Pose-guided Human Feature Aggregation for Occluded Person Re-identification

Zhe Zhang[1,2], Zongwen Bai[1,2,*],Meili Zhou[1,2]
[1]*Shaanxi Key Laboratory of Intelligent Processing for Big Energy Data, School of Physics and Electronic Information, Yan'an University, Yan'an, Shaanxi, China.*
[2]*School of Physics and Electronic Information, Yan'an University, Yan'an, Shaanxi, China*
**Corresponding Author*

Abstract: Since the appearance of most pedestrians is often obscured by various obstacles. Some existing works solve the occlusion problem by aligning the query image of the target pedestrian with the body part of the gallery image, but the body structure of the pedestrian is complicated and not easy to align. Therefore, this paper introduces a Human Feature Aggregation (HFA) approach based on Transformer without alignment, which uses pose information to separate the body parts of target pedestrians from the occlusion. This method utilizes pose information to separate the body parts of the target pedestrian from the obstructions. Initially, the Vision Transformer incorporates Convolutional Neural Network (CNN) advantages to enhance extraction more fine-grained global and local features. Subsequently, the body parts of the target pedestrian are separated from the obstructions using pose information extracted by a pose estimator. Finally, in the human feature aggregation module, local features are matched and fused with pose information to enrich the human features. It steers the model towards focus more on body parts. The experimental findings indicate that the proposed HFA approach surpasses alternative methods on multiple benchmark datasets.

Keywords: Vision Transformer; Pose Information; Pose Estimator; Feature Aggregation

1. Introduction

Person re-identification (Re-ID) is a computer vision missiom that retrieves specific pedestrians across camera scenes. In recent years, several approaches relying on deep learning [1-4] utilize local features to effectively describe pedestrian images. One such method PCB [1], which processes input feature maps by dividing them into evenly distributed local feature vectors for separate classification. Based on this work, MGN [2] combined with the global features to design the MGN network, by dividing pedestrian images into 2 and 3 blocks respectively, different fine-grained local features of pedestrians are obtained. However, all these studies assume that the body parts of the target pedestrians are not occluded. In actual scenarios (such as railway stations, hospitals, shopping center, etc.), the body parts of the target pedestrians are frequently occluded by various obstacles, thus this assumption is often not satisfied. Considering this situation, it is essential to develop a successful strategy to solve the problem of Re-ID in occlusion scenes.

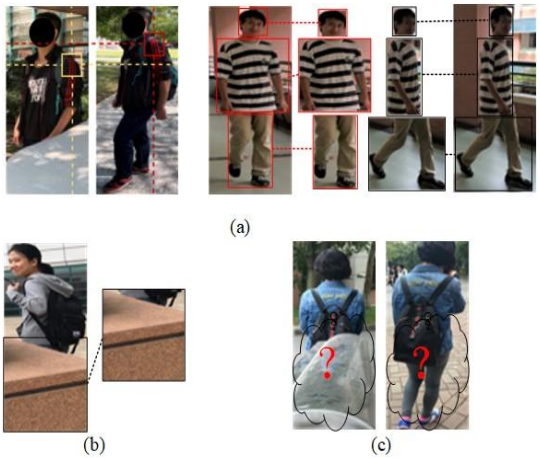

Figure 1. Examples of Three Types of Problems Caused by Occlusion in Practical Applications: (a) Misalignment in Position and Scale; (b) Noisy Information; (c) Missing Information.

In the occlusion scene of Re-ID, the following three types of problems should be considered, as shown in Figure 1: (a) The position dislocation and proportion imbalance between

the local image and the overall image; (b) The rich occlusion variation randomly occludes different body regions, changing the visual presentation of pedestrian body parts and resulting in more errors in the search results; (c) Partial occlusion of objects may have similar characteristics to pedestrian body regions, resulting in model learning failure. In early research [5], the occlusion problem was solved by manual clipping. This method involved segmenting the visible parts of a target pedestrian from the occluded sections in an image, and then using these segmented parts for matching with gallery images. However, this manual approach was not only inefficient, but also easy to introduce human errors in the cutting process. Subsequent studies [6-8] uses pose information to identify the unobscured body parts of target pedestrians, which can avoid manual image cropping. However, these efforts required a precise alignment of the query image of the target pedestrian with the body parts of the gallery image, which led to an overlooked problem (a).

To address these issues, our paper presents an innovative Pose-Guided Human Feature Aggregation Network architecture (HFA-Net) based on Transformer. HFA-Net uses the ViT (Vision Transformer) [9] architecture to avoid the complex and error-prone alignment operations mentioned in problem (1).

Specifically, the HFA-Net proposed in this paper is comprised of several key components: Occlusion data enhancement module, ViT network architecture, Pose estimator and Human Feature Aggregation module. Firstly, the occlusion data enhancement module synthesizes pedestrian images by inserting obstacles clipped on the occlusion datasets. Following this, feature extraction is performed in two stages: initially employing Convolutional Neural Networks (CNNs) for compact image representation, and subsequently utilizing the ViT architecture for the extraction of both global and local features. Finally, the pose estimator is employed to separate the visible body parts of the target pedestrian from the occluded sections, thereby mitigating occlusion-related interference mentioned in question (b). Concurrently, to address problem (c), in the human feature aggregation module, this paper matches and fuses pose information with local features to enrich the information pedestrian feature information and make the model focus on pedestrian body parts is a key objective of this study. The contributions of this paper can be summarized as follows:

1) This paper introduces a pose-guided human feature aggregation network architecture based on Transformer. Firstly, the pose information is utilized to isolate the body parts of target pedestrian from the occlusion, and the body part is matched and fused with the local features extracted by the ViT architecture.

2) This paper designs an occlusion data enhancement strategy to increase the occlusion samples in each training batch. This approach enhances the robustness of the model to diverse occlusion scenarios.

3) The experimental fingdings indicate that the proposed HFA approach surpasses alternative methods on multiple benchmark datasets, providing the empirical evidence of the proposed approach.

2. Method

This section illustrates in detail on the network architecture of our proposed Human Feature Aggregation (HFA). A depiction of our approach is illustrated in Figure 2.

2.1 Occlusion Data Enhancement

Due to the constrained quantity of occluded samples present in the ReID dataset, resulting to a reduced diversity of occlusions within each training batch, rendering the Re-ID model more vulnerable to occlusions.. Therefore, we manually extract the background and occluded objects from the training set through cropping, and these cropped background and occlusion objects as the occlusion set occlude x_{occlude} , and then select the occlusion patch $O \in R^{O_h * O_w}$ from x_{occlude} , where O_h and O_w are the height and width of the image block. In the training phase, for each training batch, augmented samples are created by randomly selecting k occlusions from x_{occlude} . For a given batch of images B and a random k occlusions, we create enhanced images $[x_{i,1}, x_{i,2}, ..., x_{i,k}]$ that are occluded by k occlusions.This increases the quantity of samples in each batch by a factor of k . Together with the original images, the augmented images are utilized for feature learning.

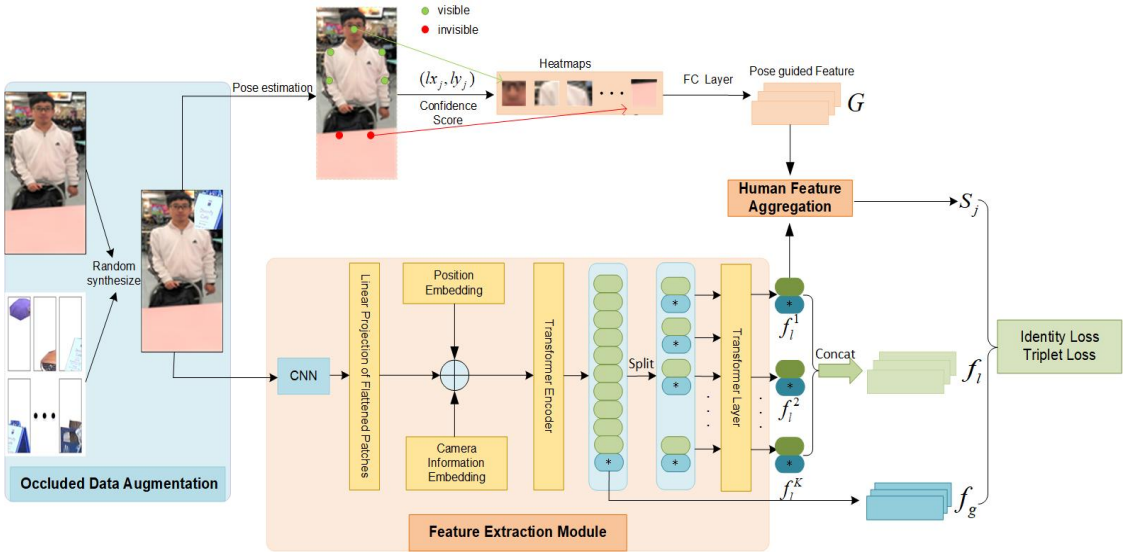

Figure 2. Pipeline of the Human Feature Aggregation method.

2.2 Backbone Network based on CNN-Transformer

In this study, we employ a pre-trained Vision Transformer (ViT) as the foundational network for extracting global and local features from images. First, input a picture $x \in R^{H \times W \times C}$ (where H, W and C represent the height, width and channel count of the image). To preserve more details and edge information of the image,we adopt a convolutional layer structure comprising 3×3 convolutions. This structure is augmented with a residual connection to stabilize the network's training, and defined as:

$$f(x) = Conv(x) + x \quad (1)$$

An input image x with a resolution of $(H \times W)$ will be converted to N image patches $x_p^i \mid i = 1, 2, ..., N$ via a sliding window mechanism. The count of patches N is computed as follows:

$$N = h \times w = \left\lfloor \frac{H + S - P}{S} \right\rfloor \times \left\lfloor \frac{W + S - P}{S} \right\rfloor \quad (2)$$

where $\lfloor . \rfloor$ denotes the floor function, S is the stride of the sliding window, and P represents the each image patch. Non-overlapping patches are created when the S matches the P. Howerve, S to be smaller than P results in overlapping patches, thereby potentially reducing the absence of spatial neighborhood information within the image. Subsequently, the patches embedding $(E(x_p^i))$ is added before the learnable embed token x_{class}. To retain the position information, we apply the position embedding P to add it with the result vector. Additionally, we incorporate the approach outlined in TransReID [10] to acquire camera angle information, addressing camera perspective changes affect feature extraction. Therefore, the final input sequence is expressed as:

$$E_{input} = \left\{ x_{class}; E(x_p^1); E(x_p^2); ...; E(x_p^N) \right\} + P + \lambda_{cm} C_{id} \quad (3)$$

Where P is positional embedding, a linear projection E that maps an patches to D dimension, $C_{id} \in R^{(N+1) \times D}$ is the camera embedding, and for the same image, C_{id} is the same. λ_{cm} is a hyper-parameter that adjusts the weighting of the camera embedding. Subsequently, the input embedding E_{input} is handled by two transformer layers. The final output of the encoder is $f_{en} \in R^{(N+1) \times D}$, which we divide into two parts (encoder global feature and local feature): $f_g \in R^{1 \times D}$ and $f_p \in R^{N \times D}$.

At the same time, fine-grained local features (such as strip features) can also bring high benefits to improve performance. So we divide f_p into K groups $f_l = [f_l^1, f_l^2, ..., f_l^L]$ in order, and each group has a size of $(N//K) \times D$. Then connect K feature sets and send them to the Transformer layer for learning.

2.3 Human Feature Aggregation

As occlusion may lead to a mixing of information between the body regions of target

pedestrian and obstacles in the final feature representation, it may lead to a decline in model performance. Therefore, this paper uses the human pose estimator to distinguish the body parts of the target pedestrians from the occlusions, so that the features acquired by the model focus on the body regions of the target pedestrians.

1) Pose estimation: Within this study, we employ a human pose estimator pretrained on the COCO dataset to identify human keypoints from pedestrian imagery, which are regarded as intermediate representations. Each keypoint, characterized by its coordinates and confidence score, is predicted by the pose estimator, generating heatmaps $M_h,\ h=1,2,...,H$ .These heatmaps are subsequently downsampled to a size of $(H/4)\times(W/4)$. Additionally, a threshold is established to discard keypoints with confidence scores below than a certain value λ. The keypoints are represented as follows:

$$M_h=\begin{cases}(lx_h,ly_h), & \text{if } s_h\geq\gamma\\ 0, & \text{else}\end{cases} \quad (4)$$

Where M_h represents the h confidence point, (lx_h,ly_h) represents the coordinates of the h confidence point, s_h is the confidence score, and λ is the hyperparameter.

The heatmap is composed of 2D Gauss centered on the position of the confidence point. When $M_h=0$, the corresponding heatmap value is set to 0.

2) Human Feature Aggregation: To integrate the pedestrian posture information with local features, we first apply a fully connected layer to the heatmaps M_h to obtain a new set of heatmaps $M_h^{'},h=1,2,3,...H$. Subsequently, through multiple mappings between this new heat map $M_h^{'},h=1,2,3,...H$ and the local feature group, we generate a rich set of pose-guided features $G=[G_1,G_2,...,G_H]$. Although G explicitly encodes information about the different parts of the body, we desire to identify the part of f_l^k that contributes the most to a particular body part. To accomplish this, we treat feature group f_l^k and human posture feature G as a set similarity measurement problem.This approach enables us to obtain the optimal pose-guided feature cluster $S=[S_1,S_2,...S_H]$ .During the matching process. We can identify the feature that is most similar to human posture features G in feature group f_l , and then fuse the two sets of features to form S_h, the expression is as follows:

$$r=\arg\max_r\left(\frac{<G_i,f_l^r>}{\|G_i\|\ \|f_l^r\|}\right) \quad (5)$$

$$S_h=G_i+\omega f_l^r \quad (6)$$

Where $i=1,2...,H$, $<\cdot,\cdot>$ represents the inner product operation, and ω is the hyperparameter set to prevent the loss of some information caused by the simple addition of features, and its value is 0.5. f_l^k represents the feature in G_i that is most similar to f_l.

2.4 Training and Inference

During the final stage of training, we employ identity loss,denoted as L_{ID} , and triplet loss, denoted as L_T . To effectively train both the global feature representation, and the pose-guided feature sets, which is expressed as:

$$L=L_{ID}(f_g)+L_T(f_g)+\frac{1}{H}\sum_{h=1}^{H}(L_{ID}(S_h)+L_T(S_h)) \quad (7)$$

In the testing process, since the global features may include information from occlusions and misleading appearance, we only concatenate the pose-guided feature cluster S_h into the final feature representation, which can express as:

$$F=[S_1,S_2,...,S_H] \quad (8)$$

3. Experiment

3.1 Datasets and Evaluation Metrics

Datasets. To demonstrate the efficacy of our approach, we evaluated its performance on three Re-ID datasets. These encompass various missions, including occluded Re-ID and standard Re-ID, as described below:

Occluded-Duke is a sub-dataset collected by DukeMTMC for the Re-ID occlusion tasks. It comprises 15,618 training images from 702 individuals, along with 2,210 occluded query images and 17,661 gallery images.

Market-1501, a conventional Re-ID dataset captured by 6 cameras. It comprises 12,936 training images depicting 751 pedestrians, alongside 19,732 gallery images of 750 pedestrians and 3,368 query images of 750 pedestrians. The images in this dataset are rarely occluded.

DukeMTMC-REID encompassess 36,411 images protraying 1,404 distinct identities, captured from 8 camera viewpoints. It consists of 16,522 training images for 702 identities, along with 17,661 gallery images, and 2,228 query images.

Evaluation Metrics. We employ two metrics, CMC (Cumulative Match Characteristic) curve and mAP (mean average precision), to assess the performance of virous Re-ID models.

3.2 Implementation Details

In our experimental setup, the initial stage involved initializing our model with a pre-trained ImageNet model. We then resized both training image and test image to 256 × 128. And we augment the training image through random horizontal flips, padding, random crops and random erasures. Next, batch size is configured to 64, 4 images per identity, and the hidden dimension D is set to 768. The initial learning rate of 0.008 is adjusted using cosine decay for learning rate optimization. For keypoints dection in images, we used the HRNet [11] pose estimator pre-trained on the COCO dataset, with the number of human keypoints H of the pose estimator was set to 17, and the threshold γ was set to 0.2. Finally, we have opted to employ identity loss and triplet loss for model training. All experiments were conducted using Nvidia Tesla T4 Gpus, utilizing PyTorch 1.6.0 training.

3.3 Comparison with State-of-the-Art Methods

This section compares our approach to previous approaches on four benchmarks, incorporate occluded Re-ID and standard Re-ID.

Results on Occluded-Duke.

To fully demonstrate the performance of HFA method, we compare it with the previous state-of-the- art Occluded-Duke methods in Table 1.

Table 1. Performance Comparison with state-of-the-art Methods on Occluded and Standard Re-ID Datasets (%)

Method	Market1501		DukeMTMC-REID		Occluded-Duke	
	R1	mAP	R1	mAP	R1	mAP
PCB	92.3	77.4	81.8	66.1	42.6	33.7
OAMN[12]	92.4	79.8	86.3	73.6	62.6	46.1
ISP[13]	95.3	88.6	89.6	80.0	62.8	52.3
FD-GAN[14]	90.5	77.7	80.0	64.5	40.8	-
PGFA[15]	91.2	76.8	82.6	65.5	51.4	37.3
PVPM[16]	93.0	84.9	86.9	75.6	47.0	37.7
PAT[17]	95.4	88.0	88.8	78.2	64.5	53.6
TransReID	94.7	88.1	88.9	80.2	65.7	58.4
(ours)	95.3	89.0	89.7	80.5	67.4	59.4

We categorize these comparison methods into three distinct groups: the first comprises standard Re-ID and manual clipping methods; the second employs pose estimation strategies; the third group consists of Transformer-based Re-ID methods, particularly addressing occlusion challenges, such as PAT and TransReID. Based on the comparison findings, it is evident that the proposed approach HFA achieves 67.4% Rank-1 accuracy and 59.4% mAP accuracy on the Occluded-Duke datasets, and is superior to all previous methods. Compared with PGFA, which utilizes keypoint information, our method enhances Rank-1 accuracy by at least 16.0% and mAP accuracy by 22.1% on the Occluded-Duke dataset. In contrast to TransReID, which based on the Transformer method, our method increases Rank-1 accuracy by 1.7% and mAP accuracy by 1.0% on Occluded-Duke.

The remarkable performance of HFA is attributable to the following points. Primarily, unlike purely Transformer-based architectures TransReID and PAT, our approach effectively integrates the advantages of Convolutional Neural Networks (CNN) into Transformer, enhancing the preservation the structural information within images. Furthermore, the

human feature aggregation module effectively separates the body parts of the target pedestrians from obstacles through pose information, and matches and fuses the body parts with the local features extracted by the ViT architecture by similarity, which highlights the body parts of the target pedestrians more prominently, so it has advantages in solving the occlusion problem.

Results on Market1501 and DukeMTMC-REID.

To assess the universality of HFA-Net across different datasets, we also apply this approach to two traditional person Re-ID datasets. The results of the comparison are presented in Table 1.The experimental findings further prove that the proposed method is not only effective in addressing the occlusion issues, but also has a strong universality to the traditional Re-ID dataset.

Table 2. Ablation Experiments were Performed on the Occluded-Duke Dataset

Index	ODA	CNN	HFA	R-1	R-5	R-10	mAP
1				65.7	80.9	85.7	58.4
2	√			66.2	81.6	86.1	58.3
3		√		66.6	79.9	86.5	58.7
4	√	√		66.9	80.9	86.6	58.9
5	√	√	√	67.4	81.0	87.1	59.4

3.4 Ablation Study

To further analyze the contributions of individual components within HFA-Net, we conducted a series of ablation experiments using the Occluded-Duke dataset. These experiments were to evaluate how each proposed module affects the overall performance of the network architecture. The findings from these experiments are presented in Table *2*.

(1) Effectiveness of Occlusion data augmentation: As depicted in *Table* 2, our finding reveal that augmenting the diversity of occlusion training samples enhances the CMC rank-1/5/10 performance in Re-ID, indicating that expanding the sample diversity through occlusion data augmentation is an effective strategy. However, despite the reasonable processing of the data, the synthesized image still exhibits discernible differences from the original image, resulting in a marginal reduction in the mAP index.

(2) Effectiveness of Convolutional Neural Networks: Distinct from the traditional pure Transformer architecture, our architecture consists of Transformer combined with CNN. and will not damage its performance on other datasets while dealing with the occlusion.

Therefore, to verify the effectiveness of this design, we compared it to the pure Transformer architecture, and the results are shown in *Table 2*. We have observed that the backbone network with CNN outperforms the pure Transformer architecture with a performance enhancement of 0.4%. This shows that combining the advantages of CNN into Transformer can protect the integrity of the internal structural information of the image during image segmentation.

(3) The Impact of Pose Estimation: This study evaluates three distinct pose estimation algorithms: Alpha-Pose [18], OpenPose[19] and HRNet. The results from *Table 3* show that with better detectors, our approach will achieve higher ReID performance. Although the accuracy of the keypoints generated in this way needs to be improved, the method in this paper can still benefit from the current level of accuracy.

Table 3. HFA Performance of Different Pose Estimation Algorithms.

Method	R-1	R-5	R-10	mAP
OpenPose	64.1	77.8	81.1	55.6
AlphaPose	65.6	78.9	89.2	57.7
HR-Net	67.4	81.0	87.1	5.4

(4) The Impact of the hyper-parameter ω : We introduce a hyperparameter w to adjust the balance between the pose-guided feature G_i and the feature set f_l, as depicted in *Figure 3*. As discussed in Section 2.3.2 , in formula(6), simply blending features may result in the loss of some information. Therefore, we evaluate the performance of the model ω across various hyper-parameter configurations. According to *Figure 3*, when is less than 0.5, both rank-1 and mAP accuracy exhibit a linear upward trend. When ω is set to 0.5, rank-1 accuracy and mAP peak. However, when ω is greater than 0.5, these two indicators begin to decline.

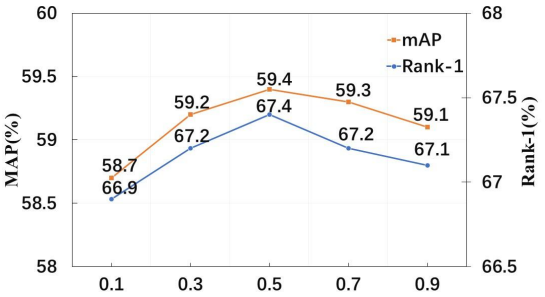

Figure 3. Parameter Analysis for the Hyperparameter ω.

(5) The Impact of the Threshold γ : The threshold γ , as defined by equation (4), it is the threshold for filtering out low confidence scores. As shown in *Figure 4*, when the value of γ is either excessively small or large, the performance of the model deteriorates. Because when γ is too small (such as 0), may lead the model to include all detected confidence points, introducing excessive noise and degrading performance. Conversely, a too high γ will result in many confidence points are dropped, potentially resulting in the loss of information about a body area.

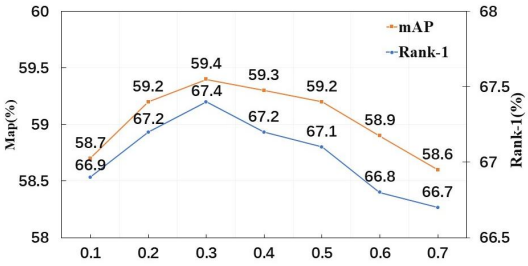

Figure 4. Parameter Analysis for the Hyper-parameter γ.

4. Conclusion

In this study, we present the Transformer-based Pose-Guided Human Feature Aggregation Network (HFA-Net), a novel frame work designed to tackle the challenges of occlusion in Re-Identification (Re-ID) tasks. Firstly, we propose a noval data augmentation strategy that not only effectively reduces the occlusion noise interference, but also augments the dataset with occlusion samples. Secondly, the advantages of CNN are combined into the network architecture of Transformer to extract both fine-grained global features and local features, these local features are sequentially grouped. Finally, we employ a pose estimator to detect key points of pedestrians, separating body part information from occlusion noise. This body part information is subsequently matched and integrated with the loacl feature group to highlight the body part. The experimental outcomes confirm the efficacy of our approach. It also indicates that with the improvement of the performance of pose estimation model, our method will have a good development prospect.

References

[1] Sun Y, Zheng L, Yang Y, et al. Beyond Part Models: Person Retrieval with Refined Part Pooling (and a strong convolutional baseline). European Conference on Computer Vision, 2018: 501-518.

[2] Wang, F., Jiang, M., Qian, C., et al. Learning Discriminative Features with Multiple Granularities for Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6199-6208), 2018.

[3] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In ICCV, 2017.

[4] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. arXiv preprint:1703.07737, 2017.

[5] Zheng L, Huang Y, Lu H, et al. Pose-invariant embedding for deep person re-identification. IEEE Transactions on Image Processing, 2019, 28(9): 4500-4509.

[6] He, S.; Luo, H.; Wang, P.; et al. Transreid: Transformer-based object re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

[7] Sun, K.; Xiao, B.; Liu, D.; et al. 2019b. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 5693–5703.

[8] Li, Y.; He, J.; Zhang, T.; et al. 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), 2898–2907, 2021.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. Zhai. An image is worth 16x16 words:

Transformers for image recognition at scale. ICLR, 2021.

[10] He, S.; Luo, H.; Wang, P.; et al. 2021. Transreid: Transformer-based object re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

[11] Sun, K.; Xiao, B.; Liu, D.; et al. 2019b. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 5693–5703.

[12] Peixian Chen, Wenfeng Liu, Pingyang Dai, et al. Occlude them all: Occlusion-aware attention network for occluded person re-id. In ICCV,pages 11833–11842, 2021. 3, 6, 7.

[13] Zhu, K.; Guo, H.; Liu, Z.; et al. 2020. Identity-guided human semantic parsing for person re- identification. In European Conference on Computer Vision (ECCV), 346–363.

[14] Ge, Y.; Li, Z.; Zhao, H.; et al. 2018. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. arXiv preprint arXiv:1810.02936.

[15] Miao, J.; Wu, Y.; Liu, P.; et al. 2019. Pose-guided feature alignment for occluded person re- identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 542–551.

[16] Shang Gao, Jingya Wang, Huchuan Lu, et al. Pose-guided visible part matching for occluded person reid. In CVPR, pages 11744–11752, 2020. 1, 6

[17] Li, Y.; He, J.; Zhang, T.; et al. 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), 2898–2907.

[18] Fang, H.-S.; Xie, S.; Tai, Y.-W.; et al. 2017. RMPE: Regional Multi-person Pose Estimation. In ICCV.

[19] Cao, Z.; Simon, T.; Wei, S.-E.; et al. 2017. Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

