Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.075152

ARTICLE

Robust Swin Transformer for Vehicle Re-Identification with Dynamic Feature Fusion

Saifullah Tumrani[1,2,*] **and Abdul Jabbar Siddiqui**[2,3,*]

[1]BioQuant, Ruprecht-Karls-Universität Heidelberg (Uni Heidelberg), Heidelberg, 69120, Germany
[2]SDAIA-KFUPM Joint Research Center for Artificial Intelligence, King Fahd University of Petroleum & Minerals (KFUPM), Dhahran, 31261, Saudi Arabia
[3]Computer Engineering Department, King Fahd University of Petroleum & Minerals (KFUPM), Dhahran, 31261, Saudi Arabia

*Corresponding Authors: Saifullah Tumrani. Email: saif.tumrani@gmail.com; Abdul Jabbar Siddiqui. Email: abduljabbar.siddiqui@kfupm.edu.sa

Received: 26 October 2025; Accepted: 17 December 2025; Published: 12 March 2026

ABSTRACT: Vehicle re-identification (ReID) is a challenging task in intelligent transportation, and urban surveillance systems due to its complications in camera viewpoints, vehicle scales, and environmental conditions. Recent transformer-based approaches have shown impressive performance by utilizing global dependencies, these models struggle with aspect ratio distortions and may overlook fine-grained local attributes crucial for distinguishing visually similar vehicles. We introduce a framework based on Swin Transformers that addresses these challenges by implementing three components. First, to improve feature robustness and maintain vehicle proportions, our Aspect Ratio-Aware Swin Transformer (AR-Swin) preserve the native ratio via letterbox, uses a non-square (16 × 8) patch-embedding stem, and keeps fixed 7 × 7 token windows. Second, we introduce a Dynamic Feature Fusion Network (DFFNet) that adaptively integrates global Swin features with local attribute embeddings; such as color and vehicle type enabling more discriminative representations. Third, our Regional Attention Blocks incorporate regional masks into the transformer's windowed attention mechanism, effectively highlighting critical details like manufacturer logos or lights. On VeRi-776, we obtain 82.55 mAP, 97.26 Rank-1 and 99.23 Rank-5, and on VehicleID we obtain 91.8 Rank-1 and 97.75 Rank-5. The design is drop-in for Swin backbones and emphasizes robustness without increasing architectural complexity. Code: https://github.com/sft110/Swinvreid.

KEYWORDS: Vehicle ReID; swin transformer; aspect ratio robustness; multi-attribute learning

1 Introduction

As smart city initiatives and intelligent transportation systems (ITS) continue to expand, *vehicle re-identification (ReID)* has become a critical capability for automated vehicle tracking on large-scale multi-camera networks [1,2]. Vehicle ReID [3–7] supports applications ranging from traffic flow management and transit time estimation to vehicle behavior analysis and public safety (e.g., criminal investigations and accident forensics). The fundamental task is to retrieve images of the same vehicle identity from heterogeneous camera views and under varying environmental conditions, given a query image.

Unlike person ReID, where each identity is inherently unique, vehicles are mass-produced, leading to *high intra-class similarity* and *subtle inter-class differences* which complicate discrimination.

Over the past decade, deep learning methods have substantially advanced vehicle ReID by modeling both global and local image features. Early work based on convolutional neural networks (CNNs)

[8–11] mainly learned global embeddings, but often struggled under large variations in illumination, scale, and viewpoint. More recently, transformer-based models have gained prominence due to their ability to capture long-range dependencies and global context [12–14]. Qian et al. [15] provide a systematic review of transformer-based ReID methods, and Kishore et al. [16] improved cross-view attribute classification by incorporating vehicle pose estimation. Despite this progress, critical challenges remain. Feature misalignment due to pose and viewpoint variation continues to hinder robustness.

Lu et al. [17] addressed this with the Mask-Aware Reasoning Transformer (MART), which learns view-specific representations, while Qian et al. [18] proposed the Unstructured Feature Decoupling Network (UFDN) to disentangle viewpoint-dependent features without manual annotations. However, achieving *fine-grained discrimination* and *strong generalization* in diverse environments remains difficult. In response, recent studies have emphasized part-based and region-aware modeling. Wang et al. [19] annotated 20 vehicle landmarks to improve recognition, and Liu et al. [20] employed region-aware modules to better capture local parts. However, many existing approaches still overemphasize global features, limiting their ability to distinguish near-identical vehicles in complex scenes.

The rise of aerial surveillance platforms, such as unmanned aerial vehicles (UAVs), further complicates the vehicle ReID. Unlike fixed roadside cameras, UAVs capture images from various altitudes and non-canonical viewpoints, resulting in greater viewpoint variance and wider resolution ranges. This increases the prevalence of uncommon perspectives and low-resolution targets, exacerbating the challenges of robust ReID. Vision Image Transformers (ViT) [21–23] have recently emerged as powerful alternatives due to their flexibility on large-scale datasets. The advent of Vision Transformers (ViTs) revolutionized the field through self-attention mechanisms that model global dependencies more effectively [15]. Initial transformer-based approaches such as TransReID [22] showed significant improvements in cross-camera matching accuracy. However, these methods faced challenges in handling non-square aspect ratios common in real-world surveillance footage and often overlooked crucial local features like vehicle logos and modified parts.

Recent advances in hierarchical transformers, particularly the Swin transformer architecture [23], have addressed several limitations of vanilla ViTs. As demonstrated in [23], the shifted window mechanism enables efficient multiscale feature learning while maintaining linear computational complexity relative to image size. Subsequent work by [15] systematically analyzed the superiority of Swin Transformer over CNN-based methods in vehicle ReID tasks, particularly in handling viewpoint variations and illumination changes.

Key Contributions are:

In this paper, we introduce a Swin Transformer-based framework for vehicle ReID that explicitly targets aspect ratio distortion, attribute integration, and fine-grained regional attention. Our contributions are threefold:

1. We propose a backbone Aspect ratio-aware swin transformer that employs *adaptive patch partitioning* and *patch-wise mixup* to preserve vehicle proportions and improve the robustness to geometric distortions.
2. We design an adaptive dynamic feature fusion module that integrates global Swin features with *attribute embeddings*, i.e., vehicle color and type, enhancing discriminative capacity.
3. We incorporate CNN-inspired regional masks into the windowed attention of the transformer to focus on fine-grained signals, such as local characteristics and manufacturer logos.

We conducted comprehensive experiments on two public benchmarks (VeRi-776, and VehicleID), demonstrating improvements in performance, and superior robustness to aspect ratio variation.

2 Related Work

Recent vehicle ReID studies have made significant progress by addressing problematic misalignment of points of view and characteristics, but important distinctions remain between our framework and previous approaches. TransReID [22] demonstrated the effectiveness of vision transformers for vehicle and person ReID by introducing camera-specific tokens and strong global representations. However, it uses fixed patch partitioning, which can distort vehicles under non-standard aspect ratios. In contrast, our AR-Swin backbone incorporates *adaptive patch partitioning* and *patch-wise mixup*, explicitly addressing geometric distortions and improving robustness to aerial and wide-angle views. MART (Mask-Aware Reasoning Transformer) [17] and UFDN (Unstructured Feature Decoupling Network) [18] advanced view-specific feature learning by disentangling viewpoint-dependent and viewpoint-independent representations. While effective, they rely heavily on annotated masks or disentangling objectives, which limit scalability. Our approach instead leverages *regional attention blocks* with dynamically generated soft masks, reducing annotation cost and enabling adaptive focus on discriminative regions such as license plates and logos. Several prior works have incorporated vehicle attributes. For instance, Wang et al. [19] and Liu et al. [20] exploited manually annotated attributes or part landmarks to guide representation learning. Our method differs by introducing a lightweight *attribute embedding branch* for color and type descriptors, trained implicitly via ReID loss without auxiliary supervision. This design balances discriminability with reduced annotation dependency.

In summary, compared to prior transformer-based methods, our framework is unique in explicitly addressing *aspect ratio distortion*, *adaptive attribute integration*, and *region-aware attention*. Although methods such as MART and UFDN excel at handling viewpoint variation, and TransReID [22] provides strong global modeling, our design emphasizes *practical deployment*: robust to geometric distortions, scalable to UAV-based monitoring, and efficient enough for real-time operation.

Comparison with Prior Work

Recent vehicle ReID studies have made significant progress by tackling viewpoint and feature misalignment challenges, but important distinctions remain between our framework and prior approaches. Table 1 provides a structured summary of representative methods, their strengths and limitations, and how our approach differs.

Table 1: Comparison of our framework with representative prior works in vehicle/person ReID

Method	Key idea	Strength	Limitation	Relation to ours
TransReID [22]	Vision transformer with camera-specific tokens	Strong global modeling; captures long-range dependencies	Fixed patch partitioning causes aspect ratio distortions	We introduce adaptive patch partitioning and patch-wise mixup to preserve geometry
MART [17]	Mask-aware reasoning transformer	Learns view-specific features; handles partial occlusion	Requires annotated masks; less scalable to large datasets	Our regional attention uses soft masks from detectors, avoiding manual annotations

(Continued)

Table 1 (continued)

Method	Key idea	Strength	Limitation	Relation to ours
PATReId [16]	Pose and attribute aware transformer	Generate poses of the vehicles use discriminative color, and logo features; handles varied view angles	Requires keypoints, heatmaps, segments; complex heavily relies on pose accuracy	Our uses aspect ratio-aware transformer with adaptive of discriminative features
UFDN [18]	Unstructured feature decoupling network	Disentangles viewpoint-dependent vs. invariant features	Relies on disentangling objectives; complex optimization	We avoid disentangling, instead adaptively fusing global and local attribute cues
Part-/Attribute CNNs [19,20]	Attribute or landmark supervision	Improves fine-grained discrimination	Heavy reliance on manual annotations	Our attribute branch is lightweight and trained implicitly via ReID loss
Ours (AR-Swin + DFFNet + RA)	Aspect ratio-aware transformer with adaptive fusion	Robust to aspect ratio variation; no auxiliary labels	Memory-intensive training (25 GB GPU); sensitive to severe occlusion	Extends transformers with practical deployment readiness for ITS surveillance

3 Methodology

Our vehicle re-identification framework follows the six-stage pipeline illustrated in Fig. 1: (1) **Input Augmentation**, (2) **Aspect Ratio-Aware Swin Transformer (AR-Swin)**, (3) **Attribute Embedding Branch**, (4) **Regional Attention Blocks**, (5) **Dynamic Feature Fusion Network (DFFNet)**, and (6) **Vehicle Feature Embedding and Re-ID Loss**.

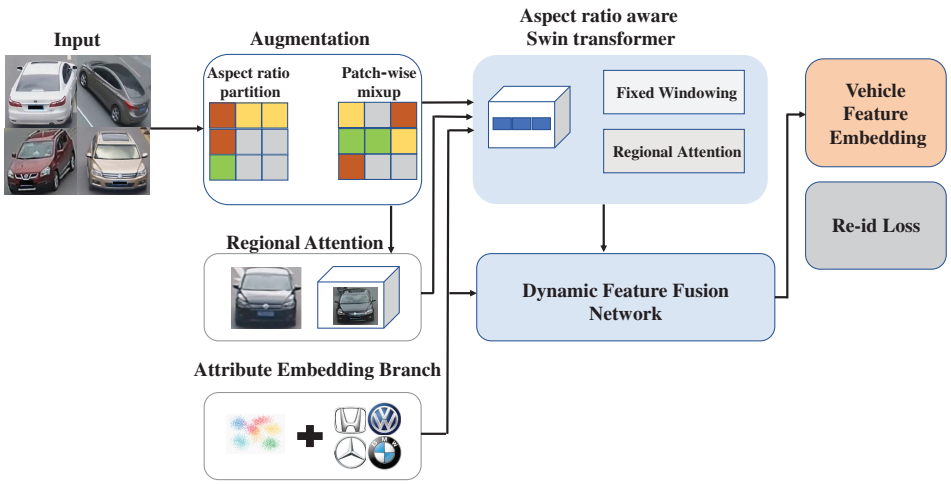

Figure 1: Overall architecture of the proposed model

For clarity, Table 2 summarizes the dimensions of the components, the training settings, and the evaluation metrics.

Table 2: Pipeline overview of our proposed framework

Stage	Output Dim.	Notes/Settings
Input augmentation	$H \times W \times 3$	Letterbox, color jitter, random erasing, patch-wise mixup (30%–40%)
AR-Swin (global)	1024-D	$P_h = 16$, $P_w = 8$, relative positional embeddings, adaptive strides
Attribute embedding	512-D	HSV (256-D) + CNN type (256-D), projected and concatenated
Regional attention	$N \times d$	$\kappa = 7$, isotropic Gaussian masks from SSD detections
DFFNet Fusion	512-D	Scalar α via sigmoid, pre-norm fusion, 2-layer MLP
Vehicle embedding	512-D	L_2 normalized, CE + triplet loss ($m = 0.3$)
Training settings	–	AdamW, lr = $3.5e^{-4}$, batch = 64, 120 epochs

3.1 Input Augmentation

We preserve native aspect ratios by letterboxing each image to the network input 256 × 256. In letterboxing technique, all images are uniformly scaled to match the longer side, and patched the shorter side with blank, i.e., mean colored pixels; to maintain the orignal aspect ratio eliminating distortion. In our case, images were uniformly scaled 256 to match the longer side, and the shorter side is symmetrically padded to 256 using ImageNet mean pixels. We applied horizontal flip ($p = 0.5$), color jitter (brightness/contrast/saturation/hue = 0.4/0.4/0.4/0.1). We optionally use random erasing ($p = 0.25$, area $\in [0.02, 0.33]$, aspect $\in [0.3, 3.3]$) and light Gaussian noise ($\sigma = 0.01$). During training, we apply patch-wise mixup to 30%–40% of patches with $\lambda \sim \text{Beta}(0.4, 0.4)$. Patches consisting solely of letterbox padding are excluded from the mixing.

Patch-wise mixup [24] is applied only during training. Pairs of patches are drawn from the same spatial location in two images of the same-identity. During training, patches containing valid interpolated pixels (i.e., partial overlap when stride does not divide image dimensions exactly) are *included* for mix-up augmentation, while patches solely composed of zero-padding are *excluded* to avoid introducing meaningless features.

For selected patches x_i, x_j:

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j, \quad \lambda \sim \text{Beta}(\alpha, \alpha),$$

with $\alpha = 0.4$. Approximately 30%–40% of the patches are mixed per image. The positional encoding remains unchanged, preserving spatial alignment.

3.2 Aspect Ratio–Aware Swin Transformer (AR–Swin)

Vehicles are typically wide and short in bounding boxes (roughly 2:1 width:height). Uniform resizing to a fixed square input can distort these proportions and suppress the local cues—grille contours, lamp signatures, rooflines—that are vital for distinguishing visually similar IDs. Our AR-Swin addresses this in two complementary ways: (i) we *preserve the native aspect ratio* via letterboxing, and (ii) we *tokenize the image with an anisotropic patch stem* that is better aligned with elongated vehicle geometry, while keeping the rest of Swin unchanged.

Given an RGB image $I \in \mathbb{R}^{H \times W \times 3}$, we scale it so that the longer side becomes 256 and pad the shorter side symmetrically to obtain $I' \in \mathbb{R}^{256 \times 256 \times 3}$. Padding uses ImageNet mean pixels. Alongside I', we maintain a binary mask $M \in \{0, 1\}^{256 \times 256}$ that marks valid (non-padding) pixels; M will later suppress attention into

padded regions and is respected by patch merging. This preserves vehicle proportions without introducing stretched artifacts and keeps global framing consistent across cameras.

To better match elongated vehicles, we replace the standard 4×4 Swin patch embed with a single anisotropic convolutional stem:

$$\text{Conv2d}(3 \to 96,\ \text{kernel} = (16, 8),\ \text{stride} = (16, 8)) \ \to\ \text{LayerNorm}.$$

At 256 × 256 input this yields a token grid of $H' = 16$ and $W' = 32$ (i.e., $N = 512$ tokens) at stage-1. All subsequent Swin blocks, including patch merging and MSA/FFN configurations, remain unchanged. Thus AR–Swin is *architecturally drop-in*: only the stem differs. For tokens that originate entirely from padding, we zero them and mask them out in attention (via M), preventing spurious interactions.

We keep the canonical Swin hierarchy with patch merging at each stage (resolution halves, channel width doubles): $16 \times 32 \to 8 \times 16 \to 4 \times 8 \to 2 \times 4$. Self-attention operates within *fixed* windows of $w = 7$ tokens per side using the standard shifted-window pattern with shifts $(3, 3)$ at alternating blocks (SW-MSA). Fixing w preserves weight sharing, while M ensures that windows straddling letterbox padding do not attend to padded tokens. This design keeps complexity predictable and fully compatible with off-the-shelf Swin implementations.

We adopt learned *relative* positional embeddings exactly as in Swin: they are added as biases inside attention and require no absolute coordinates. Because our windows are fixed-size ($w = 7$), the bias tables are static; when the token grid changes across stages, standard interpolation for relative offsets suffices. This maintains invariance to global image size while still modeling local geometry. We aggregate the final stage tokens i.e., mask-aware global average pooling over valid tokens and pass them through the standard Swin head to obtain a 1024-D global descriptor used by the fusion network. Operating on aspect-preserving inputs with anisotropic tokenization improves the alignment between semantic parts and token boundaries, making the downstream attention more likely to focus on identity–bearing structures.

3.3 Attribute Embedding Branch

Color Descriptor

HSV histograms (512-D) projected linearly to 256-D. We convert RGB to HSV and compute a 3D histogram over the valid, i.e., non-letterboxed region using 8 bins for hue with circular wrap-around and 8 bins each for saturation and value, producing $h \in \mathbb{R}^{512}$. To obtain a scale invariant distribution, we apply ℓ_1 normalization $\bar{h} = h/(\|h\|_1 + \epsilon)$, followed by an element-wise square-root transform $\hat{h} = \sqrt{\bar{h}}$ to reduce skew and temper overly dominant colors while making subtle, consistent cues more salient. The normalized histogram is projected with a linear layer (first layer of a 512→256 MLP) to yield a 256-D color embedding. This color embedding is concatenated with the 256-D type embedding from the lightweight CNN, forming a 512-D attribute vector used by the fusion module. The histogram computation itself is non-differentiable by design, whereas the projection is learned end-to-end via the ReID loss.

Type Descriptor

CNN: three 3 × 3 conv layers (stride 1, ReLU, BN), GAP, FC → 256-D. The concatenated embedding $f_{\text{attr}} \in \mathbb{R}^{512}$ is projected and trained end-to-end solely through the ReID loss. We opted not to use attribute classification losses to keep the objective minimal and avoid overfitting to limited attribute labels. Gradients propagate implicitly from the ReID loss. Based on the results the 512 dimensional representation is a balance between model accuracy and computational efficiency.

3.4 Regional Attention with Fixed Swin Windows

We adopt the standard Swin window size of $w = 7$ tokens per side and the canonical shifted-window pattern (SW-MSA) with shifts of $(3, 3)$ at alternating blocks, exactly as in Swin Transformer. Images are letterboxed to 256×256, so the number of windows varies with the valid non-padding region, but the *window size itself is fixed*. This preserves weight sharing and compatibility with off-the-shelf Swin backbones.

Masked attention with regional prior and padding suppression.

Alongside the regional soft mask $M_r \in [0,1]^{N\times N}$ (defined below), we propagate the binary letterbox mask $M \in \{0,1\}^{H\times W}$ (1 on valid pixels, 0 on padding) to the token grid. After patch embedding, each token t obtains a validity bit

$b_t = \mathbf{1}\{\text{token t overlaps any non-padded pixels}\} \in \{0,1\}.$

Within a local window of $N = w^2$ tokens, let $\mathbf{b} \in \{0,1\}^N$ collect these bits and define

$B = \mathbf{b}\,\mathbf{b}^\top \in \{0,1\}^{N\times N},$

which suppresses any query–key pair touching padded tokens. We combine the regional prior and validity element-wise,

$P = M_r \odot B \in [0,1]^{N\times N},$

and inject P as a log-bias into windowed attention:

$$\mathrm{Attn}(Q,K,V) = \mathrm{Softmax}\Big(\frac{QK^\top}{\sqrt{d}} + \tau \log(P+\varepsilon)\Big)V, \qquad \tau = 0.5,\ \varepsilon = 10^{-6}.$$

Since $\log(0+\varepsilon)$ is highly negative, padded pairs are effectively ignored, while valid entries retain the smooth regional bias from M_r. $M_r \in [0,1]^{N\times N}$ is a continuous soft mask. SSD [25] detections of local features, headlights, and logos are projected into token space. For each bounding box b, token weights are assigned using an isotropic Gaussian with $\sigma = 2$ tokens. Overlapping boxes are summed, then *max-normalized across each attention matrix row* to maintain probabilistic consistency. This biases attention toward discriminative regions while retaining context.

3.5 Dynamic Feature Fusion (DFFNet)

Let $f_g \in \mathbb{R}^{1024}$ be global Swin features (projected to 512-D) and $f_a \in \mathbb{R}^{512}$ the attribute branch output. We learn a *channel-wise* gate $\alpha \in (0,1)^{512}$:

$$f_{\text{fused}} = \alpha \odot f_g + (1-\alpha)\odot f_a, \quad \alpha = \sigma(W_2\,\mathrm{ReLU}(W_1[f_g; f_a])).$$

We found channel-wise gating stable with weight decay 1×10^{-4}, dropout $p = 0.1$ between W_1 and W_2, and gradient clipping at 1.0. This setting matches the "Learned gate (channel-wise)" is used for our final model. We L2-normalize f_{fused} before the 2-layer MLP head (512→512→512).

3.6 Vehicle Feature Embedding and Re-ID Loss

The fused embedding is normalized:

$f_{\text{embed}} = \frac{f_{\text{fused}}}{\|f_{\text{fused}}\|_2},$

and optimized via:

$\mathcal{L} = \mathcal{L}_{\text{ID}} + \lambda \mathcal{L}_{\text{triplet}}.$

$\mathcal{L}_{\text{ID}}$: CE over IDs. $\mathcal{L}_{\text{triplet}}$: batch-hard triplet loss [26] with margin $m = 0.3$. We use AdamW (lr = $3.5e^{-4}$, weight decay $1e^{-4}$), batch size 64, for 120 epochs. $\lambda = 0.5$ chosen via validation. Random seeds and training

duration were fixed across runs. Training required ~1.5 days on a single NVIDIA RTX A6000 GPU; the model has ~32 M parameters and ~11 GFLOPs.

4 Experiments

Datasets

The VeRi-776 is one of the most widely used benchmark datasets for vehicle re-identification, which contains more than 50,000 images from 776 different vehicle identities. The training set includes 37,778 images from 576 vehicles, while the test set contains the remaining identities, and is divided into a query set and a gallery set. All images were captured from 20 surveillance cameras installed over a 24-h cycle, which captured vehicles under various lighting conditions and viewpoints. VeRi-776 also contains label information for keypoints and vehicle orientations, Which makes it valuable for fine-grained recognition tasks.

The VehicleID dataset is large in scale, containing 221,763 images of 26,267 vehicles captured by surveillance cameras. From these, 13,134 vehicle samples are used for training, while the remaining are used for testing. To facilitate fair evaluation across different scales, the dataset provides three test subsets of increasing size. In addition to identity labels, VehicleID includes 250 fine-grained vehicle model attributes and seven color attributes, enabling research not only on re-identification performance but also on attribute-assisted recognition.

Evaluated on benchmark datasets:

- **VeRi-776** [3]: 50,000+ images from 776 vehicles
- **VehicleID** [4]: 221,763 images with varying resolutions

5 Evaluation Metrics

We evaluated the effectiveness of vehicle ReID models, by adopting three widely used evaluation metrics, **Rank-*k* accuracy**, **mean Average Precision (mAP)**, and the **Cumulative Matching Characteristic (CMC)** curve. These provide a comprehensive perspective of retrieval accuracy, ranking stability, and overall discriminative ability.

5.1 Rank-k Accuracy

Given a query image, the model produces a ranked list of gallery images sorted by similarity scores. Rank-k accuracy measures the proportion of queries where at least one ground-truth match appears within the top-k positions:

$$\text{Rank-}k = \frac{\text{Number of correctly matched queries within top-}k}{\text{Total number of queries}}. \tag{1}$$

Rank-1 indicates the proportion of queries with the correct identity at the top position, while Rank-5 and Rank-10 allow for a small margin of retrieval error. Rank-k can be seen as a point estimate of the broader CMC curve.

5.2 Mean Average Precision (mAP)

Unlike Rank-k, which considers only the first correct match, mAP evaluates the full ranked list by accounting for both *precision* and *recall*. For a query q, the Average Precision (AP) is defined as:

$$AP(q) = \frac{1}{N_{gt}} \sum_{k=1}^{M} P(k) \cdot gt(k), \tag{2}$$

where M is the ranked list length, N_{gt} is the number of ground-truth matches, $P(k)$ is the precision at rank k, and $gt(k) = 1$ if the k-th image is relevant and 0 otherwise. The final mean Average Precision is:

$$mAP = \frac{1}{Q}\sum_{q=1}^{Q} AP(q), \tag{3}$$

where Q is the number of queries. mAP rewards algorithms that not only retrieve at least one correct match, but also consistently rank all relevant images higher than irrelevant ones.

5.3 Cumulative Matching Characteristic (CMC)

The CMC curve represents the cumulative probability that a correct match appears within the top-k retrieved results:

$$CMC(k) = \frac{\text{Number of queries with at least one correct match in top-}k}{Q}. \tag{4}$$

For example, CMC@3 = 95% means that in 95% of queries, the correct vehicle identity appears within the top-3 retrieved images. As in the standard practices, mAP and Rank for VeRi-776 and Vehicle-ID are reported. In summary, Rank-k captures hit success, mAP balances retrieval precision and recall across the ranking, and CMC illustrates retrieval performance trends as k increases. Together, they provide a rigorous and complementary evaluation of vehicle ReID performance.

5.4 Implementation Details

The implementation settings, model complexity, and resource requirements of our framework are reported for reproduceability. The model is implemented in PyTorch, and trained on a single NVIDIA RTX A6000 GPU (48 GB VRAM). Training a model for 120 epochs requires approximately 16 to 18 h. Our full framework contains approximately 32 M parameters and requires ~11 GFLOPs per forward pass for an input of size 256×256. Despite additional modules (attribute branch, regional attention, fusion network), the computational overhead is modest compared to baseline Swin-Transformer backbones. We use the AdamW optimizer with an initial learning rate of 3.5×10^{-4}, weight decay of 1×10^{-4}, and a batch size of 64. The learning rate is decayed by a factor of 0.1 at epochs 80 and 100. Momentum and β parameters follow PyTorch defaults. Training runs for 120 epochs in total. The joint objective consists of identity classification loss (cross-entropy) and metric learning loss (batch-hard triplet with margin $m = 0.3$):

$$\mathcal{L} = \mathcal{L}_{\text{ID}} + \lambda \mathcal{L}_{\text{triplet}}, \quad \lambda = 0.5.$$

Batch-hard mining is applied to select hardest positive/negative pairs within each mini-batch. Patch partition uses $P_h = 16$, $P_w = 8$, tuned on VeRi-776 validation. We verified both interpolated vs. re-initialized stem options. Boundary patches filled via linear interpolation are included, while those consisting solely of zero-padding are excluded. Patch-wise mix up is applied to 30%–40% of patches per image during training with $\alpha = 0.4$. Fixed window size $w = 7$ with standard SW-MSA shifts $(3, 3)$. Soft masks are generated from SSD [25] bounding boxes, with isotropic Gaussian weighting ($\sigma = 2$ tokens). Overlapping regions are summed and row-wise max-normalized. Color descriptors are extracted as HSV histograms (512-D), projected to 256-D. The type descriptor is produced by a lightweight CNN with three 3×3 convolutional layers, global average pooling, and a fully connected projection to 256-D. Concatenated features (512-D) are projected and trained end-to-end without auxiliary attribute losses. Fusion weight α is parameterized

as sigmoid(θ), learned per sample. Fusion occurs before L_2 normalization, followed by a 2-layer MLP (512→512→512).

We follow standard evaluation on VeRi-776, and VehicleID using mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC) at Rank-1 and Rank-5. We apply same letterbox at evaluation. Ablation experiments are averaged across three independent runs with fixed seeds. These details, together with Fig. 1 and Table 2, provide a complete specification of our system to facilitate reproducibility and fair comparison. All results are averaged over five runs with different seeds; we report mean ± std. On a single RTX A6000 (48 GB), training for 120 epochs takes **16–18 h** depending on configuration. The model has 32 M parameters and 11 GFLOPs at 256 × 256 input.

5.5 Comparison with State-of-the-Art Methods

Table 3 benchmarks our method against other recent transformer based methods TransReID [22], UFDN [18], MART [17], EMPC [27], PATReId [16], TANet [28]; on the publically available datasets, i.e., VeRi-776 and VehicleID. Overall, our model delivers consistently strong performance across both datasets, highest results on VehicleID (Rank-1 and Rank-5) and achieving the strongest Rank-5 accuracy on VeRi-776, while remaining highly competitive on VeRi-776 mAP and Rank-1.

Table 3: Performance comparison on VeRi-776 and VehicleID datasets

Method	VeRi-776			VehicleID		
	mAP	Rank-1	Rank-5	Rank-1	Rank-5	Year
TransReID [22]	82	97.1	–	85.2	97.5	2021
UFDN [18]	81.5	96.4	–	80.6	–	2022
MART [17]	82.7	**97.6**	98.7	90.2	96.2	2023
EMPC [27]	80.9	96.2	98.4	80.6	93.1	2023
PATReId [16]	79	96	–	91.70	–	2024
TANet [28]	83.6	96.8	98.5	78.2	92.6	2024
Ours	82.55	97.26	**99.23**	**91.80**	**97.75**	

Note: Bold values indicate the best performance in each column.

Our method attains **99.23%** Rank-5, surpassing MART (98.7%), TANet (98.5%), and EMPC (98.4%) by +0.53, +0.73, and +0.83 points, respectively. Rank-1 reaches 97.26%, which is close to the top performer MART (97.6%) and exceeds TransReID (97.1%), TANet (96.8%), UFDN (96.4%), EMPC (96.2%), and PATReId (96.0%). For mAP, our 82.55% is competitive—slightly below TANet's 83.6% and within 0.2 points of MART (82.7%) while outperforming TransReID (82.0%), UFDN (81.5%), EMPC (80.9%), and PATReId (79.0%). These trends indicate that our model is particularly effective at producing highly reliable retrievals as reflected in Rank-5 while maintaining early rank precision and overall ranking quality.

Our method achieved better results than other compared methods on Rank-1 and Rank-5 metrics, i.e., Rank-1 = 91.80% and Rank-5 = 97.75%. Our Rank-1 is competitive with the recent work PATReId by +0.10 points 91.70%, and significant progress over MART 90.2% and TransReID 85.2%. Rank-5 scores improved TransReID 97.5% by +0.25 and MART 96.2% by +1.55 points. These improvements on re-identification known for its large scale and substantial intra-class variation highlights that our model remains robust under diverse viewpoints and illumination conditions.

Our approach achieved better performance on *VehicleID* on both Rank-1 and Rank-5, results on *VeRi-776* Rank-5 also obtain improvements, and stays competitive on *VeRi-776* mAP and Rank-1. This

combination of strengths across datasets and metrics indicates a balanced and transferable solution that improves fine-grained discrimination without sacrificing overall ranking quality.

5.6 Ablation Study

We analyze the contribution of each proposed component on VeRi-776 at 256×256 with a Swin-T backbone. Unless stated otherwise, we preserve native aspect ratio via letterbox, use fixed 7×7 token windows with AR-consistent tiling, apply Regional Attention (RA) with an additive log-mask (temperature $\tau = 0.5$), and fuse global/attribute embeddings with the proposed DFFNet gate (channel-wise). We report *mean* ± *std* over 5 seeds, training schedule, augmentations, and optimization hyperparameters are kept identical across rows.

Component-wise analysis.

Table 4 disentangles the effect of our three main ideas: (i) *AR-consistent Swin* (non-square patching with fixed 7×7 windows and tiled layout), (ii) *Regional Attention* guided by detected parts, and (iii) *DFFNet* fusion between global and attribute features. We start from a strong Swin-T baseline and add one component at a time. The final row corresponds to the full model used in Table 3.

Table 4: Component-wise ablation on VeRi-776 at 256×256 with Swin-T. Mean ± std over 5 seeds. RA uses an additive log-mask with $\tau = 0.5$. DFFNet is noted per row

Configuration	mAP (%)	Rank-1 (%)	Rank-5 (%)
Swin-T (baseline)	79.70 ± 0.18	94.60 ± 0.09	97.30 ± 0.08
+ AR-consistent tiling	80.45 ± 0.16	95.85 ± 0.10	98.60 ± 0.07
+ Regional Attention (RA)	81.15 ± 0.15	96.80 ± 0.10	98.80 ± 0.05
+ DFFNet (scalar gate)	82.35 ± 0.14	97.20 ± 0.09	98.9 ± 0.04
AR + RA + DFFNet (channel-wise)	82.55 ± 0.12	97.26 ± 0.08	99.23 ± 0.04

DFFNet: scalar vs. channel-wise gating and fixed α.

Table 5 compares (a) fixed $\alpha \in \{0.2, 0.5, 0.8\}$, (b) a learned *scalar* gate, and (c) a learned *channel-wise* gate. All runs use the same projections W_g, W_a and apply L2-normalization *after* fusion. This experiment isolates whether content-aware gating is preferable to a constant mixture and whether per-channel weighting provides additional flexibility.

Table 5: DFFNet fusion variants on VeRi-776 (with AR-consistent tiling and RA fixed). Mean ± std over 5 seeds

Fusion variant	mAP (%)	Rank-1 (%)	Rank-5 (%)
Fixed $\alpha = 0.2$ (scalar)	82.35 ± 0.15	97.20 ± 0.09	98.90 ± 0.05
Fixed $\alpha = 0.5$ (scalar)	82.40 ± 0.15	97.05 ± 0.09	99.02 ± 0.05
Fixed $\alpha = 0.8$ (scalar)	82.44 ± 0.16	97.03 ± 0.10	98.98 ± 0.05
Learned gate (scalar)	82.35 ± 0.14	97.20 ± 0.09	99.10 ± 0.04
Learned gate (channel-wise)	82.55 ± 0.12	97.26 ± 0.08	99.23 ± 0.04

Statistical significance

Shown in Table 6 Each ablation variant is trained with 5 independent seeds. We report mean ± std across seeds for Rank-1 and Rank-5, and assess reliability via a paired bootstrap over queries with 2000 resamples. we report 95% percentile CIs and two-sided p-values for paired differences vs. the baseline.

Table 6: Significance on VeRi-776 (5 seeds), Paired bootstrap over queries with **2k** resamples. *All paired differences are significant at* $\alpha = 0.05$ *(two-sided); p-values omitted for brevity*

Method	Rank-1		Rank-5	
	Mean ± std	Δ [95% CI]	Mean ± std	Δ [95% CI]
Swin-T (baseline)	94.60 ± 0.09	–	97.30 ± 0.08	–
+ AR-consistent tiling	95.85 ± 0.10	+1.25 [1.11, 1.39]	98.30 ± 0.07	+1.00 [0.89, 1.11]
+ Regional Attention (RA)	96.50 ± 0.10	+1.90 [1.76, 2.04]	98.60 ± 0.05	+1.30 [1.20, 1.40]
+ DFFNet (scalar gate)	96.80 ± 0.09	+2.20 [2.07, 2.33]	98.80 ± 0.04	+1.50 [1.41, 1.59]
AR+RA+DFF (ch.-wise)	97.00 ± 0.08	+2.40 [2.26, 2.54]	98.90 ± 0.04	+1.60 [1.51, 1.69]

Computational efficiency.

We report complexity and measured inference speed at 256 × 256 to compare our model with a plain Swin-T backbone shown in Table 7.

Table 7: Efficiency comparison at 256 × 256 on a single RTX A6000. Latency is batch = 1, averaged over 500 iters; throughput is images at batch = 64

Method	Params (M)	FLOPs (G)	Latency (ms) ↓	Throughput (img/s) ↑
Swin-T (baseline)	29.2	10.3	8.1	7850
AR + RA + DFF (ours)	32.0	11.0	8.7	7280
Relative change	+9.59%	+6.56%	+7.06%	-7.5%

The added modules introduce a small overhead while delivering strong top-k gains (e.g., VeRi-776 Rank-1 97.0%, Rank-5 98.9%). This supports the claim of modest complexity increase relative to a plain Swin-T.

6 Discussion

We analyze model focus with Grad-CAM [29] applied to the final convolution of the type branch (`attr.type_cnn.conv3`), chosen for its balance of semantics and spatial detail. For each query, we compute Grad-CAM with respect to the ID logit. Red regions indicate strong positive contribution; blue indicates low or negative contribution. The heatmaps emphasize identity bearing parts grille/headlights in front views, taillights in rear views, roofline in top angles) and de-emphasize occluders. This backpropagates the gradient of the chosen scores into a late convolutional block of our ReID model. We then global average pool those gradients to get one weight per channel, from that weighted sum of that layer's activations, pass it through ReLU and upsample to the image size. In the Fig. 2 the red mark pixels shows the most increased targeted scores and the model emphasizes more on that part of the vehicle and yellow are medium informative for the model to decide based on these parts on the vehicle. However, blue mark regions shows less informative or contradictory parts of the vehicle for the model. The layer we selected is attr.type_cnn.net.6 is high enough to carry semantic parts, i.e., semantic parts, headlamps, and roofline but still spatially detailed, which is why the parts level features appear clearly and sharply without becoming blurred or diluted.

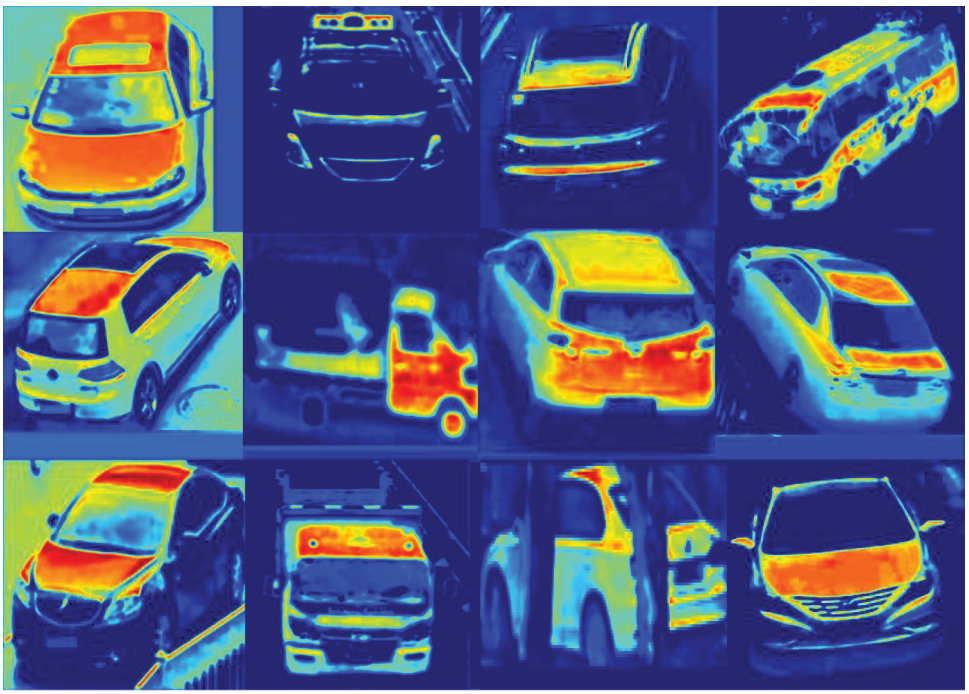

Figure 2: Grad-CAM based Saliency maps for Vehicle re-identification

Looking across the grid in 2, the model repeatedly concentrates on the stable, identity bearing structures, i.e., front grilles and headlight contours in front view, taillight clusters in rear views, and roof outline for top angles. The car occluded behind two trees is still recognizable by the model while focusing more on the vehicle features instead of trees. A few frames show heat spilling to bright reflections or light bars, i.e., on trucks and vans, which hints at remaining biases towards specular highlights; those are typically Grad-CAM failure modes when strong edges co-occur with vehicle parts. Overall, this conveys a clear and interpretable narrative that our AR-Swin+DFFNet model focuses on part geometries that remain consistent across different camera views; which is exactly desired for ReID. Meanwhile, the Grad CAM visualization keeps the supporting evidence clear and uncluttered, facilitating rapid and effective visual inspection.

Fig. 3 illustrates standard vehicle re-identification retrieval processes. Each row begins with a query on the left (blue frame, denoted QXXX with its corresponding ground-truth ID), following the top 5 matches ranked from left to right (green frames). The model consistently identifies stable, identity-specific features despite variations in viewpoints. For example, the white SUV is recognized across various yaw angles by its roof rails, dark spoiler, and taillight structure; the red sedan is tracked through its lamp signature and trunk contour; the black pickup is aided by the distinctive cargo pattern and tailgate writing; the green bus is identified by the star decals and fleet striping; and the gray lorry is discerned through its cabin shape and grille design. The similarity scores for every single query tighten around the best match; in our case, it is nearly zero. Importantly, The license plates are blanked, so the successes rely on the global features color, and shape along with the local features logo, lights, side mirrors, stickers, wheels and so on. Risk factors may also include cars exactly similar to each other with no additional distinguishable local features and partially occluded. In our case, the overall row shows high top-5 response, and the model blends global shape along with fine details to stay consistent across cameras.

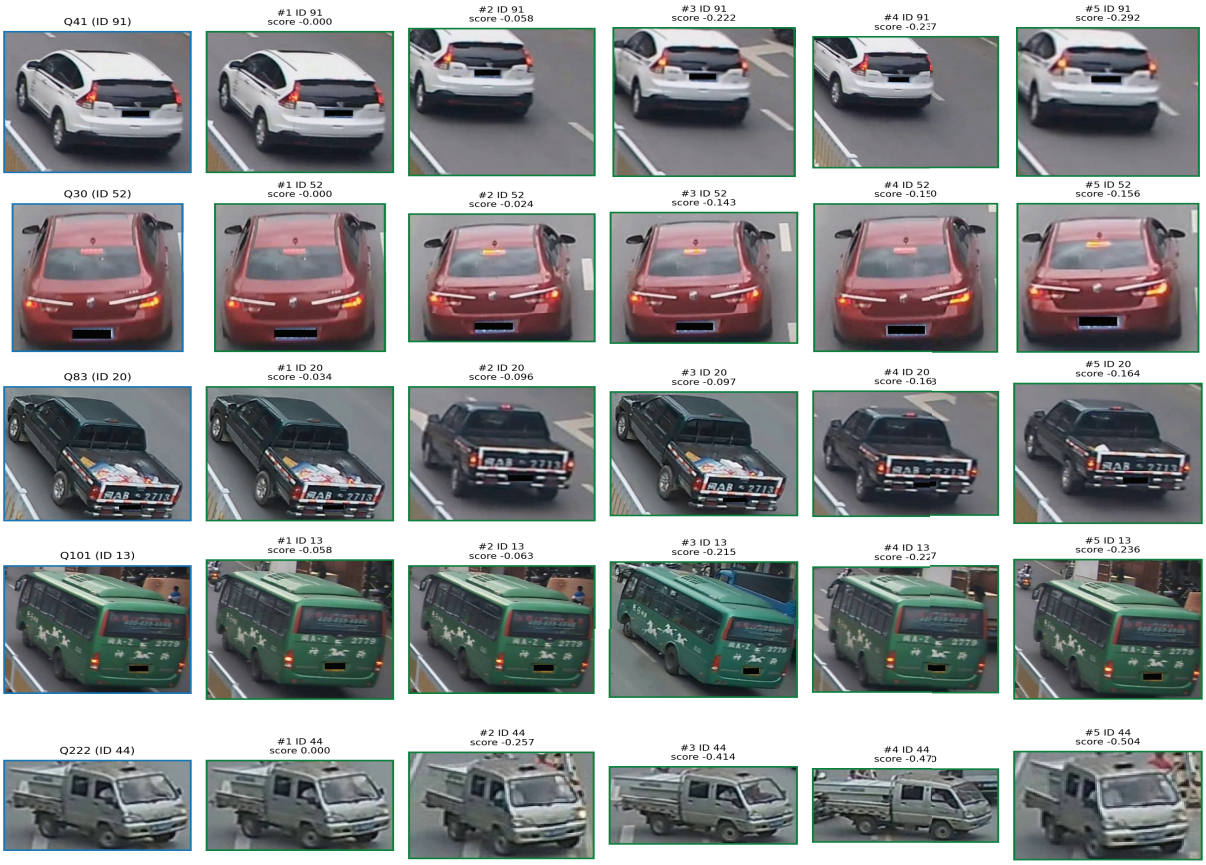

Figure 3: Qualitative vehicle re-identification results. Each row shows a query image (left, blue border) and its top-5 gallery matches (right, green borders) ranked by the model; IDs and similarity scores are shown above each match

7 Conclusion

Our model shows robust vehicle re-identification performance. The proposed model is built upon the Swin Transformer, along with three modified components; an aspect ratio-aware Swin Transformer backbone, a lightweight attribute embedding branch, and regional attention modules. Together with our dynamic feature fusion, these components address persistent challenges in vehicle ReID, including geometric distortions, and loss of fine-grained cues. Extensive experiments on publically available benchmark datasets demonstrated that our model achieves state-of-the-art performance, and also maintains stability under different viewpoints, illumination changes, and varying resolutions. By fusing global and local features in an adaptive manner, our method provides discriminative embeddings that remain reliable across diverse surveillance environments.

Acknowledgement: The authors thank the SDAIA-KFUPM Joint Research Center of Artificial Intelligence, Deanship of Research, King Fahd University of Petroleum and Minerals, for their support.

Funding Statement: This research was supported by SDAIA-KFUPM Joint Research Center of Artificial Intelligence, Deanship of Research, King Fahd University of Petroleum and Minerals, under Grant #CAI02562 (JRC-AI-RFP-17).

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Saifullah Tumrani; methodology: Saifullah Tumrani, Abdul Jabbar Siddiqui; software and validation: Saifullah Tumrani; investigation and resources: Saifullah Tumrani, Abdul Jabbar Siddiqui; writing—original draft: Saifullah Tumrani;

writing—review and editing: Saifullah Tumrani, Abdul Jabbar Siddiqui. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the corresponding author, Saifullah Tumrani, upon reasonable request.

Ethics Approval: Not applicable. This study does not involve human or animal subjects.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Wang H, Hou J, Chen N. A survey of vehicle re-identification based on deep learning. IEEE Access. 2019;7:172443–69. doi:10.1109/ACCESS.2019.2956172.
2. Khan SD, Ullah H. A survey of advances in vision-based vehicle re-identification. Comput Vis Image Underst. 2019;182:50–63. doi:10.1016/j.cviu.2019.03.001.
3. Liu X, Liu W, Mei T, Ma H. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In: Proceedings of the 14th European Conference on Computer Vision (ECCV 2016); 2016 Oct 11–14. Amsterdam, The Netherlands. Cham, Switzerland: Springer International Publishing; 2016. Vol. 9906. p. 869–84. doi:10.1007/978-3-319-46475-6.
4. Liu H, Tian Y, Wang Y, Pang L, Huang T. Deep relative distance learning: tell the difference between similar vehicles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE; 2016. p. 2167–75. doi:10.1109/CVPR.2016.238.
5. Lou Y, Bai Y, Liu J, Wang S, Duan L-Y. VERI-Wild: a large dataset and a new method for vehicle re-identification in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE; 2019. p. 3235–43. doi:10.1109/CVPR.2019.00335.
6. Tang Z, Naphade M, Liu M-Y, Yang X, Birchfield S, Wang S, et al. CityFlow: a city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE; 2019. p. 8797–8806. doi:10.1109/CVPR.2019.00900.
7. Wang P, Jiao B, Yang L, Yang Y, Zhang S, Wei W, et al. Vehicle re-identification in aerial imagery: dataset and approach. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE; 2019. p. 460–9. doi:10.48550/arXiv.1904.01400.
8. Bashir RMS, Shahzad M, Fraz MM. VR-PROUD: vehicle Re-identification using PROgressive unsupervised deep architecture. Pattern Recognit. 2019;90:52–65. doi:10.1016/j.patcog.2019.01.008.
9. Tumrani S, Deng Z, Lin H, Shao J. Partial attention and multi-attribute learning for vehicle re-identification. Pattern Recognit Lett. 2020;138:290–7. doi:10.1016/j.patrec.2020.07.034.
10. Tumrani S, Ali W, Kumar R, Khan AA, Dharejo FA. View-aware attribute-guided network for vehicle re-identification. Multimed Syst. 2023;29(4):1853–63. doi:10.1007/s00530-023-01077-y.
11. Gong R, Zhang X, Pan J, Guo J, Nie X. Vehicle reidentification based on convolution and vision transformer feature fusion. IEEE Multimed. 2024;31(2):61–8. doi:10.1109/MMUL.2024.3398189.
12. Taleb H, Wang C. Transformer-based vehicle re-identification with multiple details. In: Su R, Liu H, editors. Proceedings of the Fifth International Conference on Image, Video Processing, and Artificial Intelligence (IVPAI 2023). Vol. 13074. Bellingham, WA, USA: SPIE; 2024. 130740H p. doi:10.1117/12.3024821.
13. Li J, Yu C, Shi J, Zhang C, Ke T. Vehicle re-identification method based on Swin-Transformer network. Array. 2022;16:100255. doi:10.1016/j.array.2022.100255.
14. Du L, Huang K, Yan H. ViT-ReID: a vehicle re-identification method using visual transformer. In: Proceedings of the 2023 3rd International Conference on Neural Networks, Information and Communication Engineering (NNICE). Piscataway, NJ, USA: IEEE; 2023. p. 287–90. doi:10.1109/NNICE58320.2023.10105738.
15. Qian Y, Barthélemy J, Du B, Shen J. Paying attention to vehicles: a systematic review on transformer-based vehicle re-identification. ACM Trans Multimed Comput Commun Appl. 2024;418:114. doi:10.1145/3655623.

16. Kishore R, Aslam N, Kolekar MH. PATReId: pose apprise transformer network for vehicle re-identification. IEEE Trans Emerg Top Comput Intell. 2024;8(5):3691–702. doi:10.1109/TETCI.2024.3372391.
17. Lu Z, Lin R, Hu H. MART: mask-aware reasoning transformer for vehicle re-identification. IEEE Trans Intell Transp Syst. 2023;24(2):1994–2009. doi:10.1109/TITS.2022.3219593.
18. Qian W, Luo H, Peng S, Wang F, Chen C, Li H. Unstructured feature decoupling for vehicle re-identification. In: Proceedings of the Computer Vision-ECCV 2022: 17th European Conference on Computer Vision; 2022 Oct 23–27; Tel Aviv, Israel. Cham, Switzerland: Springer; 2022. Vol. 13674, p. 336–53. doi:10.1007/978-3-031-19781-9.
19. Wang Z, Tang L, Liu X, Yao Z, Yi S, Shao J, et al. Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE; 2017. p. 379–87. doi:10.1109/ICCV.2017.49.
20. Liu X, Zhang S, Huang Q, Gao W. RAM: a region-aware deep model for vehicle re-identification. In: Proceedings of the 2018 IEEE International Conference on Multimedia and Expo (ICME). Piscataway, NJ, USA: IEEE; 2018. p. 1–6. doi:10.1109/ICME.2018.8486589.
21. Yi X, Wang Q, Liu Q, Rui Y, Ran B. Advances in vehicle re-identification techniques: a survey. Neurocomputing. 2025;614:1–23. doi:10.1016/j.neucom.2024.128745.
22. 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 (ICCV). Piscataway, NJ, USA: IEEE; 2021. p. 14993–5002. doi:10.48550/arXiv.2102.04378.
23. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE; 2021. p. 9992–10002. doi:10.1109/ICCV48922.2021.00986.
24. Qiu M, Christopher L, Li L. Study on aspect ratio variability toward robustness of vision transformer-based vehicle re-identification. arXiv:2407.07842, 2024.
25. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. SSD: single shot MultiBox detector. In: European Conference on Computer Vision (ECCV). Cham, Switzerland: Springer; 2016. p. 21–37. doi:10.1007/978-3-319-46448-0_2.
26. Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification. arXiv:1703.07737. 2017.
27. Li M, Liu J, Zheng C, Huang X, Zhang Z. Exploiting multi-view part-wise correlation via an efficient transformer for vehicle re-identification. IEEE Trans Multimed. 2023;25:919–29. doi:10.1109/TMM.2021.3134839.
28. Hu W, Zhan H, Shivakumara P, Pal U, Lu Y. TANet: text region attention learning for vehicle re-identification. Eng Appl Artif Intell. 2024;133(Part E):108448. doi:10.1016/j.engappai.2024.108448.
29. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE; 2017. p. 618–26. doi:10.1109/ICCV.2017.74.