Pose Estimation and Occlusion Augmentation Based Vision Transformer for Occluded Person Re-identification

***Yilin Wei*[1], *Dan Niu*[12*],*Hao Gong*[3], *Yichao Dong*[1] ,*Xisong Chen*[1,2] *Ziheng Xu*[4]**

[1]*School of Automation, Southeast University, Nanjing 210096, China*
[2]*Key Laboratory of Measurement and Control of CSE, Ministry of Education Research Laboratory, China*
[3]*CloudEdge (Xuzhou) Intelligent Technology Co.,Ltd, Xuzhou, China*
[4]*Jiangyin zhixing Intelligent Technology Co.,LTD, JiangYin, China*
**Email: 101011786@seu.edu.cn*

Keywords: occluded person re-identification, Vision Transformer

Abstract

Occluded person re-identification (ReID) is a challenging task as person images suffering from occlusion of various obstacles in real surveillance camera scene. Extracting partial feature from person image is crucial for occluded person ReID. In this paper, we propose the Pose Estimation and Occlusion Augmentation Based Vision Transformer (POVT) which leverage Pose Estimation Guided Vision Transformer (PEGVT) and an Occlusion Generation Module (OGM) to extract discriminative partial features. PEGVT divides the patch embeddings into different areas by pose estimation results and guide key point tokens to interact with corresponding patch embeddings to extract key point partial features. OGM can simultaneously generate realistic occlusion data which can improve the robustness of the ReID model, and occluded mask information which can supervise the finetune of pose estimation model to alleviate the performance degradation caused by domain gap. Experimental results over occluded re-identification datasets validate the effectiveness of the proposed POVT.

1 Introduction

Person re-identification (ReID) aims to associate a particular person across different cameras with different scenes and viewpoints [24]. With the development of deep learning, the ReID methods based on convolutional neural network have achieved great success on holistic person re-identification [2, 6, 12, 16, 19, 25]. However, due to the occlusion of other persons and non-person obstacles in real scenes, it is inevitable that some body of person is invisible and unavailable which brings great challenges to alignment and feature extraction [4, 21, 26, 30].

In occluded person re-identification, occluded regions bring noise to the extraction of global feature. Generally speaking, there are two ways to solve these problems, one is how to design data enhancement methods to simulate occlusion and another is how to extract robust and discriminative partial feature from visible part of person images [5, 7]. Some previous works assisted by human pose estimate model or human semantic segmentation model [8, 13, 18, 20] leverage the person key point heatmap or parsing information to guide feature learning. However, because the training of auxiliary models (human pose estimate model or human semantic segmentation model) and ReID model uses different data, there are domain gap which affects the performance of auxiliary models and finally reduce the performance of the ReID task.

Most of the previous methods to solve person re-identification tasks are based on convolutional neural network. Recently, Transformer-based ReID methods have shown remarkable performance both on holistic and occluded ReID tasks. TransReID [9] first employs vison transformer model in ReID task using Jigsaw Branch to extract partial feature. PAT [11] proposes part prototypes and part prototype based Transformer decoder to extract partial feature. AAformer [28] introduces part token and auto alignment to extract partial feature. The method based on Transformer can improve the ability to extract features. But because its global self-attention does not have the inductive bias (i.e. locality and translation invariance) of CNN, the spatial correspondence between the output feature map and the original map is not corresponding. And the previous methods based on CNN and auxiliary models process feature map with external model output which will misleads the feature extraction.

In this paper, we propose the Pose Estimation and Occlusion Augmentation Based Vision Transformer for occluded person re-identification as shown in Fig.1. First, we build a stronger vison transformer occluded person re-identification baseline which is helpful to verify the effectiveness of subsequent algorithms. Second, we propose Pose Estimation Guided Vision Transformer which extracts discriminative key point partial features. The patch embeddings of person image is first divided into different regions according to key points region division strategy using the key points heatmap of pose estimation results. Then we proposed key point tokens to interact with specific features from corresponding regions to extract key point partial features. Third, to simultaneously generate occlusion augmentation data and alleviate the performance degradation caused by domain gap of pose estimator, occlusion generation module is proposed to generate realistic occlusion data and occluded mask information which can supervise the learning of pose estimator.

The contributions of this work are summarized as follows:
1) We propose Pose Estimation Guided Vision Transformer to divide the patch embeddings into different areas by estimation results and guide proposed key point tokens to extract features.

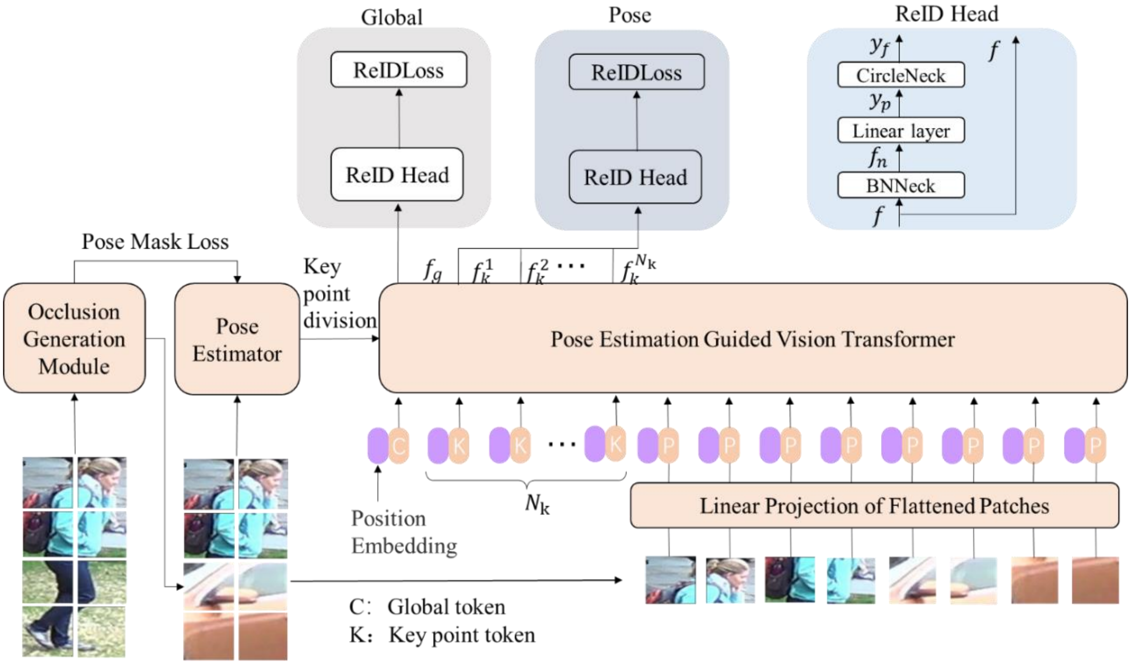

Figure 1 Framework of Pose Estimation and Occlusion Augmentation Based Vision Transformer

2) We propose Occlusion Generation Module to generate realistic occlusion data and supervise the finetune of pose estimation model to alleviate the performance degradation caused by domain gap of pose estimation simultaneously.

3) Adequate experiments show that our method brings significant improvement on occluded person re-identification.

2 Related works

2.1 *Holistic Person Re-Identification*

Person re-identification aims to retrieve images of person captured from different cameras. Existing ReID methods based on deep learning can be categorized to global feature representation [1,2,6] and partial feature representation [16,19,22]. Luo et al. [6] propose a simple but strong baseline for person re-identification only using the global features with BNNeck. Part-based methods achieved remarkable performance recently. Sun et al. [16] propose a Part-based Convolutional Baseline (PCB) which divides the pedestrian picture horizontally into several parts to extract the partial features and employs refined part pooling (RPP) to redistribute the outlier features in each block. Wang et al. [19] propose a Multiple Granularity Network (MGN) integrating discriminative information with various granularities. But these methods are based on the assumption that entire body of person is visible failing to associate person in the condition of occlusion.body.

2.2 *Occluded Person Re-Identification*

Occluded person re-identification aims to retrieve images of person existing occlusion in both query image and gallery image. The main challenge of occluded ReID is how to alleviate the influence of occlusion to extract discriminative global and partial features and how to use these features for matching. Previous methods can be divided to two categories methods assisted by auxiliary model [5, 10, 13, 18, 20] e.g. pose estimation or human parsing and adaptive feature extraction methods [11, 27]. He et al. propose a new method Pose-Guided Feature Alignment (PGFA) [13] which uses pose estimator to generate landmark of person to extract useful information with occlusion noise. However, the methods assisted with extra person-based model suffer from three problems first is performance degradation due to domains gap, second is neglect of pedestrian belongings and third is noise from other person occlusion. Recently, some adaptive feature extraction methods extract the discriminative features from occluded person images. Zhu et al. [27] propose an identity-guided human semantic parsing approach through the generation of pseudo-labels by human parts cascaded clustering. Li et al. [11] propose a Part-Aware Transformer via a transformer encoder-decoder architecture which employs part prototypes to extract partial feature.

2.3 *Vision Transformer for ReID*

Transformer [17] is proposed by Vaswani et al. to solve NLP tasks. Dosovitskiy propose pure Vision Transformer (ViT) [3] to solve vision tasks like image classification. TransReID [9] exploits the pure transformer for ReID tasks for the first time and proposes a jigsaw patches module (JPM) to facilitate perturbationinvariant and robust feature representation of objects. Auto-Aligned Transformer (AAformer) [28] introduce the “part tokens” for Transformer to learn partial features and harmoniously integrate the part alignment into the self-attention. However, the self-attention in TransReID and AAformer of the partial feature extraction is between class token or part token with a subset of patches which means it does not make full use of global information.

3 Method

This section explains Pose Estimation and Occlusion Augmentation Based Vision Transformer in this paper. Firstly, the Pose Estimation Guided Vision Transformer is introduced, and secondly the Occlusion Generation Module is introduced.

2.1 Pose Estimation Guided Vision Transformer

We first follow Vision Transformer (ViT) [3] to build our backbone and construct the main architecture to build a stronger baseline for person ReID. For Vision Transformer backbone, we employ patch overlapping to generate patches with overlapping pixels using a sliding window. Give an input image with a resolution H × W, with the patch size is P_s and step size is P_s, the number of output patches $N_p = \ N_h \times N_w$:

$$N_h \times N_w = \lfloor\frac{H+P_S-S_S}{S_S}\rfloor \times \lfloor\frac{W+P_S-S_S}{S_S}\rfloor$$

Then the input sequences are fed into Transformer encoder to get the output of CLS token which is regarded as global feature of whole person image denoted as f_g. The ReID head which consists of BNNeck and CircleNeck is applied after features. Specially, BNNeck adds a batch normalization (BN) layer and CircleNeck re-weights prediction classification logits under supervision to flexibly optimize feature learning with definite convergence target. The normalized feature f_n is acquired after f_g passing through batch normalization layer. Then a linear layer is used to map the f_n to the primary classification logits y_p = [y_p^1, y_p^2, y_p^3, ..., y_p^N,]. Then we employ CircleNeck to obtain the final prediction classification logits $y_f$and the y_f is expressed as:

$$y_f^i = \begin{cases} \gamma\alpha_p(y_p^i - \Delta_p) & if\ \hat{y}^i = 1 \\ \gamma\alpha_n(y_p^i - \Delta_n) & if\ \hat{y}^i = 0 \end{cases}$$

where y_f^i means prediction classification logits of ith category, $\alpha_n = [y_p^i + m]_+$ and $\alpha_p = [1 + m - y_p^i]_+$ are weighting factors for true and false classification, Δ_p, Δ_n and m are the margin, $\hat{y}^i$ is the label of classification.

Based on our stronger baseline, we proposed pose estimation guided vision transformer with key point tokens assisted with a pose estimation model to extract key point partial features as shown in Figure2. Specially, we use HR-Net [32] as our pose estimator to extract human pose information. The key point heatmap of pedestrian is extracted by pose estimator which contains the confidence scores of N_k human body key points. According to the pose estimation information, the patch embeddings of pedestrian image is spatially divided into N_k human body key point regions and one background region. For each pixel point in output key point heatmap, confidence scores corresponding to different human body key points are obtained. The heat map is resized to $N_h \times N_w$ through bilinear interpolation then the patch embeddings can be divided into different key point area. If the highest confidence of confidence scores is greater than the set threshold δ, divide the point into the key point area corresponding to the highest confidence, otherwise divide it into the background area. The key point token only calculates the multi-head self-attention with the patch embeddings in the corresponding key point area. During the calculation, the key point token and the block feature in the corresponding key point area are mapped to the Q, K, V spaces, and finally the human body key point feature f_k is obtained.

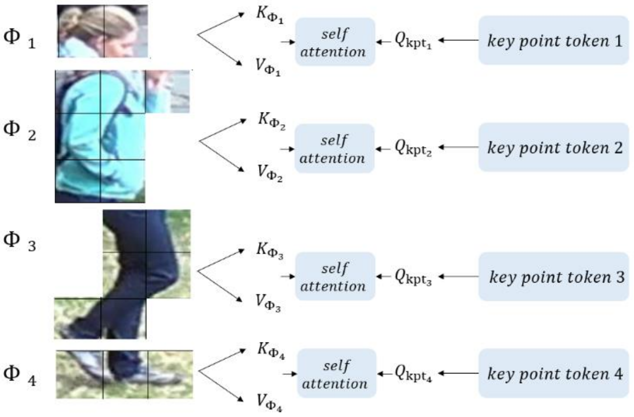

Figure 2 The pose estimation guided self-attention in PEGVT. Φ_i represent i-th human semitic region devided by pose estimation key point heatmaps.

The global feature f_g and key point partial features f_k with the number of N_k are fed into ReID head to get final logit $y_{f,g}$ and $y_{f,k}$. And the ReID loss consists of classification loss and triplet loss of global features and key point partial features, the ReID loss can be expressed as:

$$\mathcal{L}_{reid} = \lambda_{global}\mathcal{L}_{global} + \lambda_{pose}\mathcal{L}_{pose}$$

$$\mathcal{L}_{global} = \lambda_{cls}\mathcal{L}_{cls}(f_g) + \lambda_{tri}\mathcal{L}_{tri}(f_g)$$

$$\mathcal{L}_{pose} = \lambda_{cls}\sum_{i=1}^{N_k}\mathcal{L}_{cls}(f_{pose}^i) + \lambda_{tri}\sum_{i=1}^{N_k}\mathcal{L}_{tri}(f_{pose}^i)$$

where f_{pose}^i is i-th key point feature , $\mathcal{L}_{cls}$ is softmax cross entropy loss, λ_{tri} is triplet loss and λ_{global}, λ_{pose}, λ_{cls}, λ_{tri} are scale factor.

2.2 Occlusion Generation Module

Data augmentation strategy that can generate reasonable occluded person images is crucial for occluded person Re-ID. Random erasing strategy replaces specific areas with random pixel values so it is unable to produce real obstructions. Random patch strategy replaces specific areas of the person image with patches obtained from other images but it did not consider the relative relationship between occlusion patch and image because the region position for collection and replacement are both random. For example, the lower region may be replaced by the patch of sky or head of other person which is impossible to appear in real scenes. So, we propose an occlusion-oriented data augmentation strategy, a novel and elaborate strategy for patch collection and replacement that produces occlusion data with reasonable relative position relationship between occlusion patch and person image.

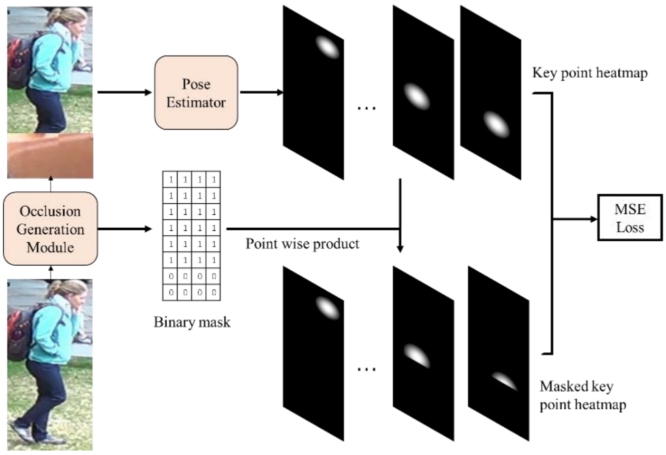

Figure 3 The calculation process of pose-mask loss in OGM.

Empirically, the occlusion objects of left-right positional relationship can be exchanged but not for up-down. So, in process of patch collection, we divide the occlusion patch into upper, middle and lower three parts according to the positional relationship between patch and person image. The occlusion patch set composed of three subsets can be represented as $S = \{\mathcal{S}_u, \mathcal{S}_m, \mathcal{S}_l\}$. Specially, we first select rectangular patch from person image same with random patch strategy. Then if the bottom edge of patch is higher than the vertical center-line of person image it is categorized to $\mathcal{S}_u$. The patch with the top edge lower than the vertical center-line is categorized into $\mathcal{S}_l$. And others are categorized to Sm. Then in process of patch replacement process, a patch is randomly selected from S with its subset label. Because different images are not exactly aligned, there is no need for accurate one-to-one correspondence in replacement. Give an image with height H and width W and a patch p with height H_0 and width W_0. Taking the upper left corner of the image as the coordinate origin, the vertical position coordinates (x_u, x_l) of the upper and lower edges of p can expressed as:

$$x_u \in \begin{cases} (0, min(H_0 - H_1, \frac{1}{2}H_0)) & if \;\; p \in \mathcal{S}_u \\ (0, H_0 - H_1) & if \;\; p \in \mathcal{S}_m \\ (max(H_1, \frac{1}{2}H_0 - H_1, H_0 - H_1)) & if \;\; p \in \mathcal{S}_l \end{cases}$$

To alleviate the performance degradation caused by domain gap, we proposed pose-mask loss as shown in Figure.3. According to the position of patch p, a binary occlusion mask can be obtained representing that pixel is masked or not. We obtain the key point heatmap $\mathrm{X}_{\mathrm{heatmap}}$of person image after occlusion-oriented data augmentation. Then $\mathrm{X}_{\mathrm{mask}}$ is obtained by performing an element-wise operation on the binary mask and $\mathrm{X}_{\mathrm{heatmap}}$, where the binary mask is obtained by resizing binary occlusion mask on the image to the same dimension as the key point heatmap. Finally, $\mathrm{X}_{\mathrm{mask}}$ supervises $\mathrm{X}_{\mathrm{heatmap}}$ by the pose-mask loss, so that the performance degradation caused by domain gap of pose estimation can be alleviate. The pose-mask loss can be formulated as:

$$\mathcal{L}_{pose\_maxk} = \sum_{i=0}^{N_k} MSE(\mathrm{X}_{\mathrm{heatmap}}^i, \mathrm{X}_{\mathrm{mask}}^i)$$

where $\mathrm{X}_{\mathrm{heatmap}}^i$ and $\mathrm{X}_{\mathrm{mask}}^i$represents i-th key point heatmap and masked heatmap.

3.4 *Training and Inference*

The objective function of our proposed methods consist of ReID loss for global feature and key point partial features. Besides, pose mask loss is employed. The loss for ReID model is $\mathcal{L}_{reid}$, and the loss for pose estimator to finetune is $\mathcal{L}_{pose\_maxk}$.

The global feature and key point partial features are concatenated to form the final representation of the person image. The distance between a query x_q image and a gallery image x_g can be expressed as:

$$d(x_q, x_g) = d(f_{x_q,g}, f_{x_g,g}) + \sum_{i=1}^{N_k} s_k^i * d(f_{x_q,k}^i, f_{x_g,k}^i)$$

where $f_{x_q,g}$ represents for global feature for x_q, $f_{x_q,k}^i$ represents for i-th key point feature for x_q and s_k^i represents confidence score of i-th key point.

4 Experiments

4.1 Dataset

Occluded-Duke [13] contains 36,441 images of 1,404 identities captured by 8 cameras. This dataset is derived from the DukeMTMC-ReID dataset where all query images are occluded by various occlusions (such as trees, cars, others) and the gallery image contains both the overall image and the occluded image.

Occluded-REID [29] contains 2,000 images of 200 identities. Each identity consists of 5 full body images and 5 occlusion images with different viewpoints and different types of severe occlusions.

Partial-REID [26] contains 900 images of 60 identities. This dataset is a specially designed partial person Re-ID benchmark where each identity five fullbody images in gallery set and five partial images in query set per person.

Following common practices, we adopt cumulative matching characteristic (CMC) curves at Rank-1 (R-1) and mean average precision (mAP) , to evaluate the performance of different Re-ID models.

4.2 Implementation Details

We adopt ViT-Base as basic backbone of our method and it is pre-trained on ImageNet21K and finetuned on ImageNet-1K. There are 12 Transformer layers where dimension of feature embedding is 768. All input person images are resize to 256×128. Patch size is 16×16 and overlapping step size is 14. For weights of loss function, we set $\lambda_{global} = 0.5$, $\lambda_{pose} = 0.5$, $\lambda_{cls} = 1$, $\lambda_{tri} = 1$, $\delta = 0.5$. And for circle neck, the set of γ is 64, m, Δ_p and Δ_n are 0.35. Besides occlusion-oriented data augment strategy, the data augment is conducted including flipping, padding and random cropping. The batch size is 64 with 16 images per identity. We use SGD optimizer with a momentum of 0.9 and weight decay of 0.0001. The Initial value of learning rate is 0.008 with cosine learning rate decay for ReID model. During the training stage, the ReID modules are trained for 120 epochs. And we finetune the pose estimation model with learning rate is 0.0001 and 5 epochs. All our methods are implemented on PyTorch 1.8.1 and we use 1 NVIDIA 3090 GPU for training.

4.3 Results on Occluded Re-ID Datasets

The performance of our methods compared with previous state-of-the-art methods on occluded and partial datasets is shown in Table 1. Our method achieves a new state-of-the-art performance 70.0% on Rank-1 and 60.4 % on mAP surpassing others by at least 2.1% and 1.2% on Occluded-Duke datasets. Because Occluded-REID dataset and Partial-REID dataset do not have corresponding training sets, following the common methods [11, 15], we simply train the model on Market-1501 for test. Our method make great achievements, the Rank-1 and mAP on Occluded-REID and Partial-REID are 87.2%/81.8% and 88.4%/83.2% respectively.

Table 1 Comparison with state-of-the-art methods on Occluded-Duke Datasets

Methods	Occluded-Duke	
	Rank-1	mAP
PCB(ECCV2018)	42.6	33.7
DSR(CVPR2018)	40.8	30.4
PGFA(ICCV2019)	51.4	37.3
HOReID(CVPR2020)	55.1	43.8
ISP(ICME2021)	62.8	52.3
PAT(CVPR2021)	64.5	53.6
TransReID(ICCV2021)	66.4	59.2
AAformer(AAAI2022)	67.0	58.2
FED(CVPR2022)	67.9	56.3
POVT(ours)	**70.0**	**60.4**

Table 2 Comparison with state-of-the-art methods on Occluded-REID and Partial-REID Datasets

Methods	Occluded-REID		Partial-REID	
	Rank1	mAP	Rank1	mAP
PCB(ECCV2018)	41.3	38.9	66.3	63.8
DSR(CVPR2018)	72.8	62.8	73.7	68.0
FPR(ICCV2019)	78.3	68.0	81.0	76.6
PGFA(ICCV2019)	-	-	69.0	61.5
HOReID(CVPR2020)	80.3	70.2	85.3	-
PAT(CVPR2021)	81.6	72.1	88.0	-
TransReID(ICCV2021)	82.1	78.0	82.3	79.4
FED(CVPR2022)	87.0	79.4	84.6	82.3
POVT(ours)	**87.2**	**81.8**	**88.4**	**88.4**

4.4 Ablation Studies

The ablation studies are performed on Occluded-Duke dataset to analyse the effectiveness of each component of our method including our stronger baseline, pose estimation guided vision transformer and occlusion generation module. The results in Table 3 are shown that stronger baseline promotes +5.5%/+3.5% on Rank-1 and mAP. The proposed PEGVT improves +3.4%/2.6% respectively. And OGM enhance Rank-1 and mAP by +1.1%/+1.2%.

Table 3 The effectiveness of each component of our method

Methods	Occluded-Duke	
	Rank-1	mAP
Vit Baseline	60.5	53.1
+ Our stronger baseline	65.5	56.6
+ PEGVT	68.9	59.2
+ OGM	70.0	60.4

4.4 Visualization of POVT

Last but not the least, we conduct the visualization experiments to show the focusing areas our methods. The results in Fig. 4 shows that key point tokens can focus on corresponding human sematic regions of person image.

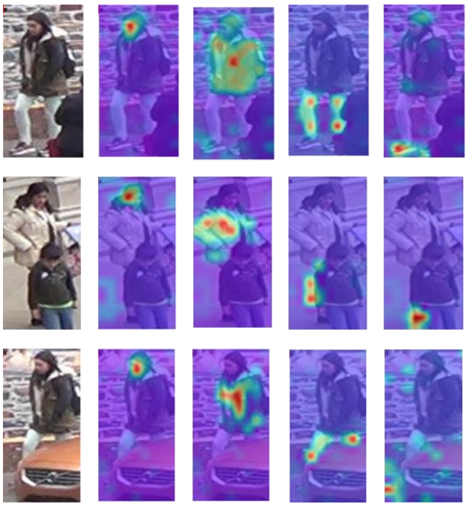

Figure 4 Visualization of the attention map of key point tokens.

5 Conclusion

In this paper, we propose the Pose Estimation and Occlusion Augmentation Based Vision Transformer to solve occluded person ReID tasks. Based on our stronger baseline, the pose estimation guided vision transformer can extract key point partial features from discriminative human semantic regions with the assist of pose estimation model. And the occlusion generation module can generate realistic occlusion data which can improve the robustness of the ReID model and weaken the performance degradation of the pose estimation model caused by gap at the same time. Comprehensive experiments on occluded datasets demonstrate the superiority of the proposed POVT.

6 References

[1] Y. Chen, X. Zhu, S. Gong, and IEEE. Person re-identification by deep learning multi-scale representations. In 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), 2017.

[2] Z. Dai, M. Chen, X. Gu, S. Zhu, and P. Tan. Batch dropblock network for person re-identification and beyond. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. 2020.

[4] Xing Fan, Hao Luo, Xuan Zhang, Lingxiao He, Chi Zhang, and Wei Jiang. Scpnet: Spatial-channel parallelism network for joint holistic and partial person reidentification. In Asian Conference on Computer Vision, 2018.

[5] S. Gao, J. Wang, H. Lu, and Z. Liu. Pose-guided visible part matching for occluded person reid. IEEE, 2020.

[6] L. Hao. Bags of tricks and a strong baseline for deep person re-identification. IEEE, 2019.

[7] L. He, J. Liang, H. Li, and Z. Sun. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[8] L. He, Y. Wang, W. Liu, X. Liao, H. Zhao, Z. Sun, and J. Feng. Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification. IEEE, 2019.

[9] S. He, H. Luo, P. Wang, F. Wang, and W. Jiang. Transreid: Transformer-based object re-identification. 2021.

[10] H. Huang, X. Chen, and K. Huang. Human parsing based alignment with multitask learning for occluded person re-identification. In 2020 IEEE International Conference on Multimedia and Expo (ICME), 2020.

[11] Y. Li, J He, T. Zhang, X. Liu, Y. Zhang, and F. Wu. Diverse part discovery: Occluded person re-identification with part-aware transformer. 2021.

[12] S. Liao and S. Z. Li. Efficient psd constrained asymmetric metric learning for person re-identification. In 2015 IEEE International Conference on Computer Vision (ICCV), 2015.

[13] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang. Pose-guided feature alignment for occluded person re-identification. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[14] E. Ristani, F. Solera, R. S. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. European Conference on Computer Vision, 2016.

[15] Y. Sun, Q. Xu, Y. Li, C. Zhang, Y. Li, S. Wang, and J. Sun. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[16] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang. Beyond part models: Person retrieval with refined part pooling. In European Conference on Computer Vision, 2017.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In arXiv, 2017.

[18] G. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, E. Zhou, and J. Sun. High-order information matters: Learning relation and topology for occluded person re-identification. 2020.

[19] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In 2018 ACM Multimedia Conference, 2018.

[20] T. Wang, H. Liu, P. Song, T. Guo, and W. Shi. Pose-guided feature disentangling for occluded person re-identification based on transformer. 2021.

[21] Z. Wang, F. Zhu, S. Tang, R. Zhao, L. He, and J. Song. Feature erasing and diffusion network for occluded person re-identification. 2021.

[22] Z. Xuan, L. Hao, F. Xing, W. Xiang, and S. Jian. Alignedreid: Surpassing humanlevel performance in person re-identification. 2017.

[23] L. Zheng, L. Shen, T. Lu, S. Wang, and T. Qi. Scalable person re-identification: A benchmark. In 2015 IEEE International Conference on Computer Vision (ICCV), 2015.

[24] L. Zheng, Y. Yang, and A. G. Hauptmann. Person re-identification: Past, present and future. 2016.

[25] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[26] W. S. Zheng, L. Xiang, X. Tao, S. Liao, J. Lai, and S. Gong. Partial person re-identification. In IEEE International Conference on Computer Vision, 2016.

[27] K. Zhu, H. Guo, Z. Liu, M. Tang, and J. Wang. Identity-guided human semantic parsing for person re-identification. 2020.

[28] K. Zhu, H. Guo, S. Zhang, Y. Wang, and M. Tang. Aaformer: Auto-aligned transformer for person re-identification. 2021.

[29] J. Zhuo, Z. Chen, J. Lai, and G. Wang. Occluded person re-identification. 2018 IEEE International Conference on Multimedia and Expo (ICME), 2018.

[30] J. Zhuo, J. Lai, and P. Chen. A novel teacher-student learning framework for occluded person re-identification. 2019

[31] Cheng B , Xiao B , Wang J , et al. HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation[C]// 2019.