Logovit: Local-Global Vision Transformer for Object Re-Identification

Conference Paper · June 2023

DOI: 10.1109/ICASSP49357.2023.10096126

9 authors, including:

Soan Duong
University of Wollongong
22 PUBLICATIONS **220** CITATIONS
SEE PROFILE

Hung Dao-Huu
Keio University
15 PUBLICATIONS **60** CITATIONS
SEE PROFILE

Trung H. Bui
Adobe Research
161 PUBLICATIONS **3,274** CITATIONS
SEE PROFILE

LOGOVIT: LOCAL-GLOBAL VISION TRANSFORMER FOR OBJECT RE-IDENTIFICATION

Nguyen Phan[†] *Sam Tran*[†] *Nguyen Tran Hoang* [†] *Ta Duc Huy*[†] *Soan T. M. Duong*[†§]
Chanh D. Tr. Nguyen[†‡] *Dao Huu Hung*[†] *Trung Bui* *Steven Q. H. Truong*[†]

[†] VinBrain JSC., Vietnam; [‡] VinUniversity, Vietnam; [§] Le Quy Don Technical University, Vietnam

ABSTRACT

Object re-identification (ReID) is prone to errors under variations in scale, illumination, complex background, and object occlusion scenarios. To overcome these challenges, attention mechanisms are employed to concentrate on interesting parts of an object to extract better discriminative features. This paper introduces local-global vision transformer (LoGoViT) for object re-identification by learning a hierarchical-level representation from fine-grained (local) to general (global) context features. It comprises two components: (i) shift and shuffle operations generate robust local features, and (ii) local-global module which aggregates the multi-level hierarchy features of an object. Extensive experiments show that our method achieves state-of-the-art on ReID benchmarks. We further investigate effective augmentation operations and discuss how patch modifications can help the model generalize under occlusion. Our code is available at `https://github.com/nguyenphan99/LoGoViT`

***Index Terms*—** object re-id, public security, vision transformer, multi-scale, patch modification augmentation

1. INTRODUCTION

ReID is a task to retrieve and assign the identities of particular objects across different cameras or viewpoints. It is prevalent in surveillance camera applications regarding object retrieval or crow behavior analysis. Approaches for ReID usually include extracting the visual features of an object to match with the visual features of its other occurrences. The previous works summarize and propose a strong baseline for person and vehicle ReID task [1–3]. Convolutional neural network (CNN) -based methods have been widely adopted as a natural choice for the feature extractors due to its efficiency [4–7]. CNN architectures always embody pooling operations that reduce the spatial dimension of the feature maps to increase the receptive field of the model. This prevents the visual representation of objects from fine-grained spatial information.

Several attempts, e.g., attention and multi-scale approaches, have been made to learn the finer granularity representation of the objects by modifying the architecture [4, 8, 9]. The attention mechanism is incorporated into CNN to exploit the long-range dependencies as well as the global context of the image [9–11] . The presence of various backgrounds, illumination variation, and occlusions are still the main challenges for those attempts.

Vision Transformer (ViT), which is a new architecture and is able to combine the structural patterns in the global scope and the fine-grained features in local patches. The self-supervised and unsupervised learning method [12, 13] was introduced to reduce the domain gap between ImageNet and ReID dataset and archive state-of-the-art results. TransReID [14] is the first ViT-based approach for the object ReID task, in which it uses the transformer instead of CNN backbone and introduces the jigsaw patch module (JPM) and side information embedding (SIE). Thus, TransReID avoids the lost of spatial information due to the pooling operation in CNN and attention on the small discriminative area due to the small receptive field, thereby producing promising ReID results. Transformer-based approaches [15–19] also investigate the possibility of applying to ReID task and then customized the improved version from the vanilla transformer. However, the variation in object scales, which could be a potential challenge for ReID, has not been carefully discussed.

This paper presents a novel object ReID framework that is able to hierarchically extract both global structure and fine-grained robust features with a high generalization. The proposed framework contains a local-global module (LGM), which is an extension of JPM, to associate multi-scale features with their respective hierarchy level in the JPM. Experimental results demonstrate the superiority of LGM compared to JPM. The proposed framework also includes the patch modification augmentation to simulate the occlusion phenomenon, aiming to address the occlusion scenario. Evaluation on both person and vehicle ReID datasets, i.e. MSMT17, Market-1501, DukeMTMC-ReID, Occluded Duke and, VeRi-776 and VehicleID, the proposed method achieves state-of-the-art among several ReID methods.

2. PROPOSED METHOD

This paper introduces a novel framework, called LoGoViT, for visual re-identification. Figure 1 illustrates an overview of the proposed LoGoViT. The novelty of LoGoViT is twofold: i) the local-global module extracts the hierarchical structure

of the object, and ii) patch modification operations are treated as the augmentation techniques to improve the generalizability over the occlusion scenario of the proposed method. In this section, we first present the transformer-based strong baseline in Section 2.1. We then introduce patch modification augmentation and the proposed LGM module in Sections 2.2 and 2.3, respectively.

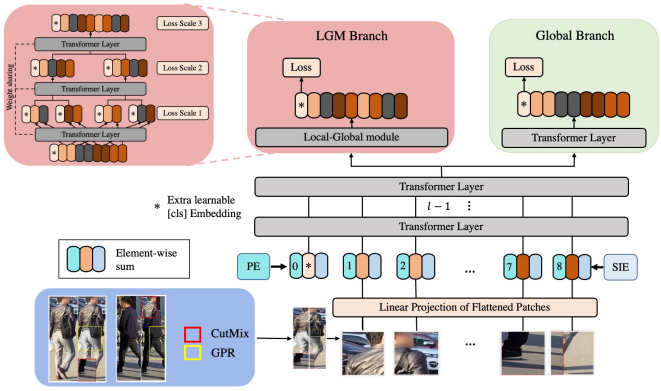

Fig. 1. Our proposed framework. CutMix and GPR and the augmentations mentioned in subsection 3.1 are applied to training images. An image is split into patches and projected into a sequence of embeddings. Each embedding, with the positional embedding and the side information embedding, is summed to feed into layers of the transformer encoder. Besides the global branch for ReID task in the last layer, the auxiliary LGM branch shuffles and regroups all patches in multiple levels. Each scale level shares the same parameter set and is penalized separately with multi-scale ReID loss. Please zoom in for better view.

2.1. Transformer-based strong baseline

Following the pipeline for person and vehicle ReID [1, 14], an input image $x \in \mathbb{R}^{H\times W\times C}$ is split into N patches, where H, W, C represent the height, width, number of channels, respectively. An extra learnable classification embedding token $x_{\text{cls}} \in \mathbb{R}^D$ is considered as a global feature representation and prepended to the input sequences. The input could be defined as $\mathcal{I}_{\text{input}_0} = [x_{\text{cls}}; x_p^1; x_p^2; ...; x_p^N]$, where $\mathcal{I}_{\text{input}} \in \mathbb{R}^{(1+N)\times D}$ is the input sequence embedding. $\{x_p^i \in \mathbb{R}^D : i = 1, 2, 3, ..., N\}$ represents the embedding of N patches after linear projection mapping to D dimensions.

Overlapping patches. To preserve the continuity of all patches in the image, we employ overlapping sliding windows to enhance patch linkage. With an input image of size $W \times H$, a patch size p, and step size s, N splitting patches can be calculated as below:

$$N = N_H \times N_W = \lfloor \frac{H+s-p}{s} \rfloor \times \lfloor \frac{W+s-p}{s} \rfloor, \quad (1)$$

where N_H and N_W are the number of patches in height and width, respectively. The operation $\lfloor . \rfloor$ represents the floor function. Step size s is set smaller than patch size p. The smaller step size s, the more overlapping patches are generated. Noted that the increased number of patches comes with higher performance and a trade-off in computational cost.

Position embedding and additional information. We follow [14] to interpolate pretrained learnable position embedding $\mathcal{P} \in \mathbb{R}^{(1+N)\times D}$ to any image size of object. Side information embedding contains camera ID or viewpoint of the object which is represented as $\mathcal{S}_C \in \mathbb{R}^{N_C\times D}$ and $\mathcal{S}_V \in \mathbb{R}^{N_V\times D}$, respectively, where N_C and N_V are the numbers of cameras and viewpoints. If the camera ID and viewpoint of an image are r and q, the embeddings could be defined $\mathcal{S}_C[r]$ and $\mathcal{S}_V[q]$ for all patches of an image. It might be counteract if we directly add two embeddings $\mathcal{S}_C[r] + \mathcal{S}_V[q]$ due to redundant or adversarial information. Thus, we follow [14] to joinly encode the camera and viewpoint as $\mathcal{S}_{C/V} \in \mathbb{R}^{(N_C\times N_V)\times D}$, and add to the input sequences. The input sequences with camera ID r and viewpoint q are now:

$$\mathcal{I}_{\text{input}_0} = [x_{\text{cls}}; x_p^1; x_p^2; ...; x_p^N] + \mathcal{P} + \alpha\mathcal{S}_{C/V}[r * N_V + q], \quad (2)$$

where α is the coefficient of the SIE. Noted that SIE are the same for each patch but may have different values for different images. In contrast, position embeddings are different for each patch but the same for all images.

2.2. Patch modification augmentation

Grayscale patch replacement in ReID. To close the gap between domains or handle the part-occluded object, we deploy grayscale patch replacement (GPR) [20]. GPR randomly selects a rectangular region and replaces it with the pixels of the same rectangular in the corresponding grayscale image. In this work, we apply it as an effective data augmentation to improve the model's generalizability.

CutMix in ReID. We propose to regularize the model while training with ID and triplet losses with strong localizable features. We apply CutMix [21] into this scheme. CutMix proposes to generate new training samples by combining two images (x_1, y_1) and (x_2, y_2). A generated example is formulated as below:

$$\begin{aligned} x_{\text{new}} &= M \odot x_1 + (1 - M) \odot x_2, \\ y_{\text{new}} &= \lambda y_1 + (1 - \lambda) y_2 \end{aligned} \quad (3)$$

where M denotes binary region dropping out to replace another image. Operators $\odot$ denote element-wised multiplication. λ denotes the combination ratio between the two images. CutMix replaces an image region with a patch from another training example that generates more locally nature images.

Supervision loss. Two losses applied in this work are ID loss and triplet loss. ID is cross-entropy loss ($\mathcal{L}_{ID}$), and triplet loss ($\mathcal{L}_{triplet}$) is defined as in Eq. 4:

$$\mathcal{L}_{\text{triplet}} = log\left[1 + exp\left(||f_a - f_p||_2^2 - ||f_a - f_n||_2^2\right)\right], \quad (4)$$

where (f_a, f_p, f_n) are triplet embedding of the triplet (a, p, n). The triplet (a, p, n) contains three samples including anchor, positive and negative. For the combination of $\mathcal{L}_{\text{ID}}$ and $\mathcal{L}_{\text{triplet}}$ losses, we reformulate the loss function $\mathcal{L}_{\text{ReID-mix}}$ regarding Equation (3):

$$\mathcal{L}_{\text{ReID}} = \mathcal{L}_{\text{ID}} + \mathcal{L}_{\text{triplet}}, \quad (5)$$

$$\mathcal{L}_{\text{ReID-mix}} = \lambda \mathcal{L}_{\text{ReID}}(x_{\text{new}}, y_1) + (1-\lambda)\mathcal{L}_{\text{ReID}}(x_{\text{new}}, y_2). \quad (6)$$

2.3. Local-global module (LGM)

Inspired by the jigsaw patch module [14], we propose a local-global module that can hierarchically capture fine-grained local features. Let $\mathcal{I}_{\text{input}_{\text{LGM}}} = [x_{l-1}^1, x_{l-1}^2, ...x_{l-1}^N]$ denote the input sequence to the last transformer layer. We first apply shift and shuffle operations to $\mathcal{I}_{\text{input}_{\text{LGM}}}$. The order of each hidden feature is shifted to the end with the value of m as $\mathcal{I}_{\text{input}_{\text{LGM}}} = [x_{l-1}^{m+1}, x_{l-1}^{m+2}, ...x_{l-1}^N, x_{l-1}^1, x_{l-1}^2, ..., x_{l-1}^m]$. Then, the shifted patches are shuffled with k groups, where m and k are hyperparameters.

Our local-global module can be configured with different hierarchical levels. Suppose three levels of scale are defined in the network after applying shift and shuffle modules; the hidden features are in the hierarchy structure flow. Each local features $[f_1^i : i = 1, 2, 3, 4]$ at scale 1 contains $N/4$ patches (the light red block in Figure 1), this will turn the output $[\mathcal{F}_{\text{last}}(f_1^i) : i = 1, 2, 3, 4]$, where $\mathcal{F}_{\text{last}}$ is the last transformer layer. The combination between two consecutive features at scale level 1 is considered as input for scale 2 $f_2^1 = [\mathcal{F}_{\text{last}}(f_1^1), \mathcal{F}_{\text{last}}(f_1^2)]$. Similarly, local features at the final scale are formed $f_3^1 = [\mathcal{F}_{\text{last}}(f_2^1), \mathcal{F}_{\text{last}}(f_2^2)]$. The other branch (the light green block as shown in Figure 1) is standard transformer, which will turn the feature output of $\mathcal{I} = [f_{\text{global}}; x_l^1, x_l^2, x_l^3, ..., x_l^N]$. It is noted that global branch features and features at every scale fed into the last transformer layer are all used to calculate the ReID loss.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ReID-mix}}(f_{\text{global}}) + \frac{1}{S}\sum_{i=1}^{S}\sum_{j=1}^{S_i}\mathcal{L}_{\text{ReID-mix}}(f_i^j), \quad (7)$$

where S and S_i represent the scale level and the number of features at each scale, respectively. f_{global} represents the global feature, and f_i^j denotes the j^{th} feature at i^{th} scale .

3. EXPERIMENTS

3.1. Datasets and settings

We evaluated the proposed method on the four person-ReID datasets, Market-1501 [22], MSMT17 [23], DukeMTMC-ReID [24] and Occlude Duke [25] and two vehicle ReID datasets, VeRi-776 [26], VehicleID [27]. Table 3.1 shows the statistic of the datasets in detail. All the person images were resized to 256×128, and the vehicle images were resized to 256×256. In addition to the patch modification augmentation, the augmentations: random horizontal flipping, padding, random cropping, and random erasing, were used in training the model. To implement the proposed LoGoViT framework, ViT-B/16 was used as the backbone, in which the initial weight was loaded from pretrained on ImageNet-21K before finetuning on ImageNet-1K. The proposed LGM was created by modifying the last transformer block of ViT-B/16. The level of hierarchy in LGM is set to 3 by default. The model was trained with a batch size of 64 for 150 epochs. The SGD optimizer was used with weight decay $1e-4$ and momentum 0.9. The learning rate was 0.008 with cosine learning rate decay. The hyperparameters $m = 5$ and $k = 4$ for the person datasets and $m = 8$ and $k = 4$ for the vehicle datasets as suggested in [14]. The coefficient of the SIE was set to 3.

Dataset	Object	#ID	#Image	#cam	#view
Market-1501	Person	1,501	32,668	6	-
MSMT17	Person	4,101	126,441	15	-
DukeMTMC-ReID	Person	1,404	36,441	8	-
Occluded Duke	Person	1,404	36,441	8	-
VeRi-776	Vehicle	776	49,357	20	8
VehicleID	Vehicle	26,328	221,567	-	2

Table 1. Information of six datasets including the figures of ID, Image, camera, view and object type.

All the experiments was conducted on a 40GB GPU in the NVIDIA A100 server. For the evaluation metrics, we measured the results on two main popular metrics to compare the performance in ReID task: mean Average Precision (mAP) and Cumulative Matching Characteristic (CMC) [14].

	Market1501		MSMT17	
	mAP (%)	R1 (%)	mAP (%)	R1 (%)
Baseline	87.0	94.5	61.0	81.8
+GPR	87.4 (+0.4)	94.4 (-0.1)	61.6 (+0.6)	81.7 (-0.1)
+CutMix	88.9 (+1.9)	94.9 (+0.4)	63.5 (+2.5)	82.3 (+0.5)
+LGM	89.5 (+2.5)	95.2 (+0.7)	65.1 (+4.1)	83.3 (+1.5)
+SIE	**89.5 (+2.5)**	**95.1 (+0.6)**	**67.3 (+6.3)**	**84.5 (+2.7)**
+GPR_OP	89.1 (+2.1)	95.1 (+0.6)	65.8 (+4.8)	84.6 (+2.8)
+CutMix_OP	90.2 (+3.2)	95.6 (+1.1)	68.6 (+7.6)	85.5 (+3.7)
+LGM_OP	90.5 (+3.5)	95.4 (+0.9)	69.3 (+8.3)	86.2 (+4.4)
+SIE_OP	**90.7 (+3.7)**	**95.5 (+1.0)**	**70.9 (+9.9)**	**87.0 (+5.2)**

Table 2. The ablation study of each component contributes to the model. GPR, LGM, SIE, and OP are abbreviated to grayscale patch replacement, local-global module, side information embedding, and overlapping patches, respectively.

3.2. Ablation study

We conducted several experiments on Market1501 and MSMT17 to determine each component's impact on the model's performance. Firstly, we opted for ViT-B/16 as the baseline. We then added on each module, including GPR, CutMix, LGM,

and SIE modules [14]. Table 2 shows the improvement when applying one-by-one components to the model.

	MSMT17		Occlude Duke		VeRi-776	
	mAP	R@1	mAP	R@1	mAP	R@1
Two Scale	71.1	87.1	62.1	67.6	**84.3**	97.1
Three Scale	70.9	87.0	61.4	67.4	84.1	97.4
Four Scale	**71.3**	**87.3**	**62.8**	**69.2**	83.5	**98.0**

Table 3. Results of different scales in LGM.

The effect of the depth in LGM. We evaluated saturation by increasing the 'deep' of the hierarchy multi-scale. The experimental results show that the deeper hierarchy might bring higher results. Particularly, we measured the changing effect of the 'deep' in the LGM at three scales $D_l \in \{2, 3, 4\}$ on three datasets MSMT17, OccludeDuke, and VeRi-776. As shown in Table 3, the performance of our proposed method on three levels of scale gains stable results on both person and vehicle datasets.

Backbone	Method	MSMT17		Market1501		DukeMTMC		Occluded-DUKE	
		mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1
CNN	SAN AAAI'20 [28]	55.7	79.2	88.0	**96.1**	75.7	87.9	-	-
	RGA-SC CVPR'20 [9]	57.5	80.3	88.4	**96.1**	-	-	-	-
	SCSN CVPR'20 [11]	58.5	83.8	88.5	95.7	79.0	**91.0**	-	-
	ABDNet ICCV'19 [10]	60.8	82.3	88.3	95.6	78.6	89.0	-	-
	CAL ICCV'21 [7]	**64.0**	**84.2**	**89.5**	95.5	**80.5**	90.0	-	-
	HOReID CVPR'20 [29]	-	-	84.9	94.2	75.6	86.9	43.8	55.1
	RGA+APRA ICIP'22 [30]	-	-	88.7	95.4	78.6	88.8	-	-
	FED CVPR'22 [31]	-	-	86.3	95.0	78.0	89.4	**56.4**	**68.1**
Transformer	TransReID ICCV'21 [14]	67.4	85.3	88.9	95.2	82.6	90.7	59.2	66.4
	PAT CVPR'21 [17]	-	-	88.0	95.4	78.2	88.8	53.6	64.5
	PFD AAAI'22 [18]	65.1	82.7	89.7	**95.5**	83.2	**91.2**	**61.8**	**69.5**
	LoGoViT (Ours)	**70.9**	**87.0**	**90.7**	**95.5**	**84.5**	91.0	61.4	67.4

Table 4. Comparison of our LoGoViT with state-of-the-art methods on Rank-1 and mAP on person ReID datasets. The top performance is highlighted in **bold** and the next best is underlined.

3.3. Comparison with state-of-the-arts

Result on the person ReID datasets. As shown in Table 4, the proposed LoGoViT consistently improves performance of baseline on all person ReID datasets. The results of the other methods are based on published numbers. Particularly, mAP and rank-1 accuracy on MSMT17 are boosted from 61.0% to 70.9% and 81.8% to 87.0%, respectively. On Market1501, our LoGoViT achieves 3.7% mAP and 1.0% rank-1 improvement better than baseline. On DukeMTMC, our proposed method enhances 5.2% mAP and 2.2% rank-1 accuracy. Similar to the Occluded-DUKE dataset, the performance is improved from 53.1% to 61.4% mAP and 60.5% to 67.4% rank-1. Compared with TransReID, our proposed method consistently performs better on all datasets. Regarding CNN-based approaches, our method outperforms CAL [7] with a large margin of 6.9% mAP and 2.8% rank-1 accuracy on MSMT17.

Result on the vehicle ReID datasets. To further demonstrate the generalization of the model, we tested our method on two vehicle ReID datasets. Table 5 illustrates the performance of our method compared to previous works on VeRi and VehicleID. On VeRi-776 dataset, we report 84.1% mAP and 97.4% rank-1 accuracy, which significantly surpasses most of the works. Specifically, LoGoViT is higher than TransReID [14] with a large margin of 3.5% of mAP. Although the rank-1 is more generalize consistent, we also achieve 97.4% compared to 95.4% of CAL [7] and 96.8% of TransReID [14]. Regarding VehicleID dataset, most previous works mainly focus on rank-1 and rank-5 accuracy; thus, we also report the two mentioned metrics as the core factors to measure the performance on this dataset. Although LoGoViT achieve 85.4%, which is slightly lower than ANet [32] 86.0% at rank-1, LoGoViT still gains better results at rank-5 in comparison with ANet [32]. LoGoViT achieves state-of-the-art performance on vehicle ReID dataset, which shows its robustness in the existing ReID challenge.

Backbone	Method	VeRi-776		VehicleID	
		mAP	R1	R1	R5
CNN	SAN AAAI'20 [28]	72.5	93.3	79.7	94.3
	PVEN CVPR'20 [33]	79.5	95.6	84.7	97.0
	SAVER ECCV'20 [34]	79.6	96.4	79.9	95.2
	CFVMNet ACM Multimedia'20 [5]	77.1	95.3	81.4	94.1
	CAL ICCV'21 [7]	74.3	95.4	82.5	94.7
	ANet Neurocomputing'21 [32]	**80.1**	**96.9**	**86.0**	**97.4**
	CAMNet ICIP'22 [35]	79.6	96.6	82.5	-
Transformer	Baseline	78.2	96.5	82.3	96.1
	TransReID ICCV'21 [14]	80.5	96.8	85.2	97.5
	LoGoViT (Ours)	**84.1**	**97.4**	**85.4**	**97.9**

Table 5. Comparison of our LoGoViT with state-of-the-art methods on Rank-1/Rank-5 and mAP on vehicle ReID datasets. The top performance is highlighted in **bold** and the next best is underlined.

4. CONCLUSION

This paper proposed an end-to-end LoGoViT framework for object ReID. The proposed LoGoViT comprises the LGM module to hierarchically extract the robust visual features from random scenes and the patch modification augmentation to handle occlusion scenarios. Extensive experiments show the superiority of our method in comparison with the existing state-of-the-art ReID methods. Furthermore, the components of the LGM module were comprehensively studied to suggest the best configuration of the proposed LoGoViT framework. Besides, the optimal metric learning loss, which is a promising factor in training the proposed framework, has not been carefully investigated. We leave it for future scope of work.

References

[1] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. CVPR Workshops*, 2019.

[2] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S.CH. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE TPAMI*, 2021.

[3] S. D. Khan and H. Ullah, "A survey of advances in vision-based vehicle re-identification," *CVIU*, 2019.

[4] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proc. ICCV*, 2019.

[5] Z. Sun, X. Nie, X. Xi, and Y. Yin, "Cfvmnet: A multi-branch network for vehicle re-identification based on common field of view," in *Proc. ACM-MM*, 2020.

[6] H. Gu, G. Fu, J. Li, and J. Zhu, "Auto-reid+: Searching for a multi-branch convnet for person re-identification," *Neurocomputing*, 2021.

[7] Y. Rao, G. Chen, J. Lu, and J. Zhou, "Counterfactual attention learning for fine-grained visual categorization and re-identification," in *Proc. ICCV*, 2021.

[8] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling," in *Proc. ECCV*, 2018.

[9] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, "Relation-aware global attention for person re-identification," in *Proc. CVPR*, 2020.

[10] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, and Z. Wang, "Abd-net: Attentive but diverse person re-identification," in *Proc. ICCV*, 2019.

[11] X. Chen, C. Fu, Y. Zhao, F. Zheng, J. Song, R. Ji, and Y. Yang, "Salience-guided cascaded suppression network for person re-identification," in *Proc. CVPR*, 2020.

[12] H. Luo, P. Wang, Y. Xu, F. Ding, Y. Zhou, F. Wang, H. Li, and R. Jin, "Self-supervised pre-training for transformer-based person re-identification," in *Proc. CVPR*, 2021.

[13] G. Cao and K. Jo, "Unsupervised person re-identification with transformer-based network for intelligent surveillance systems," in *ISIE*, 2021.

[14] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "Transreid: Transformer-based object re-identification," in *Proc. ICCV*, 2021.

[15] M. Tahir and S. Anwar, "Transformers in pedestrian image retrieval and person re-identification in a multi-camera surveillance system," *Applied Sciences*, 2021.

[16] Shengcai L. and Ling S., "Transmatcher: Deep image matching through transformers for generalizable person re-identification," in *NeurIPS*, 2021.

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

[18] T. Wang, H. Liu, P. Song, T. Guo, and W. Shi, "Pose-guided feature disentangling for occluded person re-identification based on transformer," in *AAAI*, 2022.

[19] H. Wang, J. Shen, Y. Liu, Y. Gao, and E. Gavves, "Nformer: Robust person re-identification with neighbor transformer," in *Proc. CVPR*, 2022.

[20] Y. Gong, "A general multi-modal data learning method for person re-identification," *arXiv preprint arXiv:2101.08533*, 2021.

[21] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proc. ICCV*, 2019.

[22] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. ICCV*, 2015.

[23] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proc. CVPR*, 2018.

[24] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. ECCV*, 2016.

[25] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, "Pose-guided feature alignment for occluded person re-identification," in *Proc. ICCV*, 2019.

[26] X. Liu, W. Liu, H. Ma, and H. Fu, "Large-scale vehicle re-identification in urban surveillance videos," in *Proc. ICME*. IEEE, 2016.

[27] H. Liu, Y. Tian, Y. Yang, L. Pang, and T. Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *Proc. CVPR*, 2016.

[28] X. Jin, C. Lan, W. Zeng, G. Wei, and Z. Chen, "Semantics-aligned representation learning for person re-identification," in *Proc. AAAI*, 2020.

[29] 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," in *Proc. CVPR*, 2020.

[30] Eugene P. W. A., Shan L., Rahul A., Nemath A., and Alex C. K., "Adversarial pairwise reverse attention for camera performance imbalance in person re-identification: New dataset and metrics," in *ICIP*, 2022.

[31] Z. Wang, F. Zhu, S. Tang, R. Zhao, L. He, and J. Song, "Feature erasing and diffusion network for occluded person re-identification," in *Proc. CVPR*, 2022.

[32] R. Quispe, C. Lan, W. Zeng, and H. Pedrini, "Attributenet: Attribute enhanced vehicle re-identification," *Neurocomputing*, 2021.

[33] D. Meng, L. Li, X. Liu, Y. Li, S. Yang, Z. Zha, X. Gao, S. Wang, and Q. Huang, "Parsing-based view-aware embedding network for vehicle re-identification," in *Proc. CVPR*, 2020.

[34] P. Khorramshahi, N. Peri, J. Chen, and R. Chellappa, "The devil is in the details: Self-supervised attention for vehicle re-identification," in *Proc. ECCV*. Springer, 2020.

[35] Manyu L., Mengwan W., Xin H., and Fei S., "Enhancing part features via contrastive attention module for vehicle re-identification," in *ICIP*, 2022.