Attention Based Deep Learning for Vehicle Re-Identification

Siham Hussein Ali and Eman Hato
Department of Computer Science, College of Science, Mustansiriyah University, 10011 Baghdad, Iraq
sihamhussien90@uomustansiriyah.edu.iq, emanhato@uomustansiriyah.edu.iq

Keywords: Attention Mechanisms, ResNet-50 Architecture, VeRi-776 Dataset, Mean Average Precision, Rank-1 Accuracy, Vehicle Re-Identification.

Abstract: Vehicle Re-Identification (Re-ID) is a crucial task in intelligent transportation systems, where vehicles are matched from different camera angles. While deep learning has significantly improved this process, challenges related to varying camera angles, obscuring, lighting variations, and similar factors persist. Attention mechanisms assign different weights to different input elements, enabling the model to focus more precisely on the important information that improves its performance. This paper presents the impact of a multi-level attention fusion framework for vehicle identification, integrating the ResNet-50 architecture with multiple attention mechanisms, including spatial attention mechanisms, cross-channel attention, self-attention, and mutual attention. This design allows the network to learn both precise local identity indicators and overall contextual relationships in a unified and efficient manner. Extensive experiments on the VeRi-776 dataset demonstrate that attention fusion with a two-stage convolutional neural network mechanism outperformed previous methods in terms of accuracy and robustness across the dataset's metrics, achieving the highest scores of 85.1% on the mAP scale and 94.1% on the Rank-1 scale. Therefore, these results suggest that future work should incorporate adaptive, multi-level attention.

1 INTRODUCTION

The continuous development in ITS has turned traditional surveillance and traffic monitoring into intelligent, data-driven infrastructures that can make decisions independently. In this direction, one of the important sub-research topics is Vehicle Re-Identification, which allows detecting the same vehicle in non-overlapping cameras in large urban cities. Unlike traditional methods of vehicle detection or tracking, which rely on continuity in position or general appearance cues, Re-ID resolves a much harder problem-that is, identifying vehicles under significant changes in illumination, angle, and occlusion. Early works showed that solving this challenge requires learning relational distances in the feature space rather than relying on handcrafted descriptors [1]. As deep learning matured, researchers developed discriminative embedding architectures that improved intra-class compactness and inter-class separability, allowing models to distinguish visually similar vehicles with higher precision. Further refinement came with region-aware representation learning, which emphasized localized visual cues such as headlights, license plates, and vehicle contours. This methodological evolution marked the transition from global visual representations toward spatially selective modeling, substantially improving robustness under real-world conditions. ResNet-50 remains popular for vehicle redefinition for various reasons, including stable gradients, capture of low-level and high-level features, good accuracy without excessive computation, and ease of adding attention units [2]. Attention modules enable the network to focus adaptively on the most discriminative regions, unlike traditional convolutional models that treat all visual cues equally. This is a selective perception similar to the human visual system, allowing the models to capture subtle distinctions among visually similar vehicles and preserve identity consistency across multiple camera views [3]. Meanwhile, the research went on in parallel toward illumination-adaptive and stage-wise attention mechanisms that could efficiently handle variation in the environment and transitions of lighting conditions. This process critically enhances cross-domain generalization while saving computational overhead, which has been an essential necessity for intelligent transportation and real-time surveillance systems. These innovations turned vehicle Re-ID into a context-aware reasoning

process rather than a simple image-level matching task capable of understanding spatial and semantic relationships across diverse domains [4].

2 ATTENTION MECHANISM CATEGORIES

Attention mechanisms provide different weights to different input elements so that the model can focus more on important information [5]. The following subsections provide a brief description of attention mechanism categories:

2.1 Spatial Attention Mechanism

Spatial attention has been developed to guide the neural networks towards the most informative regions within an image. It enables the network to learn where to focus by generating an adaptive spatial attention map that highlights identity-relevant regions such as headlights, license plate areas, or distinctive structural contours. The mechanism first computes channel-wise average and max pooling along the channel dimension, then concatenates and applies a convolution and sigmoid activation function to produce a location-aware weight map. The resulting attention map is then multiplied with the original feature maps, enabling the network to selectively amplify spatially informative regions while downplaying noise and uninformative areas [6].

2.2 Channel Attention Mechanism

Channel attention is a feature recalibration mechanism that enables a neural network to learn which feature channels are most relevant for identity recognition. The model first performs global average pooling across the spatial dimension to summarize each channel's contribution, followed by a lightweight fully connected transformation that generates adaptive channel-wise weights. These weights are then multiplied back into the original feature maps, allowing the model to enhance highly informative channels while suppressing noisy ones [7].

2.3 Self Attention Mechanism

Self-Attention (known as Transformer-based Attention) is a mechanism that enables the model to understand global relationships by allowing every feature in the image to attend to every other feature regardless of spatial distance. Unlike CNNs, which operate with local receptive fields, self-attention captures long-range dependencies directly, making it especially powerful for vehicle re-identification across multiple non-overlapping camera views. This mechanism computes attention scores between all pairs of patches, allowing the model to learn context-aware relationships, such as how the front lights relate to the roof structure or how the license plate relates to the rear bumper, even if they are spatially far apart [8].

2.4 Cross-Attention Mechanism

Cross-attention is an attention mechanism specifically designed to model the interaction between two different feature representations, rather than focusing on a single input. It enables a feature map from Image A (Query) to attend directly to the most relevant and identity-informative regions of Image B (Key/Value). It helps the network highlight only the mutually consistent identity cues between views while suppressing irrelevant discrepancies caused by viewpoint, illumination, or occlusion [9]. Table 1 provides the mathematical formula of attention mechanism type, while Table 2 provides a brief description of attention mechanism types [8], [9].

3 DATASETS AND EVALUATION METRICS

Performance evaluation of vehicle Re-ID models fundamentally depends on both the diversity of benchmark datasets and the rigor of the employed evaluation metrics [10].

3.1 Benchmark Datasets

Benchmark datasets play a crucial role in the development and evaluation of vehicle re-identification systems. They provide standardized testing environments that enable fair comparison of algorithms across different real-world conditions and application scenarios. The most commonly used datasets differ in scale, camera coverage, environmental conditions, and annotation richness, allowing researchers to assess model accuracy, scalability, and robustness comprehensively.

- VeRi-776 Dataset remains the most widely used benchmark, comprising over 50,000 images of 776 vehicles captured by 20 non-

overlapping cameras. Its spatio-temporal annotations support cross-camera matching and trajectory-based reasoning, making it a cornerstone for measuring both identification accuracy and cross-domain robustness [11].

- VehicleID Dataset contains more than 200,000 images of 26,000 vehicles, primarily captured from frontal and rear viewpoints. It provides scalability tests for evaluating inter-class similarity and feature generalization across massive galleries [12].
- VERI-Wild Dataset is captured from an existing large CCTV camera system consisting of 174 cameras across one month (30×24h) under unconstrained scenarios. The cameras are distributed in a large urban district of more than 200km2 [13].
- PKU-Vehicle Dataset is a vehicle dataset for a large-scale real-world scenario of vehicle re-identification. It contains images of tens of millions of vehicles taken by real surveillance cameras in several Chinese cities. The dataset contains various locations, weather conditions (e.g. sunny, rainy, foggy), lighting (e.g. day and evening), shooting angle and hundreds of vehicle brands and other information [14].

In Table 1:

- σ= Sigmoid function;
- $f^{7\times7}$= Convolution operation;
- $Q = XW_Q$;
- $K = XW_K$;
- $V = XW_V$;
- d_k= dimension of keys;
- X= input feature matrix.

But in Cross-Attention, the Q comes from one source, K, V come from another source.

3.2 Evaluation Metrics

The most widely adopted metrics in contemporary Re-ID research are:

1) Mean Average Precision (mAP). The Mean Average Precision (mAP) reflects overall retrieval quality across all ranks, which is crucial for surveillance systems that need to present multiple candidates matches to human operators. It is mathematically expressed in (1) [15]:

$$mAP = \frac{1}{Q}\sum_{i=1}^{Q}\left(\sum_{k=1}^{m_i}\frac{1}{m_i}\mathrm{P}(k)\times rel(k)\right), \quad (1)$$

where:

- Q denotes the total number of query images;
- $P(k)$ represents the precision at rank k;
- $rel(k)$ is a binary indicator of relevance, equal to 1 if the retrieved item is relevant and 0 otherwise.

A higher mAP value indicates improved retrieval stability and global ranking precision across multiple camera views [15]:

2) Rank-1 Accuracy. The Rank-1 Accuracy metric indicates the system's ability to make correct instantaneous decisions, essential for automated toll collection and law enforcement applications. This metric follows the standard definition of Rank-1 Accuracy commonly used in Re-ID systems, which is defined as (2) [16], [17]:

$$\text{Rank}-1 = \frac{N_{Correct}}{N_{total}}\times 100\%, \quad (2)$$

where $N_(correct)$ and $N_(total)$ denote the number of correctly identified queries and the total number of queries, respectively. The mAP and Rank-1metrics may not fully capture performance under extreme occlusion or severe illumination changes [18].

Table 1: Mathematical formula of attention mechanism types.

Mechanism	Focus	Formula Core
Spatial Attention	Location (H×W)	$F' = M_s = \sigma(f^{7\times7}([F^s_{avg}; F^s_{max}]) \otimes F$ $F^s_{avg} = \text{AvgPool}(F) \in \mathbb{R}^{1\times H\times W}$ $F^s_{max} = \text{MaxPool}(F) \in \mathbb{R}^{1\times H\times W}$
Channel Attention	Feature Maps (C)	$F' = \sigma(MLP(F^c_{avg}) + MLP(F^c_{max})) \otimes \text{MaxPool}(F)$
Self-Attention	Within the same feature	$\text{Self-Attention}(Q,K,V) = \text{Softmax}(\frac{QK^T}{\sqrt{d_k}})V$
Cross Attention	Between features	$\text{Cross-Attention}(Q,K,V) = \text{Softmax}(\frac{QK^T}{\sqrt{d_k}})V$

Table 2: The description of attention mechanism types.

Type	Mechanism	Best Use Case	Strength	Limitation
Spatial Attention	Generates a spatial attention map to highlight important pixel-level regions	Fine-grained region localization within a single image	Enhances discriminative local regions	Lacks semantic understanding, does not know what the features mean
Channel Attention	Recalibrates feature maps by assigning adaptive weights to each channel.	Enhancing semantic identity cues	Strong in semantic feature selection	Ignores spatial location - does not know *where* features occur
Self-Attention	Every patch attends to every other patch as in Vision Transformers	Transformer-based global feature modeling	Captures long-range dependencies and global context	Computationally expensive - requires large resources
Cross-Attention	Query of Image A attends to the Key/Value from Image B for feature alignment	Vehicle Re-ID across multiple camera views	Learns identity-level correlation across different views	Requires paired inputs, not usable for single-image processing

4 LITERATURE REVIEW

The following section highlights key studies and provides a brief description of each, illustrating the developments in this area.

Zheng et al. [19] proposed a multi-scale attention framework for vehicle re-identification. The network uses a base feature map from the first three residual blocks of ResNet-50 to create multiple feature maps with different resolutions. Each scale is propagated separately through an independent subnetwork that incorporates a spatial-channel attention module, where the spatial branch emphasizes informative regions and the channel branch dynamically re-weights discriminative feature channels. The framework relies on empirically predefined scale configurations, which may incur additional computational cost and reduce tuning flexibility in large real-time systems.

Rong et al. [20] presented a multi-branch network architecture for vehicle re-identification, used to extract complementary global and local feature representations. It consists of three parallel branches, including the Global Branch, which encodes the full vehicle appearance without spatial partitioning, and the two Local Branches, which apply horizontal feature-map segmentation. Furthermore, the proposed method incorporates a Channel Attention Module, which recalibrates channel-wise importance via Global Average Pooling (GAP), followed by two 1×1 convolutional layers, and then fuses the output back into the input feature map via residual channel superposition. The framework relies on manually predefined horizontal partitioning schemes and statically assigned weighting coefficients rather than learning these configurations adaptively.

A dual-stage vehicle re-identification framework called DRA-OVG was presented by Li et al. [21] and was specifically created to improve viewpoint robustness and discriminative region awareness. It is composed of two tightly integrated components. First, in Discriminative Region Attention (DRA), the network automatically identifies and strengthens vehicle-specific identity regions, such as vehicle headlights, emblem logos, or even windshield markings, by adaptively learning where the most discriminative spatial cues are located. The second stage, Orthogonal View Generation (OVG), addresses the viewpoint dependency issue without the need for additional physical camera viewpoints. Incorporating the DRA and the OVG into a unified architecture ensures that the extracted representation captures both the fine-grained local distinctiveness and the robust multi-view adaptability. However, the orthogonal view synthesis relies on a manually constrained and implicitly linear transformation.

Pan et al. [22] present a Progressively Hybrid Transformer (PHT) framework for multi-modal vehicle re-identification by using long-range global dependency modelling. The framework integrates multi-modal information in two complementary dimensions: feature-level fusion via a multi-modal hybrid transformer with the feature hybrid mechanism, where each modality is first processed independently before fusing through cross-modal attention layers, and image-level fusion via a Random Hybrid Augmentation (RHA) module, which randomly mixes various modalities before feature extraction to inject complementary visual cues early. The fusion policy is still manually predefined and restricted to fixed dual-modality configurations.

Shen et al. [23] proposed a graph interactive transformer for vehicle re-identification. Each node in the suggested framework's vehicle part-level graph represents a localized semantic region. The edges are updated dynamically based on cross-part similarity to

model the structural correlations. The graph is then injected into a Transformer backbone by a bidirectional interaction mechanism that refines the feature tokens of the Transformer globally by self-attention, while the graph branch reinforces relational geometric priors. The proposed hybrid architecture leads to much more robust identity representations against viewpoint shifting and occlusion. The cross-modal and temporal priors have yet to be integrated, and there is a further opportunity to extend graph-transformer co-learning to multi-sensor applications.

We et al. [24] proposed a Two-Stage Progressive Learning (TSPL) framework, which was designed to address the severe illumination discrepancy and fine-grained similarity problem in vehicle re-identification. The framework begins with an illumination-aware metric learning stage where images are first coarsely classified as daytime or nighttime, and two separate feature spaces are learned via dual CNN branches with no shared weights. At the second stage, TSPL comes up with a detail-aware local feature extraction mechanism, incorporating an attention-guided module focusing on the identity-critical region, while adopting a model-level hard negative strategy to discriminate between visually similar vehicles of the same model. The learning strategy still essentially lives in RGB-only modalities.

Guo et al. [25] introduced a dual-pooling attention framework for UAV-based vehicle re-identification. The proposed framework uses a ResNet-50 CNN backbone and adds a Dual-Pooling Attention (DPA) module that uses both channel pooling to selectively amplify identity-relevant feature channels and spatial pooling to concentrate on informative structural regions. This joint spatial-channel recalibration allows the network to enhance both fine-grained local identity cues and global semantic consistency. The framework cannot model global dependencies and conduct long-range cross-view reasoning compared to transformer-based architectures.

Huang et al. presented the transformer-enhanced vehicle re-identification framework [26]. The proposed network presents the Relation Attention (RA) module and the Nuance-Disparity Mask (N&D Mask) mechanism. The RA module learns relational dependencies that align multi-view features within a shared embedding space using a lightweight transformer encoder to capture the structural relationships between vehicle components. The N&D masks are automatically generated using feature activations. The model's robustness across various dataset scales, but is still limited to RGB-only modalities and relies on manually adjusted parameters for mask generation.

Holla et al. [27] presented the Multi-Scale Feature Fusion Transformer (MSFFT) to capture both local structural cues (such as headlights, rooftops, and logos) and global semantic relationships through self-attention. The proposed MSFFT combines Transformer encoders and Inception layers into a single architecture. A two-phase training approach was used: first, cross-view augmentations between CCTV and UAV images were used to apply self-supervised semantic learning; after that, 500 epochs of fine-tuning were carried out using the Adam optimizer. The lack of spatial-temporal modeling and the high computational cost still hinder the use of MSFFT framework.

Bai and Rong [28] introduced a deep learning framework for vehicle re-identification to deal with the challenge of distinguishing between vehicles of the same model. To strengthen the ResNet-50 backbone, three non-local attention blocks are used after convolution layers to enable the network to capture long-range spatial dependencies across distant image regions, thereby allowing it to understand structural relationships among vehicle components such as license plates, roofs, and headlights. The authors used a multi-objective loss function that combined Softmax loss, Center loss, and Weighted Regularized Triplet loss to improve feature discrimination. This combined maximizes inter-class separation while minimizing intra-class variance.

Wang et al. [29] combined CNNs employing transformer-based self-attention for vehicle-based Re-identification in a hybrid model. For reducing the effects of illumination variations, the model employs a ResNet50-IBN as a backbone, which integrates instance and batch normalization. For outpatient alleviation of inter-branch conflict between classification loss and metric learning, as well as to maintain a balance between them. Even though the model has completed impressive work for cross-view ReID, it's only efficient for an RGB-based image.

A multi-axis compression fusion net is proposed by Ma et al. [30] for car re-identification, focusing on resolving the issue of poor discriminatory capabilities in current deep learning-based approaches. To handle both horizontal and vertical relationships between features, a multi-axis attention mechanism is incorporated. Later, to adaptively compress and amplify features, a compression fusion module is formed. The performance is promising, yet the proposed network has difficulty in complex occlusions and dynamic camera angles. Table 3 presents the literature review, its techniques, the limitations, and the results obtained.

Table 3: Comparative summary of literature review.

Study	Technique	Strength	Weakness	Dataset	mAP	Rank-1
[19]	CNN	Multi-scale representation	Fixed scales	VeRi-776	62.89	92.1
[20]	Multi-CNN	Fine-grained local cues	Static partitions	VeRi-776	77.12	96.3
[21]	Dual-stage CNN	Region & viewpoint robustness	Limited view synthesis	VeRi-776	85.1	94.1
[22]	Hybrid	Multi-modal fusion	Manual fusion depth	RGBN100	79.9	92.7
[23]	Hybrid	Global + relational modeling	RGB-only	VeRi-776	83.9	95
[24]	Two-stage CNN	Day-night adaptation	Manual labeling	VERI-DAN	97.8	99.3
[25]	CNN	UAV viewpoint handling	No global modeling	VeRi-UAV	81.7	96.6
[26]	Transformer	Relational context modeling	Manual mask tuning	VeRi-776	83.5	97.7
[27]	Hybrid	Cross-platform learning	High cost	VeRi-776	82.8	95.9
[28]	CNN	Long-range dependency	High GPU usage	VeRi-776	80.33	97.1
[29]	CNN + Transformer	Efficient hybrid	RGB-only	VeRi-776	84.9	97.7
[30]	CNN	Lightweight & robust	No temporal info	VeRi-776	83.7	97.5

5 FRAMEWORK RESULTS DISCUSSION

Vehicle Re-Identification (ReID) aims to find a specific vehicle identity across multiple nonoverlapping cameras. The main challenge of this task is the large intra-class and small inter-class variability of vehicle appearance, illumination changes, different camera resolutions and complex backgrounds.

Therefore, a good vehicle detection algorithm is essential. Deep learning and the attention mechanisms have revolutionised Vehicle Re-Identification systems by focusing on different aspects of feature representation and overcoming the limitations of traditional methods. To visually represent the results in Table 3, a bar graph was used in Figure 1 for easier comprehension. Regarding Table 3, several important observations have been made. The Spatial-Channel Attention Module, which combines spatial and channel attention, effectively captures both global cues and local cues (including headlights and logos). Additionally, multi-resolution learning improves robustness against variations in viewpoint and illumination.

However, the module lacks adaptive scale learning and efficiency optimization, which negatively impacts identification accuracy; one study reported an average precision of 62.89 mAP. Spatial attention at the regional level improves focus on key vehicle areas, reduces background noise, and determines where to focus and how to generalize.

Although ResNet-50 is widely adopted as a backbone for vehicle re-identification, it primarily extracts global convolutional features and treats all spatial regions and feature channels equally. For example, the architecture in [20] consisted of:

- Backbone. ResNet-50.
- Attention. Channel Attention Module (Global Average Pooling is 1×1, conv is residual weighting).
- Results on VeRi-776 was mAP = 77.12% and Rank 1 = 96.3%.

This leads to suboptimal performance under viewpoint changes, occlusion, illumination variation, and high inter-class similarity. In Figure 2, the performance of ResNet-50 with simple attention mechanism on the VeRi-776 Dataset for Method 1 [19], Method 2 [20] and Method 3 [25] the poor performance of these structures is evident from the results obtained.

The integration of multi-level attention mechanisms with ResNet-50 has significantly enhanced Vehicle Re-Identification (Re-ID) performance on the VeRi-776 dataset. ResNet-50 serves as a strong backbone due to its deep residual learning, which mitigates vanishing gradients and enables effective feature extraction. When augmented with attention modules, the model can focus more precisely on discriminative vehicle regions while suppressing irrelevant background noise. For example, the architecture in [21] consisted of:

- Backbone. ResNet-50.
- Attention. Discriminative Region Attention automatically locates and enhances key vehicle regions (headlights, logos, plates).
- Results on VeRi-776 was mAP = 85.1% and Rank -1 = 94.1%.

The high result is reported in Table 4 within a hybrid converter that focuses on multi-mode fusion.

Table 4: Performance of ResNet-50 with multi-level attention on the VeRi-776 Dataset.

Study	Attention Mechanism	mAP	Rank-1
Method 4 [21]	Discriminative Region Attention (DRA) + Orthogonal View Generation + Dual-stage CNN	85.1	94.1
Method 5 23]	Graph Interactive Transformer (GiT) + part-level graph attention	83.9	95.0
Method 6 [26]	Hybrid	83.5	97.7
Method 7 [29]	Relation Attention (RA) + Nuance-Disparity Mask (Transformer)	84.9	97.7
Method 8 [30]	Efficient Self-Attention CNN + Transformer Hybrid	83.7	97.5

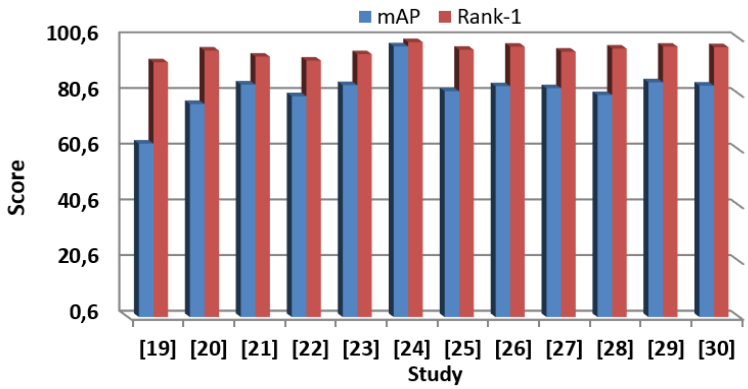


Figure 1: The mAP and Rank-1 score of previous studies.

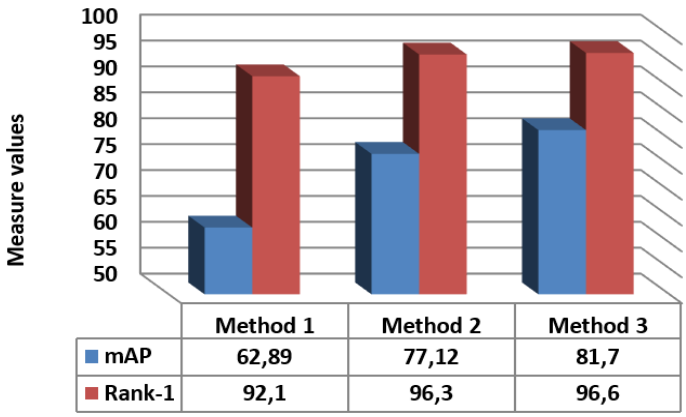


Figure 2: Performance of ResNet-50 with attention mechanism on the VeRi-776 Dataset.

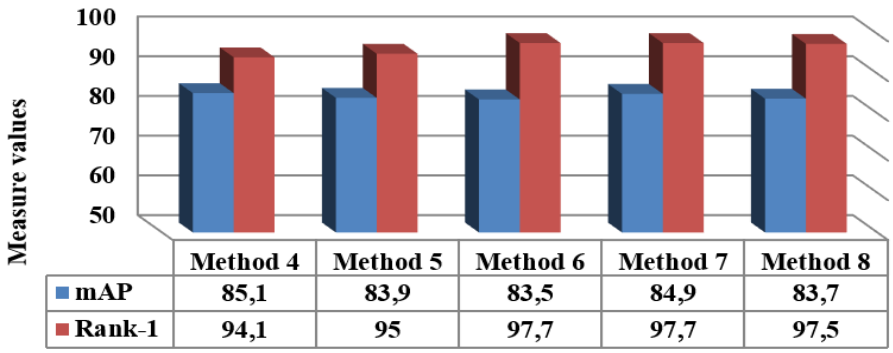
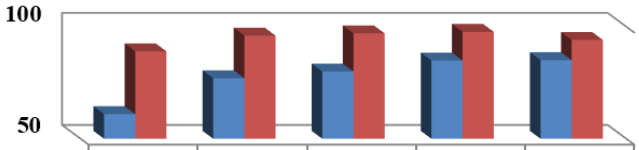


Figure 3: Results of ResNet-50 with multi-level attention on the VeRi-776 Dataset.



	ResNet-50 Baseline	ResNet-50 + Channel Attn	ResNet-50 + Spatial Attn	ResNet-50 + Transformer Attn	ResNet-50 + DRA (Region Attn)
mAP	61	77	80	84,9	85,1
Rank-1	89	96	97	97,7	94,1

Figure 4: Summary of ResNet-50 enhancement.

Figure 4 presents a visual comparison diagram illustrating the enhancement of Resnet-50 with different attentional mechanisms in vehicle reorientation

This has led to improved precision, reaching levels seen in some studies of 85% mAP, as illustrated in Figure 3. The main difficulties have been adapting to complex, real-world multi-camera setups and the associated computational cost. Relation attention using Multi-scale Transformer learns multi-view relational dependencies and provides high accuracy and robustness across dataset scales. The impact of using these techniques was reflected in the accuracy, which was recorded at 97.7% in terms of the Rank-1 accuracy metric. But these systems are very time-consuming. Earlier CNN-only attention performed lower mAP. Hybrid ResNet-50 and Transformer attention models (e.g., Relation Attention, Hybrid Transformer) now dominate, offering better global context and robustness. Therefore, attention mechanisms significantly boost ResNet-50 performance on VeRi-776.

However, when multiple attention mechanisms are combined, prior work uses simple concatenation or predefined fusion weights.

This static approach cannot adapt to varying input conditions for example, emphasizing spatial attention more under occlusion, or cross-attention more when matching across extreme viewpoint changes. Additionally, most studies report final performance without systematically quantifying each attention component's contribution. This makes it impossible to determine whether gains come from genuine architectural improvements or simply increased parameter counts.

6 CONCLUSIONS

The present work has comprehensively compared different vehicle identification methods based on attentional mechanisms, with the aim of explaining how various attentional strategies contribute to enhancing accuracy and effectiveness in the recognition of vehicles. This paper primarily focuses on the capabilities of attention mechanisms in extracting discriminatory features with better representations and higher identification rates in complex variations of viewpoints, lighting, and partial occlusion conditions. From the results, it is clear that attention mechanisms have a pivotal part to play in better vehicle identification rates, which is only going to further reinforce their importance in influencing the development of the next-generation intelligent transport systems.

ResNet-50 with modern attention mechanisms (region-based, transformer-based, illumination-aware) achieves state-of-the-art performance on VeRi-776, with results up to ~90% mAP and ~99% Rank 1 in specialized settings. The integration of attention is a proven and effective strategy for boosting Vehicle Re-ID accuracy.

Through the analysis of spatial, channel, and self-attention mechanisms in hybrid transformer-based models, it can be observed that attention modules enable the model to focus on the most informative vehicle regions and discriminative attributes.

Among the methods considered, transformer-based and multi-level attention models generally demonstrate superior performance compared to convolutional attention approaches across most benchmark datasets, including VeRi-776. This advantage can be attributed to their enhanced ability to capture complex spatial relationships and model global contextual information.

Despite these advances, several challenges remain open for research: computational complexity, scalability of large datasets, and sensitivity to background noise. Overcoming these limitations necessitates that future research address simple attentional structures, effective training strategies, and large-scale datasets with richer contexts. Analysis of inference time has shown that cross-attention is computationally expensive. For future work, a lightweight attention design can be introduced to develop efficient attention variables. Furthermore, adaptive attentional routing demonstrates that different attentional mechanisms excel under different conditions, dynamically selecting attentional units based on input characteristics.

REFERENCES

[1] A. Farid, F. Hussain, K. Khan, M. Shahzad, U. Khan, and Z. Mahmood, "A fast and accurate real-time vehicle detection method using deep learning for unconstrained environments," Applied Sciences, vol. 13, no. 5, 2023, doi: 10.3390/app13053059.

[2] S. H. Mousa, N. M. Shati, and N. Sakthivadivel, "DeepRing: Convolution neural network based on blockchain technology," Al-Mustansiriyah Journal of Science, vol. 35, no. 2, pp. 61-69, Jun. 2024.

[3] S. Abdulhaleem and E. Hato, "Video skimming using object detection and relation-based importance score," International Journal of Intelligent Engineering Systems, vol. 18, no. 8, pp. 485-498, 2025, doi: 10.22266/ijies2025.0930.30.

[4] G. Szűcs, R. Borsodi, and D. Papp, "Multi-camera trajectory matching based on hierarchical clustering and constraints," Multimedia Tools and Applications, vol. 83, no. 15, pp. 44879-44902, 2024.

[5] M. R. Shuvo, M. S. Mekala, and E. Elyan, "Deep learning and attention-based methods for human activity recognition and anticipation: A comprehensive review," Cognitive Computation, 2025, doi: 10.1007/s12559-025-10513-2.

[6] C. M. Mylonakis, P. Velanas, P. I. Lazaridis, P. Sarigiannidis, S. K. Goudos, and Z. D. Zaharis, "Deep learning framework using spatial attention mechanisms for adaptable angle estimation across diverse array configurations," Technologies, vol. 13, no. 2, 2025, doi: 10.3390/technologies13020046.

[7] Y. Wang, W. Wang, Y. Li, Y. Jia, Y. Xu, Y. Ling, and J. Ma, "An attention mechanism module with spatial perception and channel information interaction," Complex & Intelligent Systems, vol. 10, no. 4, pp. 5427-5444, 2024.

[8] B. Alnasyan, M. Basheri, M. Alassafi, and K. Alnasyan, "Kanformer: An attention-enhanced deep learning model for predicting student performance in virtual learning environments," Social Network Analysis and Mining, vol. 15, no. 1, 2025.

[9] M. Gheini, X. Ren, and J. May, "Cross-attention is all you need: Adapting pretrained transformers for machine translation," in Proc. 2021 Conference on Empirical Methods in Natural Language Processing, pp. 1754-1765, 2021.

[10] Y. Qian, J. Barthelemy, E. Karuppiah, and P. Perez, "Identifying re-identification challenges: Past, current and future trends," SN Computer Science, vol. 5, no. 7, 2024.

[11] X. Sun, Y. Chen, Y. Duan, Y. Wang, J. Zhang, B. Su, and L. Li, "Vehicle re-identification method based on multi-attribute dense linking network combined with distance control module," Frontiers in Neurorobotics, vol. 17, 2023, doi: 10.3389/fnbot.2023.1294211.

[12] L. Fei and B. Han, "Multi-object multi-camera tracking based on deep learning for intelligent transportation: A review," Sensors, 2023.

[13] Y. Lou, Y. Bai, J. Liu, S. Wang, and L.-Y. Duan, "VERI-wild: A large dataset and a new method for vehicle re-identification in the wild," in Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 3230-3238, 2019.

[14] H. Wang, J. Hou, and N. Chen, "A survey of vehicle re-identification based on deep learning," IEEE Access, 2019, doi: 10.1109/ACCESS.2019.2956172.

[15] Y. Bai, Y. Lou, F. Gao, S. Wang, Y. Wu, and L.-Y. Duan, "Group-sensitive triplet embedding for vehicle re-identification," IEEE Transactions on Multimedia, vol. 20, no. 9, pp. 2385-2399, 2018.

[16] Z. Sun, X. Wang, Y. Zhang, Y. Song, J. Zhao, J. Xu, W. Yan, and C. Lv, "A comprehensive review of pedestrian re-identification based on deep learning," Complex & Intelligent Systems, vol. 10, no. 2, 2024, doi: 10.1007/s40747-023-01229-7.

[17] S. Liu and S. S. Agaian, "VERI-D: A new dataset and method for multi-camera vehicle re-identification of damaged cars under varying lighting conditions," APL Machine Learning, vol. 2, no. 1, 2024.

[18] R. S. M. Saad, M. M. Moussa, N. S. Abdel-Kader, and H. Farouk, "Deep video-based person re-identification (Deep Vid-ReID): Comprehensive survey," EURASIP Journal on Advances in Signal Processing, vol. 2024, no. 1, 2024.

[19] A. Zheng, X. Lin, J. Dong, W. Wang, J. Tang, and B. Luo, "Multi-scale attention vehicle re-identification," Neural Computing and Applications, 2020, doi: 10.1007/s00521-020-05108-x.

[20] L. Rong, Y. Xu, X. Zhou, L. Han, L. Li, and X. Pan, "A vehicle re-identification framework based on the improved multi-branch feature fusion network," Scientific Reports, vol. 11, no. 1, 2021.

[21] H. Li, Y. Wang, Y. Wei, L. Wang, and L. Ge, "Discriminative-region attention and orthogonal-view generation model for vehicle re-identification," Applied Intelligence, vol. 53, no. 1, pp. 186-203, 2023.

[22] W. Pan, L. Huang, J. Liang, L. Hong, and J. Zhu, "Progressively hybrid transformer for multi-modal vehicle re-identification," Sensors, vol. 23, no. 9, 2023.

[23] F. Shen, Y. Xie, J. Zhu, X. Zhu, and H. Zeng, "GiT: Graph interactive transformer for vehicle re-identification," IEEE Transactions on Image Processing, vol. 32, 2023.

[24] Z. Wu, Z. Jin, and X. Li, "Two-stage progressive learning for vehicle re-identification in variable illumination conditions," Electronics, vol. 12, no. 24, 2023.

[25] X. Guo, J. Yang, X. Jia, C. Zang, and Z. Chen, "A novel dual-pooling attention module for UAV vehicle re-identification," Scientific Reports, vol. 14, no. 1, 2024, doi: 10.1038/s41598-024-52225-x.

[26] Y. Huang, H. Sheng, and W. Ke, "RANDnet: Vehicle re-identification with relation attention and nuance-disparity masks," Applied Sciences, vol. 14, no. 11, 2024.

[27] A. Holla B., M. M. Pai, U. Verma, and R. M. Pai, "MSFFT: Multi-scale feature fusion transformer for cross platform vehicle re-identification," Neurocomputing, vol. 582, 2024.

[28] L. Bai and L. Rong, "Vehicle re-identification with multiple discriminative features based on non-local-attention block," Scientific Reports, vol. 14, no. 1, 2024, doi: 10.1038/s41598-024-82755-3.

[29] Y. Wang, R. Li, and Y. Shao, "Vehicle re-identification method based on efficient self-attention CNN-transformer and multi-task learning optimization," Sensors, vol. 25, no. 10, 2025, doi: 10.3390/s25102977.

[30] T. Ma, K. Sun, X. Pang, W. Si, T. Liu, and C. Wang, "Multi-axis compression fusion network for vehicle re-identification," Scientific Reports, vol. 15, no. 1, 2025, doi: 10.1038/s41598-025-15854-4.