Pattern Recognition 178 (2026) 113476

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

DiPerceiveNet: A bidirectional cross-scale perception network for vehicle re-identification

Jihao Cai[a], Zhiqiang He[b], Zhi Liu[b], Yangjie Cao[a,*]

[a] Zhengzhou University, Zhengzhou, 450002, China
[b] The University of Electro-Communications, Tokyo, 1828585, Japan

ARTICLE INFO

Keywords:
Vehicle re-identification
Cross-scale feature fusion
Bidirectional interaction
Attention-guided aggregation
Unified representation learning

ABSTRACT

Vehicle Re-identification (ReID) aims to match the same vehicle across non-overlapping cameras, yet it remains highly challenging due to strong inter-vehicle similarity and severe appearance variations caused by viewpoint changes. Most existing methods treat feature extraction and fusion as two disjoint stages - they first extract hierarchical features and then perform vector-level fusion in a one-way manner. This paradigm ignores the rich semantic cues and structural correlations within and across feature layers, leading to weak cross-scale consistency and limited representational alignment. To address these issues, a Dual Interaction Perception Network (DiPerceiveNet) is designed to establish a unified and iterative information flow among hierarchical features. DiPerceiveNet consists of three components: (i) a Residual Multi-Scale Abstraction Pathway (ReMAP) that integrates intermediate feature maps and refines multi-scale information via attention; (ii) a Bidirectional Information Flow (X-Flow) module that enables two-way information propagation and cross-layer interaction for effective multi-scale fusion; and (iii) a Global-Local Attention Mixer (GLoAM) that adaptively re-weights hierarchical features to generate a discriminative embedding. Experiments on VeRi-776, VehicleID, and CityFlow-ReID demonstrate that DiPerceiveNet consistently outperforms recent state-of-the-art methods while maintaining competitive efficiency. Extensive ablation studies and retrieval visualizations further validate the contribution of each component and confirm the effectiveness of the proposed unified fusion paradigm for fine-grained vehicle ReID.

1. Introduction

Vehicle tracking plays a crucial role in Intelligent Transportation Systems (ITS), supporting traffic management and public safety [1]. As a key subtask, vehicle Re-identification (ReID) aims to associate instances of the same vehicle captured by spatially and temporally disjoint cameras. Despite notable advances in deep ReID architectures [2], vehicle ReID still faces challenges inherent to the visual characteristics of vehicles. In practice, a vehicle's appearance can vary drastically across viewpoints [3], while vehicles of identical models may appear nearly indistinguishable when viewed from the same angle [4]. Consequently, two images of the same vehicle captured from distinct viewpoints may appear less similar than two different vehicles observed under an identical viewpoint (Fig. 1). This ambiguity highlights the necessity for features that balance global semantics with fine-grained local cues.

The central challenge lies in learning discriminative representations that remain robust under such ambiguities. The relevance of specific cues depends on visual similarity: for visually distinct categories (e.g., buses vs. sedans), coarse global attributes such as overall shape or size may suffice, whereas for vehicles sharing nearly identical global appearances, subtle local details-such as body decals, license plate regions, or windshield stickers-become indispensable [2]. These observations highlight the necessity of joint global-local representation learning to maintain discriminative power under intra- and inter-vehicle ambiguities.

Most existing vehicle ReID methods share a common two-stage paradigm: they first extract hierarchical features with a backbone and then perform fusion/aggregation at a later stage. Under this paradigm, feature extraction and fusion are largely treated as disjoint processes. As shown in Fig. 2, Existing methods can be broadly categorized into three groups. The first group [5,6] consists of externally supervised approaches that rely on detectors or manual annotations to localize discriminative regions, but they incur heavy annotation costs and scale poorly. The second group [7,8] includes part-based methods that partition images into predefined semantic regions to extract local features; however, rigid partitioning often disrupts semantic consistency and is sensitive to viewpoint changes and occlusion. The third group [9,10] employs multi-branch architectures that separately process global and local cues after backbone extraction, which may under-exploit multi-

[*] Corresponding author.
E-mail address: caoyj@zzu.edu.cn (Y. Cao).

https://doi.org/10.1016/j.patcog.2026.113476
Received 1 November 2025; Received in revised form 4 March 2026; Accepted 9 March 2026
Available online 10 March 2026

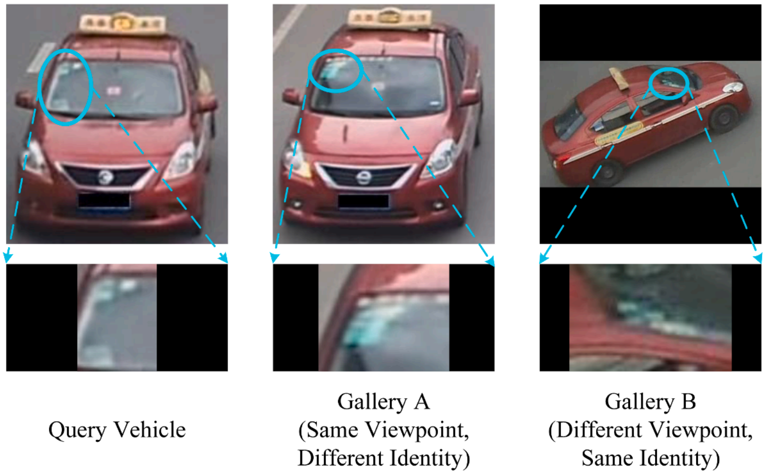
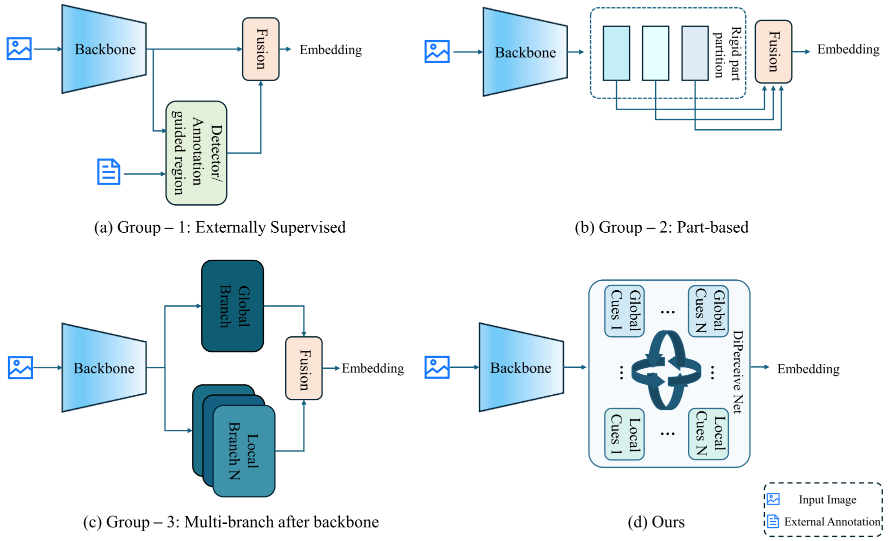


Fig. 1. Challenges in vehicle ReID. The query (left) is compared with: (Gallery A) a different vehicle under the same viewpoint and (Gallery B) the same vehicle under a different viewpoint. The former exhibits high inter-identity similarity, whereas the latter shows large intra-identity variation. These cases highlight the difficulty of distinguishing vehicles based solely on global appearances and the necessity of jointly leveraging both global and local cues. (All images are from the VeRi-776.)



(a) Group – 1: Externally Supervised

(b) Group – 2: Part-based

(c) Group – 3: Multi-branch after backbone

(d) Ours

Fig. 2. Structural comparison of the three vehicle ReID paradigms discussed in the Introduction, highlighting how and where global/local cues are extracted and fused in each design.

level (low-/high-level) complementary cues across the backbone hierarchy [11] and limit cross-branch interaction. As a result, cross-scale correlations cannot be sufficiently propagated during feature extraction [12], motivating a unified framework that enables bidirectional cross-scale interaction during feature learning rather than only after it.

A promising perspective arises from human visual perception, where recognition emerges through dynamic top-down and bottom-up interactions [13]. For instance, a blurry object may first be perceived as “a vehicle” at a coarse level and then refined to a specific identity through attention to details such as contours or logos. This analogy inspires our design, emphasizing dynamic bidirectional information exchange across feature hierarchies. Under the prevalent disjoint extraction–fusion paradigm, however, most existing ReID methods neglect the explicit modeling of such interactions [14], instead enforcing a rigid separation between global and local representations and performing fusion only at the semantic level, which limits hierarchical interplay across scales [15].

To address these limitations, we propose the Dual Interaction Perception Network (DiPerceiveNet), which unifies feature extraction and fusion into a dynamic cross-scale process. DiPerceiveNet comprises three key components: (i) a Residual Multi-Scale Abstraction Pathway (ReMAP) that aggregates intermediate features and refines both fine-grained details and high-level semantics through attention; (ii) a Bidirectional Information Flow (X-Flow) module that establishes two-way propagation and cross-layer interaction, ensuring consistent alignment during both upsampling and downsampling; and (iii) a Global-Local Attention Mixer (GLoAM) that introduces compact self-attention with positional encoding and channel compression to enhance semantic consistency with minimal overhead. Unlike prior multi-branch designs, DiPerceiveNet performs interaction within the backbone hierarchy, allowing feature extraction and fusion to co-evolve rather than remain decoupled.

The main contributions are summarized as follows:

- We present DiPerceiveNet, a unified framework that integrates multi-scale feature extraction and fusion into a single interactive process. Unlike conventional architectures that separately extract and fuse features, DiPerceiveNet enables feature representations to co-evolve through dynamic cross-scale interaction, thus achieving consistent semantic alignment across hierarchical levels.
- The proposed network comprises three components. ReMAP fully exploits shallow layers of the backbone and emphasizes their contribution to vehicle discrimination, refining multi-scale feature ab-

straction through attention-driven enhancement. X-Flow establishes bidirectional information propagation across feature hierarchies, facilitating interaction during both upsampling and downsampling. GLoAM introduces a lightweight global-local attention mechanism with positional encoding and channel compression at the final stage to enhance semantic consistency with minimal computational cost.
- Extensive experiments on VeRi-776 and VehicleID benchmarks demonstrate that DiPerceiveNet achieves superior accuracy and robustness without auxiliary annotations, validating the effectiveness of its cross-layer and cross-scale design.

The remainder of this paper is organized as follows. Section 2 reviews related works on vehicle ReID, focusing on local-global feature extraction, attention mechanisms, and fusion strategies. Section 3 introduces the proposed DiPerceiveNet and its three core modules. Section 4 presents the experimental setup and evaluation on VeRi-776 and VehicleID. Section 4.4 provides ablation studies and visual analyses. Finally, Section 5 concludes the paper.

2. Related work

2.1. Local-global feature extraction

Convolutional neural networks (CNNs) are effective at extracting coarse-grained global features. However, they often fail to capture fine-grained local details [16]. Such details are crucial for distinguishing visually similar vehicles in ReID tasks. To address this limitation, some methods employ external detectors or rely on part-level annotations. They identify key regions such as headlights or license plates [5,17]. Although these approaches improve sensitivity to subtle variations, they require additional modules and supervision signals, which increase computational cost, hinder end-to-end optimization, and make the system vulnerable to detection errors. An alternative is to avoid detectors and annotations altogether. Instead, shallow backbone feature maps can be directly leveraged. These maps inherently encode fine-grained spatial cues associated with structural details. This design eliminates the overhead and fragility of detector-based schemes. At the same time, it preserves sensitivity to local details within a unified framework.

Another line of research adopts parallel branches to extract local and global features separately [15,18]. The goal is to exploit their complementary strengths. Typically, these branches operate only on the backbone's final-layer feature map. As a result, local and global representations are learned in isolation and merged only after extraction. This strategy captures complementary information to some extent. However, it creates a semantic bottleneck by relying solely on final-layer representations. This design restricts access to intermediate and shallow spatial cues [19,20]. Furthermore, treating local and global streams as independent pathways prevents effective interaction and alignment. A more promising strategy is to integrate multi-depth features early in the extraction process. For example, channel and pixel attention can refine shallow, intermediate, and deep feature maps. This enables the capture of both fine-grained details and high-level semantics. Unlike conventional parallel-branch methods, this approach leverages backbone features across multiple depths. It thus obtains local and global semantics in a more comprehensive and hierarchical manner.

These strategies provide a potential solution to the limitations of detector-based and parallel-branch architectures. By eliminating external detection, they reduce unnecessary complexity. By avoiding reliance on the final layer alone, they also improve robustness. Consequently, they yield semantically richer, more spatially aware, and more discriminative representations. This conclusion is supported by recent visualization results and quantitative analyses.

2.2. Attention mechanism for ReID

Current attention-based methods in vehicle ReID can be grouped into three types: channel attention, pixel/spatial attention, and self-attention. Channel attention determines what to focus on by strengthening informative feature channels. Pixel attention determines where to focus by enhancing spatially salient regions. For example, the module [21] in combines channel pooling and spatial pooling to amplify local feature responses. ADPRP-Net [22] applies attention to emphasize key vehicle parts such as emblems and lights. Many of these modules are applied only to the backbone's output feature map. They act as static weighting mechanisms, selectively increasing or suppressing activations according to learned importance scores. By the output stage, aggressive downsampling and semantic abstraction have already reduced fine-grained spatial cues. The low spatial resolution makes these details difficult to recover. In addition, the weighted features are processed by pooling, aggregation, and projection heads. These operations dilute the enhancements and limit their contribution to the final embedding. A more effective strategy is to integrate channel and pixel attention into multi-scale fusion. This allows refinement across shallow, intermediate, and deep features. In this way, locally enhanced cues can propagate to the final representation and remain aligned with global semantics.

Unlike channel or pixel attention, self-attention can model long-range dependencies and capture global context [23]. Recent studies have combined self-attention with convolutional networks to model both local details and global semantics [24,25]. For example, LKA [24] couples self-attention with convolutions. TANet [25] employs text-region self-attention to guide the network toward vehicle body text regions. Transformer-style self-attention has also been explored as a complementary direction in vehicle ReID, e.g., MsKAT [26], which leverages knowledge-aware attention to interact visual features with semantic cues. More recently, masked autoencoding (MAE) [27] has been explored as a pre-training paradigm for vehicle ReID to obtain stronger representations. Although such pre-training-driven methods can yield large improvements, they usually depend on additional resources (e.g., large external pre-training data and/or foundation backbones) beyond the standard supervised setting in this paper. Many existing self-attention modules are introduced in the early stages of the network or within isolated blocks [28]. They operate only on partial features, with positional awareness provided mainly by the backbone or predefined encoding [29]. As a result, the global dependencies they capture do not propagate reliably to the final embedding. Our GLoAM addresses this issue. It introduces a lightweight self-attention mechanism at the final fusion stage. At this point, features are semantically enriched and globally aggregated. This placement enables GLoAM to model holistic dependencies in the final representation. It also incorporates explicit spatial awareness through 2D sine–cosine positional encoding.

Another limitation is computational overhead. Many existing self-attention modules are applied to high-resolution intermediate feature maps. This design greatly increases cost. By contrast, applying self-attention only at the final stage reduces spatial resolution and computation. A lightweight design further mitigates cost while preserving effective global semantic modeling.

2.3. Feature fusion

Bidirectional interaction between top-down and bottom-up flows plays a critical role in the human visual system [30]. The brain first forms a global overview of an object to capture salient cues (coarse perception). It then shifts attention to finer structural details (fine-grained analysis) [31]. During this process, global semantics and local details interact continuously. They complement each other and form stable, robust cognitive representations.

Modern convolutional neural networks (CNNs) often rely on pyramid-like structures for feature construction and fusion [32]. These designs enable multi-scale feature extraction. They also allow limited cross-level information flow through sequential sampling. To improve propagation, some works augment pyramid structures with bidirectional paths [20]. These add connections between layers to support top-down and bottom-up interactions. Such methods enhance information flow to some extent. However, they still treat feature fusion as a static process performed after features are extracted independently. As a result, the integration of global semantics and local details remains fragmented. The fusion process also lacks dynamic adaptability and semantic alignment, unlike human visual perception.

In most existing vehicle ReID methods [33–35], local-global fusion is performed only at the final output stage. Local and global features are extracted independently and then combined by simple concatenation or weighted summation. This paradigm overlooks interactions across depths. It leads to early loss of fine-grained details and semantic misalignment [32]. Consequently, the compressed representation space hinders the recovery of structural cues. It also limits the expressiveness and robustness of learned representations [36].

Several recent works have attempted to address these issues. PEFN [37] enhances feature patches and introduces hierarchical interactions across convolutional layers. Overlock [32] adopts a top-down design to model the transition from coarse perception to fine understanding. Nevertheless, these approaches remain essentially static. They do not explicitly integrate bidirectional interactions during feature construction. A promising direction is to incorporate dynamic bidirectional interactions into feature generation. This design would allow hierarchical features from different depths to complement each other. It could also produce representations that preserve fine-grained details while maintaining global semantic consistency.

3. Method

To integrate feature extraction and fusion into a single dynamic process, we design a unified framework named Dual Interaction Perception Network (DiPerceiveNet). DiPerceiveNet consists of three key modules: ReMAP, X-Flow, and GLoAM. ReMAP captures hierarchical multi-scale features from different backbone depths. It preserves low-level structural details while abstracting high-level semantics. X-Flow enables bidirectional interaction between scales, ensuring cross-scale semantic consistency and spatial correspondence. GLoAM applies a lightweight self-attention mechanism at the final stage. It adaptively balances the contributions of local details and global semantics.

The following subsections present the overall architecture of DiPerceiveNet and the detailed design of each module.

3.1. Dual interaction perception network

DiPerceiveNet is inspired by the hierarchical processing of the human visual system: a coarse global impression is progressively refined through the interplay of high-level semantics and fine-grained evidence. Semantic context supports both global understanding and local interpretation, while fine-grained cues enhance local discrimination and simultaneously refine the global view. Accordingly, DiPerceiveNet unifies feature extraction and fusion into a dynamic process, enabling bidirectional interaction between global semantics and fine-grained details across backbone depths for robust vehicle ReID embeddings. We further interpret DiPerceiveNet as a three-stage functional process: (i) selective encoding of discriminative information from multi-depth features, (ii) bidirectional refinement between high-level semantics and lower-level structures, and (iii) global integration that consolidates the refined evidence into a spatially consistent identity embedding. In our framework, ReMAP, X-Flow, and GLoAM mainly correspond to these three stages, respectively.

As shown in Fig. 3, ReMAP first extracts hierarchical features at multiple depths. X-Flow then aligns semantics across scales and couples information through bidirectional interactions. Finally, GLoAM models long-range dependencies with a lightweight self-attention mechanism and explicit positional encoding. This module adaptively balances the contributions of local and global cues. Through their collaboration, DiPerceiveNet produces semantically enriched and highly discriminative representations.

3.2. Residual multi-scale abstraction pathway

From a perceptual perspective, ReMAP implements the selective encoding stage by extracting multi-depth features and preserving discriminative fine-grained cues via residual multi-path abstraction and lightweight attention. In CNN-based visual tasks, deep layers provide strong semantic abstraction, while shallow layers retain fine-grained structural information. To leverage their complementarity, ReMAP captures multi-scale features from the last three backbone stages through a multi-path residual design. This structure preserves both semantic and structural representations. Residual connections further promote gradient flow and stabilize training.

To suppress redundancy and strengthen discriminability, ReMAP integrates two lightweight attention modules: Residual Pixel Attention (R-PA) and Residual Channel Attention (R-CA), as shown in Fig. 4. R-PA computes pixel-level weights through convolution and ReLU-sigmoid activations. In contrast, R-CA generates channel-wise weights using global pooling followed by convolution and sigmoid activation. Their operations are defined as:

$$x_{\text{R-PA}} = x \otimes \sigma\big(w_2 * \delta(w_1 * x)\big), \tag{1}$$

$$x_{\text{R-CA}} = x \otimes \sigma\big(w * \text{AvgPool}(x)\big), \tag{2}$$

where $*$ denotes convolution, $\delta(\cdot)$ is ReLU, $\sigma(\cdot)$ is sigmoid, w, w_1, w_2 are convolution kernels, and $\otimes$ indicates element-wise multiplication. As illustrated in Fig. 5, shallow layers highlight local details such as headlights and license plates, whereas deeper layers capture semantic structures and overall vehicle shape. This visualization confirms their complementary roles.

For efficient multi-scale aggregation, ReMAP adopts the Efficient Depthwise Context Aggregation Module (EDCAM, Fig. 6). EDCAM combines depthwise separable convolutions with spatial pyramid pooling. This design compresses channels and enhances hierarchical abstraction. It reduces parameters and computational cost while preserving resolution. As a result, EDCAM enables robust fusion of part-level features with minimal overhead-an essential property for vehicle ReID where local scales vary considerably.

In summary, ReMAP leverages residual multi-path connections and attention mechanisms at different backbone depths to capture complementary features. Shallow layers focus on spatial details, while deeper layers emphasize semantic abstraction. This design ensures that both local structural cues and global semantics are effectively preserved for subsequent fusion.

3.3. Bidirectional information flow

From a perceptual perspective, X-Flow implements the bidirectional refinement stage by coupling top-down semantic guidance with bottom-up structural refinement across scales to improve semantic consistency and spatial correspondence. For vehicles with highly similar appearances, subtle local structures are critical for discrimination. In contrast, for vehicles with distinct appearances, global semantic cues dominate recognition. After ReMAP extracts multi-scale features, the key challenge is to fuse them effectively. Simple concatenation or summation is inadequate. Efficient fusion requires dynamic cross-scale interaction to ensure both cross-scale semantic consistency and spatial correspondence.

To address this issue, the X-Flow module establishes bidirectional information flow between feature levels (Fig. 3). In the top-down path, high-level semantics cues guide shallow features, enhancing sensitivity to discriminative regions. In the bottom-up path, shallow details refine high-level semantics by introducing fine-grained cues. Lateral residual connections maintain semantic consistency and spatial correspondence, enabling stable two-way propagation.

To improve efficiency, X-Flow integrates the Dual-Path Feature Extraction and Compression (DPFEC) module (Fig. 7). DPFEC contains two parallel paths. The main path employs residual 3×3 convolutions,

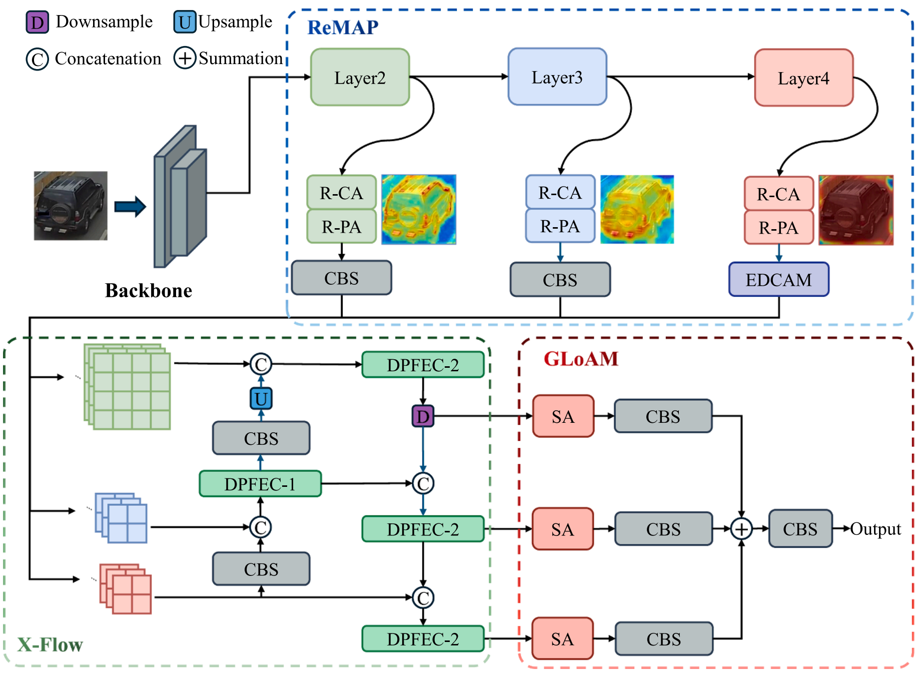
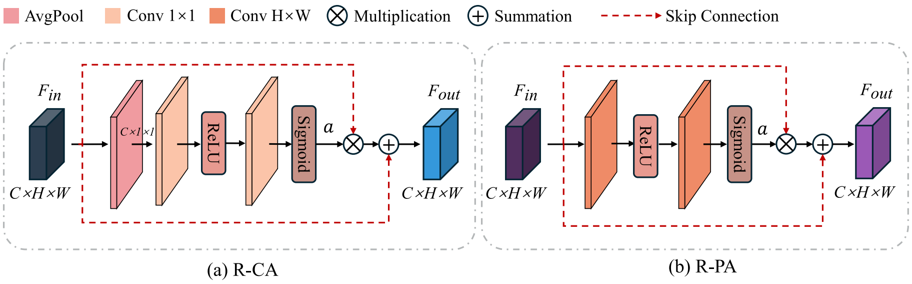


Fig. 3. Architecture of the proposed dual interaction perception network (DiPerceiveNet). The framework consists of three modules: ReMAP, X-Flow, and GLoAM. ReMAP extracts hierarchical multi-scale features from different backbone depths while preserving structural details. It integrates Residual Channel Attention (R-CA), Residual Pixel Attention (R-PA), Convolution-BatchNorm-SiLU (CBS), the Efficient Depthwise Context Aggregation Module (EDCAM), and the Dual-Path Feature Extraction and Compression (DPFEC) block. X-Flow establishes semantic correspondences across scales and enables bidirectional interaction through top-down and bottom-up pathways. GLoAM applies a lightweight self-attention mechanism with positional encoding to enhance global semantic consistency. Together, these components progressively fuse local and global features, yielding semantically rich and discriminative vehicle representations. From a mechanistic perspective, ReMAP performs selective detail encoding, X-Flow conducts bidirectional cross-scale refinement (top-down guidance and bottom-up correction), and GLoAM integrates the refined evidence via lightweight global dependency modeling to produce the final embedding.



(a) R-CA

(b) R-PA

Fig. 4. Structures of the Residual Channel Attention (R-CA) and Residual Pixel Attention (R-PA) modules. R-CA applies global pooling, followed by convolution and sigmoid activation, to generate channel-wise weights. R-PA employs convolutional layers and ReLU and Sigmoid activations to compute pixel-level attention weights. Both modules enhance discriminative capability by emphasizing salient channels or spatial regions.

whose outputs are concatenated and compressed with a 1×1 convolution to capture features across multiple receptive fields. The auxiliary path applies a direct 1×1 convolution for low-cost channel compression and information distillation. The two paths are merged through residual addition, which enhances discriminability while keeping computational cost low.

To accommodate different semantic levels, two DPFEC variants are designed. DPFEC-1, used in the top-down path, adopts deeper residual stacks to preserve rich semantics from higher layers. DPFEC-2, applied in the bottom-up path, uses a lighter residual stack to suppress redundancy in shallow features while retaining fine-grained structural cues. By combining bidirectional interaction with these tailored DPFEC variants, X-Flow achieves cross-scale semantic consistency and robust cross-scale fusion for vehicle ReID. Importantly, this dual-variant design highlights the adaptive nature of X-Flow, as each path is optimized for the distinct roles of semantic abstraction and structural refinement across different feature levels.

3.4. Global-local attention mixer

From a perceptual perspective, GLoAM implements the global integration stage by modeling long-range dependencies over the fused feature maps to form a more coherent and semantically consistent embedding. To balance the roles of local and global cues in the final vehicle embedding, we introduce the Global-Local Attention Mixer (GLoAM). GLoAM employs a lightweight single-head self-attention mechanism with 2D sine-cosine positional encoding. This design enables the modeling of long-range spatial dependencies across fused feature maps, as illustrated in Fig. 8. These maps carry complementary information: some highlight fine-grained local details, while others emphasize coarse global semantics. By attending jointly to all positions, GLoAM adaptively integrates local and global features according to each vehicle instance, yielding a more discriminative and semantically consistent representation.

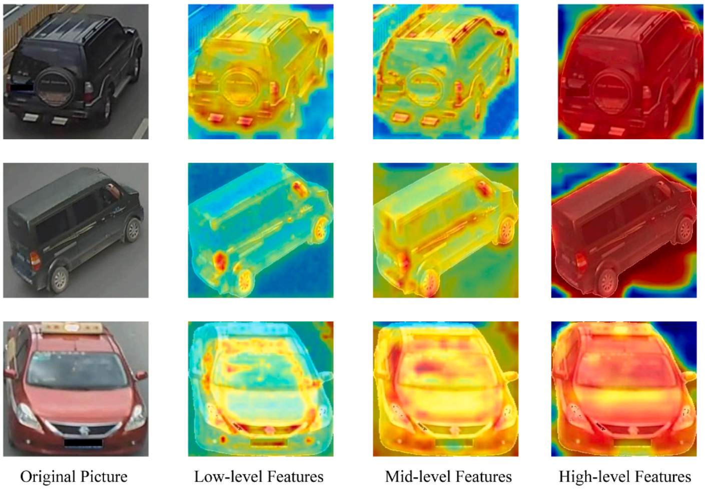


Fig. 5. Heatmaps of attention-enhanced feature maps at different backbone depths. Shallow layers focus on fine-grained details such as headlights, license plates, and contours. Deeper layers emphasize semantic structures and global shape. These visualizations demonstrate the complementarity of shallow and deep features in ReMAP.

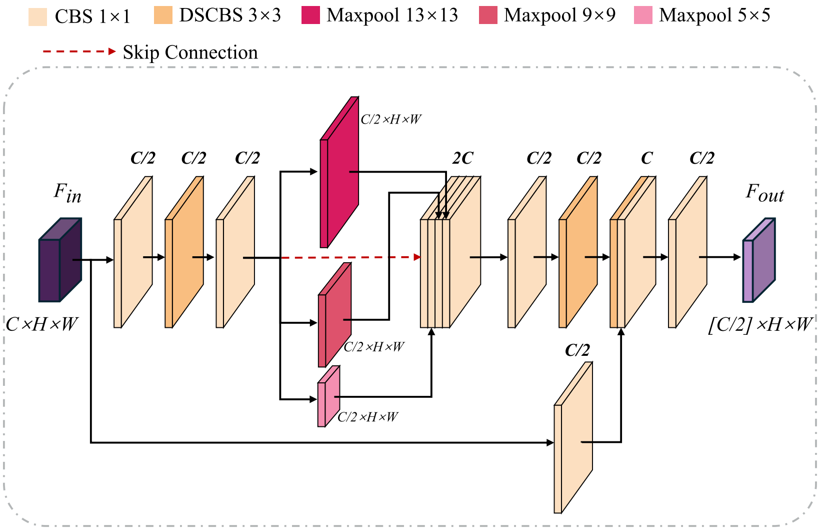
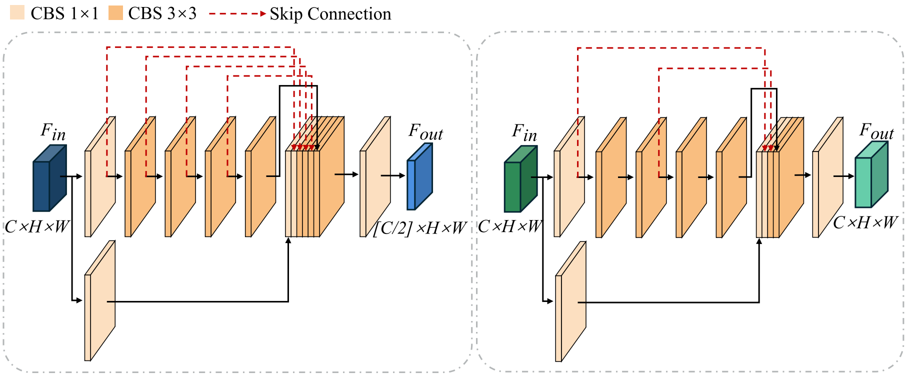


Fig. 6. Structure of the Efficient Depthwise Context Aggregation Module (EDCAM). EDCAM integrates depthwise separable convolution with spatial pyramid pooling (denoted as DSCBS) to efficiently compress, enhance, and fuse multi-scale features while preserving resolution. This design achieves effective hierarchical aggregation with minimal computational cost, thereby supporting robust feature learning for vehicle ReID.



(a) DPFEC-1

(b) DPFEC-2

Fig. 7. Structures of the Dual-Path Feature Extraction and Compression (DPFEC) module and its variants. DPFEC employs a dual-path design: a main path with residual 3×3 convolutions, concatenation, and compression, and an auxiliary path with a direct 1×1 convolution for low-cost information distillation. Two variants are instantiated: DPFEC-1 processes deep semantic features in the top-down path, while DPFEC-2 refines shallow spatial features in the bottom-up path. Each variant is tailored to the characteristics of its respective semantic level. This design supports compact, efficient, and discriminative feature enhancement for multi-scale fusion.

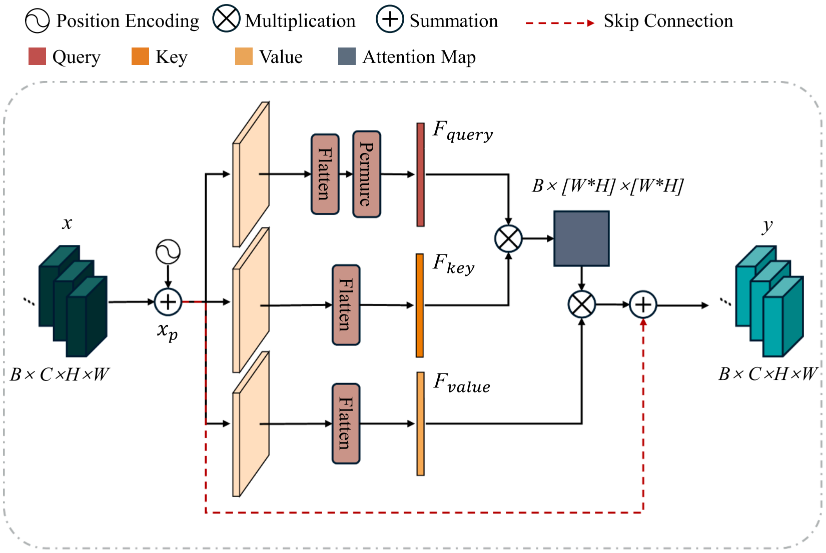


Fig. 8. Structure of the self-attention mechanism in the Global-Local Attention Mixer (GLoAM). GLoAM applies a lightweight single-head self-attention mechanism to the fused feature maps, incorporating 2D sine-cosine positional encoding for spatial awareness. This design models long-range dependencies across feature maps and adaptively balances the contributions of local and global information in the final embedding.

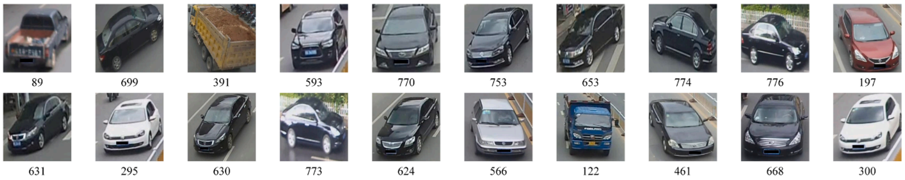
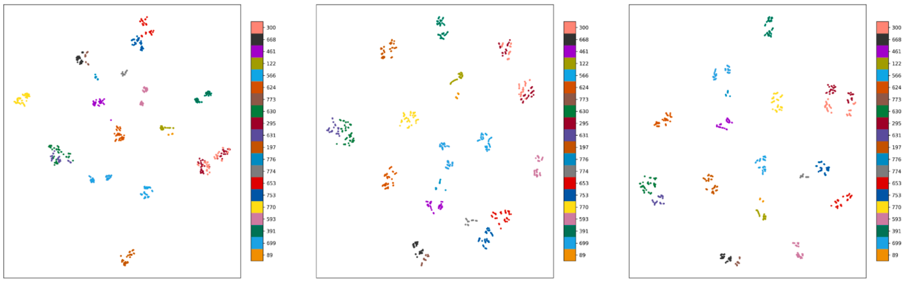


Fig. 9. Vehicle samples from 20 identities in the VeRi-776 dataset, including several visually similar vehicles.



(a) Resnet-50

(b) Baseline

(c) DiPerceiveNet

Fig. 10. t-SNE visualization of feature distributions for 892 samples from 20 PIDs in VeRi-776. (a) ResNet-50, (b) Baseline, and (c) DiPerceiveNet. All models output 2,048-dimensional features.

Given an input feature map $x \in \mathbb{R}^{B\times C\times H\times W}$, where B denotes the batch size, C the number of channels, and H, W the spatial dimensions, we first add 2D sine-cosine positional encoding p to provide explicit positional awareness:

$$x_p = x + p, \quad p \in \mathbb{R}^{1\times C\times H\times W}. \tag{3}$$

Three 1×1 convolutions are then applied to generate the query, key, and value feature maps:

$$q = w_q * x_p, \quad k = w_k * x_p, \quad v = w_v * x_p, \tag{4}$$

where $q, k, v \in \mathbb{R}^{B\times C\times H\times W}$.

The spatial dimensions are flattened, and the attention matrix is computed:

$$q' = \text{reshape}(q) \in \mathbb{R}^{B\times N\times C}, \quad k' = \text{reshape}(k) \in \mathbb{R}^{B\times C\times N}, \tag{5}$$

$$a = \text{Softmax}(q' \cdot k'), \tag{6}$$

where $N = H \times W$ is the number of spatial positions and $a \in \mathbb{R}^{B\times N\times N}$ is the spatial attention matrix. Since $N = H \times W$ and the attention map is of size $N \times N$, the computational and memory cost of self-attention increases substantially (approximately quadratically) as the input resolution grows. Although our attention operation is applied only at the final stage with a relatively small spatial size, higher-resolution inputs may still incur additional overhead in practice.

The output feature map is obtained by applying the attention matrix to the reshaped value features:

$$v' = \text{reshape}(v) \in \mathbb{R}^{B\times C\times N}, \tag{7}$$

$$o = \text{reshape}(v' \cdot a^{\top}) \in \mathbb{R}^{B\times C\times H\times W}. \tag{8}$$

Finally, a residual connection scaled by a learnable parameter γ produces the enhanced output:

$$y = \gamma \cdot o + x_p, \tag{9}$$

where γ controls the strength of attention refinement. In summary, GLoAM integrates lightweight self-attention with positional encoding to capture long-range dependencies across fused features. By adaptively weighting local details and global semantics, it generates compact,

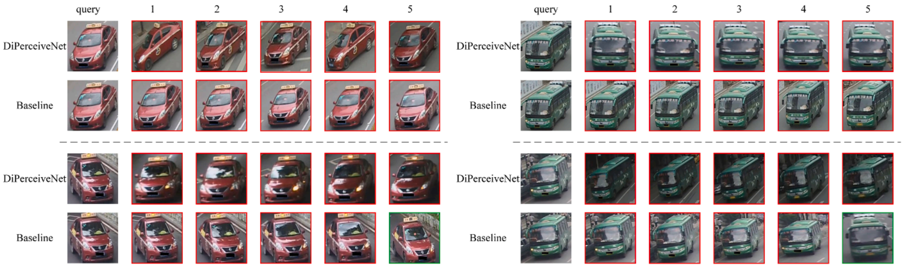


Fig. 11. Qualitative retrieval results on VeRi-776. DiPerceiveNet accurately distinguishes vehicles with subtle appearance differences.

Fig. 12. Visualization of the constructed hard subset on VeRi-776. The hard subset is a query-only subset composed of visually confusing vehicle images selected by a small positive–negative margin ($\Delta = d^{-} - d^{+}$) under the standard evaluation protocol, while the gallery remains unchanged.

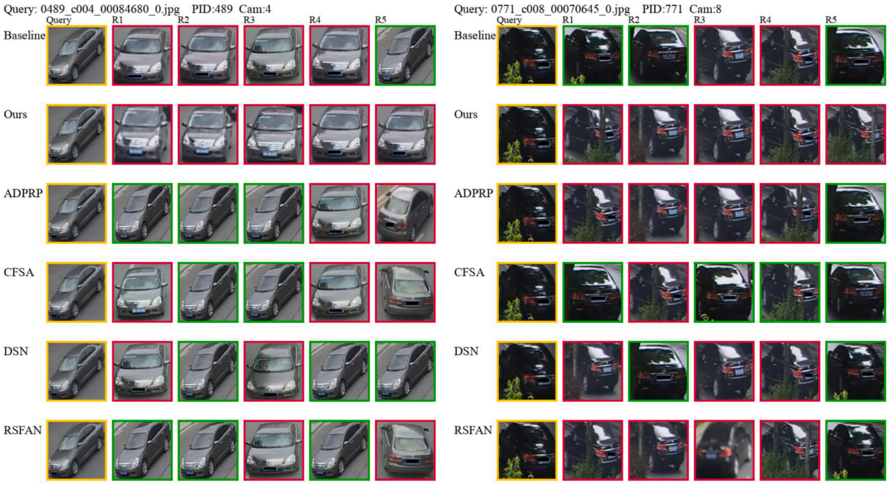

Query: 0489_c004_00084680_0.jpg PID:489 Cam:4
Query
R1
R2
R3
R4
R5
Baseline
Ours
ADPRP
CFSA
DSN
RSFAN
Query: 0771_c008_00070645_0.jpg PID:771 Cam:8
Query
R1
R2
R3
R4
R5
Baseline
Ours
ADPRP
CFSA
DSN
RSFAN

Fig. 13. Representative Rank-10 retrieval examples on the VeRi-776 hard subset (PID 295/142/489). Red boxes indicate correct matches and green boxes denote incorrect matches. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

discriminative, and semantically consistent embeddings without incurring significant computational cost.

4. Experiments

4.1. Datasets and evaluation metrics

Datasets. We evaluate on three standard vehicle ReID benchmarks: VeRi-776 [38], VehicleID [39], and CityFlow-ReID [40]. VeRi-776 contains 50K+ images of 776 vehicle IDs captured by 20 non-overlapping cameras (train/test: 37,781/576 and 11,579/200), with limited color labels that increase fine-grained ambiguity. VehicleID is a large-scale dataset with 221,763 images of 26,267 IDs, evaluated on three subsets (Small/Medium/Large: 800/1600/2400 IDs). CityFlow-ReID includes 52,717 images of 440 IDs with trajectory information; following prior practice, we perform an offline split with 360 IDs for training and 80 for testing.

Evaluation protocols. Following standard practice [41], we report CMC/Rank-k and mAP. For VeRi-776 and CityFlow-ReID, we adopt image-to-track evaluation with mAP/Rank-1/Rank-5 [38]; for VehicleID, we use image-to-image evaluation and report Rank-1/Rank-5 on Small/Medium/Large, averaging over 10 random splits. Unless otherwise stated, we report results without re-ranking or spatiotemporal constraints.

Table 1
Comparison with state-of-the-art methods on VeRi-776. Best results among reproduced CNN-based methods are highlighted in **bold**. Methods marked with †/‡ denote different training assumptions: † CLIP-based methods with large-scale diffusion-generated external pre-training data, and ‡ ViT-based methods. Results marked with * are reported results from the corresponding papers. Under the standard supervised single-modality setting without external pre-training data, DiPerceiveNet achieves the best performance among the reproduced CNN-based methods.

Method	mAP (%)	Rank-1 (%)	Rank-5 (%)
AIVR-Net [43]	64.3	87.1	94.3
URR-Net [5]	73.9	93.7	96.5
GLSIPNet [44]	81.0	95.7	97.6
SRaCR [45]	82.3	97.1	98.4
TANet [25]	83.6	96.8	98.5
RSFAN [17]	83.9	97.2	98.4
DSN [46]	78.6	95.1	97.3
VehicleGAN [47]	79.4	94.7	97.5
LKA [24]	84.1	96.1	98.4
CFSA [48]	79.1	94.7	97.3
ADPRP-Net [22]	82.8	95.6	98.7
DiPerceiveNet (Ours)	**85.0**	**97.7**	**98.9**
VehicleMAE [27]†	87.6*	97.4*	N/R
Git [49]‡	78.9*	95.8*	N/R
SeCap [50]‡	81.3	96.1	97.9

4.2. Implementation details

We use ImageNet-pretrained ResNet-50 (last stride = 1 [42]) as backbone. Images are resized to 384×384 with random horizontal flip and Random Erasing (p=0.5; area ratio $[0.02, 0.33]$, aspect ratio $[0.3, 3.3]$; erase value = $[0.485, 0.456, 0.406]$). Mini-batch: P=4 IDs and K=8 images per ID (32 images). Loss: cross-entropy + triplet (margin 0.3). Optimizer: AdamW (betas $(0.9, 0.999)$; weight decay 7×10^{-4}, bias decay 5×10^{-4}) for 120 epochs; LR warms from 3×10^{-5} to 3×10^{-3} in 15 epochs, then decays by 5× at epochs 50 and 90. We extract 2048-D features; for VeRi-776 we use official tracklets and mean-pool gallery features per tracklet, with Euclidean distance for matching. We use SiLU in CBS blocks; replacing it with ReLU gives similar but slightly lower results on VeRi-776 (85.0/97.7 vs. 84.8/97.2 mAP/Rank-1). Params/FLOPs are computed on the full model using fvcore at 384×384; higher input resolutions will introduce additional computation and memory overhead, especially for attention-based components. Implementation uses PyTorch 2.2.2 with CUDA 12.1 on an NVIDIA RTX 3090 (24GB).

4.3. Comparison with state-of-the-art methods

Results on VeRi-776. Table 1 shows that DiPerceiveNet attains **85.0%** mAP, **97.7%** Rank-1, and **98.9%** Rank-5, surpassing recent approaches (e.g., +2.2 mAP / +2.1 Rank-1 over ADPRP-Net). Compared with methods requiring auxiliary cues (e.g., attributes or detectors), our model is annotation-free yet more accurate.

Results on VehicleID. On Small/Medium/Large, DiPerceiveNet achieves **84.6/80.1/77.8** Rank-1 and **97.5/96.5/95.3** Rank-5 (Table 2), consistently outperforming prior methods, including attention- and multi-branch-based designs. This confirms good scalability under increasing gallery sizes.

Table 2
Comparison with state-of-the-art methods on VehicleID. Best results are highlighted in **bold**.

Method	Small		Medium		Large	
	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5
AIVR-Net [43]	67.7	87.9	61.5	82.7	54.5	77.2
URR-Net [5]	83.5	96.7	80.8	96.3	77.4	95.0
VAAG [51]	77.8	91.8	74.4	89.0	73.1	84.5
SRaCR [45]	83.1	96.5	77.6	95.7	73.8	93.3
TANet [25]	83.9	96.7	78.4	96.1	75.2	94.3
RSFAN [17]	84.0	96.7	78.3	96.1	76.8	94.5
DSN [46]	82.9	96.3	77.1	95.3	73.2	93.5
MIMA-Net [28]	82.9	95.6	78.3	90.8	75.2	90.6
HCINet [52]	83.1	95.8	78.7	91.1	75.6	90.9
GLSIPNet [44]	83.7	95.3	78.6	92.7	76.1	91.3
ADPRP-Net [22]	83.9	96.5	79.8	93.7	76.7	91.9
DiPerceiveNet (Ours)	**84.6**	**97.5**	**80.1**	**96.5**	**77.8**	**95.3**

Table 3
Comparison with state-of-the-art methods on CityFlow-ReID. Best results are highlighted in **bold**.

Method	mAP (%)	Rank-1 (%)	Rank-5 (%)
GLSIPNet [44]	74.6	83.7	89.3
SRaCR [45]	73.2	82.7	87.2
RSFAN [17]	76.1	85.3	90.4
GLSIPNet [44]	75.8	84.3	89.8
LKA [24]	76.6	84.9	91.7
ADPRP-Net [22]	76.3	85.1	90.7
DiPerceiveNet (Ours)	**80.7**	**88.7**	**92.5**

Table 4
Ablation study of DiPerceiveNet on VeRi-776.

Variant	mAP (%)	Rank-1 (%)	Rank-5 (%)
Baseline	78.3	93.9	96.7
+ ReMAP	80.8	95.6	97.8
+ ReMAP + X-Flow	83.2	97.1	98.3
+ ReMAP + X-Flow + GLoAM (Ours)	**85.0**	**97.7**	**98.9**

Table 5
Component-level ablation of ReMAP.

Variant	mAP (%)	Rank-1 (%)	Rank-5 (%)
Full ReMAP (ours)	**85.0**	**97.7**	**98.9**
w/o multi-depth feature paths	83.9	97.3	98.4
w/o R-C&PA module	83.7	97.2	98.4
w/o EDCAM	84.3	97.5	98.6

Table 6
Component-level ablation of X-Flow.

Variant	mAP (%)	Rank-1 (%)	Rank-5 (%)
Full X-Flow (ours)	**85.0**	**97.7**	**98.9**
w/o DPFEC	84.5	97.5	98.7
w/o bidirectional interaction paths	83.1	97.1	98.2

Results on CityFlow-ReID. Table 3 shows that DiPerceiveNet achieves **80.7%** mAP, **88.7%** Rank-1, and **92.5%** Rank-5 on CityFlow-ReID, outperforming recent methods (e.g., +4.4 mAP / +3.6 Rank-1 over ADPRP-Net). Since CityFlow-ReID contains more severe occlusions and larger viewpoint variations, these gains indicate that our model learns more discriminative identity representations rather than overfitting to a particular viewpoint pattern, leading to more robust cross-camera retrieval.

4.4. Ablation study

DiPerceiveNet contains three core modules: the Residual Multi-Scale Abstraction Pathway (ReMAP), the Bidirectional Information Flow (X-Flow), and the Global-Local Attention Mixer (GLoAM). To validate their effectiveness, we conduct ablation studies on the VeRi-776 dataset with ResNet-50 as the backbone, evaluating both module-level and component-level contributions.

Table 7
Sensitivity to the number of interacting scales in ReMAP/X-Flow. In our design, the number of ReMAP paths is coupled with the number of cross-scale interaction loops in X-Flow; thus we vary the number of participating backbone stages for analysis.

Setting	Stages	Params (M)	GFLOPs	mAP (%)	Rank-1 (%)
2-scale	Layer3 + Layer4	148.28	93.44	82.5	95.8
3-scale (Ours)	Layer2 + Layer3 + Layer4	201.47	126.78	**85.0**	**97.7**
4-scale	Layer1 + Layer2 + Layer3 + Layer4	252.13	156.19	83.9	97.2

Table 8
Component-level ablation of GLoAM. FPS and Peak GPU memory are measured with forward-only inference (excluding data loading and preprocessing) under the same hardware/software setting.

Model	Params (M)	GFLOPs	FPS	Peak Mem (MB)	mAP (%)	Rank-1 (%)
w/o GLoAM	194.59	121.47	39.63	957.8	83.2	97.1
Full (ours)	201.47	126.78	38.19	999.4	**85.0**	**97.7**
w/ ViT-style Block	260.69	161.95	28.9	1289.4	84.8	97.7

Table 9
Comparison with state-of-the-art methods on hard-subset.

Method	mAP (%)	Rank-1 (%)	Rank-5 (%)
Baseline	54.2	74.9	85.1
ADPRP-Net [22]	75.1	90.5	91.7
CFSA [22]	73.7	89.1	91.3
DSN [22]	72.3	88.6	90.6
RSFAN [22]	77.5	91.9	93.1
DiPerceiveNet (Ours)	**82.3**	**92.6**	**95.2**

Table 10
Ablation study of DiPerceiveNet on hard-subset.

Variant	mAP (%)	Rank-1 (%)	Rank-5 (%)
Baseline	54.2	74.9	85.1
+ ReMAP	71.7	81.3	90.3
+ ReMAP + X-Flow	78.4	89.6	93.2
+ ReMAP + X-Flow + GLoAM (Ours)	**82.3**	**92.6**	**95.2**

4.4.1. Effect of ReMAP, X-Flow, and GLoAM

Table 4 presents the progressive integration of the proposed modules. Starting from the baseline, ReMAP raises mAP by 2.5%, X-Flow further improves Rank-1 to 97.1%, and GLoAM achieves the best results (85.0% mAP, 97.7% Rank-1). Each module consistently contributes to performance gains, and their combination yields the highest accuracy.

4.4.2. Ablation within the ReMAP module

We further analyze ReMAP's internal designs under the full DiPerceiveNet. As shown in Table 5, removing multi-depth feature paths (*w/o multi-depth*) decreases mAP by 1.1%, verifying the value of hierarchical features. Excluding the R-C&PA module (*w/o R-C&PA*) or replacing ED-CAM with standard pyramid pooling both reduce accuracy, confirming that ReMAP's attention-driven refinement and contextual aggregation jointly enhance discriminative capability.

4.4.3. Ablation within the X-Flow module

As shown in Table 6, replacing DPFEC with conventional Conv-BN-ReLU layers (*w/o DPFEC*) slightly reduces mAP to 84.5%. Disabling bidirectional interaction paths (*w/o bidirectional*) causes a larger drop (83.1% mAP, 97.1% Rank-1), indicating the importance of vertical information exchange for robust cross-scale alignment.

4.4.4. Sensitivity to the number of interacting scales

In DiPerceiveNet, the number of ReMAP paths is coupled with the number of interaction loops in X-Flow; thus, involving more backbone stages naturally expands the bidirectional interaction structure. We therefore perform an architecture-level sensitivity analysis by varying the participating stages (2/3/4-scale), as summarized in Table 7. Using 2-scale (Layer3 + Layer4) reduces parameters and computation but causes a clear mAP/Rank-1 drop, suggesting that relying only on deeper features weakens fine-grained cues. Expanding to 4-scale (Layer1 + Layer2 + Layer3 + Layer4) substantially increases complexity, and the additional shallow details may introduce noise, leading to degraded performance. Overall, 3-scale (Layer2 + Layer3 + Layer4) offers the best accuracy–complexity balance and is adopted as default.

4.4.5. Ablation of the GLoAM module

Table 8 compares GLoAM with alternative designs in terms of accuracy and efficiency. Removing GLoAM leads to a clear performance drop (83.2% mAP), confirming the importance of final-stage global–local mixing.

To benchmark against standard heavy-weight designs, we compare GLoAM with a “ViT-style Block” (consisting of full Multi-Head Self-Attention, FFN, and LayerNorm). Results show that the ViT-style Block incurs significantly higher computational costs (260.69M params, 161.95 GFLOPs, and only 28.9 FPS) due to its redundant components (e.g., FFN). In contrast, our lightweight GLoAM achieves higher accuracy (85.0% mAP) with much lower overhead (201.47M params, 38.19 FPS). This confirms that GLoAM provides a superior accuracy–efficiency trade-off by stripping away unnecessary complexity while retaining essential global dependency modeling.

4.4.6. Qualitative analysis of feature representation and retrieval

To further illustrate the discriminative power of DiPerceiveNet, we visualize both the selected vehicle samples and their feature distributions on the VeRi-776 dataset. Fig. 9 shows 20 representative vehicle identities, intentionally including many visually similar categories (e.g., vehicles with the same color or nearly identical models). The corresponding t-SNE distributions in Fig. 10 compare ResNet-50, the baseline, and DiPerceiveNet using 892 samples from these 20 PIDs.

Compared with ResNet-50 and the baseline, DiPerceiveNet forms more compact intra-class clusters and larger inter-class margins, even among visually similar vehicles shown in Fig. 9. This confirms its superiority in handling small inter-class variation and viewpoint changes, where conventional methods tend to confuse nearly identical vehicles.

Additionally, Fig. 11 presents retrieval examples under challenging cases such as visually identical taxis and buses. While the baseline often ranks samples with similar viewpoints, DiPerceiveNet correctly identifies the target by capturing subtle yet discriminative local cues (e.g., window reflections, roof attachments). These results highlight DiPerceiveNet's ability to integrate global and local features for fine-grained vehicle ReID.

4.4.7. Hard-subset analysis on VeRi-776

Beyond standard retrieval metrics, we provide a qualitative analysis to examine how effectively our method addresses the core challenge in Fig. 1, i.e., distinguishing vehicles with highly similar global appearances where identity discrimination relies on subtle local differences. We construct a hard subset on the VeRi-776 test split and visualize representative retrieval results.

Hard subset construction. As shown in Fig. 12, the hard subset is a query-only subset consisting of visually confusing queries, while the gallery remains unchanged to ensure a fair and reproducible comparison. We first use a baseline model to extract features for all queries and gallery images and compute the query-to-gallery distance matrix. Under the standard VeRi-776 protocol, gallery images that share the same identity and camera with the query are discarded. For each query, we compute (i) the distance to the closest positive sample, d^+, and (ii) the distance to the closest negative sample, d^-, and define the hardness margin:

$$\Delta = d^- - d^+. \tag{10}$$

We sort queries by ascending Δ and select the top-K smallest-margin queries. We set $K = \lfloor r \cdot Q_{\text{valid}} \rfloor$ with $r = 0.10$, where Q_{valid} denotes the number of valid queries after protocol-based filtering (thus $K = 167$ for VeRi-776). This procedure is fully automatic and deterministic given the baseline distance matrix, directly targeting the ambiguity pattern in Fig. 1.

Quantitative results and qualitative observations. We evaluate different methods on this hard subset using the same gallery and protocol, including the baseline, ours, ADPRP [22], CFSA [48], DSN [46], and RSFAN [17]. In addition to visualization, we report quantitative results (mAP/Rank-1/Rank-5) on the constructed hard subset under the standard VeRi-776 protocol, where the gallery remains unchanged for fair comparison. As shown in Table 9, DiPerceiveNet achieves clear gains over the baseline on hard queries. Moreover, Table 10 shows that each proposed module consistently improves Hard-mAP/Hard-R1/Hard-R5, further confirming their contributions on challenging cases. These results align with our motivation: by unifying feature extraction and fusion via iterative cross-scale interaction, DiPerceiveNet better preserves and aligns complementary cues across depths, improving robustness on fine-grained hard cases. Fig. 13 further provides a representative Rank-5 example, where our method retrieves the correct match at Rank-1 while competing methods exhibit confusions under highly similar global cues (e.g., body color and silhouette).

5. Conclusion

This paper presents DiPerceiveNet for vehicle Re-Identification, drawing inspiration from the dynamic perceptual processes of the human visual system. DiPerceiveNet unifies top-down semantics and bottom-up details within a single perception framework, jointly modeling global cues and fine-grained local information. The framework consists of three key modules: ReMAP for enhancing multi-scale feature extraction, X-Flow for enabling bidirectional cross-layer interaction, and GLoAM for compact global-local attention modeling. Experiments on standard vehicle Re-Identification benchmarks demonstrate that the proposed perception-inspired architecture yields more unified and discriminative representations, alleviating fine-grained confusion among visually similar vehicles.

Although our primary goal is to enhance discriminability in these challenging scenarios, we recognize that practical applicability necessitates a careful balance between accuracy and efficiency. Consequently, future efforts will target leaner, more efficient variants that preserve the core perception-driven behavior at lower computational overhead. Ultimately, we hope this work inspires future designs building on our perception-inspired representation learning for vehicle Re-Identification, with deployability as a key next step.

CRediT authorship contribution statement

Jihao Cai: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Conceptualization; **Zhiqiang He:** Writing – review & editing, Data curation; **Zhi Liu:** Writing – review & editing, Supervision, Project administration; **Yangjie Cao:** Writing – review & editing, Supervision, Resources, Funding acquisition.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] X. Yi, Q. Wang, Q. Liu, Y. Rui, B. Ran, Advances in vehicle re-identification techniques: a survey, Neurocomputing 614 (2025) 128745.

[2] T. Ma, K. Sun, X. Pang, W. Si, T. Liu, C. Wang, et al., Multi-axis compression fusion network for vehicle re-identification: T. Ma et al., Sci. Rep. 15 (1) (2025) 30541.

[3] B. Li, P. Liu, L. Fu, J. Li, J. Fang, Z. Xu, H. Yu, Enhancing vehicle re-identification by pair-flexible pose guided vehicle image synthesis, Green Energy Intell. Transp. 4 (5) (2025) 100269.

[4] Y.-J. Sun, L.-W. Qiao, S. Ji, AG-GCN: vehicle re-identification based on attention-guided graph convolutional network, Comput. Mater. Continua 84 (1) (2025) 1769-1785.

[5] J. Qian, M. Pan, W. Tong, R. Law, E.Q. Wu, URRNet: a unified relational reasoning network for vehicle re-identification, IEEE Trans. Veh. Technol. 72 (9) (2023) 11156–11168.

[6] B. Ashutosh Holla, M.m. Manohara Pai, U. Verma, R.M. Pai, MSFFT: multi-scale feature fusion transformer for cross platform vehicle re-identification, Neurocomputing 582 (2024) 127514.

[7] J. Lian, D.-H. Wang, Y. Wu, S. Zhu, Multi-branch enhanced discriminative network for vehicle re-identification, IEEE Trans. Intell. Transp. Syst. 25 (2) (2023) 1263–1274.

[8] Z. Sun, X. Nie, X. Bi, S. Wang, Y. Yin, Detail enhancement-based vehicle re-identification with orientation-guided re-ranking, Pattern Recognit. 137 (2023) 109304.

[9] Y. Wang, R. Li, Y. Shao, Vehicle re-identification method based on efficient self-attention CNN-transformer and multi-task learning optimization, Sensors 25 (10) (2025) 2977.

[10] X. Pang, X. Tian, X. Nie, Y. Yin, G. Jiang, Vehicle re-identification based on grouping aggregation attention and cross-part interaction, J. Vis. Commun. Image Represent. 97 (2023) 103937.

[11] H. Zhang, X. Chen, H. Yu, K.L. Teo, Diversified distillation fusion network for vehicle re-identification, Expert Syst. Appl. 283 (2025) 127708.

[12] H. Zheng, M. Zhang, M. Gong, A.K. Qin, T. Liu, F. Jiang, Multi-scale hierarchical feature fusion network for change detection, Pattern Recognit. 161 (2025) 111266.

[13] Q. Li, B. Cui, P. Liu, Y. Zhu, Image quality assessment characteristics for super-resolution, in: 2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML), IEEE, 2024, pp. 294–298.

[14] S. Liu, S.S. Agaian, VERI-D: a new dataset and method for multi-camera vehicle re-identification of damaged cars under varying lighting conditions, APL Mach. Learn. 2 (1) (2024).

[15] B. Ashutosh Holla, M.M.M. Pai, U. Verma, R.M. Pai, Vehicle re-identification and tracking: algorithmic approach, challenges and future directions, IEEE Open J. Intell. Trans. Syst. 6 (2025) 155–183.

[16] J.-J. Li, S.-B. Chen, C. Ding, B. Luo, Multi-scale feature sharing and collaborative sampling for unsupervised vehicle re-identification, Pattern Recognit. 172 (2026) 112353.

[17] Y. Xiong, J. Peng, Z. Tao, H. Wang, Region-guided spatial feature aggregation network for vehicle re-identification, Eng. Appl. Artif. Intell. 139 (2025) 109568.

[18] X. Dong, P. Shi, T. Liang, A. Yang, CTAFFNet: CNN–transformer adaptive feature fusion object detection algorithm for complex traffic scenarios, Transp. Res. Rec. 2679 (1) (2025) 1947–1965.

[19] D. Bolya, P.-Y. Huang, P. Sun, J.H. Cho, A. Madotto, C. Wei, T. Ma, J. Zhi, J. Rajasegaran, H. Rasheed, J. Wang, M. Monteiro, H. Xu, S. Dong, N. Ravi, D. Li, P. Dollár, C. Feichtenhofer, Perception encoder: the best visual embeddings are not at the output of the network, 2025, arXiv:2504.13181

[20] G. Zhang, Z. Li, C. Tang, J. Li, X. Hu, CEDNet: a cascade encoder–decoder network for dense prediction, Pattern Recognit. 158 (2025) 111072.

[21] X. Guo, J. Yang, X. Jia, C. Zang, Y. Xu, Z. Chen, A novel dual-pooling attention module for UAV vehicle re-identification, Sci. Rep. 14 (1) (2024) 2027.

[22] X. Zhou, X. Li, H. Zhou, X. Pang, J. Tian, X. Nie, C. Wang, Y. Yin, Adaptive division and priori reinforcement part learning network for vehicle re-identification, Pattern Recognit. 163 (2025) 111453.

[23] S. Lafrance, S. Bernadin, Evaluating visual transformers for wafer defect detection: a feasibility study, in: SoutheastCon 2025, IEEE, 2025, pp. 1288–1293.

[24] X. Xiang, Z. Ma, L. Zhang, D. Ombati, H. Himu, X. Zhen, LKA-ReID:vehicle re-identification with large kernel attention, in: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025, pp. 1–5.

[25] W. Hu, H. Zhan, P. Shivakumara, U. Pal, Y. Lu, TANet: text region attention learning for vehicle re-identification, Eng. Appl. Artif. Intell. 133 (2024) 108448.

[26] H. Li, C. Li, A. Zheng, J. Tang, B. Luo, MsKAT: multi-scale knowledge-aware transformer for vehicle re-identification, IEEE Trans. Intell. Transp. Syst. 23 (10) (2022) 19557–19568.

[27] Q. Wang, Z. Zhang, D. Wang, D. Gai, X. Xiong, J. Xu, R. Zhou, VehicleMAE: view-asymmetry mutual learning for vehicle re-identification pre-training via masked autoencoders, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 4701–4711.

[28] X. Pang, Y. Zheng, X. Nie, Y. Yin, X. Li, Multi-axis interactive multidimensional attention network for vehicle re-identification, Image Vis. Comput. 144 (2024) 104972.

[29] X. Chen, H. Yu, F. Zhao, Y. Hu, Z. Li, Global–local discriminative representation learning network for viewpoint-aware vehicle re-identification in intelligent transportation, IEEE Trans. Instrum. Meas. 72 (2023) 1–13.

[30] H. Kankam, Perception and attention, in: H. Kankam (Ed.), A Brief Excursion into Human Cognition: The Evolving Influence of Social Media & Artificial Intelligence, Springer Nature Switzerland, Cham, 2025, pp. 17–31.

[31] X. Zou, Z. Ji, T. Zhang, T. Huang, S. Wu, Visual information processing through the interplay between fine and coarse signal pathways, Neural Netw. 166 (2023) 692–703.

[32] M. Lou, Y. Yu, OverLoCK: an overview-first-look-closely-next ConvNet with context-mixing dynamic kernels, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025, pp. 128–138.

[33] T.Q. Nguyen, O.D.A. Prima, S.A. Irfan, H.D. Purnomo, R. Tanone, CORE-ReID V2: advancing the domain adaptation for object re-Identification with optimized training and ensemble fusion, AI Sens. 1 (1) (2025) 4.

[34] B. Zhu, H. Sang, Vehicle re-identification based on wavelet feature enhancement and global-local differential attention fusion, J. Comput. Sci. Artif. Intell. 2 (1) (2025) 53–60.

[35] Z. Wu, H. Zhen, X. Zhang, X. Bai, X. Li, SEMA-YOLO: lightweight small object detection in remote sensing image via shallow-layer enhancement and multi-scale adaptation, Remote Sens. 17 (11) (2025) 1917.

[36] K. Nai, G. Li, Visual object tracking via adaptive feature fusion and two-stage channel selection, Pattern Recognit. 172 (2026) 112682.

[37] W. He, F. Wang, Y. Bai, N.N. Xiong, G. Xu, F. Guo, PEFN: a patches enhancement and hierarchical fusion network for robust vehicle reidentification, IEEE Internet Things J. 12 (14) (2025) 26898–26910.

[38] X. Liu, W. Liu, H. Ma, H. Fu, Large-scale vehicle re-identification in urban surveillance videos, in: 2016 IEEE International Conference on Multimedia and Expo (ICME), 2016, pp. 1–6.

[39] H. Liu, Y. Tian, Y. Yang, L. Pang, T. Huang, Deep relative distance learning: tell the difference between similar vehicles, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2167–2175.

[40] Z. Tang, M. Naphade, M.-Y. Liu, X. Yang, S. Birchfield, S. Wang, R. Kumar, D. Anastasiu, J.-N. Hwang, CityFlow: a city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification, in: Proc. CVPR, Long Beach, CA, USA, 2019, pp. 8797–8806.

[41] Y. Shen, T. Xiao, H. Li, S. Yi, X. Wang, Learning deep neural networks for vehicle re-ID with visual-spatio-temporal path proposals, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1900–1909.

[42] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, J. Gu, A strong baseline and batch normalization neck for deep person re-identification, IEEE Trans. Multimedia. 22 (10) (2020) 2597–2609.

[43] H. Zhang, Z. Kuang, L. Cheng, Y. Liu, X. Ding, Y. Huang, AIVR-Net: attribute-based invariant visual representation learning for vehicle re-identification, Knowl. Based Syst. 289 (2024) 111455.

[44] R.K. Nath, D. Mitra, Learning part-based features for vehicle re-identification with global context, Appl. Sci. 15 (13) (2025) 7041.

[45] M. Liu, W. Min, Q. Han, H. Xiang, M. Zhu, Learning super-resolution and pyramidal convolution residual network for vehicle re-identification, Sci. Rep. 14 (1) (2024) 26531.

[46] W. Zhu, Z. Wang, X. Wang, R. Hu, H. Liu, C. Liu, C. Wang, D. Li, A dual self-attention mechanism for vehicle re-identification, Pattern Recognit. 137 (2023) 109258.

[47] B. Li, P. Liu, L. Fu, J. Li, J. Fang, Z. Xu, H. Yu, VehicleGAN: pair-flexible pose guided image synthesis for vehicle re-identification, in: 2024 IEEE Intelligent Vehicles Symposium (IV), IEEE, 2024, pp. 447–453.

[48] F. Huang, X. Lv, L. Zhang, Coarse-to-fine sparse self-attention for vehicle re-identification, Knowl. Based Syst. 270 (2023) 110526.

[49] F. Shen, Y. Xie, J. Zhu, X. Zhu, H. Zeng, GiT: graph interactive transformer for vehicle re-identification, IEEE Trans. Image Process. 32 (2023) 1039–1051.

[50] S. Wang, Y. Wang, R. Wu, B. Jiao, W. Wang, P. Wang, SeCap: self-calibrating and adaptive prompts for cross-view person re-Identification in aerial-ground networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025, pp. 22119–22128.

[51] S. Tumrani, W. Ali, R. Kumar, A.A. Khan, F.A. Dharejo, View-aware attribute-guided network for vehicle re-identification, Multimedia Syst. 29 (4) (2023) 1853–1863.

[52] K. Sun, X. Pang, M. Zheng, X. Nie, X. Li, H. Zhou, Y. Yin, Heterogeneous context interaction network for vehicle re-identification, Neural Netw. 169 (2024) 293–306.