Computer Vision and Image Understanding
journal homepage: www.elsevier.com

Learning transformer-based attention region with multiple scales for occluded person re-identification

Zhi Liu[a], Xingyu Mu[a], Yunhua Lu[a], Tingting Zhang[b], Yingli Tian[c,**]

[a]*School of Artificial Intelligence, Chongqing University of Technology, No.459 Pufu Ave., Yubei, Chongqing 401120, P.R. China*
[b]*College of Computer Science, Sichuan University, No.24 South Section 1, Yihuan Road, Chengdu, 610065, P.R. China*
[c]*Dept. of Electrical Engineering, City College of New York, 138th Street & Convent Ave., New York, NY 10031, USA*

ABSTRACT

Occluded person re-identification(Re-ID), with the aim of matching occluded person pairs under cross-camera, remains challenging due to incomplete information and spatial misalignment. The state-of-the-art (SOTA) methods usually include a two-stage architecture based on the existing pose estimation models or the attention mechanism to generate human masks to extract features, which complicate the model and introduce additional biases. To address this issue, we propose a novel end–to-end transformer-based occluded person Re-ID model. Specifically, our model contains two crucial components: (1) the features of global and non-occluded person regions are extracted by two independent Transformer-based feature extraction networks respectively; (2) the distribution of common non-occluded human regions is learnt via a multiheaded self-attention mechanism, and then the Minimized Character-box Proposal (MCP) is utilized to generate accurate shared non-occluded crops. In our model, non-occluded human regions are not annotated and only weakly-supervision of ID labels with multiheaded self-attention are employed to jointly learn the distribution. Further, the human feature contains multi-scale information which is extracted from our dual-branch architecture. Extensive experiment results on four benchmarks of person Re-ID for two tasks (occluded, partial) demonstrate the effectiveness of our proposed framework which achieves the SOTA or the comparable performance on all benchmarks.

Keywords: Person Re-identification; Transformer; Deep Learning; Multiple Scales

1. Introduction

Person Re-ID refers to the technology of retrieving a given person in a non-overlap multi-camera system(Zheng et al., 2016). With the fierce increase of social demand for public security, person Re-ID has made a breakthrough both in academic research and industrial applications, which has been widely applied in security systems, intelligence monitoring, criminal investigation and other fields(Ye et al., 2021). Recently, many excellent methods have been proposed to study the *Holistic* person Re-ID problem(Sun et al., 2017; Zhao et al., 2017; Wang et al., 2018; Fu et al., 2019) under a strong assumption of a complete torso must be visible. However, persons in real applications are frequently occluded by obstacles (such as vehicles, plants, pillars, other pedestrians, etc.)(Zheng et al., 2016). These pose a greater challenge to person Re-ID. Therefore, it is necessary to design an effective mechanism to solve the above-mentioned occluded person Re-ID problem.

To deal with the occluded issue, some preliminary studies are carried out based on Partial Re-ID. In order to reduce the interference of occluding obstacles in the feature extraction process, the occluded query images are manually cropped, and then the cropped images are used as new query images(Zheng et al., 2015b). However, these methods not only result in excessive workload and human bias but also cannot guarantee the completeness of all gallery images at all the time. Therefore, recent Occluded Re-ID methods have remedied the above shortcomings, that is all query images are occluded, and the person in gallery images are not complete, even partially occluded(Zhuo

[**]Corresponding author: Tel.:+1-212-650-5389; fax: +1-212-650-8249;
e-mail: ytian@ccny.cuny.edu (Yingli Tian)

et al., 2018). As a result, at least one occluded image exists when retrieving a person. This setting is more in line with the real world applications and attracts more research attention(Miao et al., 2019).

Recent research studies on occluded person Re-ID are mainly divided into three categories. (1) The hard partition strategy(He et al., 2018; Fan et al., 2019; Yang et al., 2021) directly divides the image into fixed patches through hand-craft partitioning, and then matches the person by comparing the patch-level features of the image pair. However, this method is unfavorable to be applied in challenging scenes because of coarse division and occluded noise information. (2) Additional semantic models are established for pose estimation(Miao et al., 2019; Gao et al., 2020; Guan'an et al., 2020; Yang et al., 2020), and semantic information is used to assist in locating visible human regions and learn accurate feature representations. However, this method heavily relies on a powerful semantic model, and the introduction of additional models cannot enable end-to-end training of the entire structure, even bring unconstrained deviations to the original task, and thus hindering further improvement of model performance. (3) Attention-based method(Sun et al., 2019; Xu et al., 2021; Li et al., 2021; He et al., 2021b; Wang et al., 2021) obtains the response map of person regions through the attention mechanism, in which the human body is given more weights than other unrelated regions. In summary, it can be seen that the core of the most existing methods is to locate the visible regions of persons and then construct saliency features to reduce the impact of occlusion on matching. The SOTA Part-aware Transformer (PAT)(Li et al., 2021) method based on Transformer(Vaswani et al., 2017) attention uses the Encoder and Decoder to locate a human of the interested-region. Transreid proves that the pure Vision Transformer (ViT)(Dosovitskiy et al., 2020) is advanced in the Re-ID field. In addition, the multiheaded self-attention mechanism(Vaswani et al., 2017) can better capture the characteristics of human, but the feature scale is single, leading to the poor expression.

In this paper, we propose a novel transformer-based occluded person Re-ID structure. Inspired from (Hu et al., 2021), combining imformation of different scales can construct more complete feature representation in image representation learning. Our architecture incorporates two similar ViT branches for constructing multi-scale features, and a common non-occluded region localization module for localizing common visible person regions. In the localization module, the multiheaded self-attention mechanism is utilized to search different modes so that it can find those person patches with subtle discrepancy and the most distinguishing ability from the original image as the initial distribution of humans in the Transformer layer. Then the fine-scale images are cropped by the proposed Minimized Character-box Proposal (MCP). We constrain the previous human patch search by only using weakly supervised signals generated base on ID labels. In the inference stage, the features at different scales generated by the two branches are directly concatenated as the final human representation. The main contributions of our work are summarized as follows:

- A novel transformer-based end-to-end architecture for occluded person Re-ID is proposed. This dual-branch structure can extract multi-scale features of humans and improve the expressiveness of features.
- A module for locating non-occluded regions of humans is designed, in which the multiheaded self-attention mechanism is employed to learn the distribution of visible human regions, and then non-occluded cropping is generated by MCP. In the locating process, only ID labels are used for weakly supervised learning to complete the preliminary distribution estimation instead of additional manual coordinate annotation.
- The proposed architecture is evaluated on two occluded datasets and two partial datasets. To our best knowledge, our model obtains the SOTA performance on OccludedREID, and also achieves competitive results on the remaining datasets.

2. Related Work

Most of the early research has focused on holistic person Re-ID, which aims to correctly match full-body person pairs. Existing methods can be summarized as hand-crafted feature representations and deep learning methods. In the traditional hand-crafted feature methods, the Bag-of-Words (BoW) model(Zheng et al., 2015a) proposed by Zheng et al. aggregates the 11-dimensional color name descriptors extracted by local patches into a global vector. In (Dong et al., 2018), Liao et al. proposed a Local Maximal Occurrence (LOMO) descriptor, which includes color and Scale Invariant Local Ternary Pattern (SILTP) histogram, and then uses Cross-view Quadratic Discriminant Analysis (XQDA) for scale learning to measure the similarity between images. Traditional hand-crafted feature has limited expression ability, resulting in the difficulty to adapt the tasks in large-scale and changing scenes. Deep networks have achieved great success in the field of computer vision due to their strong robustness, so many researchers also adopt them to address the person Re-ID problem. Sun et al. directly divided the original image horizontally into fixed-size blocks(Sun et al., 2017), and used ResNet(He et al., 2016) as the backbone to learn the block-level features for person feature representation. Wang et al. horizontally divided pedestrian images into different sizes(Wang et al., 2018), and learned multi-granularity feature information from the divisions of various sizes. Zhao et al. used human body keypoints to generate region proposals(Zhao et al., 2017) for different body parts and learned their features according to the proposals to form accurate pedestrian features. These methods have achieved remarkable results in holistic person Re-ID, but they are not suitable to be directly applied to occluded person Re-ID. This is because obstacles can cause interference and some crucial parts may be invisible due to occlusion.

Partial person Re-ID has been extensively studied to solve the occluded person Re-ID in the early time. He et al. proposed Deep Spatial feature Reconstruction (DSR)(He et al., 2018) to match fixed-size feature maps generated by full convolution network, which could alleviate the impact of the misalignment. Inspired by the PCB(Sun et al., 2017), Sun et al. proposed the Visible-aware Part Model (VPM)(Sun et al., 2019), which locates the visible regions of pedestrian images through weak supervision, and computes the similarity of person pairs by comparing shared block-level features. Luo et al. proposed STNReID(Luo et al., 2020), which utilizes Spatial Transformer Network (STN) to solve the misalignment problem in Partial Re-ID. However, partial Re-ID requires manually cropping the occluded query image, which not only introduce additional bias but also is cost-consuming. Therefore, this method is inappropriate for large-scale datasets and real-world applications.

Different from partial Re-ID, occluded person Re-ID directly matches occluded images. The Pose-Guided Feature Alignment (PGFA)(Miao et al., 2019) proposed by Miao et al. utilized the existing pose estimation model(Fang et al., 2017) to determine the visible person region, which is conducive to reduce the noise interference of obstacles, and then combined the idea of horizontal hard division to compare visible person blocks. Pose guided Visible Part Matching (PVPM)(Gao et al., 2020) learned discriminative person regions using pose-guided attention. HOReID(Guan'an et al., 2020) proposed by Wang et al. further improved the performance by utilizing person landmarks and graph convolution matching. The above methods all introduced existing pose estimation models to help find the key points of persons, which may cause bias and increase the complex of the model. Transformer(Vaswani et al., 2017) has achieved remarkable performance in NLP(Devlin et al., 2019; Dai et al., 2020), and then many researchers attempt to apply it to computer vision, and the results indicate that a large number of downstream visual tasks (such as object detection(Carion et al., 2020), semantic segmentation(Xie et al., 2021; Chen et al., 2021; Yun et al., 2021), object Re-ID(He et al., 2021b) and so on) have achieved excellent performance. For example, Li et al. designed the PAT(Li et al., 2021) using the transformer encoder and decoder to discover non-occluded human regions and obtain more accurate person features. Different from these methods, in order to improve the robustness of person features, we construct a two-branch network to extract multi-scale person features. Compared to the full transformer structure, our feature extraction branch constructed with pure ViT(Dosovitskiy et al., 2020) is simpler, which does not contain convolutional layers and only has encoder modules. Further, our method only use the person ID label to weakly supervise the human distribution learning instead of relying on an additional pose estimation model. Beyond that, our model can be trained end-to-end with higher efficiency.

3. Proposed Method

As shown in Figure 1, our framework includes two branches to extract features in multi-scales and a module for common non-occluded region localization. Specifically, the augmented image x is fed into the global scale branch on the left to learn global features, and then the attention map of its L encoding layers are used as the input of the common non-occluded region localization module, and the visible human region cropping is generated by MCP. Finally, the cropped and scaled human image x1 is input into the fine-scale branch to learn subtle features. Note that we only use ID labels to weakly supervise the human distribution learning. The batch-hard sampling triple loss and cross-entropy loss are calculated together for training the proposed architecture.

In this section, we first introduced the transformer-based multi-scales Re-ID framework and the robust feature extraction process in Sec. 3.1. Then, in order to accurately determine non-occluded human region and measure the distance of the image pairs, a concrete elaboration is carried out on the common non-occluded region locating in Sec. 3.2. Finally, the overall training and inference process is introduced in Sec. 3.3.

3.1. Transformer-based multi-scale Re-ID framework

To enrich the single-scale features learned by the pure ViT structure, we design a structure with two similar pure ViT branches as shown on the left and right sides of Figure 1, respectively. The left branch is used to extract global information and generate an index of the distribution of person's regions. The right branch can extract accurate fine-scale information to reduce the interference of irrelevant information, and simultaneously generates a weakly supervised signal to guide better learning of the distribution index. These two structures adopt independent parameter settings for learning different scale features and matching their respective tasks. The original image and the cropped image are sent to the feature extraction branches at different scales.

For the global scale branch, given an input image $x \in R^{H\times W\times C}$, where H, W, C are the height, width, and the number of channel, respectively. We utilize overlapping sampling to divide the input x to obtain a better local neighbor representation. Denote the sampling step as s, the sampled patch size as p, and the input image x is divided into N fixed-size patches $[x^i|i = 1, 2, \cdots, N]$, where N can be easily calculated by Eq. 1:

$$N = X_H \times X_W = \lfloor (H - P)/S + 1 \rfloor \times \lfloor (W - P)/S + 1 \rfloor \quad (1)$$

Where $\lfloor\cdot\rfloor$ is the lower integer bound of the result. X_H and X_W indicate the number of blocks on the height and width axis respectively. When s is less than p, the result of overlapping sampling can be obtained, and the smaller s is, the more overlap and the higher computational cost.

Following the setting of ViT, after linear projection of N patches, we add a classification token for capturing global information before the first patch, and then add a learnable position embedding E_p to maintain spatial information, and finally as input of the first transformer encoding layer:

$$Z_0 = \left[X_{CLS}, x^1E, x^2E, \cdots, x^NE\right] + E_P \quad (2)$$

Where $X_{CLS} \in R^{1\times D}$ represents the classification token, $E \in R^{P\times P\times C\times D}$represents the matrix that performs linear projection on the sampled $x^i(i \in [1, N])$, and $E_P \in R^{(N+1)\times D}$represents position embedding.

Denote the Transformer code contains L layers, each of which is composed of a multiheaded self-attention mechanism (MSA) and a multi-layer perceptron (MLP) module. LayerNorm (LN) is used before the MSA and MLP modules, and

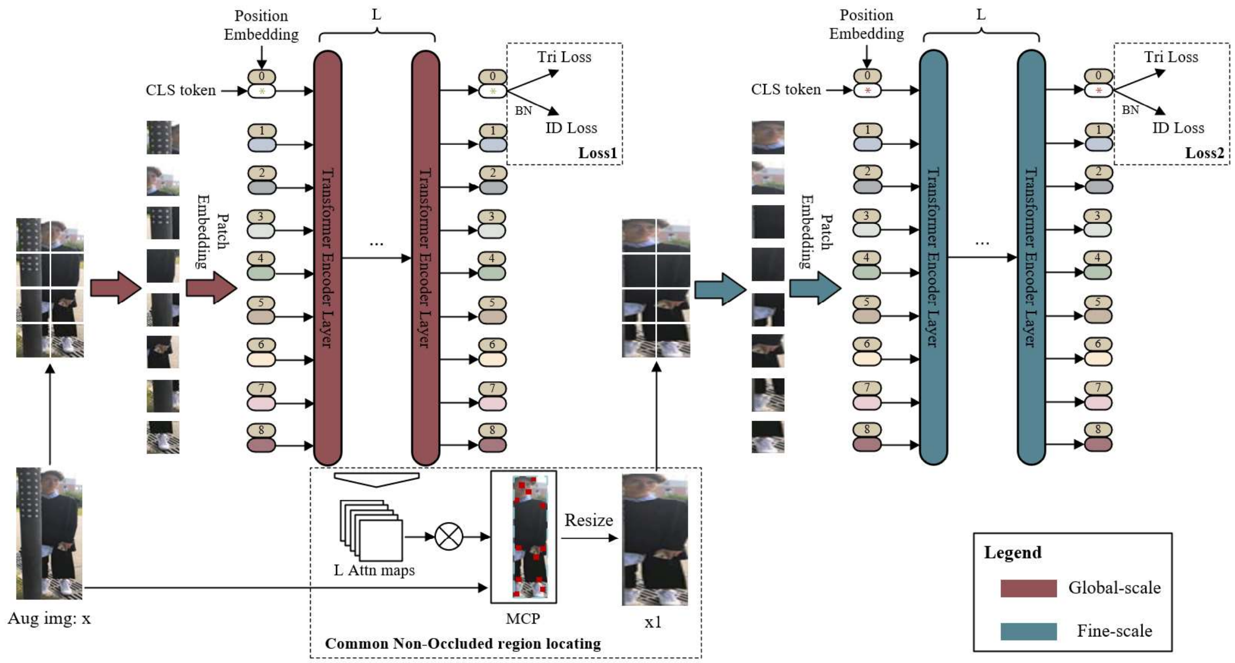

Fig. 1. The illustration of the proposed framework. The left global scale branch is used to extract the global features of human and generate more accurate human region cropping. The right fine-scale branch is used to extract more accurate human features after cropping. The common non-occluded region localization module at the bottom consists of three steps. First, we multiply the attention maps in the L Transformer coding layers to obtain the approximate pedestrian distribution, then use the MCP to generate accurate human cropping, and finally perform the resize operation. We directly concatenate the global classification token of each scale as the final human representation in inference stage.

residual connection are added to each module. Each layer consists of two stages, the input of the previous layer pass through the multiheaded self-attention mechanism, which is shown in Eq. 3. The output of the l-th Transformer coding layer is shown in Eq. 4, and the final output is shown in Eq. 5:

$$z_l{}' = z_{l-1} + MSA(LN(z_{l-1})) \qquad l \in 1, 2, \cdots, L \tag{3}$$

$$z_l = z_l{}' + MLP(LN(z_l{}')) \qquad l \in 1, 2, \cdots, L \tag{4}$$

$$F = LN(z_L) \tag{5}$$

For fine-scale branches, the input x_1 is cropped from x, and the rest of the process is consistent with the Eq. 1-5 in the original scale branch above.

In the pure ViT Re-ID model, only the first token z_l^0 in the final output F is used, and the triple loss and ID loss are combined to train the model. However, a single global information is unable to represent the occluded image, thus our model uses the global classification token of various scales and the attention map in the transformer encoding layer to compute the accurate common human region, which can generate input images x_1 of fine-scale, and then obtains multi-scale representations through two ViT branches.

3.2. *Common non-occluded region locating*

Although the pure ViT can be directly applied to conventional person Re-ID with a stunning impression, it cannot better distinguish the obstacles from the human body with an efficient mechanism. In this sub-section, we preliminarily determine the correct human body distribution by searching non-occluded interest patches, and a Minimized Character-box Proposal (MCP) is devised to construct an accurate shared non-occluded human body region. The details are as follows.

3.2.1. *Non-occluded interest patch search*

Obviously, finding the preliminary distribution of the human body is helpful to satisfy the purpose of determining an accurate non-occluded body region for extracting a finer scale feature. Besides, when ViT is used to solve visual problems, the utilization of local information in the intermediate transformer encoding layers is generally ignored. Inspired from TransFG(He et al., 2021a), by aggregating local information of different depths, the strongest attention blocks in ViT can be obtained. After training, these blocks are approximately distributed in the visible region of the human body. This method can roughly determine the distribution of the visible human body without introducing an additional pose estimation model. We denote the original attention matrix in the L Transformer coding layers as $[A_i | i = 1, 2, \cdots, L]$, where A_i can be easily obtained by Eq. 6:

$$A_i = softmax(\frac{Q_i K_i^T}{\sqrt{d}}) \tag{6}$$

Where Q_i,K_i refer to the query set and key set of the i-th layer, and d is the dimensions of Q and K.

The attention matrix in the model is divided according to different heads. Assuming that there are K self-attention heads in our model, each original attention matrix can be expressed as:

$$A_i = [A_i^1, A_i^2, \cdots, A_i^K] \qquad i = 1, 2, \cdots, L \tag{7}$$

Where $A_i^k \in R^{(N+1)\times(N+1)}$ represents the original matrix of each head, which contains amount of local interaction weights, and we aggregate the attention matrices of different layers by using matmul product:

$$A = \prod_{i=1}^{L} A_i \tag{8}$$

Matrix A contains the weight information in all layers, and is divided according to different heads. Then, the block with the largest response value from the global classification token in each head is selected as the human body distribution. The index of these blocks is denoted as $[I_j | j = 1, 2, \cdots, K]$, which are important basis for generating common non-occluded region.

3.2.2. Minimized Character-box Proposal

According to the obtained index distribution, we propose a very convenient strategy called Minimized Character-box Proposal (MCP) to generate the bounding box of the human body. This rectangular bounding box covers all the indexes mentioned above and has the smallest size. Using the number of blocks X_W in the width axis obtained by Eq. 1, we can find the boundary of the index value in the two dimensions of height and width. The equations are as follows:

$$M_H = \left\lfloor \frac{F(I_j)}{X_W} \right\rfloor \qquad j = 1, 2, \cdots, K \tag{9}$$

$$M_W = F(I_j \% X_W) \qquad j = 1, 2, \cdots, K \tag{10}$$

When $F(\cdot)$ represents the minimum and maximum functions, M_H and M_W are the lower and upper boundaries of the height and width dimensions, respectively.

Combining the overlap sampling step size s and the block size p, the cropped x_1 can be obtained through the original input x:

$$x_1 = x[m, n] \tag{11}$$

Where $m \in [M_{W-min} \times s, M_{W-max} \times s + p], n \in [M_{H-min} \times s, M_{H-max} \times s + p] \quad m, n \in Z$.

3.3. Training and Inference

The two branches in our model are composed of the same loss function. In each branch, the first global classification token of the final output F is taken, which is similar to the loss setting of the general Re-ID model. The batch-hard sampling triplet loss(Hermans et al., 2017) and cross-entropy loss(Zheng et al., 2017) are used to train each global token, and before using cross-entropy loss, the BNNeck strategy(Luo et al., 2019) is added to synchronize the convergence of the two losses. The loss of each branch can be expressed as:

$$L_i = L_{CE} + L_{Tri} \qquad i = 1, 2 \tag{12}$$

Since the two branches contain multi-scale feature learning, and the original scale branch also needs to learn the distribution information of the human body in the intermediate transformer encoding layer, multi-task loss(Ruder, 2017) for the two branches is used to control its contribution to the overall loss degree, specifically expressed as:

$$L = \lambda_1 Loss_1 + \lambda_2 Loss_2 \tag{13}$$

Where $Loss_1$ and $Loss_2$ are the original scale and fine scale loss respectively, and λ_1,λ_2 are the hyperparameter of the multi-task loss. In all of our experiments, $\lambda_1 + \lambda_2 = 1$ is guaranteed.

In the inference stage, we directly concatenate the global classification tokens in the two branches as a representation of person:

$$f = [F_1^0, F_2^0]. \tag{14}$$

Where $[\cdot]$ represents the concatenate operation of vectors.

4. Experiments

4.1. Datasets and Evaluation Metrics

To demonstrate the effectiveness of the proposed framework, we verify its performance on two occlusion datasets and two partial datasets: Occluded-Duke, Occluded-REID, Partial-REID, Partial-iLIDS.

(1) Occluded-Duke(Miao et al., 2019) is a subset of DuckMTMC-REID with the characteristic of the query set having all occluded images. It contains 15,618 training images, 17,661 gallery images and 2,210 occluded query images. It is one of the most challenging datasets in the occluded person Re-ID.

(2) Occluded-REID(Zhuo et al., 2018) is taken by a mobile device, which contains 2,000 pictures of 200 people. Each identity contains 5 full-body images and 5 occluded images. It also provides different capture angles and occlusion types.

(3) Partial-REID(Zheng et al., 2015b) is designed for partial pedestrian Re-ID. It contains 600 pictures of 60 people, each of which has 5 full-body images and 5 partial images.

(4) Partial-iLIDS(Zheng et al., 2011) is constructed based on the iLIDS and taken at the airport. It contains 238 pictures of 119 people, each of which has 5 full-body images and 5 partial images.

In the above four datasets, only Occluded-Duke completely contains the training set, gallery set and query set, and the other three datasets only contain the gallery set and the query set. For these three datasets, following the previous setting in person Re-ID(Miao et al., 2019; Gao et al., 2020; Sun et al., 2019; Li et al., 2021), we train our model on the Market-1501(Zheng et al., 2015a).

For evaluation metrics, we use cosine distance to measure the similarity between image pairs, and use cumulative matching characteristic (CMC) and mean Average Precision(mAP) to verify the quality of the results.

4.2. Implementation details

We employ ViT-B/16 pre-trained on imganet21k and fine-tuned on imagenet1k as the backbone of our two branches. The patch embedding stride in ViT uses overlapping sampling with stride of 12. All images are resized to 256 × 128. The training images are augmented with random horizontal flipping, padding 10 pixels, random cropping and random erasing(Zhong et al., 2020). The batch size is set to 32 with 4 images per person. In the training stage, all modules are jointly trained for 120 epoches. The set of optimizer and scheduler is referred to Transreid(He et al., 2021b). The SGD optimizer is deployed with momentum 0.9, weight decay 1e-4. The initial learning rate is set to 0.008 and decays at the cosine learning rate. For all datasets, we set $\lambda_2/\lambda_1 = 2$. All experiments are implemented on one Nvidia RTX 3090 GPU by using Pytorch.

Table 1. The comparison results of Rank-1 and mAP between the SOTA methods and ours on Occluded-Duke(O-Duke) and Occluded-REID(O-REID). The best performance is shown in bold

Method	O-Duke		O-REID	
	R-1	mAP	R-1	mAP
FPR(He et al., 2019)	-	-	78.3	68.0
PGFA(Miao et al., 2019)	51.4	37.3	57.1	56.2
PVPM(Gao et al., 2020)	47.0	37.7	70.4	61.2
GASM(He and Liu, 2020)	-	-	74.5	65.6
HOReID(Guan'an et al., 2020)	55.1	43.8	80.3	70.2
ISP(Zhu et al., 2020)	62.2	46.3	-	-
PAT(Li et al., 2021)	**64.5**	53.6	81.6	72.1
RFCnet(Hou et al., 2022)	63.9	54.5	-	-
RACNN(Fu et al., 2017)	41.8	32.5	52.3	44.3
ViT-B/16(He et al., 2021b)	61.4	54.8	84.2	80.2
Ours	62.4	**54.9**	**85.8**	**81.5**

4.3. Effectiveness on Occluded Re-ID datasets

Our model is compared with eight SOTA occluded Re-ID methods (top group in Table 1) and two relevant methods (middle group in Table 1) on two occluded Re-ID datasets, and the comparison results are shown in Table 1. The Rank-1/mAP of our model achieves 85.8%/81.5% on Occluded-REID, with an increase of 4.2%/9.4% compared with the SOTA PAT method. Further, 54.9% mAP on the Occluded-Duke is obtained, with an enhance of 0.4% than that of RFCnet. The Rank-1(62.4%) of ours is slightly lower than that of PAT(Li et al., 2021)(64.5%), this is because it can learn a set of part-aware masks, leading to learning better feature representations on some complex images in Occlude-Duke. However, our model and operating mechanism are simpler, which can show better performance in a balanced manner on different datasets. Therefore, when retrieving a given image, the correct ID in the result is near the front, so the mAP index is better. This shows that our model has better global optimality than PAT and RFCnet.

Compared with the SOTA methods that use pose estimation models such as PGFA(Miao et al., 2019), PVPM(Gao et al., 2020), and HOReID(Guan'an et al., 2020), the results of Rank-1/mAP of ours on Occluded-Duke and Occluded-REID have significantly improved by 7%-15%/11%-17% and 5%-28%/11%-25%, respectively. This is because our model is data-driven, which does not heavily rely on additional models to provide person cues, and thus can avoid deviations from disturbing the results. On the other hand, because of the robustness of the multiheaded self-attention mechanism to different occlusions and viewpoints, the distribution of the human region can be accurately estimated.

Experiments of two related methods are also carried out on each dataset. Recurrent Attention convolution Neural Network (RACNN)(Fu et al., 2017), which was first proposed for fine-grained classification, is a multi-branch CNN backbone based structure that can capture the multi-scale features of objects. The proposed structure has similarities with RACNN. Based on this, we designed an experiment on Re-ID of RACNN to evaluate the performance of the proposed two-branch structure. We use VGG19 as its backbone and initialize it with parameters pre-trained on Imagenet. The Attention Proposal Network (APN) is initialized with random parameters. Triplet loss and cross-entropy loss are used to train the network. The setting of hyperparameters such as optimizer and learning rate refers to Sec. 4.2. It is obvious that the proposed method has superior performance improvement over RACNN on both datasets, which proves that the proposed method is more advanced. In addition, we also implemented experiments on ViT-B/16 of the proposed approach. In the experiment, the pre-trained ViT-B/16 on Imagenet21k was used. The setting of hyperparameters refers to Sec. 4.2. As with the proposed method, overlapping sampling with step size of 12 is adopted. Our model also improves on both datasets, which proves the effectiveness of the proposed method.

Table 2. The comparison results of Rank-1 and Rank-3 between the SOTA methods and ours on Partial-REID(P-REID) and Partial-iLIDS(P-iLIDS). The best performance are shown in bold

Method	P-REID		P-iLIDS	
	R-1	R-3	R-1	R-3
PGFA(Miao et al., 2019)	68.0	80.0	69.1	80.9
FPR(He et al., 2019)	81.0	-	68.1	-
PVPM(Gao et al., 2020)	78.3	-	-	-
HOReID(Guan'an et al., 2020)	85.3	**91.0**	72.6	**86.4**
DSR(He et al., 2018)	43.0	60.3	54.6	64.5
VPM(Sun et al., 2019)	64.3	83.6	67.2	76.5
STNReID(Luo et al., 2020)	66.7	80.3	54.6	71.3
PPCL+(He et al., 2021c)	83.7	88.7	71.4	85.7
RACNN(Fu et al., 2017)	54.3	60.7	42.9	63.9
ViT-B/16(He et al., 2021b)	78.0	85.7	64.7	83.2
Ours	**86.0**	89.3	**73.9**	84.0

4.4. Effectiveness on Partial Re-ID datasets

To further verify our model, four occluded Re-ID methods (first group in Table 2), four partial Re-ID SOTA methods (second group in Table 2) and two relevant methods(third group in Table 2) are compared on two partial Re-ID datasets. The Rank-1/Rank-3 of our model achieves 86.0%/89.3% and 73.9%/84.0% on Partial-iLIDS and partial-REID, respectively.

Compared with the SOTA methods as shown in the table, we obtain the best Rank-1 results on both datasets with an increase of 0.7% and 1.3% than that of the most advanced SOTA method. In addition, Rank-3 results are also competitive.

Experiments on RACNN and ViT-B/16 are implemented on two occlusion datasets, and the experimental details are the same as in Section 4.2. Compared with RACNN, our results show significant improvement on both datasets. Compared with ViT-B/16, the improvement of Rank-1 on the two datasets is more obvious, which shows that this dual-branch structure can well construct human features and improve the probability of successful one-time matching.These results further verify the effectiveness and robustness of our proposed model.

5. Ablation studies and Visualization

In this section, we conduct extensive ablation study on Occluded-REID and Partial-REID to evaluate effectiveness on each component in Sec.5.1. To verify the effectiveness of multi-task loss and explore the contribution of different scales to human representation, we set different multi-task hyperparameters on Occluded-REID and Partial-REID to conduct extensive experiments in Sec. 5.2. The time complexity of the model is analyzed in Sec. 5.3. Finally, we visualize the experimental results to further demonstrate the working mechanism of the proposed model in Sec. 5.4.

Table 3. Comparison results of different module settings. In particular, 'CD' refers to the common non-occluded human distribution.

I	Backbone	CD	Proposal	O-REID		P-REID	
				R-1	mAP	R-1	R-3
1	ViT		w/o	80.2	84.2	78.0	85.7
2	Ours		Random	81.5	84.8	83.7	88.7
3		√	Center	79.7	85.4	76.2	82.4
4		√	MCP	81.5	85.8	86.0	89.3

5.1. *Effectiveness on each component*

Effectiveness of our two-branch model. A set of experiments using only one independent ViT are designed in index 1. And a set of experiments using randomly cropped proposals on our multi-scale Re-ID model are designed in Index 2. Taking the random cropping proposal as fine scales can minimize the impact on multi-scale model. From the comparison of indexes 1, 2 in Table 3, it can be found that the results of Rank-n and mAP of the multi-branch model on the two datasets are better than independent branch, demonstrating that this multi-scale feature extraction mechanism can more accurately construct person feature representations without using specially designed proposals, and validates the effectiveness and robustness of our proposed dual-branch structure.

Effectiveness of common non-occluded region locating. In the experiments of indexes 3 and 4, we utilize the preliminary human distribution information generated by the model. The detailed calculation method of these distribution information is described in Section 3.2.1. A center clipping strategy is set in index 3, that is, the center point of the distribution is selected and a proposal with height and width as H/2, W/2 is constructed respectively. Our proposed method is presented in index 4, which uses MCP to construct fine-scale input. In comparison between indexes 3 and 4, it can be obviously found that the Rank-n and mAP results of MCP mechanism are better than those of the center cropped ones. Moreover, it can be seen that using CD information to construct MCP has achieved the best results in Rank-n, mAP by comparing indexes 2 and 4. All these results not only prove the effectiveness of non-occluded distribution information, but also demonstrate the rationality of using non-occluded distribution information through the MCP mechanism. In addition, from indexes 1,3 and indexes 2,3 in Table 3, we could find that some results of center cropping are not only lower than the random cropping using the double-branch structure, but also lower than the single-branch model, suggesting that inappropriate proposal strategies may exert detrimental effect on the construction of multi-scale inputs, and even seriously damage the original robustness of the model.

5.2. *Hyper Parameter*

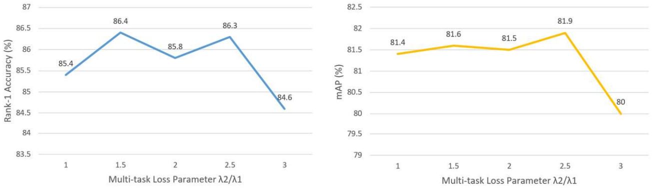

Fig. 2. Comparison of Rank-1 and mAP on Occluded-REID on the multi-tasking loss λ of different settings.

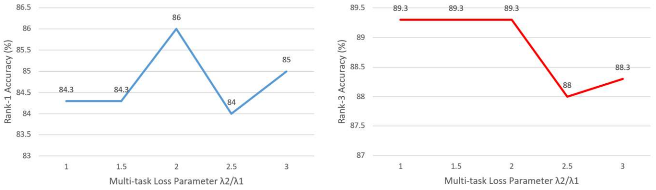

Fig. 3. Comparison of Rank-1 and Rank-3 on Partial-REID on the multi-tasking loss λ of different settings.

We chose different λ settings to observe the effect of multi-task loss on our model. The experimental results on Occluded-REID and Partial-REID are shown in Figure 2 and Figure 3, respectively. We can find that different λ settings can exert a maximum impact of about 2% on the final result, and the experimental results are better when the proportion of fine-scale branches is larger, indicating that the fine-scale branch contributes more than the global-scale one in our two-branch model. This not only shows a better representation of person features from cropped image, but also proves the effectiveness of our proposed method.

5.3. *Analysis of time complexity*

Table 2 records the amount of parameters, training and inference speeds in the ViT-B/16, RACNN and our proposed method

Table 4. Comparison of training and inference time on Market-1501

Model	Parameters	Train (images/s)	Inference (fps)
ViT-B/16	115M	193.2	67.36
RACNN	638M	41.2	14.71
Ours	174M	101.1	37.42

on Market-1501. Overlapping sampling with stride 12 is used for all ViTs in the ViT-B/16 and our method.

Compared with the ViT-B/16 and ours, our model has a larger number of parameters. So our training and inference are relatively slow. For large-scale Re-ID datasets such as Market-1501 (gallery: 19732 images)(Zheng et al., 2015a), our method also achieves more than 30fps when using an Nvidia RTX3090 for inference. For most surveillance systems, the inference speed and equipment cost are acceptable in order to obtain higher recognition quality.

Compared with RACNN, which is also a multi-branch structure, our method has smaller parameters and faster training and inference speed. Because RACNN is more suitable for capturing information of multiple different scales in fine-grained classification, and the feature information of two scales is sufficient when excluding the occlusion situation, RACNN has some disadvantages in pedestrian Re-ID. But this recurrent structure of RACNN inspired us to design a dual-branch structure based on ViT, which balances the time efficiency and improves the ability of Re-ID.

5.4. Visualization of human distribution

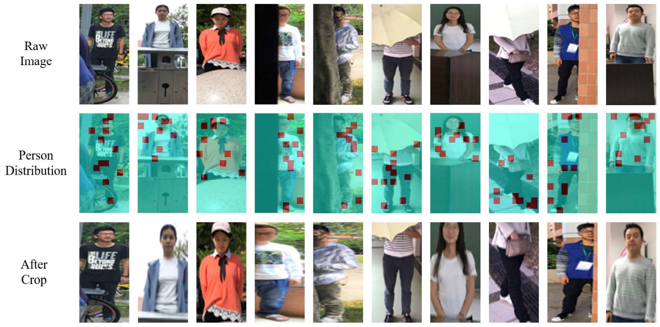

Fig. 4. The visualization of common non-occluded interest block search and MCP cropping. The top row is the raw image, the middle row is a visualization of the non-occluded region distribution of human, and the bottom row is the image after MCP cropping. The red patches in the distribution map are the human body that is highly concerned by our model. Best view in color.

The results of common non-occluded interest block search and MCP cropping is visualized as shown in Figure 4. It can be seen that for severe horizontal occlusion similar to the head and legs, or side occlusion in the vertical direction, our attention can still be well distributed around the non-occluded human body. Therefore, the image cropped according to this distribution hardly contains obstructing obstacles. In this way, when the cropped image is input into the model to extract multi-scale features, more accurate features of people rather than obstacles can be extracted.

6. Conclusion

To address the problem of occluded person Re-ID, a novel end-to-end two-branch model based on Transformer was proposed in this paper. Extensive experimental results of 4 datasets on the occlusion/partial Re-ID task demonstrate the effectiveness of the proposed model. In summary, we argued that proposed two-branch model can effectively extract multi-scale features of global and non-occluded person regions. And by weakly-supervision of ID label combination with multiheaded self-attention, the proposed MCP can generate accurate shared non-occluded crops by which the learned representation contributes more than the global one. More experiments also suggest that inappropriate proposal strategies may exert detrimental effect on the original robustness of the model. Future research should be devoted to simplifying the model, filtering outliers in the preliminary human distribution to obtain more advanced cropping.

References

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers, in: European Conference on Computer Vision, Springer. pp. 213–229.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y., 2021. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. arXiv preprint arXiv:2102.04306 .

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q.V., Salakhutdinov, R., 2020. Transformer-XL: Attentive language models beyond a fixed-length context. ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference , 2978–2988.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference 1, 4171–4186.

Dong, H., Lu, P., Zhong, S., Liu, C., Ji, Y., Gong, S., 2018. Person re-identification by enhanced local maximal occurrence representation and generalized similarity metric learning. Neurocomputing 307, 25–37.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 .

Fan, X., Luo, H., Zhang, X., He, L., Zhang, C., Jiang, W., 2019. SCPNet: Spatial-Channel Parallelism Network for Joint Holistic and Partial Person Re-identification, in: Jawahar, C.V., Li, H., Mori, G., Schindler, K. (Eds.), Computer Vision – ACCV 2018, Springer International Publishing, Cham. pp. 19–34.

Fang, H.S., Xie, S., Tai, Y.W., Lu, C., 2017. Rmpe: Regional multi-person pose estimation, in: Proceedings of the IEEE international conference on computer vision, pp. 2334–2343.

Fu, J., Zheng, H., Mei, T., 2017. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4438–4446.

Fu, Y., Wei, Y., Zhou, Y., Shi, H., Huang, G., Wang, X., Yao, Z., Huang, T., 2019. Horizontal pyramid matching for person re-identification, in: Proceedings of the AAAI conference on artificial intelligence, pp. 8295–8302.

Gao, S., Wang, J., Lu, H., Liu, Z., 2020. Pose-guided visible part matching for occluded person ReID, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11744–11752.

Guan'an, W., Yang, S., Liu, H., Wang, Z., Yang, Y., Wang, S., Yu, G., Zhou, E., Sun, J., 2020. High-order information matters: Learning relation and topology for occluded person re-identification. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition , 6448–6457.

He, J., Chen, J.N., Liu, S., Kortylewski, A., Yang, C., Bai, Y., Wang, C., Yuille, A., 2021a. TransFG: A Transformer Architecture for Fine-grained Recognition. arXiv preprint arXiv:2103.07976 .

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2016-Decem, 770–778.

He, L., Liang, J., Li, H., Sun, Z., 2018. Deep Spatial Feature Reconstruction for Partial Person Re-identification: Alignment-free Approach. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2, 7073–7082.

He, L., Liu, W., 2020. Guided saliency feature learning for person re-identification in crowded scenes, in: European Conference on Computer Vision, Springer. pp. 357–373.

He, L., Wang, Y., Liu, W., Zhao, H., Sun, Z., Feng, J., 2019. Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 8450–8459.

He, S., Luo, H., Wang, P., Wang, F., Li, H., Jiang, W., 2021b. Transreid: Transformer-based object re-identification. arXiv preprint arXiv:2102.04378 .

He, T., Shen, X., Huang, J., Chen, Z., Hua, X.S., 2021c. Partial Person Re-identification with Part-Part Correspondence Learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9105–9115.

Hermans, A., Beyer, L., Leibe, B., 2017. In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 .

Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., Chen, X., 2022. Feature completion for occluded person re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence 44, 4894–4912. doi:10.1109/TPAMI.2021.3079910.

Hu, Y., Jin, X., Zhang, Y., Hong, H., Zhang, J., He, Y., Xue, H., 2021. Rams-trans: Recurrent attention multi-scale transformer for fine-grained image recognition, in: Proceedings of the 29th ACM International Conference on Multimedia, pp. 4239–4248.

Li, Y., He, J., Zhang, T., Liu, X., Zhang, Y., Wu, F., 2021. 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), pp. 2898–2907.

Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W., 2019. Bag of tricks and a strong baseline for deep person re-identification. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops 2019-June, 1487–1495.

Luo, H., Jiang, W., Fan, X., Zhang, C., 2020. Stnreid: Deep convolutional networks with pairwise spatial transformer networks for partial person re-identification. IEEE Transactions on Multimedia 22, 2905–2913.

Miao, J., Wu, Y., Liu, P., DIng, Y., Yang, Y., 2019. Pose-guided feature alignment for occluded person re-identification. Proceedings of the IEEE International Conference on Computer Vision 2019-Octob, 542–551.

Ruder, S., 2017. An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098 .

Sun, Y., Xu, Q., Li, Y., Zhang, C., Li, Y., Wang, S., Sun, J., 2019. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2019-June, 393–402.

Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S., 2017. Beyond Part Models: Person Retrieval with Refined Part Pooling. European Conference on Computer Vision , 1–17.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need, in: Advances in neural information processing systems, pp. 5998–6008.

Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X., 2018. Learning discriminative features with multiple granularities for person re-identification. MM 2018 - Proceedings of the 2018 ACM Multimedia Conference , 274–282.

Wang, L., Zhou, Y., Sun, Y., Li, S., 2021. Occluded person re-identification based on differential attention siamese network. Applied Intelligence , 1–13.

Xie, E., Wang, W., Wang, W., Sun, P., Xu, H., Liang, D., Luo, P., 2021. Trans2Seg: Transparent Object Segmentation with Transformer. Onikle .

Xu, Y., Zhao, L., Qin, F., 2021. Dual attention-based method for occluded person re-identification. Knowledge-Based Systems 212, 106554.

Yang, J., Zhang, J., Yu, F., Jiang, X., Zhang, M., Sun, X., Chen, Y., Zheng, W.s., 2021. Learning to Know Where to See : A Visibility-Aware Approach for Occluded Person Re-identification. ICCV , 11885–11894.

Yang, Q., Wang, P., Fang, Z., Lu, Q., 2020. Focus on the visible regions: Semantic-guided alignment model for occluded person re-identification article. Sensors (Switzerland) 20, 1–15.

Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C., 2021. Deep learning for person re-identification: A survey and outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence .

Yun, B., Wang, Y., Chen, J., Wang, H., Shen, W., Li, Q., 2021. Spectr: Spectral transformer for hyperspectral pathology image segmentation. arXiv preprint arXiv:2103.03604 .

Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X., Tang, X., 2017. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 , 907–915.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q., 2015a. Scalable Person Re-identification : A Benchmark. Iccv , 1116–1124.

Zheng, L., Yang, Y., Hauptmann, A.G., 2016. Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984 .

Zheng, W.S., Gong, S., Xiang, T., 2011. Person re-identification by probabilistic relative distance comparison. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition , 649–656.

Zheng, W.S., Li, X., Xiang, T., Liao, S., Lai, J., Gong, S., 2015b. Partial person re-identification. Proceedings of the IEEE International Conference on Computer Vision 2015 Inter, 4678–4686.

Zheng, Z., Zheng, L., Yang, Y., 2017. A discriminatively learned cnn embedding for person reidentification. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 14, 1–20.

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

Zhu, K., Guo, H., Liu, Z., Tang, M., Wang, J., 2020. Identity-guided human semantic parsing for person re-identification, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16, Springer. pp. 346–363.

Zhuo, J., Chen, Z., Lai, J., Wang, G., 2018. Occluded Person Re-Identification, in: 2018 IEEE International Conference on Multimedia and Expo (ICME), IEEE. pp. 1–6.