GlintNet: A Lightweight Global-Local Integration Network with Spatial-Channel Mixed Attention for ReID

Xingbiao Zhou[1,2], Tianze Zhang[2,3], YiXin Zhang[1,2], and Gang Shi[1,2(✉)]

[1] School of Computer Science and Technology, Xinjiang University, Urumqi 830017, China
shigang@xju.edu.cn
[2] Xinjiang Key Laboratory of Signal Detection and Processing, Urumqi 830046, China
[3] Faculty of Science, The University of Melbourne, Parkville, VIC 3010, Australia

Abstract. Existing object re-identification models struggle with parameter inefficiency and computational overhead under complex scenarios involving viewpoint variations or occlusions. To address this, we propose a lightweight Global-Local Feature Integration Network (GlintNet) which integrating CNN-extracted local features and attention-based global representations through a simplified fusion framework. Its bottleneck architecture ensures parameter efficiency, while the novel Spatial-Channel Mixed Attention (SCMA) mechanism—comprising Spatial Embedding Attention (SEA) and Channel Mixed Attention (CMA)—enhances discriminative feature learning. SEA captures spatial patterns using depthwise separable convolutions, while CMA refines channel-wise dependencies, both maintaining linear complexity via grouped computation. Evaluations across multiple benchmarks demonstrate GlintNet's superiority achieving state-of-the-art accuracy, particularly excelling in occlusion and cross-view scenarios.

Keywords: Re-identification · Lightweight NetWork · Global-Local Feature Integration · Spatial-Channel Mixed Attention

1 Introduction

Object re-identification (ReID) addresses the challenge of matching object instances across camera views, as illustrated in Fig. 1. The process involves feature extraction from query and gallery images, followed by similarity computation against a predefined threshold to establish identity correspondence [1]. Practical deployments face inherent limitations due to viewpoint variations, occlusions, and heterogeneous image quality [2], compounded by the demand for resource-efficient models on embedded devices [3].

Current mainstream approaches leverage ResNet's CNN architecture for ReID feature extraction [2, 5, 7] or Transformer-based frameworks for global context modeling [8, 12, 13]. While ResNet mitigates gradient degradation through residual learning and benefits from depth scalability, both paradigms suffer from parameter-intensive designs, limiting deployment in resource-constrained environments. Lightweight alternatives [9, 10] often compromise feature discriminability under complex scenarios.

D.-S. Huang et al. (Eds.): ICIC 2025, CCIS 2565, pp. 383–394, 2025.
https://doi.org/10.1007/978-981-96-9946-9_33

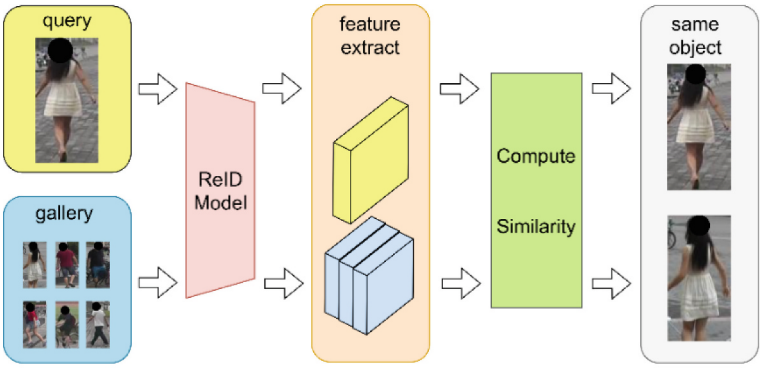

Fig. 1. ReID identifies objects by comparing feature similarities extracted by a model. The images in the query and gallery may come from different cameras and may vary in quality

To bridge this gap, we propose GlintNet, a lightweight backbone with only 2.1M parameters synergizing CNN-localized patterns and attention-driven global cues via a simplified fusion strategy. Complementing this, the Spatial-Channel Mixed Attention (SCMA) mechanism employs depthwise separable convolutions and grouped computations to jointly optimize spatial-channel relationships with linear complexity, enhancing robustness to occlusions and viewpoint shifts.

Extensive evaluations in Market1501, MSMT17 and vehicle ReID benchmarks validate the benefits of GlintNet, which achieves SOTA accuracy with only 1/40 of the parameters compared to its ViT based counterparts. Our main contributions are as follows:

1. The lightweight backbone network, GlintNet, has a parameter size of only 2.1M and can effectively integrate global and local features;
2. SCMA's hierarchical spatial-channel attention with linear time complexity in both spatial and channel dimensions of the features;
3. A large number of experimental verifications conducted on the pedestrian and vehicle ReID dataset.

2 Related Work

2.1 ResNet-Based ReID

ResNet-based approaches dominate feature extraction in ReID research due to their residual learning framework, which mitigates gradient degradation while enhancing depth scalability. Chen et al. [11] developed a Bidirectional Interaction Network using inter-layer bilinear pooling to improve feature discrimination without part annotations. Wang et al. [5] proposed HOReID, aggregating multi-layer features for pose alignment robustness, while Somers et al. [6] integrated attention maps to address occlusion challenges. He et al. [7] further extended ResNet's utility through a modular toolbox supporting diverse ReID tasks. Despite their strong performance, these methods rely on parameter-intensive ResNet architectures, limiting deployment in resource-constrained environments.

2.2 Transformer-Based ReID

Vision Transformers have gained traction in ReID for global context modeling. Zhu et al. [8] enhanced fine-grained discrimination via dual cross-attention learning. He et al. [12] introduced TransReID, leveraging patch embeddings and camera-aware modules to reduce viewpoint bias, and Jia et al. [13] decoupled semantic components with DRL-Net for occlusion robustness. While Transformer-based methods excel in capturing long-range dependencies, their quadratic complexity and large parameter counts hinder practical applications, offering limited insights for lightweight ReID optimization.

2.3 Lightweight ReID

General lightweight techniques like depthwise separable convolutions [14] have inspired domain-specific adaptations. Zhou et al. [9] designed scale-aware residual blocks for pedestrian ReID, while Li et al. [15] proposed a neural architecture search space CDS to identify pedestrian-specific patterns. Gu et al. [16] advanced cross-domain ReID via MSINet's multi-scale interaction. However, these methods often sacrifice discriminability under complex scenarios like occlusion. Our work bridges this gap by integrating CNN's local precision with attention's global awareness through a unified lightweight framework, GlintNet, optimized for both parameter efficiency and cross-scenario robustness.

3 Method

GlintNet is a lightweight network with dual branches: a CNN for local feature extraction and an attention branch for global feature capture. It combines SCMA, a two-stage attention mechanism for spatial-channel feature fusion, to enhance computational efficiency and feature representation. Details are provided in later sections.

3.1 Global-Local Integration Network

GlintNet combines CNN-based local feature extraction with attention-driven global modeling to address feature representation challenges. As illustrated in Fig. 2, the architecture consists of three main components: a Stem module for initial feature extraction, a backbone with six Stage modules and two downsample layers, and a network head for ReID-oriented feature generation.

Each Stage module contains parallel CNN and attention branches. The CNN branch employs three 3×3 DSCNN layers and an SE block [17] to optimize spatial-channel feature extraction with reduced parameters, enhancing computational efficiency and training convergence. This design preserves detailed local patterns (e.g., footwear or hand features) through hierarchical processing. The attention branch incorporates a lightweight self-attention mechanism to model long-range pixel dependencies, dynamically suppress background noise, and capture multi-scale contextual information, thereby compensating for CNN's locality constraints (Fig. 3).

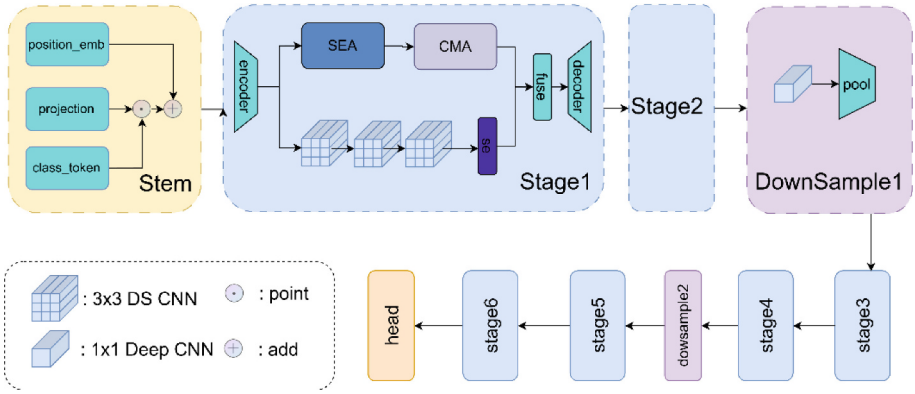

Fig. 2. GlintNet sequentially processes features through Stage and Downsample modules. The Stage integrates global/local features via CNN-attention fusion, and Downsample employs CNN-pooling operations for dimension reduction.

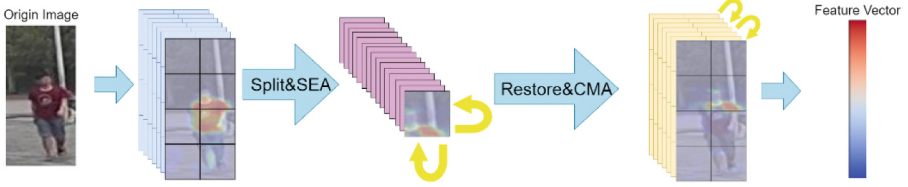

Fig. 3. Extracted features undergo block splitting and rearrangement. SEA generates local spatial attention, followed by block reconstruction and CMA-derived grouped channel attention, culminating in SCMA's final output.

Feature fusion uses additive integration to linearly combine global and local features in a unified semantic space [4], retaining dimensionality for both branches while balancing attention-based contextual guidance and CNN-derived details without significant complexity. The network employs a bottleneck structure with encoder-decoder layers, compressing intermediate dimensions to reduce computation before recovering features for subsequent modules.

3.2 Spatial-Channel Mixed Attention

Traditional self-attention mechanisms are effective for global features but struggle with fine details and incur high computational costs. Inspired by ViT, we propose SCMA (Spatial-Channel Mixed Attention), a lightweight two-stage model combining Spatial Embedding Attention (SEA) and Channel Mixed Attention (CMA). SCMA replaces linear layers with Depthwise Separable CNN (DSCNN) for attention extraction, enhancing spatial features without positional encoding while cutting computational costs (Fig. 4).

SCMA is a two-stage lightweight model that combines spatial and channel attention mechanisms. In the SEA, the input feature map $X \in R^{H \times W \times d}$ is partitioned into blocks

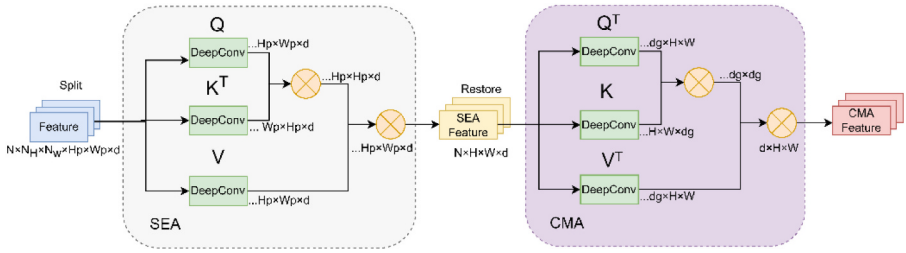

Fig. 4. The SCMA structure consists of two modules: SEA processes features by spatial-level reshaping followed by attention computation, while CMA operates through channel-level reshaping before calculating attention, where $\otimes$ indicates matrix multiplication.

$X_p \in R^{(H/P)\times(W/P)\times d}$ via spatial splitting. Query, key, and value matrices are generated through learnable projections:

$$Q_p = X_p W_Q, K_p = X_p W_K, V_p = X_p W_V \tag{1}$$

Block-wise attention is then computed using Eq. 2, which reduces parameters by a factor of P^2 while enhancing local spatial features.

$$X_{sea} = SEA(X_p) = Softmax\left(\frac{Q_p K_p^T}{\sqrt{d}}\right) V_p \tag{2}$$

In the second stage (CMA), the SEA-processed features X_{sea} are restored to $X_{res} \in R^{H\times W\times d_g\times G}$. Grouped channel attention is applied within G channel groups via Eq. 3, which grouped approach improves feature diversity and memory efficiency.

$$Y = CMA(X_{res}) = Softmax\left(\frac{Q_{res}^T K_{res}}{\sqrt{d_g}}\right) V_{res}^T \tag{3}$$

The complete SCMA formulation integrates both stages:

$$Y = CMA(Restore(SEA(Split(X)))) \tag{4}$$

This dual-stage architecture synergizes spatial-block and grouped-channel attention, enabling efficient local-global feature interaction with lower computational cost than conventional self-attention methods.

3.3 SCMA Time Complexity Analysis

SCMA's complexity analysis proceeds as follows: Standard self-attention exhibits $O\left(4Nd^2 + 2N^2 d\right)$ complexity, where N is sequence length and d is embedding dimension. The $O\left(4Nd^2\right)$ term corresponds to linear projections, while $O\left(2Nd^2\right)$ arises from attention computation. Comparatively, CNN operations scale as $O\left(K^2 Md\right)$ with kernel size K and feature map size M, where $M^2 = N$. SCMA decomposes complexity into two stages: In the SEA Stage, replaces linear projections with CNN layers and feature

slicing, yielding $O\left(K^2M_p^2d + 2M_p^4d_g\right)$. Here, $O\left(K^2M_p^2d\right)$ represents convolutional costs on sliced features M_p, and $(2M_p^4d_g)$ denotes sliced attention computation.

In the CMA Stage, processes global features via CNN-enhanced attention, contributing $O\left(4K^2M^2d + 2N^2d_g\right)$, where $4K^2M^2d$ covers CNN operations and $2N^2d_g$ accounts for channel-wise attention.

Aggregating both stages, SCMA achieves linear spatial-channel complexity:

$$O\left(4K^2d\left(N_p + N\right)\right) + O\left(2Nd\left(N_p + d_g\right)\right) \tag{5}$$

This two-stage design first partitions features into blocks for localized SEA, then reconstructs them for CMA, balancing discriminative power and memory efficiency. In addition to the time complexity analysis in this part, we also verify the inference speed of the model in the experimental part.

4 Experiments

We systematically evaluate GlintNet through three phases: establishing the experimental protocol, conducting comparative analysis with SOTA methods including ablation studies, and verifying feature extraction stability under viewpoint variations and occlusion conditions.

4.1 Experimental Environment

The research platform comprised an Ubuntu 20.04 operating system powered by an NVIDIA RTX3090 GPU featuring 24 GB VRAM. Implementation utilized the PyTorch deep learning framework with Python 3.9 and CUDA 11.3 computational architecture.

4.2 Dataset and Experimental Details

Experiments were conducted on four ReID benchmarks: pedestrian ReID datasets Market1501and MSMT17 [20], and vehicle ReID datasets VeRi-776 [21] and VehicleID [22]. Dataset statistics are summarized in Table 1. These datasets evaluate different capabilities—Market1501 and MSMT17 test pedestrian ReID under occlusion and viewpoint changes, while VeRi-776 and VehicleID assess vehicle ReID across cameras, covering both local and global feature learning.

Training settings: All models were trained for 350 epochs with a batch size of 64. Input resolutions were 256 × 128 for pedestrians and 256 × 256 for vehicles. Optimization used triplet loss and softmax loss with SGD (initial LR = 0.065, warm-up scheduling [23]). Data augmentation included random flipping, cropping, and erasing.

4.3 Evaluation Metrics

In the experiments of this paper, we use Cumulative Matching Characteristics (CMC) and mean Average Precision (mAP) as the main evaluation metrics for the ReID model's performance. Model size is evaluated by the number of parameters, inference speed by inference time and FPS, and feature similarity by cosine similarity.

Table 1. The overiew of the dataset

Dataset	Ids	Images	Camera
Market1501	751	32,217	6
MSMT17	4,101	126,441	15
VeRi-776	776	51,035	20
VehicleID	26,267	221,763	1

Table 2. Comparison with SOTA on Market1501 and MSMT17

Methods	Backbone	Params (M)	Market1501		MSMT17	
			mAP	r1	mAP	r1
ResNet [4]	ResNet50	24+	68.3	85.7	25.7	48
HOReID [5]	ResNet50	24+	84.9	94.2	–	–
MVI2P [24]	ResNet50	24+	87.9	95.3	61.4	83.9
LTReID [25]	ResNet50	24+	89	95.9	58.6	81
HAT [26]	ResNet50	24+	89.5	95.6	61.2	82.3
FastReID [7]	ResNet50	24+	90.3	96.3	63.3	85.1
Nformer [27]	ResNet50	24+	93	95.7	62.2	80.8
DRL-Net [13]	Transformer	41	86.9	94.7	55.3	78.4
DCAL [8]	DeiT-s	88.4	87.5	94.7	64	83.1
PFD [28]	ViT-B	86	89.6	95.5	65.1	82.7
CPT [30]	ViT-B	86	91.9	96.7	68	84.6
MobileNetV4 [16]	MobileNet	2.1	69.5	87	27	50.9
AutoReID [29]	AutoReID	11.4	72.7	89.7	–	–
OSNet [9]	OSNet	2.2	81	93.6	43.3	71
LightMBN [3]	OSNet	9	91.5	96.3	66	80.6
CDNet [15]	CDNet	1.8	83.7	93.7	48.5	73.6
MSINet [16]	MSINet	2.3	89.9	95.5	59.6	81
GlintNet	GlintNet	2.1	91.8	96.4	63.3	82.5

4.4 Compared with the State-of-the-Art Methods

To validate GlintNet, we benchmarked it against SOTA methods on multiple datasets. According to Table 2, GlintNet achieves efficient performance with only 2.1M parameters. It has less than 1/10 of the number of parameters of the ResNet backbone network model and less than 1/40 of the ViT backbone network, yet it achieves 91.8% mAP and 96.4% R1 on Market1501, which is close to the CPT model with 40 times larger

parameters. At MSMT17, it outperforms most lightweight models and even outperforms some of the larger models with 63.3% mAP and 82.5% R1.

Table 3. Comparison with different methods on the VeRi-766 and VehicleID dataset

Methods	VeRi			VehicleID	
	mAP	r1	r5	r1	r5
PNVD [31]	74.3	94.3	98.3	74.2	86.4
SAVER [32]	79.6	96.4	98.6	75.3	88.3
TransReID [12]	82	97.1	–	85.2	97.5
DSAM-GN [33]	82.2	97.4	98.8	81.2	93.9
PartFormer [34]	82.9	97.7	–	86.6	98
GlintNet	82.7	97.3	98.8	84.9	96.8

Table 4. A comparison of the inference time and FPS across different models

Model	Inference time (ms)	FPS
ResNet	10.7	93.17
OSNet	7.9	126.58
CDNet	6.7	149.25
MSINet	7.9	126.87
GlintNet	8.8	112.82

According to Table 3, GlintNet has balanced performance in vehicle ReID. The mAP reaches 82.7% on VeRi and 98.8% on R5. On VehicleID, R1 reaches 84.9% and R5 reaches 96.8%. GlintNet's performance in vehicle ReID is close to the level of heavyweight models, only worse than PartFormer.

According to Table 4, GlintNet balances real-time capability and accuracy With 8.8ms inference time and 112.82 FPS, meeting resource-constrained deployment needs. These results confirm that its lightweight architecture effectively combines discriminative local feature extraction with memory efficiency, achieving broad applicability across person and vehicle ReID tasks.

4.5 Viewpoint Variation and Occlution

As shown in Fig. 5 a multi-view case from Market1501 demonstrates feature consistency across camera perspectives. The target is captured from the rear by Camera 1, the front by Camera 2, and the side by Camera 6. Feature heatmaps reveal consistent attention to shoulder regions, with cosine similarity scores of 81.99% between Camera 1 and Camera 2, and 78.65% between Camera 1 and Camera 6. These results indicate GlintNet's ability to extract viewpoint-robust local features by focusing on stable anatomical regions.

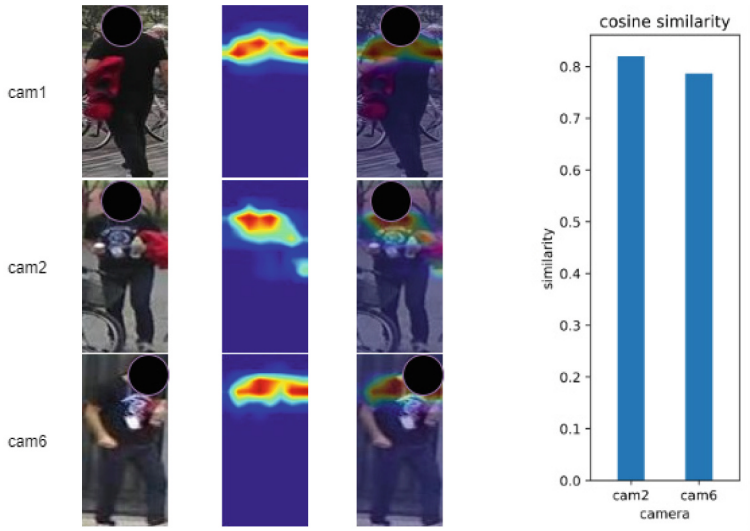

Fig. 5. The left shows original images, feature heatmaps, and their overlays. The right bar chart compares feature similarities between Camera 1 and 2, and Camera 1 and 6.

Fig. 6. Analysis of feature extraction under occlusion. The figure shows feature heatmaps from GlintNet's layers. The first two rows display the same person under varying occlusions.

Figure 6 analyzes the occlusion robustness using the front view image and its occluded counterpart. The model achieves 82.18% feature similarity between the two images. The shallower layers 1–3 prioritize the global context reconstruction to discriminate the main target. While the deeper layers 4–6 refine the discriminative local features to finally confirm the target identity. This global-to-local layered processing ensures accurate identification under occlusion through complementary feature extraction.

4.6 Ablation Experiments

Ablation experiments on Market1501 show that replacing 3 × 3 CNNs with DSCNN reduces parameters by 40.8% but slightly decreases accuracy (mAP by 2.1%, R1 by 0.7%, R5 by 1.3%). Adding CMA further cuts parameters by 12% while improving R1 by 0.7% and R5 by 1.1%, confirming its global modeling ability. Combining SEA and CMA enhances mAP by 1.4% and R5 by 0.8% but slightly reduces R1 by 0.1%, suggesting a trade-off between local feature enhancement and global fusion. Optimizing this balance could improve single-instance matching (Table 5).

Table 5. Ablation experiments

Ablation Experiments	params	flops	mAP	r1	r5
Only CNN	7.1	3.53	91.3	96.5	98.2
conv+conv+ DSCNN	4.2	2.09	89.2	95.8	96.9
conv+CMA+ DSCNN	3.2	1.56	90.4	96.5	98
SEA+CMA+ DSCNN	2.1	1.03	91.8	96.4	98.8

5 Conclusion

This paper presents GlintNet, a lightweight network. It has an efficient SCMA mechanism to deal with occlusion and viewpoint variations in ReID. GlintNet combines local feature extraction based on CNN and global enhancement driven by attention through SCMA with linear complexity. Experiments on four datasets indicate that GlintNet outperforms other lightweight models, getting close to the accuracy of heavyweight models. We hope this work can be helpful for ReID's work.

Acknowledgments. This research was funded by the Key R&D projects of Xinjiang Uyqur Autonomous Region, grant number 2022B01006.

References

1. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.H.: Deep learning for person reidentification: a survey and outlook. IEEE Trans. Pattern Anal. Mach. Intell. **44**(6), 2872–2893 (2022)
2. Zahra, A., Perwaiz, N., Shahzad, M., Fraz, M.M.: Person reidentification: a retrospective on domain specific open challenges and future trends. Pattern Recognit. **142**, 109669 (2023)
3. Herzog, F., Ji, X., Teepe, T., et al.: Lightweight multi-branch network for person reidentification. In: 2021 IEEE International conference on image processing (ICIP), pp. 1129–1133. IEEE (2021)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR abs/1512.03385 (2015)

5. Wang, P., Zhao, Z., Su, F., Zu, X., Boulgouris, N.V.: HOReID: deep high-order mapping enhances pose alignment for person re-identification. IEEE Trans. Image Process. **30**, 2908–2922 (2021)
6. Somers, V., Vleeschouwer, C.D., Alahi, A.: Body part-based representation learning for occluded person re-identification. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, pp. 1613–1623. IEEE (2023)
7. He, L., Liao, X., Liu, W., Liu, X., Cheng, P., Mei, T.: FastReID: a Pytorch toolbox for general instance re-identification. In: Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, pp. 9664–9667. ACM (2023)
8. Zhu, H., Ke, W., Li, D., et al.: Dual cross-attention learning for fine-grained visual categorization and object re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4692–4702 (2022)
9. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Omni-scale feature learning for person re-identification. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, pp. 3701–3711. IEEE (2019)
10. Li, J., Wang, M., Gong, X.: Transformer based multi-grained features for unsupervised person re-identification. In: IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), Waikoloa, pp. 1–9. IEEE (2023)
11. Chen, X., Zheng, X., Lu, X.: Bidirectional interaction network for person re-identification. IEEE Trans. Image Process. **30**, 1935–1948 (2021)
12. He, S., Luo, H., Wang, P., Wang, F., Li, H., Jiang, W.: TransReID: transformer-based object re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15013–15022. IEEE (2021)
13. Jia, M., Cheng, X., Lu, S., Zhang, J.: Learning disentangled representation implicitly via transformer for occluded person re-identification. IEEE Trans. Multimed. **25**, 1294–1305 (2023)
14. Howard, A.G., et al.: MobileNets: efficient convolutional neural networks for mobile vision applications. CoRR abs/1704.04861 (2017)
15. Li, H., Wu, G., Zheng, W.: Combined depth space based architecture search for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, pp. 6729–6738. IEEE (2021)
16. Gu, J., et al.: MSINet: twins contrastive search of multi-scale interaction for object ReID. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, pp. 19243–19253. IEEE (2023)
17. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. CoRR abs/1709.01507 (2017)
18. Chen, H., et al.: AdderNet: do we really need multiplications in deep learning? In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, pp. 1465–1474. IEEE (2020)
19. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, pp. 1116–1124. IEEE (2015)
20. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer GAN to bridge domain gap for person re-identification. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, pp. 79–88. IEEE (2018)
21. Liu, X., Liu, W., Mei, T., Ma, H.: A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 869–884. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_53
22. Liu, H., Tian, Y., Wang, Y., Pang, L., Huang, T.: Deep relative distance learning: tell the difference between similar vehicles. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, pp. 2167–2175. IEEE (2016)

23. Ma, J., Yarats, D.: On the adequacy of untuned warmup for adaptive optimization. In: Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI), pp. 8828–8836. AAAI Press (2021)
24. Dong, N., et al.: Multi-view information integration and propagation for occluded person re-identification. Inf. Fus. **104**, 102201 (2024)
25. Wang, P., et al.: LTreID: factorizable feature generation with independent components for long-tailed person re-identification. IEEE Trans. Multimed. **25**, 4610–4622 (2022)
26. Zhang, G., et al.: HAT: hierarchical aggregation transformers for person re-identification. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 1–10. ACM (2021)
27. Wang, H., et al.: NFormer: Robust person re-identification with neighbor transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–10. IEEE (2022)
28. Wang, T., et al.: Pose-guided feature disentangling for occluded person re-identification based on transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 3, pp. 1–10 (2022)
29. Asghar, H.A., Khan, B., Zafar, Z., Sabri, A.Q.M., Fraz, M.M.: PakVehicle-ReID: a multi-perspective benchmark for vehicle reidentification in unconstrained urban road environment. Multimed. Tools Appl. **83**(17), 53009–53024 (2024)
30. Zhang, Z., He, D., Liu, S., et al.: Completed part transformer for person re-identification. IEEE Trans. Multimed. **26**, 2303–2313 (2023)
31. He, B., Li, J., Zhao, Y., Tian, Y.: Part-regularized near-duplicate vehicle re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, pp. 3997–4005. IEEE (2019)
32. Khorramshahi, P., Peri, N., Chen, Jc., Chellappa, R.: The devil is in the details: self-supervised attention for vehicle re-identification. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) ECCV 2020. LNCS, vol. 12359, pp. 369–386. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58568-6_22
33. Jiao, Y., Qiu, S., Chen, M., Han, D., Li, Q., Lu, Y.: DSAM-GN: graph network based on dynamic similarity adjacency matrices for vehicle re-identification. In: Liu, F., Sadanandan, A.A., Pham, D.N., Mursanto, P., Lukose, D. (eds.) PRICAI 2023. LNCS, vol. 14325, pp. 353–364. Springer, Singapore (2024). https://doi.org/10.1007/978-981-99-7019-3_33
34. Tan, L., et al.: PartFormer: awakening latent diverse representation from vision transformer for object re-identification. arXiv preprint arXiv:2408.16684 (2024)