ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446014

PARALLEL AUGMENTATION AND DUAL ENHANCEMENT FOR OCCLUDED PERSON RE-IDENTIFICATION

Zi Wang[1], *Huaibo Huang*[2], *Aihua Zheng*[3], *Chenglong Li*[3], *Ran He* [2]

[1] School of Computer Science and Technology, Anhui University [2] MAIS & CRIPAC, CASIA
[3] IMIS Laboratory of Anhui Province, Anhui Provincial Key Laboratory of MCC, Anhui University

ABSTRACT

Occluded person re-identification (Re-ID), the task of searching for the same person's images in occluded environments, has attracted lots of attention in the past decades. Recent approaches concentrate on improving performance on occluded data by data/feature augmentation or using extra models to predict occlusions. However, they ignore the imbalance problem in this task and can not fully utilize the information from the training data. To alleviate these two issues, we propose a simple yet effective method with **P**arallel **A**ugmentation and **D**ual **E**nhancement (**PADE**), which is robust on both occluded and non-occluded data and does not require any auxiliary clues. First, we design a parallel augmentation mechanism (**PAM**) to generate more suitable occluded data to mitigate the negative effects of unbalanced data. Second, we propose the global and local dual enhancement strategy (**DES**) to promote the context information and details. Experimental results on three widely used occluded datasets and two non-occluded datasets validate the effectiveness of our method. The code is available at PADE (GitHub).

Index Terms— Person Re-identification, Data Augmentation, Feature Enhancement

1. INTRODUCTION

Occluded person Re-ID, which incorporates the data obscured by various obstacles, has recently gained popularity. And the occlusions are uncommon in the training set [1, 2] while abundant in the test set (especially in query), as illustrated in Fig. 1 (a). Training with such unbalanced data increases the challenge for the network while testing on unknown data. Efforts in data and feature augmentation are emerging to eliminate the imbalance between training and testing. Most methods [3, 4, 5, 6] employ standard data augmentation such as random flipping, random deleting, random cropping, and so on. Furthermore, FED [7] provides feature augmentation strategies to improve the network's adaptability to occluded data. The widely used data/feature augmentation mechanisms take one image/feature as the input and output only one changed image/feature to the subsequent network for training. However, as illustrated in Fig.1 (b), practically all occlusions occur in the query, and the gallery images almost have no obstructions in occluded Re-ID datasets [8, 9]. The methods mentioned that focus on data/feature augmentations ignore the unbalanced occlusion between query and gallery. To increase the robustness of the network on both the non-occluded data (in the gallery) and the occluded data (in the query), we propose a data augmentation method called the Parallel Augmentation Mechanism (PAM). Our PAM consists of three independent components: Base Augmentation (BA), Erasing Augmentation (EA), and Cropping Augmentation (CA). In our parallel augmentation mechanism, EA only implements the erase operation, and CA only crops the original image. We will obtain an image triplet after the PAM, as shown in Fig. 2 (left). Then the ViT-based feature extractor takes the image triplet as the input.

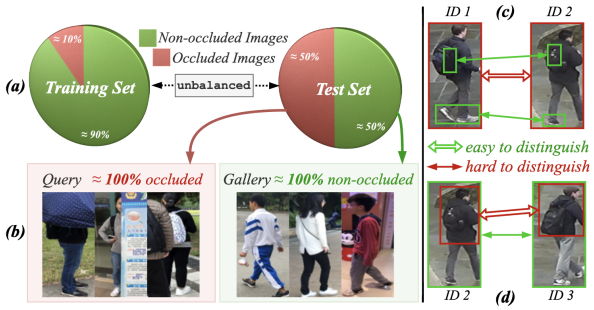

Fig. 1. (a) & (b): Imbalance problem. (c) & (d): Global and local information have their advantages, respectively.

Additionally, both details and context information are crucial for the Re-ID task. As illustrated in Fig. 1 (c), we can simply identify *ID1* and *ID2* by local details while finding it hard to distinguish them based on their outward appearance. [10, 11, 12] propose using additional clues by leveraging foreground segmentation and pose estimation models. [4, 13, 14] propose to split the global feature into several parts and use finer features with detailed information for training. In some cases, the global information becomes more crucial when the body is hindered by unknown impediments or the details are

Corresponding author: Aihua Zheng (ahzheng214@foxmail.com)

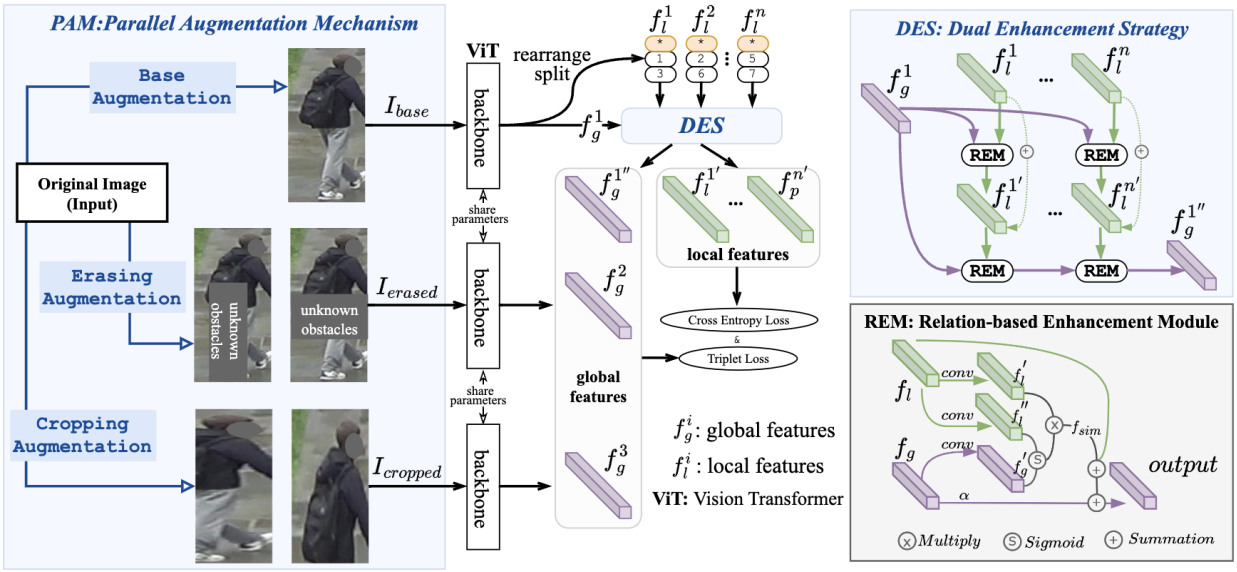

Fig. 2. Overall structure of **PADE**. First, we implement erase and crop operations on original inputs to form the image triplet (Original, Erased, and Cropped images). The image triplet will be sent to the ViT-based backbone to extract global features. Then the global and local features from the non-occluded image branch will be interactively enhanced by each other.

similar, as shown in Fig.1 (d). ViT-based approaches [3, 7, 12] propose using ViT as a feature extractor due to the extreme sensitivity of global context information. However, when encountering new datasets, the methods using auxiliary clues are susceptible and require extra annotations or fine-tuning. While the different persons have similar appearances, only using extracted global information is not discriminative enough. To make full use of global and local features, we propose the Dual Enhancement Strategy (DES) to enhance context information and details by forcing them to promote each other. As shown in Fig. 2, the global and local features will be enhanced in two sequential steps. First, each local feature can be boosted by the context information in the global feature, and then the global feature can absorb the detailed information in the enhanced local features. It is worth noting that our DES does not require additional annotation or model assistance. Our contributions are as follows: (1) We design the Parallel Augmentation Mechanism to form image triplets that will improve the robustness of the network. (2) We propose to enhance the global and local features in an interactive way according to the Dual Enhancement Strategy. (3) Our method does not need any auxiliary clues, and experimental results on five Re-ID datasets demonstrate the effectiveness and generality of the proposed methods.

2. PROPOSED METHODS

2.1. Parallel Augmentation Mechanism

The widely used data augmentations are performed randomly and in a serial manner. However, occluded person Re-ID aims to match non-occluded and occluded images of the same person. To solve the unbalanced testing problem, we design the parallel augmentation mechanism (PAM), which generates augmented images in a parallel manner. For each original input image I, we implement three augmentations on it to obtain the image triplet $[I_{base}, I_{erased}, I_{cropped}]$ for training. This process can be formulated as:

$$I_{base} = BA(I), I_{erased} = EA(I), I_{cropped} = CA(I), \quad (1)$$

where BA(·), EA(·), and CA(·) denote Base Augmentation, Erasing Augmentation, and Cropping Augmentation, respectively. Specifically, we keep the resize and normalization operations in all three augmentations. Compared to traditional data augmentation, the BA(·) only changes the size of the input image, the EA(·) only adds obstacles at random locations on the image, and the CA(·) only crops the image irregularly. After PAM, we obtain an image triplet, which consists of one image similar to the non-occluded image and two augmented images with different types of occlusion. Then, all three images will be sent to a parameter-shared multi-branch network:

$$f_g^1, f_g^2, f_g^3 = \theta(I_{base}, I_{erased}, I_{cropped}), \quad (2)$$

where $\theta(\cdot)$ denotes the feature extractor, we choose ViT-base [15] as our backbone due to its powerful ability.

2.2. Dual Enhancement Strategy

Many approaches extract critical global and local features in the training phase and optimize them together. However, they ignore the context information, and the details may not be

valid at the same time. To solve this problem, we design the dual enhancement strategy (DES) to enhance both global and local features in an interactive way inspired by [14].

After feature extraction, we can obtain the global features f_g^1 from I_{base} and the local features (f_l^1, f_l^2 ... f_l^n) by splitting the f_g^1. The n denotes the number of local features and is set to 4 in our experiment. First, each local feature will be enhanced by the global feature, and the original local features are treated as residuals and added back to the enhanced local features. This process can be formulated as follows:

$$f_l^{i'} = REM(f_l^i, f_g^1) + f_l^i, \quad i = 1, ..., n. \tag{3}$$

Then the global features are also enhanced similarly to absorb the details in the local features, as shown in Fig. 2. This process can be formulated as:

$$f_g^{1''} = \begin{cases} REM(f_g^1, f_l^{i'}), \text{ if } i = 1 \\ REM(f_g^{1''}, f_l^{i'}), \text{ if } i = 2, ..., n. \end{cases} \tag{4}$$

where the REM($\cdot$) denotes the relation-based enhancement module and can be formulated as:

$$f_{sim} = \sigma(f_l^{'} \odot f_g^{'}) * f_l^{''}, \quad f_g^{'} = f_{sim} + f_g + f_l, \tag{5}$$

where σ denotes sigmoid, the $f_l^{'}$, $f_l^{''}$ and $f_g^{'}$ are the features after $Conv1 \times 1$, the $\odot$ denotes transpose multiply. The structure of REM is shown in Fig. 2, and we use only one local feature and the global feature to illustrate. Then global features (f_g^1, f_g^2, f_g^3) and local features (f_l^1, ... f_l^n) will be concatenated to form the final person description.

2.3. Loss Function

We choose widely used cross-entropy loss (L_{id}) and triplet loss (L_{tri}) to train our model. All the global and local features are under the constraint of L_{id} and L_{tri}. The final loss function can be formulated as:

$$\begin{aligned} L_{final} = &\sum_{i=1}^{3} L_{id}(p_g^i, y) + \sum_{j=1}^{4} L_{id}(p_l^j, y) \\ &+ \sum_{i=1}^{3} L_{tri}(f_g^i) + \sum_{j=1}^{4} L_{tri}(f_l^j) \end{aligned} \tag{6}$$

where the p denotes the predicted results from global features f_g and local feature f_l. The y denotes the ground truth.

2.4. Implementation Details

The implementation platform of our experiment is Pytorch with RTX 3090Ti GPUs. The original learning rate is set as 0.008 and will be reduced in epochs 40 and 70. The max epoch is 170, the batch size is set to 32. We concatenate one global feature and four local features for testing, the dimension of the final person description is $768 * (1 + 4) =$ 3840-dim. The Stochastic Gradient Descent (SGD) with a weight decay of 0.0004 is used in our experiment to fine-tune the whole network. Evaluation metrics are widely adopted Cumulated Matching Characteristics (CMC) curves and the mean Average Precision (mAP).

Methods	Occluded-Duke		Partial-REID		Occluded-ReID	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
HOReID* [16]	43.8	55.1	-	85.3	70.2	80.3
PVPM* [11]	-	-	72.3	78.3	61.2	70.4
PFD* [12]	61.8	69.5	-	-	**83.0**	81.5
PCB [4]	33.7	42.6	63.8	66.3	28.9	41.3
PGFA [17]	37.3	51.4	61.5	69.0	-	-
OAMN [18]	46.1	62.6	77.4	86.0	-	-
MoS [19]	49.2	61.0	-	-	-	-
ISP [20]	52.3	62.8	-	-	-	-
ViT Base [15]	52.3	59.9	74.0	73.3	76.7	81.2
PAT [21]	53.6	64.5	-	88.0	72.1	81.6
FED [7]	56.4	68.1	80.5	83.1	79.3	**86.3**
TransReID [3]	59.2	66.4	-	-	-	-
LoGoViT [22]	61.4	67.4	-	-	-	-
PADE (ours)	**63.0**	**72.3**	**84.8**	**89.3**	79.9	83.7

Table 1. Experimental results on Occluded-Duke, Partial-REID, and Occluded-ReID (in %). The best results are shown in **blod**, and the second results are shown in blue. The superscript * denotes the method needs auxiliary clues.

3. EXPERIMENT

3.1. Results on Occluded Re-ID

We compared our method with state-of-the-art occluded Re-ID methods on Occluded-Duke[2], Partial-REID[8] and Occluded-ReID[9]. As shown in Table 1, our PADE achieves 63.0%/84.8% Rank-1 accuracy and 72.3%/89.3% mAP on Occluded-Duke and Partial-REID, respectively, and the results outperform all compared methods. Moreover, the results of our PADE on Occluded-ReID are better than all the methods with no auxiliary clues. Because of the additional keypoint detection model, PFD [12] performs slightly better than ours on mAP evaluation, but the Rank-1 accuracy of PFD is still 2.2% lower than ours. Our PADE considers the data unbalanced problems in testing and uses the parallel augmentation mechanism to improve the robustness of the network in the training phase. More visualization of datasets and ranking lists can be found at PADE (GitHub).

3.2. Results on non-occluded Re-ID

We choose Market-1501 [1] and DukeMTMC-reID [23] to evaluate our PADE and several state-of-the-art methods, and the results are shown in Table 2. The dual enhancement strategy helps our PADE to enhance and utilize both context and detailed information and achieve high accuracy, we can observe PADE outperforms all the methods without using auxiliary clues and obtain 89.8%/95.8% mAP and Rank-1 on

Methods	*Auxiliary Clues*	Market-1501		DukeMTMC-reID	
		mAP	Rank-1	mAP	Rank-1
HOReID [16]	✓	84.9	94.2	75.6	86.9
PFD† [12]	✓	89.7	95.5	**83.2**	91.2
PGFA [17]	×	76.8	91.2	65.5	82.6
OAMN [18]	×	79.8	92.3	72.6	86.3
MoS [19]	×	86.8	94.7	77.0	88.7
ISP [20]	×	88.6	95.3	80.0	89.6
JMLFNet [24]	×	89.2	95.7	80.6	89.7
PAT† [21]	×	88.0	95.4	78.2	88.8
FED† [7]	×	86.3	95.0	78.0	89.4
TransReID† [3]	×	88.9	95.2	82.0	90.7
PADE† (ours)	×	**89.8**	**95.8**	82.8	**91.3**

Table 2. Experimental results on Market-1501 and DukeMTMC-reID (in %). The superscript † denotes the backbone of the method is ViT. The best results are shown in **blod**, and the second results are shown in blue.

Ablation Study	*Modules*			Occluded-Duke	
	PAM		DES	mAP	Rank-1
	Cropping	*Erasing*			
(a)	×	×	×	57.3	67.8
(b)	✓	×	×	60.1	69.5
(c)	×	✓	×	61.2	70.0
(d)	✓	✓	×	62.7	71.8
(e)	✓	✓	✓	**63.0**	**72.3**

Table 3. Ablation study of the proposed method on Occluded-Duke (in %). **PAM**: Parallel Augmentation Mechanism. **DES**: Dual Enhancement Strategy.

Market-1501. PADE even achieves the best results on Rank-1 accuracy compared with the methods using the extra pre-trained model on DukeMTMC-reID.

3.3. Ablation Study

To prove the effectiveness of the proposed PAM and DES, we conduct the ablation study on Occluded-Duke by progressively introducing every component, as shown in Table 3. As shown in lines (b), (c), and (d) in Table 3, the model trained with PAM various occluded data achieves better results than the baseline (line (a)). In addition, DES utilizes the context and details information, further improving the accuracy of the proposed method, as Table 3 (e) shows. Moreover, to evaluate the adaptability of the proposed PAM, we replaced the traditional augmentation in OSNet [5], ViT base [15], and TransReID [3] with PAM, the experimental results are shown in Table 4. We can observe that all methods gain improvement after using the parallel augmentation mechanism (PAM).

3.4. Robustness Evaluation on Occluded Data

To evaluate the robustness of PADE, we gradually increase the number of occlusion data by manually adding occlusions to the test data of DukeMTMC-reID [23]. The crop operation

Methods	Occluded-Duke	
	mAP	Rank-1
OSNet [5] (base)	29.5	38.1
OSNet [5] + **PAM**	32.7 (**+3.2**)	42.5 (**+4.4**)
ViT [15] (base)	52.3	59.9
ViT [5] + **PAM**	57.9 (**+5.6**)	66.2 (**+6.3**)
TransReID [3] (base)	59.2	66.4
TransReID [3] + **PAM**	62.7 (**+3.5**)	71.8 (**+5.4**)

Table 4. Results of combining Parallel Augmentation Mechanism **(PAM)** with baseline on Occluded-Duke (in %).

Percentages		20%	40%	60%	80%	100%
mAP	ViT base [15]	57.2	53.6	50.9	48.0	45.5
	TransReID [3]	60.6	58.4	56.2	54.2	50.2
	PADE (ours)	**71.8**	**68.9**	**65.8**	**63.5**	**62.2**
Rank-1	ViT base [15]	79.7	75.7	73.5	71.5	68.7
	TransReID [3]	82.2	79.2	79.4	76.5	74.5
	PADE (ours)	**87.3**	**86.8**	**84.2**	**83.2**	**81.2**

Table 5. Experimental results on different percentages of occluded data in DukeMTMC-reID (in %).

and the erase operation are implemented on the original input to simulate occluded data. We can observe from Table 5 that our PADE always achieves the best results and outperforms ViT base/TransReID on both mAP and Rank-1 accuracy. It means that our method suffers less when the occlusion data increases and our model is more robust on unexpected occlusions than the other two methods.

4. CONCLUSION

In this paper, we propose a simple yet effective method with Parallel Augmentation and Dual Enhancement for robust performance on both occluded and non-occluded person Re-ID. The parallel augmentation mechanism (PAM) can generate the image triplet (including non-occluded, erased, and cropped images), and help the network gain better ability on occluded-agnostic test data. The context information and details in global and local features will promote each other according to the dual enhancement strategy (DES). Both PAM and DES in our method can be flexibly embedded into other Re-ID methods, and they do not rely on any additional data annotations or models. In the future, we will focus on adaptive data augmentation and dynamic feature enhancement to deal with more complex occlusion environments.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China (Grants 62372003), the University Synergy Innovation Program of Anhui Province (Grant GXXT-2022-036), the Natural Science Foundation of Anhui Province (No. 2208085J18, No. 2308085Y40), the Natural Science Foundation of Anhui Higher Education Institution (No. 2022AH040014).

5. REFERENCES

[1] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015, pp. 1116–1124.

[2] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang, "Pose-guided feature alignment for occluded person re-identification," in *ICCV*, 2019, pp. 542–551.

[3] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang, "Transreid: Transformer-based object re-identification," in *ICCV*, 2021, pp. 15013–15022.

[4] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *ECCV*, 2018, pp. 480–496.

[5] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang, "Omni-scale feature learning for person re-identification," in *ICCV*, 2019, pp. 3702–3712.

[6] Zhen Lei, Shengcai Liao, Ran He, Matti Pietikainen, and Stan Z Li, "Gabor volume based local binary pattern for face representation and recognition," in *FG*, 2008, pp. 1–6.

[7] Zhikang Wang, Feng Zhu, Shixiang Tang, Rui Zhao, Lihuo He, and Jiangning Song, "Feature erasing and diffusion network for occluded person re-identification," in *CVPR*, 2022, pp. 4754–4763.

[8] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shaogang Gong, "Partial person re-identification," in *ICCV*, 2015, pp. 4678–4686.

[9] Jiaxuan Zhuo, Zeyu Chen, Jianhuang Lai, and Guangcong Wang, "Occluded person re-identification," in *ICME*, 2018, pp. 1–6.

[10] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang, "Mask-guided contrastive attention model for person re-identification," in *CVPR*, 2018, pp. 1179–1188.

[11] Shang Gao, Jingya Wang, Huchuan Lu, and Zimo Liu, "Pose-guided visible part matching for occluded person reid," in *CVPR*, 2020, pp. 11744–11752.

[12] Tao Wang, Hong Liu, Pinhao Song, Tianyu Guo, and Wei Shi, "Pose-guided feature disentangling for occluded person re-identification based on transformer," in *AAAI*, 2022, vol. 36, pp. 2540–2549.

[13] Aihua Zheng, Zi Wang, Zihan Chen, Chenglong Li, and Jin Tang, "Robust multi-modality person re-identification," in *AAAI*, 2021, vol. 35, pp. 3529–3537.

[14] Zi Wang, Chenglong Li, Aihua Zheng, Ran He, and Jin Tang, "Interact, embed, and enlarge (ieee): Boosting modality-specific representations for multi-modal person re-identification," in *AAAI*, 2022, vol. 36, pp. 2633–2641.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[16] Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun, "High-order information matters: Learning relation and topology for occluded person re-identification," in *CVPR*, 2020, pp. 6449–6458.

[17] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang, "Pose-guided feature alignment for occluded person re-identification," in *ICCV*, 2019, pp. 542–551.

[18] Peixian Chen, Wenfeng Liu, Pingyang Dai, Jianzhuang Liu, Qixiang Ye, Mingliang Xu, Qi'an Chen, and Rongrong Ji, "Occlude them all: Occlusion-aware attention network for occluded person re-id," in *ICCV*, 2021, pp. 11833–11842.

[19] Mengxi Jia, Xinhua Cheng, Yunpeng Zhai, Shijian Lu, Siwei Ma, Yonghong Tian, and Jian Zhang, "Matching on sets: Conquer occluded person re-identification without alignment," in *AAAI*, 2021, vol. 35, pp. 1673–1681.

[20] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang, "Identity-guided human semantic parsing for person re-identification," in *ECCV*. Springer, 2020, pp. 346–363.

[21] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu, "Diverse part discovery: Occluded person re-identification with part-aware transformer," in *CVPR*, 2021, pp. 2898–2907.

[22] Nguyen Phan, Ta Duc Huy, Soan TM Duong, Nguyen Tran Hoang, Sam Tran, Dao Huu Hung, Chanh D Tr Nguyen, Trung Bui, and Steven QH Truong, "Logovit: Local-global vision transformer for object re-identification," in *ICASSP*, 2023, pp. 1–5.

[23] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *ECCV*, 2016, pp. 17–35.

[24] Yunzuo Zhang, Weili Kang, Yameng Liu, and Pengfei Zhu, "Joint multi-level feature network for lightweight person re-identification," in *ICASSP*, 2023, pp. 1–5.