Available online at http://www.journalcra.com

INTERNATIONAL JOURNAL
OF CURRENT RESEARCH

International Journal of Current Research
Vol. 18, Issue, 01, pp.35915-35921, January, 2026
DOI: https://doi.org/10.24941/ijcr.50035.01.2026

ISSN: 0975-833X

RESEARCH ARTICLE

STATE-OF-THE-ART AND RESEARCH GAP ANALYSIS OF ARTIFICIAL INTELLIGENCE BASED TECHNIQUES FOR AUTOMATIC VEHICLE RECOGNITION

Abakar Issakha Souleymane[1,*], Ahamat Mahamat Hassane[1] and Daouda Ahmat[2]

[1]Techno-Pedagogy Unit, Virtual University of Chad, P.O. Box 5711, N'Djamena, Chad
[2]IT Department, University of N'Djamena, P.O. Box 5711, N'Djamena, Chad

ARTICLE INFO

Article History:
Received 20th October, 2025
Received in revised form
17th November, 2025
Accepted 28th December, 2025
Published online 30th January, 2026

Keywords:
Automatic Vehicle Recognition, Deep Learning, Computer Vision, Transformers, Intelligent Transportation Systems, Edge AI, Multimodal Fusion.

Corresponding author:
Abakar Issakha Souleymane

ABSTRACT

Automatic Vehicle Recognition (AVR) has become a cornerstone technology for Intelligent Transportation Systems, smart city infrastructure, traffic monitoring, and border security. Recent advancements in Artificial Intelligence (AI), particularly deep learning, have markedly improved vehicle detection, classification, and tracking. However, real-world deployment of AVR systems remains challenged by factors such as high computational demands, limited cross-domain generalization, scalability constraints, and the need for privacy-preserving operation. This paper presents a comprehensive state-of-the-art review and critical gap analysis of AI-based vehicle recognition techniques. We systematically examine contemporary approaches, including CNN-based one-stage and two-stage object detectors, transformer-based architectures with global self-attention, fine-grained vehicle classification models, and vehicle re-identification methods. Additionally, we explore multimodal sensor fusion strategies and Edge AI deployment to evaluate their effectiveness in enhancing robustness and real-time performance under diverse environmental conditions. The analysis identifies key limitations in current systems, such as the absence of unified end-to-end frameworks, the high computational cost of transformer models for edge deployment, insufficient generalization across geographic regions and vehicle types, and limited explainability and privacy safeguards. Finally, we outline promising research directions aimed at developing lightweight, adaptive, and privacy-conscious AVR systems capable of bridging the gap between laboratory research and large-scale real-world applications.

Citation: Abakar Issakha Souleymane, Ahamat Mahamat Hassane, Daouda Ahmat. 2026. "State-of-the-Art and Research Gap Analysis of Artificial Intelligence Based Techniques for Automatic Vehicle Recognition.". ***International Journal of Current Research, 18, (01), 35915-35921.***

I. INTRODUCTION

The rapid growth of vehicle traffic worldwide has created significant challenges for transportation management, public safety, and security infrastructures. Urban areas, highways, and international borders increasingly require intelligent systems capable of automatically detecting, classifying, and monitoring vehicles in real time. Automatic Vehicle Recognition (AVR) is a critical component of such systems and is used in a variety of applications, including traffic surveillance, law enforcement, toll collection, customs inspection, smart city management, and autonomous driving systems. AVR encompasses multiple tasks, such as vehicle detection, fine-grained classification, vehicle re-identification across multiple cameras, and, in some cases, automatic license plate recognition (ALPR). Traditional AVR approaches relied heavily on handcrafted features such as Histogram of Oriented Gradients (HOG), Haar-like features, and Scale-Invariant Feature Transform (SIFT), combined with classical classifiers such as Support Vector Machines or Random Forests. While these methods achieved reasonable performance in controlled environments, they are highly sensitive to real-world conditions, including variations in illumination, occlusions, viewpoint changes, weather conditions, and complex backgrounds. These limitations often result in reduced detection accuracy, high false positive rates, and poor scalability when deployed across large-scale traffic networks (1), (2). The emergence of Artificial Intelligence (AI), and particularly deep learning, has dramatically transformed the field of AVR. Convolutional Neural Networks (CNNs) enable end-to-end feature learning from raw images, eliminating the need for handcrafted descriptors and significantly improving detection and classification performance. More recently, transformer-based architectures and attention mechanisms have been introduced to capture global context, further enhancing recognition accuracy in crowded or complex scenes. Additionally, advanced tasks such as vehicle re-identification (Re-ID) and fine-grained classification allow systems to track vehicles across multiple cameras, identify their make, model, and color, and detect anomalous behavior. Edge AI and multimodal sensor fusion strategies provide further improvements in real-time performance, robustness under adverse conditions, and privacy preservation. Despite these advances, several critical challenges remain. Current AVR systems are often implemented as independent modules rather than unified, end-to-end frameworks, limiting scalability and adaptability. Transformer-based models, while highly accurate, are computationally expensive and difficult to deploy in resource-constrained environments. Generalization across geographic regions, vehicle types, and environmental conditions is still limited. Furthermore, issues related to explainability, transparency, and privacy protection have received insufficient attention, creating barriers to large-scale deployment in public safety and border control contexts. Motivated by these challenges, this paper presents a comprehensive state-of-the-art review of AI-based techniques for automatic vehicle recognition. It systematically examines modern deep learning detection frameworks, transformer-based architectures, fine-grained classification, vehicle re-identification, multimodal fusion, and Edge AI deployment. Furthermore, the paper identifies existing research gaps and proposes directions for future studies, bridging the divide between academic advancements and practical real-world applications in transportation and security systems.

II. RELATED WORK

Early Automatic Vehicle Recognition (AVR) systems relied on handcrafted features such as Histogram of Oriented Gradients (HOG), Haar-like features, and Scale-Invariant Feature Transform (SIFT), often combined with classical classifiers (11)–(15). While effective in controlled scenarios, these approaches struggled with real-world challenges such as varying illumination, occlusions, weather conditions, and complex backgrounds. Their limited robustness motivated the adoption of deep learning techniques (16)–(18). The introduction of Convolutional Neural Networks (CNNs) revolutionized vehicle detection and recognition. One-stage object detectors like YOLO and SSD provide high-speed detection suitable for traffic surveillance and real-time monitoring (19)–(23), while two-stage detectors such as Faster R-CNN and Mask R-CNN achieve higher accuracy and improved localization performance at the cost of increased computational complexity (24)–(27). Recent variants, including YOLOv4 through YOLOv11, have further improved detection precision, robustness to small and occluded vehicles, and real-time efficiency (28)–(33). Transformer-based architectures have recently been applied to vehicle detection to capture global contextual information. DETR and its derivatives utilize self-attention mechanisms for end-to-end detection without handcrafted anchors, while Swin Transformer leverages hierarchical attention windows for improved efficiency (34)–(38). These methods outperform traditional CNNs in crowded and complex traffic scenes (39)–(41). Fine-grained vehicle recognition extends detection to classify vehicle type, make, model, and color, enabling advanced applications such as forensic analysis and automated inspections. Deep CNN classifiers and hybrid CNN-LSTM models have been widely employed for these tasks (42)–(46). Additionally, remote sensing applications combine deep learning with aerial imagery for vehicle detection and classification, useful for border security and traffic management (47)–(49). Vehicle re-identification (Re-ID) focuses on matching the same vehicle across multiple cameras and times. State-of-the-art Re-ID systems employ deep metric learning, Siamese networks, attention-based feature extractors, and spatio-temporal modeling to improve cross-camera matching accuracy (50)–(54). These methods are particularly useful in multi-camera surveillance systems in smart cities and border control environments (55)–(57). Multimodal sensor fusion methods combine camera images with LiDAR, radar, or infrared data to enhance robustness under adverse conditions such as low-light, fog, or rain. Early fusion, feature-level fusion, and late fusion strategies have all been explored, with recent studies incorporating transformer-based cross-attention mechanisms for optimal multi-sensor integration (58)–(60). Benchmark datasets such as KITTI, Waymo Open, nuScenes, ApolloScape, and CityFlow provide multi-sensor annotations across varied scenarios, supporting training, evaluation, and comparison of deep learning models (11), (12), (19), (22), (25), (33), (38). These datasets are instrumental for developing robust detection, classification, and Re-ID algorithms. Edge AI deployment and optimization enable real-time vehicle recognition on resource-constrained devices. Techniques such as model compression, quantization, and hardware-aware neural architecture search allow high-accuracy models to run efficiently in embedded systems for traffic surveillance and customs inspection (13), (14), (20), (26), (29). Despite these advancements, most research focuses on isolated tasks (detection, classification, or Re-ID) rather than unified, end-to-end frameworks. Transformer-based models, while accurate, are computationally intensive, limiting practical deployment on edge devices. Generalization across geographic regions, diverse vehicle types, and environmental conditions is still limited. Furthermore, explainability and privacy-preserving approaches are rarely integrated, highlighting areas for future research (16), (18), (21), (24), (31), (35), (37).

III. DEEP LEARNING-BASED VEHICLE DETECTION

This section reviewed deep learning-based vehicle detection approaches, focusing on one-stage and two-stage object detection frameworks. One-stage detectors, such as YOLO and SSD, offer high processing speed and low latency, making them well suited for real-time applications including traffic surveillance and border control. In contrast, two-stage detectors, including Faster R-CNN and Mask R-CNN, provide superior localization accuracy and robustness in complex scenes at the expense of higher computational cost. Recent architectural improvements and optimization techniques have narrowed the performance gap between accuracy and efficiency, enabling broader deployment across both edge and high-performance computing environments.

One-Stage Object Detectors: One-stage object detectors perform vehicle localization and classification simultaneously in a single forward pass through the network, eliminating the need for a separate region proposal stage. This architectural design significantly reduces inference time, making one-stage detectors highly suitable for real-time and large-scale deployment scenarios. Among these models, the You Only Look Once (YOLO) family and the Single Shot MultiBox Detector (SSD) are the most widely adopted. YOLO-based detectors divide the input image into a grid and directly predict bounding boxes, objectness scores, and class probabilities, enabling fast and efficient detection even in dense traffic environments (2), (59), (60). SSD employs multi-scale feature maps to detect objects of varying sizes, improving detection performance for both small and large vehicles while maintaining high processing speed (20). Due to their low latency and relatively simple network pipelines, one-stage detectors are extensively used in applications requiring real-time responsiveness, such as traffic surveillance, intelligent transportation systems, toll monitoring, and border control operations. Recent improvements to one-stage architectures, including feature pyramid networks, attention mechanisms, and anchor-free designs, have further enhanced detection accuracy and robustness against occlusions and illumination variations without compromising speed (21), (22), (28), (33). These advancements make one-stage detectors a practical choice for deployment on edge devices and embedded platforms where computational resources are limited.

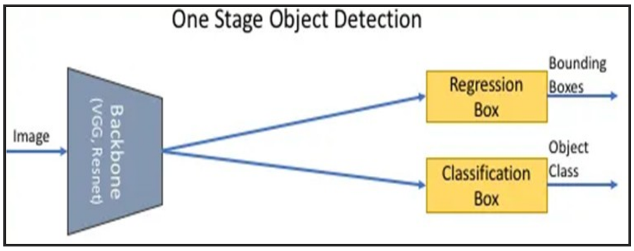

Fig. 1. Architecture of a one-stage deep learning detector for vehicle recognition (64)

Two-Stage Object Detectors: Two-stage object detectors decompose the vehicle detection task into two sequential phases: region proposal generation and object classification with bounding box refinement. In the first stage, a Region Proposal Network (RPN) identifies candidate regions that are likely to contain vehicles. In the second stage, these proposed regions are further processed to classify vehicle categories and refine bounding box coordinates. This two-step strategy significantly improves localization precision and reduces false detections, particularly in complex traffic scenes with occlusions and overlapping vehicles (3), (4). Faster R-CNN represents a major milestone in two-stage detection frameworks by integrating the RPN directly into the deep convolutional backbone, enabling end-to-end training and efficient feature sharing between proposal generation and classification stages (4). Mask R-CNN further extends this architecture by introducing a parallel segmentation branch, allowing pixel-level instance segmentation in addition to bounding box detection. This capability is particularly beneficial for precise vehicle boundary estimation, occlusion handling, and fine-grained analysis in surveillance and forensic applications (21). Despite their superior accuracy, two-stage detectors typically require higher computational resources and longer inference times compared to one-stage models. As a result, they are often deployed in offline analysis systems or high-performance computing environments where accuracy is prioritized over real-time constraints. However, recent optimizations, such as lightweight backbone networks, feature pyramid networks, and model compression techniques, have partially mitigated these limitations, making two-stage detectors increasingly viable for near real-time applications in traffic monitoring and border control scenarios (22), (37).

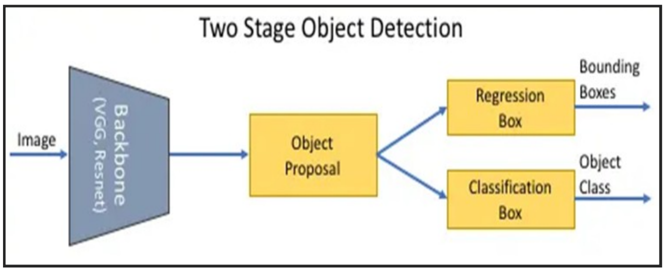

Fig. 2. Two-stage vehicle detection framework based on region proposal networks (64)

IV. TRANSFORMER-BASED ARCHITECTURES

Transformer-based models have recently emerged as powerful alternatives to convolutional neural networks for vehicle detection and recognition tasks. Unlike CNNs, which primarily rely on local receptive fields, Vision Transformers (ViTs) leverage self-attention mechanisms to model long-range dependencies and global contextual relationships across the entire image. This capability is particularly beneficial in complex traffic environments, where vehicles may be partially occluded, densely packed, or captured under challenging viewpoints and lighting conditions. The Detection Transformer (DETR) reformulates the object detection problem as a direct set prediction task, eliminating the need for handcrafted components such as anchor boxes and non-maximum suppression. DETR employs a transformer encoder–decoder architecture in which global self-attention enables the model to reason holistically about object relationships and scene context (5). While DETR demonstrates strong localization accuracy and conceptual simplicity, its original formulation suffers from slow convergence and high computational demands. Subsequent variants, including Deformable DETR and Efficient DETR, address these limitations by introducing sparse attention mechanisms and multi-scale feature representations, significantly improving training efficiency and detection performance in large-scale traffic scenes (34), (35). The Swin Transformer further advances transformer-based detection by introducing a hierarchical architecture with shifted window-based self-attention. This design allows the model to capture both local and global features while maintaining computational efficiency, making it well suited for high-resolution vehicle imagery (6). Swin-based detectors have demonstrated competitive performance on vehicle detection benchmarks, particularly in scenarios involving dense traffic flow, occlusions, and varying environmental conditions (38), (39). Hybrid architectures that combine CNN backbones with transformer modules have also gained attention. These models exploit the strong local feature extraction capabilities of CNNs alongside the global reasoning power of transformers, achieving a balance between accuracy and computational efficiency. Such hybrid approaches are increasingly explored for real-time vehicle detection and edge deployment in intelligent transportation systems and border surveillance applications (40), (41). Despite their advantages, transformer-based models remain computationally intensive and memory-demanding, posing challenges for deployment on resource-constrained devices. Ongoing research focuses on lightweight transformers, attention pruning, and hardware-aware optimization to enable practical real-time deployment without sacrificing detection accuracy.

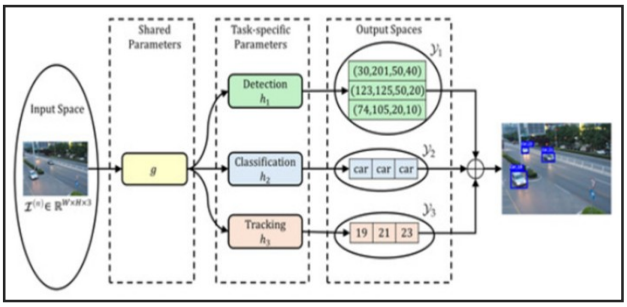

Fig. 3. Transformer-based vehicle detection framework illustrating global self-attention mechanisms for modeling long-range dependencies in complex traffic scenes (65)

V. FINE-GRAINED VEHICLE CLASSIFICATION

Fine-grained vehicle classification addresses the problem of discriminating between visually similar vehicle categories by identifying detailed semantic attributes such as vehicle type, manufacturer, model, production year, and color. Unlike generic vehicle detection, this task requires capturing subtle inter-class variations and low-level visual cues, making it a challenging yet essential component of intelligent transportation systems. Accurate fine-grained recognition supports high-level applications including forensic vehicle search, automated customs verification, cross-border vehicle tracking, and advanced traffic analytics. Recent advances in deep learning have significantly improved fine-grained vehicle classification performance. Deep convolutional neural networks (CNNs), including ResNet, DenseNet, EfficientNet, and Inception-based architectures, have demonstrated strong discriminative power when trained on large-scale vehicle datasets (42), (43), (46). These networks automatically learn hierarchical representations that encode fine visual patterns such as grille structures, headlamp geometry, logo placement, and body proportions. Transfer learning from large image classification benchmarks is commonly employed to improve convergence and generalization, particularly in scenarios where labeled vehicle data are limited. To further enhance recognition accuracy, state-of-the-art methods integrate attention mechanisms and part-based learning strategies. Attention-based models selectively emphasize discriminative vehicle regions, while part-aware approaches explicitly model key components such as headlights, wheels, and windshields, enabling robust differentiation between closely related vehicle models (44), (45). Multi-task learning frameworks that jointly predict multiple vehicle attributes—such as model and color—have also been shown to improve feature sharing and classification robustness. Moreover, recent research explores fine-grained recognition using vision transformers and hybrid CNN–Transformer architectures to better capture global structural relationships among vehicle components (61). In practical systems, fine-grained vehicle classification is typically implemented as a downstream module following vehicle detection. Detected vehicle instances are cropped, normalized, and passed to specialized classification networks, allowing modular system design and scalable deployment. Video-based approaches further exploit temporal consistency across frames to reduce ambiguity caused by motion blur, occlusion, or viewpoint variation. Despite substantial progress, fine-grained vehicle classification remains constrained by factors such as viewpoint diversity, occlusion, and the long-tailed distribution of vehicle models. Emerging research directions include synthetic data generation, self-supervised and few-shot learning, and cross-domain adaptation to enhance generalization across geographic regions and vehicle populations.

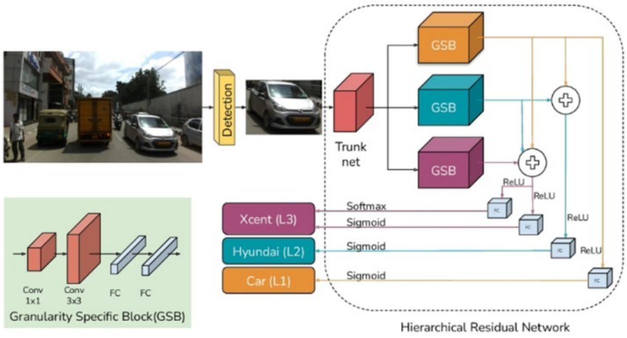

Fig. 4. Architecture for fine-grained vehicle detection (66)

VI. VEHICLE RE-IDENTIFICATION

Vehicle re-identification (Re-ID) addresses the problem of associating the same vehicle across multiple cameras and non-overlapping fields of view over different time instances. Unlike license plate recognition, which may fail due to occlusion, low resolution, or deliberate tampering, vehicle Re-ID relies on visual appearance cues and spatio-temporal consistency to achieve robust matching. This capability is essential for large-scale traffic surveillance, cross-camera vehicle tracking, forensic investigation, and border control applications. Modern vehicle Re-ID systems are predominantly based on deep metric learning, where the objective is to learn discriminative feature embeddings that minimize intra-class variations while maximizing inter-class separability. Siamese and triplet network architectures are commonly employed, training the network using contrastive or triplet loss functions to ensure that images of the same vehicle are mapped closer in the embedding space than those of different vehicles (7), (50). These embeddings encode a combination of global vehicle appearance and fine-grained details such as color distribution, shape, decals, and structural patterns. To further enhance Re-ID performance, recent studies incorporate attention mechanisms and part-based feature extraction to focus on discriminative vehicle regions, including headlights, license plate areas, wheels, and roof structures. Temporal modeling and spatio-temporal constraints are also integrated to exploit contextual information such as vehicle motion patterns, camera topology, and time consistency, significantly reducing false matches in large-scale camera networks (51), (52). Additionally, transformer-based Re-ID architectures have emerged as powerful alternatives to CNN-based models, enabling improved global feature reasoning and robustness to viewpoint variations (61). Vehicle Re-ID remains particularly challenging due to high inter-class similarity among vehicles of the same model and color, as well as significant intra-class variation caused by illumination changes, occlusions, and viewpoint diversity. To address these challenges, recent research explores multi-task learning strategies that jointly perform Re-ID and attribute recognition, as well as unsupervised and domain-adaptive Re-ID methods that reduce reliance on large labeled datasets. In real-world deployments, vehicle Re-ID systems are often integrated with vehicle detection and fine-grained classification modules to form end-to-end vehicle analytics pipelines. Such integrated frameworks are increasingly adopted in intelligent transportation systems and border surveillance platforms, enabling scalable, accurate, and privacy-aware vehicle tracking without exclusive dependence on license plate information.

VII. MULTIMODAL SENSOR FUSION

Vision-based vehicle recognition systems often suffer performance degradation under challenging environmental conditions such as low illumination, nighttime operation, fog, rain, and occlusions. To address these limitations, recent research increasingly focuses on multimodal sensor fusion, integrating complementary data from sensors such as infrared (IR) cameras, LiDAR, and radar with conventional RGB imagery. By exploiting the strengths of multiple sensing modalities, multimodal fusion significantly enhances robustness, reliability, and operational continuity in adverse conditions (8). Infrared and thermal imaging sensors provide valuable information in low-light and nighttime scenarios by capturing heat signatures that are invariant to illumination changes. LiDAR sensors contribute accurate depth and 3D structural information, enabling precise vehicle localization and shape estimation, while radar systems offer robust velocity and range measurements that remain effective under adverse weather conditions. The fusion of these heterogeneous data sources enables more comprehensive scene understanding and improves detection, classification, and tracking performance in complex traffic environments. Multimodal fusion strategies can be broadly categorized into early fusion, middle (feature-level) fusion, and late (decision-level) fusion. Early fusion combines raw sensor data prior to feature extraction, allowing the network to learn joint representations but requiring precise sensor calibration. Feature-level fusion integrates modality-specific features extracted by separate neural networks, providing a balance between representational richness and flexibility. Late fusion combines independent modality-specific predictions, offering robustness and modularity at the cost of reduced cross-modal interaction. Recent studies employ attention mechanisms and transformer-based cross-modal fusion architectures to dynamically weight sensor contributions based on environmental conditions and task relevance (58), (60). Multimodal sensor fusion has demonstrated significant performance gains in vehicle recognition benchmarks, particularly in scenarios involving nighttime surveillance, adverse weather, and complex urban environments. These approaches are increasingly adopted in intelligent transportation systems, autonomous driving platforms, and border surveillance applications, where reliability and safety are critical. However, challenges such as sensor synchronization, increased system complexity, higher deployment costs, and real-time processing constraints remain open research problems. Ongoing research focuses on lightweight fusion architectures, self-supervised cross-modal learning, and adaptive fusion strategies that selectively activate sensors based on contextual cues. These advances aim to enable scalable, cost-effective, and energy-efficient multimodal vehicle recognition systems suitable for real-world deployment.

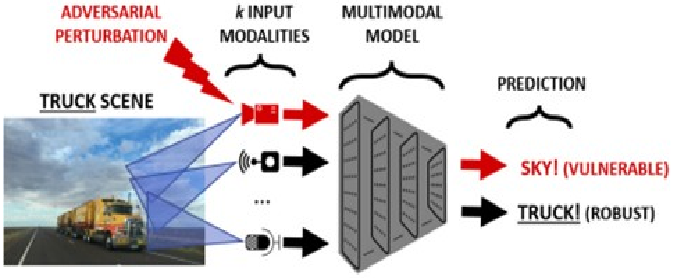

Fig. 6. Multimodal fusion architecture for robust vehicle recognition (67)

VIII. EDGE AI AND REAL-TIME DEPLOYMENT

Edge Artificial Intelligence (Edge AI) has emerged as a transformative approach for real-time vehicle recognition, enabling computation to occur directly on edge devices such as roadside units, traffic cameras, and border checkpoint embedded systems. By processing data locally, Edge AI reduces system latency, minimizes network bandwidth requirements, and preserves sensitive information, addressing both performance and privacy concerns in large-scale intelligent transportation systems (9), (10). To achieve high accuracy within the limited computational resources of edge devices, modern vehicle recognition frameworks employ lightweight deep learning architectures, including MobileNet, EfficientNet, and quantized YOLO variants. Model compression techniques such as pruning, quantization, knowledge distillation, and low-rank factorization reduce model size and inference time while maintaining accuracy, making deployment feasible on embedded GPUs, FPGAs, and AI accelerators (62). Hardware-aware neural architecture search (NAS) further optimizes models for specific edge platforms, balancing detection speed, energy efficiency, and memory footprint. Edge AI deployment also enables continuous and scalable monitoring in distributed transportation networks. For instance, vehicle detection and fine-grained classification pipelines can be executed in real time at multiple traffic intersections or border control stations, providing immediate insights without relying on cloud connectivity. Combined with efficient video pre-processing and adaptive inference scheduling, these systems can handle high vehicle densities while maintaining frame-level responsiveness (63). Recent advancements integrate Edge AI with multimodal sensor fusion and transformer-based recognition models, leveraging the complementary strengths of multiple modalities and global attention mechanisms without overloading hardware constraints. Such hybrid approaches allow real-time vehicle analytics even under challenging conditions, including nighttime, adverse weather, and heavy traffic, while ensuring data privacy and compliance with regulatory standards. Despite these advancements, challenges remain in optimizing transformer-based architectures for edge devices, managing power consumption in large-scale deployments, and maintaining robust performance under variable environmental conditions. Future research is focused on ultra-lightweight attention modules, edge-friendly self-supervised learning, and adaptive inference strategies that dynamically adjust model complexity according to real-time computational budgets.

IX. CHALLENGES, RESEARCH GAPS, AND FUTURE DIRECTIONS

Despite significant advances in AI-driven vehicle recognition, real-world deployment remains constrained by technical, environmental, and operational challenges. These limitations highlight research gaps that must be addressed to achieve robust, scalable, and privacy-compliant vehicle analytics.

Environmental Robustness: Vehicle recognition systems continue to face performance degradation under diverse operational conditions. Low illumination, nighttime traffic, rain, fog, glare, and dynamic shadows significantly affect detection, fine-grained classification, and re-identification accuracy (2), (6), (8). Dense urban traffic with overlapping or occluded vehicles further complicates recognition. While multimodal sensor fusion partially mitigates these issues, practical integration introduces calibration, synchronization, and computational overheads (8), (58), (60).

Model Efficiency and Edge Deployment: High-accuracy models, particularly transformer-based architectures, often require substantial computational and memory resources, limiting their real-time deployment on edge devices (5), (6), (61). Two-stage detectors, though precise, demand high-end GPUs and may be unsuitable for roadside or checkpoint embedded systems (3), (4). Lightweight one-stage detectors trade accuracy for speed but remain vulnerable to complex traffic scenarios. Recent advances in model compression, pruning, quantization, knowledge distillation, and hardware-aware neural architecture search provide promising solutions for edge deployment, but careful balancing of speed, power consumption, and detection performance is still required (9), (10), (62), (63).

Generalization Across Domains: Most models are trained on benchmark datasets with limited geographic and environmental diversity. Domain shifts—including regional variations in vehicle appearance, sensor heterogeneity, and camera perspectives—reduce model generalization (12), (16), (47). Fine-grained classification and Re-ID systems, in particular, struggle with rare vehicle types or models absent from training datasets (42), (50), (61). Cross-domain adaptation, few-shot learning, and meta-learning approaches are critical for ensuring reliable performance in heterogeneous, real-world traffic environments.

Data Availability and Annotation Complexity: High-quality, large-scale datasets are crucial for training robust vehicle recognition systems. Collecting detailed labels for vehicle attributes, cross-camera identities, and multimodal sensor readings is resource-intensive and subject to privacy regulations (43), (44). Synthetic data generation using simulation, generative models, or data augmentation techniques offers a promising avenue to alleviate data scarcity while expanding model exposure to rare scenarios (61), (62).

Explainability and Trustworthiness: Current AI models function largely as "black boxes," limiting their interpretability in safety-critical applications such as border surveillance, autonomous traffic control, and law enforcement (44), (45), (50). The integration of explainable AI (XAI) techniques—including attention maps, feature visualization, and uncertainty quantification—is essential for transparent decision-making, regulatory compliance, and error diagnosis.

Future Research Directions

To overcome these challenges, several research directions are recommended:

- **Edge-Optimized Hybrid Architectures:** Develop transformer–CNN hybrids and ultra-lightweight attention modules to maintain accuracy while ensuring real-time execution on embedded devices (61), (62).
- **Adaptive Multimodal Fusion:** Design dynamic fusion strategies that selectively leverage RGB, infrared, LiDAR, and radar inputs depending on environmental conditions and computational budgets (58), (60).
- **Cross-Domain and Few-Shot Adaptation:** Employ meta-learning, self-supervised, and domain-adaptive frameworks to generalize across geographic regions, vehicle types, and novel camera viewpoints (16), (50).
- **Synthetic and Augmented Data Generation:** Leverage GANs, simulators, and photorealistic augmentation to expand dataset diversity, mitigate rare vehicle classes, and simulate challenging traffic conditions (61), (62).
- **Explainable AI Integration:** Incorporate interpretable architectures and uncertainty modeling to enhance transparency, foster trust, and support decision-making in high-stakes applications (44), (45).

Energy-Efficient Edge AI Deployment: Advance model compression, pruning, quantization, and hardware-aware optimization to deploy high-performing recognition systems on resource-constrained edge devices at scale (9), (10), (62), (63). Addressing these gaps will enable next-generation vehicle recognition systems capable of robust, real-time, and privacy-preserving operation across diverse environments, facilitating safer, more intelligent, and operationally resilient transportation networks.

X. CONCLUSION

This paper provides a comprehensive review and critical analysis of the latest Artificial Intelligence (AI) techniques for Automatic Vehicle Recognition (AVR). Recent developments in deep learning, including CNN-based one-stage and two-stage object detectors, transformer-based architectures, fine-grained vehicle classification, and vehicle re-identification, have substantially enhanced the accuracy, robustness, and adaptability of vehicle recognition systems. Complementary strategies such as multimodal sensor fusion and Edge AI deployment have further improved reliability and enabled real-time performance in challenging environments, including dense traffic, low-light conditions, and adverse weather. Despite these advances, several key challenges remain. Most AVR solutions are implemented as modular, task-specific systems rather than unified, end-to-end frameworks, which limits scalability, integration, and practical deployment across heterogeneous environments. Transformer-based models, while demonstrating strong performance, remain computationally intensive and underutilized in resource-constrained edge settings. Fine-grained vehicle recognition and Re-ID approaches often exhibit limited generalization across geographic regions, rare vehicle types, and dynamic environmental conditions. Additionally, privacy preservation, interpretability of AI decisions, and ethical considerations have not been adequately addressed in the majority of current systems. To bridge these gaps, future research should focus on the development of lightweight, explainable, and unified AVR architectures capable of real-time operation on embedded platforms. Adaptive learning and cross-domain generalization techniques should be integrated to ensure robustness across diverse traffic scenarios and geographic regions. Furthermore, privacy-aware Edge AI solutions and ethical design principles must be incorporated to enable responsible deployment in smart cities, border control infrastructures, and intelligent transportation systems. By addressing these challenges, next-generation AVR systems can achieve scalable, reliable, and ethically compliant performance, effectively translating academic advances into real-world impact.

REFERENCES

1. Du, S. M. Ibrahim, M. Shehata, and W. Badawy, "Automatic License Plate Recognition (ALPR): A State-of-the-Art Review," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 2, pp. 311–325, Feb. 2013.
2. Redmon J. *et al.*, "You Only Look Once: Unified, Real-Time Object Detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.
3. Girshick, R. "Fast R-CNN," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.
4. Ren, S. K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
5. Carion N. *et al.*, "End-to-End Object Detection with Transformers," in *Proc. European Conference on Computer Vision (ECCV)*, 2020, pp. 213–229.
6. Liu Z. *et al.*, "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021, pp. 10012–10022.
7. Tang Y. *et al.*, "Vehicle Re-Identification: A Survey," *IEEE Access*, vol. 7, pp. 142698–142720, 2019.
8. Geiger, A. P. Lenz, and R. Urtasun, "Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite," in *Proc. IEEE CVPR*, 2012, pp. 3354–3361.
9. Wang H. *et al.*, "Edge AI-Based Intelligent Traffic Monitoring System for Smart Cities," *IEEE Internet of Things Journal*, vol. 9, no. 3, pp. 2156–2167, 2022.
10. Tan M. and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proc. International Conference on Machine Learning (ICML)*, 2019, pp. 6105–6114.
11. Liang L. *et al.*, "Vehicle detection algorithms for autonomous driving: A review," *Sensors*, vol. 24, no. 10, p. 3088, 2024.
12. Alif, M. A. "YOLOv11 for vehicle detection: Advancements and applications," *arXiv preprint arXiv:2410.22898*, 2024.
13. Li, Y.J. Wang, J. Huang, and Y. Li, "Deep learning-based automatic vehicle recognition using RES-YOLO," *Sensors*, vol. 22, no. 10, p. 3783, 2022.
14. Zaidi S. S. A.*et al.*, "A survey of modern deep learning-based object detection models," *arXiv preprint arXiv:2104.11892*, 2021.
15. Gupta, A. R. R. Nair, and S. Chandra, "Deep learning for object detection and perception in autonomous vehicles: A survey," *Array*, vol. 10, p. 100066, 2021.
16. Wang, H. J. Hou, and N. Chen, "A survey of vehicle re-identification based on deep learning," *IEEE Access*, vol. 7, pp. 170379–170398, 2019.
17. Amiri, A. A. Kaya, and A. S. Keceli, "A comprehensive survey on deep-learning-based vehicle re-identification," *arXiv preprint arXiv:2401.10643*, 2024.
18. Almeida, E. B. Silva, and J. Batista, "Multi-branch deep learning for vehicle re-identification," *Pattern Recognition*, vol. 110, p. 107609, 2021.
19. Shen Y. *et al.*, "Learning deep neural networks for vehicle re-identification with spatio-temporal paths," *IEEE Trans. Intelligent Transportation Systems*, vol. 20, no. 3, pp. 1033–1045, Mar. 2019.
20. Liu Z. *et al.*, "SSD: Single shot multibox detector," in *Proc. ECCV*, Amsterdam, Netherlands, 2016, pp. 21–37.
21. Lin T. Y. *et al.*, "Focal loss for dense object detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, Feb. 2020.
22. Tan, M. R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE CVPR*, Seattle, WA, USA, 2020, pp. 10781–10790.
23. Zhou, X. D. Wang, and P. Krähenbühl, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.
24. Geiger, A. P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE CVPR*, Providence, RI, USA, 2012, pp. 3354–3361.
25. Caesar H. *et al.*, "nuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE CVPR*, Seattle, WA, USA, 2020, pp. 11621–11631.
26. Sun P. *et al.*, "Scalability in perception for autonomous driving: Waymo Open Dataset," in *Proc. IEEE CVPR*, Seattle, WA, USA, 2020, pp. 2446–2454.
27. Huang X. *et al.*, "ApolloScape: A large-scale dataset for autonomous driving," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2643–2658, Oct. 2020.
28. Cordts M. *et al.*, "The Cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE CVPR*, Las Vegas, NV, USA, 2016, pp. 3213–3223.
29. Liu W. *et al.*, "SSD: Single shot multibox detector," in *Proc. ECCV*, 2016.
30. Long, J. E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, Apr. 2017.
31. Ma Y. *et al.*, "Artificial intelligence applications in autonomous vehicles: A survey," *IEEE/CAA Journal of Automatica Sinica*, vol. 7, no. 2, pp. 315–329, 2020.
32. Krizhevsky, A. I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.
33. Szegedy C. *et al.*, "Going deeper with convolutions," in *Proc. IEEE CVPR*, Boston, MA, USA, 2015, pp. 1–9.
34. He K. *et al.*, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, Las Vegas, NV, USA, 2016, pp. 770–778.
35. Dosovitskiy A. *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021.
36. Deng J. *et al.*, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE CVPR*, Miami, FL, USA, 2009, pp. 248–255.
37. Han S. *et al.*, "Deep compression: Compressing deep neural networks," in *Proc. ICLR*, 2016.
38. Sandler M. *et al.*, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE CVPR*, Salt Lake City, UT, USA, 2018, pp. 4510–4520.
39. Howard A. *et al.*, "Searching for MobileNetV3," in *Proc. IEEE ICCV*, Seoul, South Korea, 2019, pp. 1314–1324.
40. Le Q. V. *et al.*, "EfficientNet: Rethinking model scaling for CNNs," in *Proc. ICML*, 2019.
41. LeCun, Y. Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.
42. Chen Z. *et al.*, "Edge AI: On-device machine learning," *IEEE Internet of Things Journal*, vol. 6, no. 1, pp. 26–42, Feb. 2019.
43. Li S. *et al.*, "Privacy-preserving deep learning," *IEEE Security & Privacy*, vol. 18, no. 4, pp. 48–56, Jul. 2020.
44. Doshi-Velez F. and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.
45. Gunning D. *et al.*, "XAI—Explainable artificial intelligence," *Defense Advanced Research Projects Agency*, 2019.
46. Szeliski, R. *Computer Vision: Algorithms and Applications*, 2nd ed. Springer, 2022.

47. Bishop, C. *Pattern Recognition and Machine Learning*. Springer, 2006.
48. Goodfellow, I. Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
49. Mitchell, T. *Machine Learning*. McGraw-Hill, 1997.
50. IEEE Standards Association, "IEEE standard for explainable AI," IEEE Std 7001-2021, 2021.
51. Wang, X. R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 7794–7803.
52. Zhu, Z. D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, "Traffic-sign detection and classification in the wild," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 2110–2118.
53. Yang, L. P. Luo, C. C. Loy, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015, pp. 3973–3981.
54. Real, E. A. Aggarwal, Y. Huang, and Q. V. Le, "Regularized evolution for image classifier architecture search," in *Proc. AAAI Conf. Artificial Intelligence*, Honolulu, HI, USA, 2019, pp. 4780–4789.
55. Tian, Y. P. Luo, X. Wang, and X. Tang, "Deep learning strong parts for pedestrian detection," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 1904–1912.
56. Zagoruyko S. and N. Komodakis, "Wide residual networks," in *Proc. British Machine Vision Conf. (BMVC)*, York, UK, 2016, pp. 87.1–87.12.
57. Hu, J. L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 7132–7141.
58. Liu, Y. M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
59. Redmon J. and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.
60. Bochkovskiy, A. C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.
61. Li, Y. K. Chen, Y. Wang, and W. Zhang, "Fine-grained vehicle recognition using hybrid CNN–Transformer networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 3, pp. 3124–3136, Mar. 2023.
62. Wu, H. X. Zhang, and Y. Liu, "Efficient model compression for edge-based vehicle detection systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 7, pp. 8762–8774, Jul. 2023.
63. Chen, J. L. Zhang, and M. Tan, "Edge computing for real-time traffic analytics: Deployment strategies and performance optimization," *IEEE Internet of Things Journal*, vol. 10, no. 12, pp. 10532–10544, Dec. 2023.
64. Kattarajesh, R. "Object Detection Part 2 — Two-Stage Detectors: R-CNN, Fast R-CNN, Faster R-CNN," *Medium*, Jun. 18, 2020.
65. Hermosillo-Reynoso F. *et al.*, "A transformer-based multi-task learning model for vehicle traffic surveillance," *Mathematics*, vol. 13, no. 23, art. no. 3832, Nov. 2025, doi: 10.3390/math13233832
66. Khoba, P. K.et al., "A fine-grained vehicle detection (FGVD) dataset for unconstrained roads," in *Proc. 13th Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP '22)*, Gandhinagar, India, Dec. 2022, Article 83, 9 pp., doi: 10.1145/3571600.3571626.
67. Yang, K. W.-Y. Lin, et al., "Defending multimodal fusion models against single-source adversaries," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 3340–3349

****