24], PGFA [8], ISP [25], and HOReID [12]. Methods based on Vision Transformer mainly include PAT [13] and TransReID [26].

2.2. Partial Person Re-ID

In the Partial Re-ID task, the pedestrian's body information is incomplete because part of the pedestrian's body is out of the camera or there is a random lack of body information (for example, the picture of the upper body and the picture of the right half of the body), which may result in poor model transfer in the Holistic Person Re-ID task training and the misalignment of pedestrian information.

To solve these problems, Sun et al. proposed a framework DSR [19] combining deep feature learning and sparse reconstruction learning, which can directly extract feature information from the original size of images. Luo et al. proposed an end-to-end learning framework SCPNet [27] that utilizes locally guided global feature extraction. Han et al. proposed a feature matching framework KBFM [28] based on key points of attitude. Sun et al. proposed a framework VPM [20] for sharing local areas through self-supervised learning perception. Luo et al. proposed a framework combining spatial transformation network, STNReID [29], which combined Spatial Transformer Network (STN) [30] for the first time to solve the problem of Person Re-ID. He et al. proposed a self-supervised learning framework, PPCL [31], which achieves the matching between parts without additional local supervision.

2.3. Occluded Person Re-ID

The Occluded Re-ID task is more challenging than the Holistic Person Re-ID task, and the difficulties are reflected in the following two aspects. (1) Misalignment caused by random occlusion of parts and incomplete body information. (2) The occlusion is similar to the pedestrian, thus introducing noise of occlusion. Existing methods can be basically divided into three categories, which are hand-crafted splitting-based methods, methods using additional clues, and methods based on the Transformer.

Methods based on hand-crafted splitting handle the occlusion problem by measuring the similarity relationship of the aligned patches. Sun et al. proposed a network called Part-based Convolutional Baseline (PCB) [18], which can uniformly divide feature graphs and directly learn local features. Jia et al. proposed MoS [32] to use the Jaccard similarity coefficient between the corresponding partial sets to take the shaded person Re-ID as the set matching problem.

Some methods use additional information to locate body parts, such as segmentation, pose estimation, or body parsing. Song et al. proposed a mask-guided contrastive attention model [33] to learn features from the body separately. [8] introduced Pose-Guided Feature Alignment (PGFA) that uses attitude information to mine the recognition part. Gao et al. put forward a Pose-guided Visible Part Matching (PVPM) [10] model to learn features of discriminating parts with pose-guided attention. Wang et al. proposed HOReID [12] that introduces high-order relations and human topological information to learn robust features.

Recently, Transformer-based design methods have emerged, and Transformer has been proved to have strong feature extraction capability. Li et al. were the first to propose a Part Aware Transformer (PAT) [13] for occluded person Re-ID, which opened up research to explore the performance of Transformer on the partial and occluded person Re-ID. He et al. studied a pure ViT framework called TransReID [26], which combines camera perspective information, and proposed a JPM module to enhance ViT ability to extract local features.

3. A Transformer-Based Anti-Occlusion Person Re-ID Network

In this section, first, the overall structure of the proposed approach is introduced. Then, we briefly introduce the ViT process and our proposed Dual-Transformer multi-scale Re-ID architecture in Section 3.1. The TAPS and constraints of search are presented in Section 3.2. In Section 3.3, the fine-scale visible region cropping method AVPC is described in detail. Finally, we introduce the training and inference procedures in Section 3.4, including the setting of the loss function and multi-task loss.

Our proposed model contains two main branches based on ViT. As shown in Figure 1. The upper part of the figure contains two sub-forks. The upper sub-fork is used for global-scale feature extraction, and the lower sub-fork is used for visible area distribution estimation and attention-scale feature extraction. In the middle of the figure is the AVPC module, which is used to generate precise cropped images adaptively. The branches in the lower half of the figure are used to extract finer-grained crop-scale features.

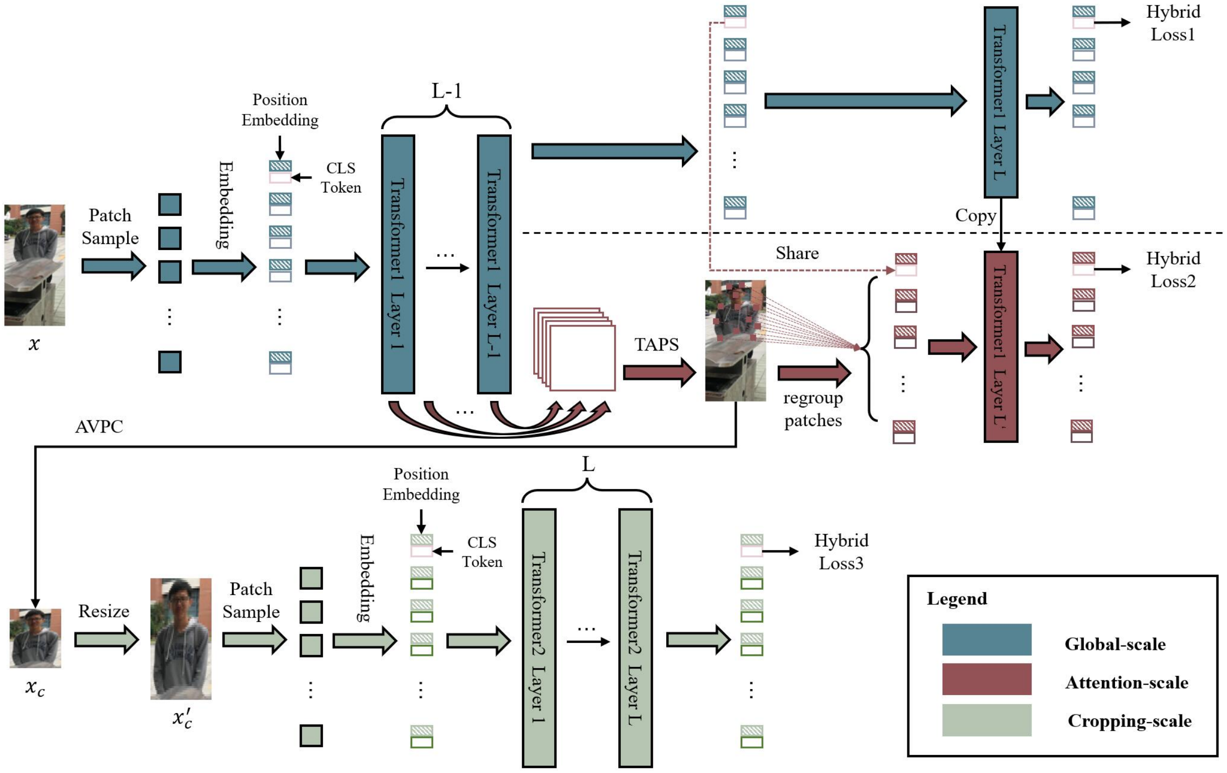

Figure 1. The Architecture of the proposed approach.

First, given an augmented image x, which is sent into the standard ViT branch above. After the $L-1$ layers encoding is completed, a part of the sub-fork in the upper side continues to complete the last encoding to generate global-scale features, and another part of the sub-fork in the lower side feeds the generated $L-1$ attention maps to the TAPS module to obtain the approximate visible area distribution. Then, on the one hand, the obtained distribution patches share the previous classification token to form a new encoding input to the last transformer encoding layer to obtain Attention-scale features.

On the other hand, accurate cropping of visible pedestrian regions x_c is generated by AVPC and fed into another standard ViT branch below to extract cropping-scale features.

3.1. Dual-Transformer Multi-Scale Re-ID Architecture

Transformer has achieved remarkable success in the Nature Language Processing (NLP) [34,35]. With its in-depth research in computer vision, Transformer-based image processing methods surpass traditional convolutional neural networks, such as in object detection [36], image segmentation [37–39]and other fields. ViT is a Transformer-based network structure commonly used for image classification and feature extraction [17,26].

For an ordinary pure ViT structure, as shown in Figure 2, given an input image $I \in R^{C\times H\times W}$, the first token output by the last encoding layer is generally used as the final output, which is the classification token. In detail, the image is firstly divided into multiple fixed-size patches $[I_x \mid x \in [1, N], x \in Z]$ by sampling. Then the different sampling patches are integrated into a vector through a linear transformation. A classification token and a learnable positional encoding are added, as shown in Equation (1):

$$I_x^{'} = E(I_x) + p_x \quad x \in [0, N], x \in Z \tag{1}$$

where $E(\cdot)$ refers to the process of a linear transformation, and p_x is the learnable position embedding for each patch.

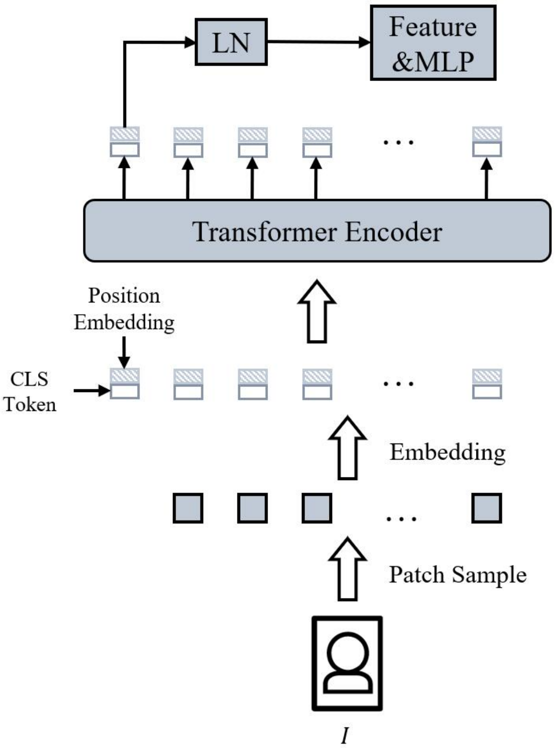

Figure 2. The overview of ViT.

The integrated matrix is sent to the Transformer encoder layer for encoding learning. Each encoder layer has the same structure, and a total of L iterations of learning operations are performed. Assume that the input vector matrix of the i-th layer is V_{i-1}, the output is V_i, and the output of the i-th layer is the input of the $(i+1)$-th layer. The operation of the i-th encoder layer is shown in Equations (2) and (3):

$$V_i' = MSA(LN(V_{i-1})) + V_{i-1} \tag{2}$$

$$V_i = MLP(LN(V_i')) + V_i' \tag{3}$$

Among them, V_i represents the output of the i-th layer, $LN(\cdot)$ represents the Layer Normalization, MSA represents the multiheaded self-attention mechanism operation, and MLP represents the full connection of the multilayer perceptron. These two sub-operations include a residual connection. The final output takes the first classification token and performs a layer normalization operation, as shown in Equation (4):

$$V_{Final} = LN\left(V_L^0\right) \tag{4}$$

The output vector can be directly used as the feature vector of the image and can also be used to calculate the classification loss for backpropagation.

It can be seen that the image features learned by ViT only contain one scale, which is too single for the person Re-ID with increasingly complex scenes, especially for the situation of occlusion that contains interfering objects.

The Dual-Transformer multi-scale Re-ID structure formed by combining two pure ViTs can make up for the above shortcomings. For each branch or sub-fork, the different scales of input information are used, leading to the final extracted pedestrian features being composed of multi-scale feature vectors. The ViT in our structure adopts the strategy of overlapping sampling; that is, the sampling step size s is smaller than the sampling size p so that more interactive information between patches can be obtained. We can calculate the number N of sampling patches through Equation (5):

$$N = \left\lfloor \frac{(H-p)}{s} + 1 \right\rfloor \times \left\lfloor \frac{(W-p)}{s} + 1 \right\rfloor \tag{5}$$

where H and W are the height and width of the input image, respectively.

3.2. Transformer-Aware Patch Searching

To accurately locate the visible area of pedestrians, the first step is to search for the approximate distribution of the visible part of the pedestrian in the original input image. To make better use of the multiheaded self-attention mechanism in Transformer, inspired by [40], it designs a method that can effectively capture the most discriminative regions of objects in an image. This is similar to our goal of seeing that the approximate distribution of pedestrians also contains some of the most discriminating regions in the person, such as heads, hands, etc. To obtain the distribution information of these regions, we first perform an aggregate operation on the attention matrix in the Transformer encoder layer. The attention matrix of each layer is calculated by Equation (6):

$$a_i = softmax\left(\frac{Q(v_{i-1}) \times K(v_{i-1})}{dim(Q(v_{i-1}))}\right) \tag{6}$$

Among them, $Q(\cdot)$, $K(\cdot)$ refers to the method of a linear transformation of Query and Key matrix in the Transformer coding layer, and $dim(\cdot)$ refers to the integration dimension of Query or Key matrix.

We aggregate the attention matrices of each layer using matrix multiplication, as shown in Equation (7):

$$A = \prod_{i=0}^{L-1} a_i \tag{7}$$

The aggregated matrix A is divided into H two-dimensional matrices $[A_i \mid i \in [1, H], i \in Z]$ according to the multiheaded self-attention mechanism, and H refers to the number of heads in multiheaded self-attention.

According to the principle of the multiheaded self-attention mechanism, different heads generally pay attention to a different pattern. Therefore, from the classification token of each head, we select the most active index in each head as the pedestrian's active attention points $[Att_i \mid i \in [1, H], i \in Z]$ and H attention points will form the final distribution. In order to better capture the character feature points, we only summarize the $L-1$ attention matrix before the last Transformer encoding layer, as shown in Equation (7).

Then we shuffle the order of the collected attention points to improve the robustness of the network. These points are stacked to form a new encoding matrix, which feeds the sub-fork, as shown to the right of the upper branch in Figure 1. For the two sub-fork parameters, the classification token parameters are shared, and the last Transformer encoder layer parameters are copied for independent calculation. The upper sub-fork continues to complete the global learning operation, and the lower sub-fork is used to learn the pedestrian distribution information. This branch of attention scale can construct the most fine-grained attention feature vector, and on the other hand, it is used for weakly supervised learning of pedestrian distribution. We only use the ID labels of the pedestrian to weakly supervise the learning of the distribution of the visible region so that the selected distribution points are mostly covered on the characters and not on other irrelevant regions.

3.3. Adaptive Visible Part Cropping

An adaptive visible part cropping is devised for precise building anti-occlusion cropping, as shown in Figure 3. The image on the left contains all the distribution point information, the image in the middle contains only four key distribution point patches, and the image on the right is the cropped image.

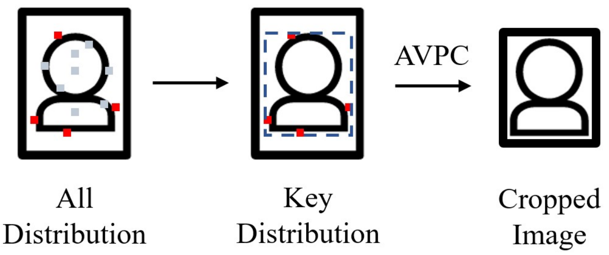

Figure 3. The process of cropping the visible pedestrian area.

After obtaining the estimated visible person distribution information, the image is cropped into a new fine-scale rectangular image along the edge of the H distribution point information, as shown in Figure 3. In all the distribution information, four key distribution points (the most top, bottom, left, and right points, respectively) are found to determine the positions of the four sides. This method of adaptive rectangular cropping according to distribution information is relatively simple and efficient and can largely exclude irrelevant image information around pedestrians. The cropped image is input to the second branch in the lower part of Figure 1, which is used to extract the cropping scale feature. The granularity of the cropping scale is between the global and the attention scale, which further refines and completes the scale space of pedestrians.

We supervise the learning of this branch only using ID labels. The supervision signal can not only constrain the current branch to learn better crop scale features but also generate weak supervision signals to constrain the learning of distribution information because our crop input is adaptively cropped according to the distribution information.

3.4. Training and Inference

During the training phase, the same combination of loss functions L_{Hybrid} to the final branch or sub-fork is applied. For each final output V_{Final}, batch-hard sampling triplet loss [41] and cross-entropy loss [42] are used as in general re-identification models. The BNNeck strategy introduced in Luo's article [43] is used for reference so that triplet loss and cross-entropy loss can converge simultaneously in different optimization spaces. The specific loss combination is shown in Equation (8):

$$L_{Hybrid-i} = L_{CE}(BN(V_{Final})) + L_{Tri}(V_{Final}) \quad i = 1,\ 2,\ 3 \tag{8}$$

Among them, $BN(\cdot)$ refers to Batch Normalization,L_{CE} refers to cross-entropy loss, and L_{Tri} refers to triplet loss.

Because the contribution of each branch or sub-fork to the model is different, to balance the proportion of their loss values in the whole model, we use a multi-task loss [44] to train our model, which is expressed in detail as:

$$L = \lambda_1 L_{Hybrid-1} + \lambda_2 L_{Hybrid-2} + \lambda_3 L_{Hybrid-3} \tag{9}$$

where λ_1, λ_2, λ_3 are hyperparameters that control multi-task loss, we need to ensure $\lambda_1 + \lambda_2 + \lambda_3 = 1$.

In the inference stage, we directly concatenate the feature vectors of the three scales as the final feature information of pedestrians, as shown in Equation (10):

$$V = [V_{Final-1}, V_{Final-2}, V_{Final-3}] \tag{10}$$

where $[\cdot]$ represents the concatenation operation of vectors.

4. Experiments

In this section, to verify the effectiveness of our proposed method, experiments are conducted on two occluded person re-ID datasets. In Section 4.1 we introduce the dataset and evaluation metrics used for the experiments. The experimental-related implementation settings and details are introduced in Section 4.2. In Section 4.3, the experimental results with recent SOTA methods are compared and analyzed. In Section 4.4, ablation experiments are carried out to verify the effectiveness of the proposed module.

4.1. Datasets and Evaluation Metrics

Our experiments involve three-person re-ID datasets. Details are shown below.

(1) Market-1501 [45] contains a total of 32,668 person images, including 19,732 gallery images, 3368 query images, and 12,936 training images, which are collected from 6 different cameras. Contains 751 IDs in the training pictures and 750 and IDs in the test pictures. This is a widely used large-scale conventional person re-ID dataset.

(2) Occluded-DukeMTMC [8] contains 15,618 training images, 17,661 gallery images, and 2210 query images. This dataset is a subset of DukeMTMC-reID, assembled by manual re-screening, specifically for Occluded person re-identification research.

(3) Occluded-REID [9] contains 2000 images of people with a total of 200 IDs captured by mobile devices. Each of these pedestrians contains 5 full images and 5 occlusion images.

About evaluation metrics, cumulative matching characteristic (CMC) and mean average precision (mAP) are adopted to validate the proposed approach. The experimental results we show include mAP, Rank-1, and a total of 2 indicators. All experiments use a single-query setting, and the post-processing methods like re-ranking are not used in our experiments.

4.2. Implementation Details

The ViT-B/16 model pre-trained on imagenet21k and fine-tuned on imagenet1k is employed as the backbone of our dual-branch network, and the ViT parameters in the two branches are independent and not shared. We resized the input and cropped the image to 256 × 128 and used random flipping and random erasing to enhance the image. The sampling step size s in ViT is set to 12, and the sampling size p is set to 16 so that it can achieve the overlapping sampling purpose. The batch and epoch size is set to 32 and 120, respectively. Following the settings in [26], the learning rate is set to 0.008, the warmup learning rate strategy and the cosine learning rate decay are used. The optimizer uses SGD, momentum is set to 0.9, and weight decay is set to 1×10^{-4} . The multi-task loss hyperparameters are set to $\lambda_1 : \lambda_2 : \lambda_3 = 1 : 1 : 2$. All experiments are deployed on Pytorch and finished on an Nvidia RTX 3090 for training and inference.

4.3. Performance on Occluded Person Re-ID Benchmarks

The results of our method and 7 previous works with backbone ViT are shown in Table 1 the first group is the recent SOTA methods, the second group is the results of the backbone ViT used in our model under different stride length settings, and the third group is our method. The best results are shown in bold. In both Occluded-DukeMTMC and Occluded-REID datasets, our model achieves state-of-the-art results in most of the mAP and Rank-1 metrics, which are 56.5% in mAP on Occluded-DukeMTMC, 81.6%/86.1% in mAP/Rank-1 on Occluded-REID, respectively. Compared with methods such as PGFA, PVPM, etc., which use existing pose estimation models to provide auxiliary information, the results of our method are significantly superior to them. Compared with the Transformer-based SOTA method PAT, our model also has advantages; it can be seen that our model completely suppresses it on Occluded-REID and is only 0.9% lower on the Rank-1 metric of Occluded-DukeMTMC. This is because the pedestrian features extracted by our method contain a variety of scales, which can well describe the subtle features of pedestrians. Compared with our adopted Backbone ViT-B, our method completely outperforms them regardless of whether the backbone uses the overlapping sampling strategy, demonstrating the effectiveness of our strategy for generating multi-scale features with a Dual-Transformer structure.

Table 1. Comparison with the recent SOTA methods on Occluded-Duke and Occluded-REID.

Approach	**Occluded-Duke**		**Occluded-REID**	
	mAp	**r-1**	**mAp**	**r-1**
PGFA [8]	37.3%	51.4%	56.2%	57.1%
FPR [46]	—	—	68.0%	78.3%
HOReID [12]	43.8%	55.1%	70.2%	80.3%
ISP [25]	46.3%	62.2%	—	—
PVPM+Aug [10]	37.7%	47.0%	61.2%	70.4%
DAReID [14]	53.2%	63.4%	—	—
PAT [13]	53.6%	**64.5%**	72.1%	81.6%
ViT-B/16 (s = 16)	53.6%	61.4%	77.9%	81.9%
ViT-B/16 (s = 12)	54.8%	61.4%	80.2%	84.2%
Ours	**56.5%**	63.6%	**81.6%**	**86.1%**

4.4. Alation Study

We conduct ablation experiments on the proposed model, as shown in Table 2. Among them, G-Scale, C-Scale, and A-Scale represent three different scale branches of global, cropping, and attention, respectively. Disruption indicates whether the input tokens of the Attention-scale branch are out of order. The ViT model presented in all experiments is based on ViT-B/16, and the step size is set to 12 for overlapping sampling. In experiments, two scales (global and cropping) are involved, and the multi-task loss ratio of the global scale and cropping scale are kept as 1:2. If the dual-branch structure finally contains three scales, the three-scale branches have a global, attention, and clipping ratio of 1:1:2.

Table 2. Ablation experiments of our method are performed using different modules.

Approach	**G-Scale**	**C-Scale**	**A-Scale**	**Disruption**	**Occluded-Duke**		**Occluded-REID**	
					mAp	**r-1**	**mAp**	**r-1**
ViT-B/16(s = 12)					54.8%	61.4%	80.2%	84.2%
Dual-Transformer	✓	✓			54.9%	62.4%	81.5%	85.8%
	✓	✓	✓		56.5%	63.4%	81.4%	85.2%
	✓	✓	✓	✓	**56.5%**	**63.6%**	**81.6%**	**86.1%**

A set of experiments are designed to compute only two scales (global, cropping) in our proposed Dual-Transformer model. The performance exceeds that of the independent ViT model, with 1–2% improvement in both two metrics in both Occluded-DukeMTMC and Occluded-REID datasets, which directly proves the effectiveness of the Dual-Transformer structure. When calculating the three scales(global, cropping, attention), 1–2% improvement in both metrics in the Occluded-DukeMTMC dataset can be achieved when the token order of the attention scales are not shuffled. But there is a slight fluctuation in the Occluded-REID dataset, showing a downward trend of less than 1%. To suppress the instability of the model after adding the attention scale, the tokens in the attention scale are shuffled to improve the robustness of the model. After adding the token disruption operation, the experimental results show that compared with the Dual-Transformer structure with only two scales (global, cropping), both indicators are improved, proving that the disruption operation can improve the stability of a small number of token branches.

5. Conclusions

In this paper, a novel Dual-Transformer structure for person Re-ID is proposed to settle the occluded issue, which can construct multi-scale feature descriptions of people. Extensive experiments are conducted on two occluded benchmarks to verify the effectiveness of our approach. The proposed model can achieve the SOTA effect except for Occluded-Duke (r-1). We argued that the proposed Transformer-Aware Patch Searching (TAPS) module could determine the approximate distribution information of visible pedestrian regions, and the Adaptive Visible-Part Cropping (AVPC) strategy can build non-occluded person image. However, the approach cannot deal well with multiple people in the original input image. In future research, we will continue to explore the working mechanism of the multi-headed self-attention, further improve the accuracy of cropping and the anti-interference ability of the model, design an effective mechanism to settle multiple person issues, and strengthen the robustness of the Dual-Transformer structure.

Author Contributions: Data curation, X.M.; Methodology, Z.L., X.M. and M.J.; Project administration, S.D. and Y.L.; Writing–original draft, X.M. and M.J.; Writing–review and editing, Z.L., X.M, S.D. and Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: The APC was funded partly by the Natural Science Foundation of Chongqing, China (Grant Nos. cstc2019jcyj-msxmX0487, cstc2021jcyj-msxmX0605). The APC was funded partly by the National Natural Science Foundation of China (Grant Nos. 61971078, 61501070). The APC was funded partly by the Science and Technology Foundation of Chongqing Education Commission (Grant Nos. KJQN202001137, CQUT20181124). The APC was funded partly by the Scientific Research Foundation of Chongqing University of Technology (2020ZDZ026).

Institutional Review Board Statement: The authors declare no conflict of interest.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The Natural Science Foundation of Chongqing, China (Grant Nos. cstc2019jcyj-msxmX0487, cstc2021jcyj-msxmX0605). The National Natural Science Foundation of China (Grant Nos. 61971078, 61501070). The Science and Technology Foundation of Chongqing Education Commission (Grant Nos. KJQN202001137, CQUT20181124). Scientific Research Foundation of Chongqing University of Technology (2020ZDZ026).

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

References

1. Xu, W.; Yuan, K.; Li, W.; Ding, W. An Emerging Fuzzy Feature Selection Method Using Composite Entropy-Based Uncertainty Measure and Data Distribution. *IEEE Trans. Emerg. Top. Comput. Intell.* **2022**, 1–13 . [CrossRef]
2. Xu, W.; Li, W. Granular computing approach to two-way learning based on formal concept analysis in fuzzy datasets. *IEEE Trans. Cybern.* **2014**, *46*, 366–379. [CrossRef] [PubMed]

3. Yuan, K.; Xu, W.; Li, W.; Ding, W. An incremental learning mechanism for object classification based on progressive fuzzy three-way concept. *Inf. Sci.* **2022**, *584*, 127–147. [CrossRef]
4. Xu, W.; Yuan, K.; Li, W. Dynamic updating approximations of local generalized multigranulation neighborhood rough set. *Appl. Intell.* **2022**, *52*, 9148–9173. [CrossRef]
5. Zheng, L.; Yang, Y.; Hauptmann, A.G. Person re-identification: Past, present and future. *arXiv* **2016**, arXiv:1610.02984.
6. 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]
7. Zheng, W.S.; Li, X.; Xiang, T.; Liao, S.; Lai, J.; Gong, S. Partial person re-identification. *Proc. IEEE Int. Conf. Comput. Vis.* **2015**, *2015*, 4678–4686.
8. Miao, J.; Wu, Y.; Liu, P.; DIng, Y.; Yang, Y. Pose-guided feature alignment for occluded person re-identification. *Proc. IEEE Int. Conf. Comput. Vis.* **2019**, *2019*, 542–551.
9. Zhuo, J.; Chen, Z.; Lai, J.; Wang, G. Occluded Person Re-Identification. *Proc. IEEE Int. Conf. Multimedia Expo.* **2018**, *2018*, 1–6.
10. 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.
11. Yang, Q.; Wang, P.; Fang, Z.; Lu, Q. Focus on the visible regions: Semantic-guided alignment model for occluded person re-identification article. *Sensors* **2020**, *20*, 4431. [CrossRef]
12. Guan'an, W.; 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. *Proc. IEEE Conf. Comput. Vis. Pattern Recog.* **2020**, *2020*, 6449–6458.
13. 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 (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 2898–2907.
14. Xu, Y.; Zhao, L.; Qin, F. Dual attention-based method for occluded person re-identification. *Knowl.-Based Syst.* **2021**, *212*, 106554. [CrossRef]
15. Hu, Y.; Jin, X.; Zhang, Y.; Hong, H.; Zhang, J.; He, Y.; Xue, H. RAMS-Trans: Recurrent Attention Multi-scale Transformer for Fine-grained Image Recognition. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event, China, 20–24 October 2021; pp. 4239–4248.
16. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
17. 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 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
18. Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; Wang, S. Beyond Part Models: Person Retrieval with Refined Part Pooling. *Eur. Conf. Comput. Vis.* **2017**, 1–17.
19. He, L.; Liang, J.; Li, H.; Sun, Z. Deep Spatial Feature Reconstruction for Partial Person Re-identification: Alignment-free Approach. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* **2018**, *2*, 7073–7082.
20. 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. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* **2019**, *2019*, 393–402.
21. Sun, H.; Chen, Z.; Yan, S.; Xu, L. Mvp matching: A maximum-value perfect matching for mining hard samples, with application to person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 6737–6747.
22. Luo, C.; Chen, Y.; Wang, N.; Zhang, Z. Spectral feature transformation for person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 4976–4985.
23. Hou, R.; Ma, B.; Chang, H.; Gu, X.; Shan, S.; Chen, X. Interaction-and-aggregation network for person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 9317–9326.
24. Guo, J.; Yuan, Y.; Huang, L.; Zhang, C.; Yao, J.G.; Han, K. Beyond human parts: Dual part-aligned representations for person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 3642–3651.
25. Zhu, K.; Guo, H.; Liu, Z.; Tang, M.; Wang, J. Identity-guided human semantic parsing for person re-identification. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part III 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 346–363.
26. He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; Jiang, W. Transreid: Transformer-based object re-identification. *arXiv* **2021**, arXiv:2102.04378.
27. 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 Computer Vision—ACCV 2018*; Jawahar, C.V., Li, H., Mori, G., Schindler, K., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 19–34.
28. Han, C.; Gao, C.; Sang, N. Keypoint-based feature matching for partial person re-identification. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 25–28 October 2020; pp. 226–230.

29. 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]
30. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial transformer networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 2017–2025
31. He, T.; Shen, X.; Huang, J.; Chen, Z.; Hua, X.S. Partial Person Re-identification with Part-Part Correspondence Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 9105–9115.
32. 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 Proceedings AAAI Conference Artificial Intelligence, Virtual, 2–9 February 2021; pp. 1673–1681.
33. 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.
34. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *Naacl Hlt 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol.Proc. Conf.* **2019**, *1*, 4171–4186.
35. Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q.V.; Salakhutdinov, R. Transformer-XL: Attentive language models beyond a fixed-length context. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019 ; pp. 2978–2988.
36. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
37. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv* **2021**, arXiv:2102.04306.
38. Xie, E.; Wang, W.; Wang, W.; Sun, P.; Xu, H.; Liang, D.; Luo, P. Trans2Seg: Transparent Object Segmentation with Transformer. *arXiv* **2021**, arXiv:2101.08461v2.
39. Yun, B.; Wang, Y.; Chen, J.; Wang, H.; Shen, W.; Li, Q. Spectr: Spectral transformer for hyperspectral pathology image segmentation. *arXiv* **2021**, arXiv:2103.03604.
40. He, J.; Chen, J.N.; Liu, S.; Kortylewski, A.; Yang, C.; Bai, Y.; Wang, C.; Yuille, A. TransFG: A Transformer Architecture for Fine-grained Recognition. *arXiv* **2021**, arXiv:2103.07976.
41. Hermans, A.; Beyer, L.; Leibe, B. In defense of the triplet loss for person re-identification. *arXiv* **2017**, arXiv:1703.07737.
42. Zheng, Z.; Zheng, L.; Yang, Y. A discriminatively learned cnn embedding for person reidentification. *ACM Trans. Multimed. Comput. Commun. Appl.* **2017**, *14*, 1–20. [CrossRef]
43. Luo, H.; Gu, Y.; Liao, X.; Lai, S.; Jiang, W. Bag of tricks and a strong baseline for deep person re-identification. *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.* **2019**, *2019*, 1487–1495.
44. Ruder, S. An overview of multi-task learning in deep neural networks. *arXiv* **2017**, arXiv:1706.05098.
45. Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; Tian, Q. Scalable Person Re-identification: A Benchmark. *Iccv* **2015**, 1116–1124.
46. 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, Korea, 27 October–2 November 2019; pp. 8450–8459.