

Article

Fine-Grained Vehicle Make and Model Recognition for Smart City Environmental Monitoring: A YOLO11-Based Two-Stage Framework

Aya Elouali * and Antonio J. Jara 

Libelium, 50018 Zaragoza, Spain; aj.jara@libelium.com

* Correspondence: a.elouali@libelium.com

Abstract

Accurate, real-time vehicle identification is essential for data-driven urban planning, enabling applications like emissions monitoring and the enforcement of environmental regulations. However, identifying a vehicle's model, generation, and production year remains a significant challenge for VMMR systems. This is especially true when cameras capture multiple vehicles simultaneously under suboptimal imaging conditions. This challenge is amplified in Europe, as most existing VMMR datasets are designed for non-European markets. To address this, we present two contributions: a newly curated dataset of 84,732 images across 625 classes, and a robust two-stage YOLO11 system trained on this data. The dataset focuses on the European market and realistic viewpoints like front and rear angles. The system, comprising a Vehicle Localization Module (VLM) and a Fine-Grained Classification Module (FGCM), performs detailed model classification without relying on license plates or additional sensors. When tested on real European traffic footage, our system achieved 80% accuracy and outperformed models trained on U.S.-centric datasets.

Keywords: Vehicle Make and Model Recognition (VMMR); computer vision; YOLO; sustainable transportation; smart cities; emissions monitoring

1. Introduction

The ability to automatically detect and identify vehicles from images and video has become increasingly important in modern cities. Among the technologies in this domain, Vehicle Make and Model Recognition (VMMR) has garnered significant attention, with key applications in traffic monitoring, urban planning, environmental policy enforcement, and intelligent transportation systems. While contemporary systems are highly capable of recognizing vehicle makes, many still struggle to accurately identify specific models, differentiate between generations, or estimate a vehicle's production period. This is a critical limitation, as achieving this level of fine-grained recognition is essential for advanced applications like emissions monitoring, law enforcement investigations, and regulatory compliance. For instance, vehicles from different production years often have vastly different performance and emission standards.

A key obstacle to fine-grained classification is the limited availability of richly annotated datasets. Although large-scale datasets like Stanford Cars and CompCars exist, they often lack the detail needed to distinguish between different versions of the same model. Moreover, many of these datasets focus on specific markets (most notably North America and China), which introduces a geographical bias. Consequently, when these models are



Academic Editors: Nenad Ruškić and
Valentina B. Mirović

Received: 17 December 2025

Revised: 20 January 2026

Accepted: 23 January 2026

Published: 1 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

applied to European roads, they tend to underperform, particularly with vehicles from makes like SEAT, Renault, or Peugeot, which are common in Europe but underrepresented in existing data.

Another important challenge arises from the mismatch between real-world imaging conditions and those represented in most datasets. In practical applications, vehicles are typically recorded from the front or rear by surveillance or traffic management cameras, as these positions are ideal for reading license plates or monitoring traffic flow. However, most available datasets favor side views, which, while more visually informative, are less representative of typical deployment environments. This discrepancy can lead to models that perform well under testing conditions but fail to generalize effectively in deployment.

Some systems address these challenges by incorporating additional technologies like license plate recognition, LiDAR, or magnetic sensors. While these can enhance accuracy, they introduce significant trade-offs. License plate recognition raises privacy and legal concerns, particularly under strict data protection laws. Meanwhile, LiDAR and magnetic sensors require significant infrastructure investments and may be unsuitable for large-scale urban deployment due to their cost and complexity.

To address these gaps, we propose a scalable, privacy-respecting system for fine-grained vehicle recognition in Europe that operates without license plate data. This study makes three key contributions:

- **EuroVMMR Dataset:** A novel dataset of 84,732 images across 625 classes targeting the European market, providing generation-level annotations essential for emission enforcement.
- **YOLO11-Based Pipeline:** A streamlined two-stage pipeline using the standard YOLO11 detection and classification models, fine-tuned on our custom dataset to provide a solution suitable for real-time deployment.
- **Real-World Validation:** Validation on real European traffic footage achieving 80% accuracy, confirming that domain-specific training data outweighs model complexity for effective deployment.

2. State of the Art

Vehicle Make and Model Recognition (VMMR) has advanced significantly with the advent of deep learning, enabling more detailed classification across diverse vehicle types and real-world conditions. This section reviews the state of the art in three main areas: (1) Advanced Architectures and Mechanisms, (2) Dataset Contributions, and (3) Real-Time Detection and Open-World Scenarios.

2.1. Advanced Architectures and Mechanisms for VMMR

Recent advances in deep learning have led to architectures designed to distinguish between visually similar vehicle models by focusing on fine-grained visual differences. Semiromizadeh et al. introduced a 3D attention module integrated into convolutional neural networks (CNNs) that enhances feature extraction by focusing on critical vehicle details, which they tested on the Stanford Cars dataset [1]. Yang et al. used attention mechanisms targeted at specific parts like wheels and headlights, improving the discrimination of closely related models [2]. Wang et al. combined CNNs with temporal convolutional networks (TCNs) to capture spatial and temporal information, enabling analysis of vehicle behavior in traffic and thereby expanding the scope of VMMR [3]. To address challenges such as image noise and intra-class variation, Liu proposed the Progressive Multi-task Anti-Noise Learning (PMAL) framework, improving robustness on datasets including Stanford Cars, CompCars, and BIT-Vehicle [4].

Earlier work by Fang et al. developed coarse-to-fine convolutional neural network architectures that progressively refined recognition performance by focusing on fine-grained features [5]. Sochor et al. introduced BoxCars, which leverages 3D bounding box representations to encode spatial and geometric vehicle information, enhancing fine-grained classification [6]. Demonstrating the importance of perspective, Llorca et al. combined rear emblem features with appearance-based descriptors, a relevant technique for models analyzing the front or rear views commonly captured by traffic cameras [7].

Part-based recognition approaches have also proven effective. Biglari applied latent Support Vector Machines (SVM) and Histogram of Oriented Gradients (HOG) features to vehicle parts like headlights and grilles to improve classification accuracy [8]. Bularz et al. developed CNN architectures focused on rear-lamp patterns, demonstrating robustness in challenging lighting and occlusion conditions [9].

Beyond traditional CNNs, transformer-based architectures have recently set new benchmarks in computer vision and fine-grained recognition. Models like the Vision Transformer (ViT) [10] and its variants, such as the Swin Transformer [11], have demonstrated exceptional performance by capturing global relationships within an image through self-attention mechanisms. In the context of vehicle recognition, these models have been used to overcome the limitations of CNNs in modeling long-range dependencies, which is crucial for distinguishing between models with subtle but globally distributed feature differences [12]. For instance, recent studies show that transformer backbones can improve accuracy in challenging fine-grained tasks by focusing on a holistic representation of the vehicle rather than just localized features like headlights or grilles [13]. While computationally more intensive, their state-of-the-art performance makes them an important benchmark for future VMMR systems.

2.2. Dataset Contributions

The Stanford Cars dataset [14] provides over 16,000 images spanning 196 car classes, annotated by make, model, and production year across multiple viewpoints. The CompCars dataset [15] includes more than 136,000 images across 1716 car models, with detailed annotations of viewpoints, parts, and attributes to facilitate fine-grained recognition. The VehicleID dataset [16] comprises over 200,000 images collected from surveillance cameras, supporting vehicle re-identification research under realistic conditions. The Car-1000 dataset [17] features 1000 vehicle classes with frontal and rear images, emphasizing practical scenarios aligned with surveillance camera angles.

More recent datasets aim for broader geographic and contextual diversity. The Diverse Large-scale VMM (DVMM) dataset [18] includes 23 vehicle makes and 326 models focusing on European vehicles to enhance cross-region applicability. The Global License Plate Dataset [19] contains over five million images from 74 countries, richly annotated with license plate, make, color, and model information, enabling studies that integrate license plate recognition and vehicle classification. Semi-automatic annotation methods developed by Zwemer et al. [20] have improved dataset scalability and labeling consistency, facilitating the creation of large datasets necessary for effective model training.

2.3. Real-Time Detection and Open-World Scenarios

Real-time VMMR systems are crucial for traffic management and urban mobility applications. Manzoor et al. demonstrated a system capable of live vehicle classification with efficient processing suitable for deployment in traffic monitoring [21]. Maurya et al. further developed a real-time classification system that distinguishes between multiple vehicle classes and was tested under diverse traffic conditions [22].

A critical consideration for deploying VMMR in real-world traffic systems is the trade-off between model accuracy and computational efficiency. While large models achieve high performance, their resource requirements often make them unsuitable for deployment on edge devices like traffic cameras or roadside units. This has driven research into lightweight architectures designed for real-time processing. Models such as MobileNetV3 [23] and EfficientNetV2 [24] have been specifically engineered to minimize parameters and floating-point operations (FLOPs) while preserving competitive accuracy. In the vehicle recognition domain, lightweight variants of the YOLO family have become particularly popular, offering a robust balance between detection speed and precision, making them ideal for on-the-fly traffic analysis without requiring high-end GPU hardware [25]. Our work builds on this trend by leveraging a recent YOLO variant, aiming to provide a solution that is both accurate for fine-grained tasks and practical for large-scale, sustainable urban deployment.

Handling open-world scenarios, where new or unseen vehicle models appear during deployment, is an emerging challenge. Muñoz et al. proposed Veri-Car, an integrated system combining YOLO5-based license plate detector, a fine-tuned TrOCR model for plate recognition, and multi-similarity loss functions to adapt to novel classes dynamically [26]. Zhang et al. explored non-visual modalities for VMMR, employing magnetic fingerprint recognition combined with adversarial autoencoders and AdaBoostSVM. This offers an alternative approach when visual data is insufficient or unavailable [27].

2.4. Motivation

Despite meaningful progress in Vehicle Make and Model Recognition (VMMR), many existing systems still face significant limitations when applied in real-world conditions, particularly in European settings. Our work is motivated by addressing several key challenges observed in prior research:

- **Geographical Bias in Datasets:** Most widely used datasets, such as Stanford Cars, are largely focused on vehicles from the U.S. market. As a result, models trained on these datasets often struggle when applied to European roads, where the variety of makes, models, and design details can differ significantly.
- **Lack of Fine-Grained Classification:** Many VMMR systems are limited to identifying a vehicle's make but do not distinguish between specific models, their generations, or approximate production years. This level of detail is essential for several applications, including emissions estimation, regulatory compliance, vehicle taxation, fleet monitoring, market analysis, and fraud detection.
- **Limited Viewpoint Diversity:** Existing datasets and methods often rely on side-view images, which offer more distinct visual cues. However, in practical applications, vehicles are usually captured from the front or rear by traffic or surveillance cameras. These angles present fewer distinguishing features, making it more difficult to recognize the exact model and production range.
- **Privacy Concerns Related to License Plate Use:** Some approaches use license plate recognition or associated metadata to improve classification. While this can increase accuracy, it also raises significant privacy and legal concerns, especially in regions with strict data protection laws.
- **Dependence on Specialized Hardware:** Certain methods rely on additional sensors, such as LiDAR or magnetic detectors, which can improve recognition but also increase both the cost and complexity of the system. This makes them less scalable for widespread public deployment.

3. Two-Stage VMMR Framework

Accurate vehicle make and model classification requires high-resolution images that capture the detailed features of individual cars. In real-world traffic scenes, vehicles appear at different distances and sizes, making it challenging to classify them directly from a full image. To address this, our system separates detection and classification into two distinct yet connected steps. First, the Vehicle Localization Module (VLM) locates and crops each vehicle to produce close-up images. Then, the Fine-Grained Classification Module (FGCM) uses these cropped images to identify the vehicle's make, model, and generation. This two-stage pipeline provides the FGCM with clearer, more detailed inputs, enabling it to distinguish between similar models and different versions, even in multi-vehicle scenes. Figure 1 presents the architecture of the framework.

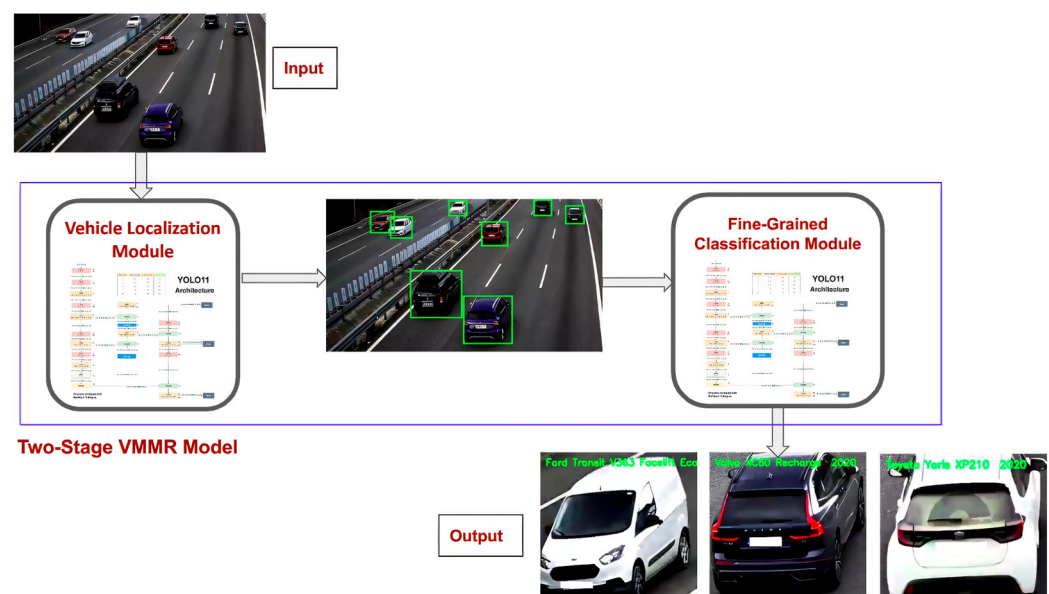


Figure 1. Overview of the two-stage VMMR framework. The process begins with the VLM stage, where the YOLO11m detector generates green spatial bounding boxes to isolate individual vehicles from the input scene. These localized crops are then passed to the FGCM for fine-grained identification of make, model, and generation.

3.1. Vehicle Localization Module (VLM)

The first stage of our pipeline is designed to locate vehicles within the input images. To accomplish this, we use YOLO (You Only Look Once), a widely adopted object detection model known for its balance of speed and accuracy. Its performance makes it suitable for applications that require real-time or near-real-time processing, such as traffic monitoring. YOLO processes the entire image in a single pass, simultaneously predicting both bounding boxes and object classes. Specifically, we use the YOLO11m (Medium) variant, configured with an input resolution of 640×640 pixels. The model's improved architecture results in faster and more precise detections compared to previous versions. When compared to YOLOV8, YOLO11 achieves higher accuracy on standard benchmarks like the COCO dataset while using 22% fewer parameters (approx. 20.1 million). This reduction in model complexity helps improve computational efficiency, a critical factor for deployment on devices with limited resources.

The YOLO11 model we use is pretrained on the COCO dataset, which contains a variety of common street-level objects, including several vehicle classes such as cars, buses, trucks, and motorcycles. Since our focus is on traffic analysis, we limit the model's detection to only these vehicle categories. This restriction simplifies the task for the detector by filtering out irrelevant objects. The output from the VLM consists of bounding boxes

around all identified vehicles in each frame. These bounding boxes serve as inputs for the next stage of the pipeline, where the vehicles are classified by their make, model, and generation.

3.2. Fine-Grained Classification Module (FGCM)

Once vehicles are detected in a frame, the bounding boxes for each detected vehicle are passed to the Fine-Grained Classification Module. This module utilizes the YOLO11m-cls architecture, designed for single-label image classification tasks and accepting standard 224×224 pixel inputs. We selected this architecture specifically to optimize the trade-off between recognition granularity and operational speed. With approximately 20.1 million parameters, the model maintains a compact memory footprint and achieves high-throughput inference, significantly outperforming heavier state-of-the-art alternatives that often struggle to meet the low-latency requirements of live video processing. While the VLM identifies where objects are located, this classification model assigns a specific label to each vehicle, determining its make, model, and generation. To adapt the classifier, we use transfer learning through fine-tuning. We begin with a YOLO11 model pre-trained on ImageNet and fine-tune it on our specialized European vehicle dataset. This process allows the model to adjust its learned features to the specific visual details required to distinguish between similar vehicle generations.

3.3. Inference Protocol

To ensure the reproducibility of the framework, the inference pipeline follows a strict operational sequence defined in the system's implementation. First, input video frames are processed by the Vehicle Localization Module (VLM) utilizing the YOLO11m architecture. To minimize false positives and resolve overlapping detections in multi-vehicle scenes, the detector applies a confidence threshold of 0.5 and a standard Non-Maximum Suppression (NMS) threshold of 0.7 (Intersection over Union). Valid objects are extracted directly from the bounding box coordinates without additional padding to strictly localize object features. These crops are passed to the Fine-Grained Classification Module (FGCM), which applies a confidence threshold of 0.5 to assign the make, model, and generation based on the highest Top-1 probability score. Finally, these classification labels are mapped back to the original frame coordinates using the stored bounding box IDs, enabling the visualization of the specific vehicle model and generation directly on the output video stream.

4. EuroVMMR Dataset

Recognizing that high-quality and diverse data is essential for training accurate classification models, we built a custom dataset specifically designed for fine-grained vehicle recognition. This dataset includes a wide range of vehicle classes, where each class is defined by a unique combination of make, model, generation, and year, as illustrated in Figure 2. For instance, it includes detailed examples such as the Volvo 360c Concept (2018), the Škoda Enyaq iV First Generation (2020), and the Chevrolet Trax J600 (2023). By including different generations and variants of the same vehicle model, the dataset helps the model learn to distinguish the subtle differences in design and features that change over time.

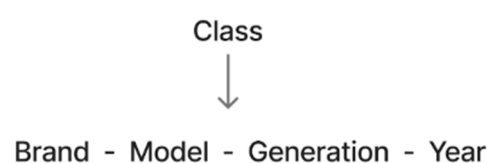


Figure 2. Class naming convention in the dataset, including make, model, generation, and production year.

To ensure the dataset is representative of real-world traffic conditions, we focused on collecting images of vehicles commonly seen on European roads. As a result, the dataset includes a strong presence of European makes such as Mercedes-Benz, Peugeot, Renault, Volkswagen, BMW, Opel, and Škoda; key Asian manufacturers like Kia, Nissan, and Hyundai; and several prominent American makes like Ford, Tesla, and Jeep. The dataset also includes commercial logistics vehicles (e.g., DAF, Scania) and modern electric platforms. This balanced mix ensures the model can handle the variety of vehicles typically found in European cities and highways.

The dataset contains 84,732 total images across 625 different vehicle classes, curated from publicly accessible web sources. The images were divided using a stratified 80/20 train/validation split and arranged in class-specific folders following YOLO classification standards. Vehicle images were collected from multiple angles—front, rear, side, and partial views—to reflect the real-world conditions in which ideal camera angles are not always possible. This variety helps improve the model's predictive performance, even when vehicles are partially blocked or captured from unusual viewpoints. For privacy, any visible license plates in the images were blurred as a pre-processing step, and label quality was ensured via a multi-pass verification process.

Figure 3 shows an example of distribution of image in vehicle classes within the dataset, while Figure 4 provides example images, highlighting the range of models and viewpoints represented.

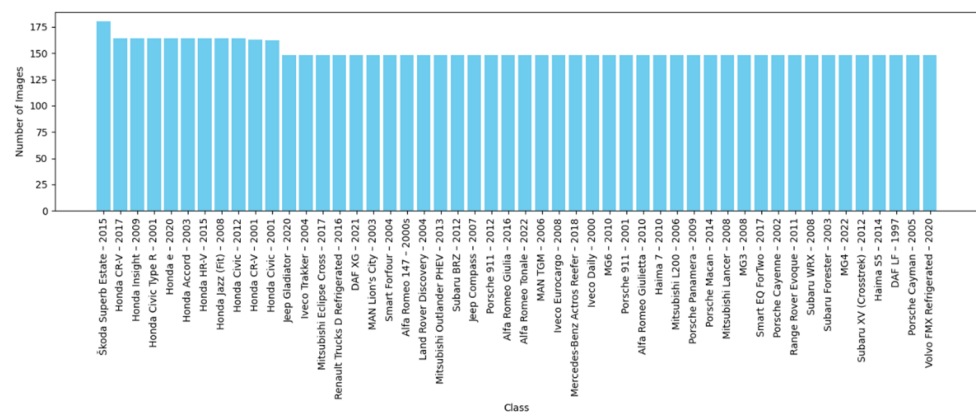


Figure 3. Distribution of images across 625 vehicle classes in the EuroVMR dataset.

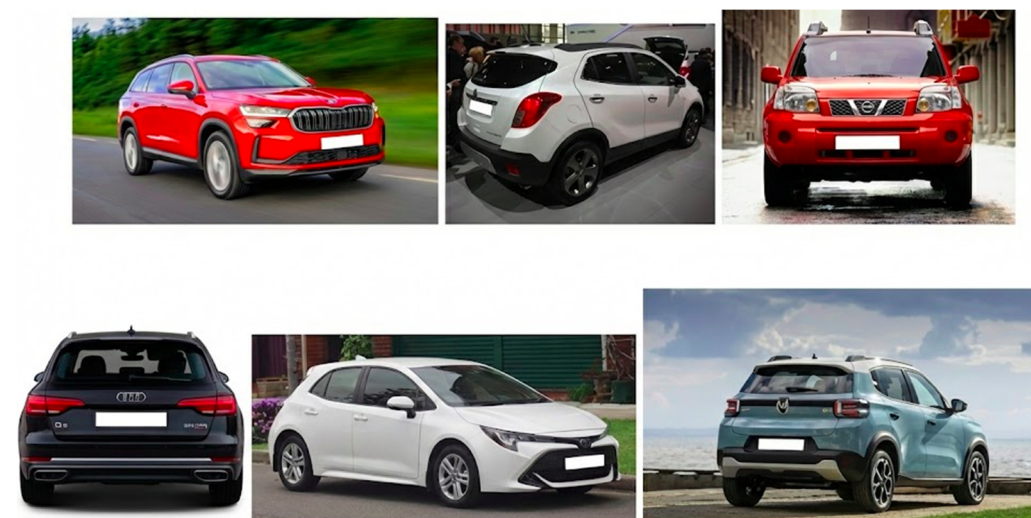


Figure 4. Sample images from the EuroVMR dataset, illustrating multiple viewpoints (front, rear, side).

Table 1 presents a comparison between our dataset and several established benchmarks in vehicle make and model recognition (VMMR). While datasets such as Stanford Cars, CompCars, and Car-1000 have advanced the field, they often face limitations when applied to real-world European scenarios. Most focus on U.S. or Chinese markets and often lack fine-grained annotations like vehicle generation or production details. Our dataset addresses these gaps by including detailed class labels (make, model, generation, and year) and a diverse range of vehicles commonly seen across Europe. It also offers broader viewpoint coverage (front, rear, and side), which better reflects the realistic deployment conditions where side views are not always available.

Table 1. Characteristics of Existing VMMR Datasets Compared to Our Dataset.

Dataset	Num of Classes	Num of Images	Geographic Focus	Fine-Grained Labels	Includes Generation Info?	Viewpoint Diversity
Our Dataset	625	84,732	Europe	Make, model, generation, production year	Yes	Front, rear and side views
Stanford Cars [14]	196	16,185	USA	Make, model, production year	Model year, no clear generation	Mostly side views
CompCars [15]	1716	136,726	China	Make, model	No	Front, rear and side views
Car-1000 [17]	1000	140,312	China	Make, model	No	Mostly side views

5. Model Training and Data Augmentation

The YOLO-based classification model was trained on our custom fine-grained vehicle dataset via transfer learning, using ImageNet-pretrained weights. Training proceeded for up to 200 epochs with an early stopping mechanism triggered after 10 consecutive epochs without an improvement in validation loss, helping to prevent overfitting. The training used a batch size of 6 and the auto optimizer (which selects an adaptive optimizer such as AdamW) with a weight decay of 0.0005. We employed a linear learning rate schedule with a 3-epoch warmup. The initial learning rate (lr0) was set to 0.0001, which decayed to a final value of 0.000001 (based on lrf: 0.01). The entire training process was performed on GPU hardware to efficiently handle the computational demands.

To improve robustness and generalization, we employed a comprehensive data augmentation pipeline during training. For the initial 10 epochs, Mosaic augmentation was enabled to expose the model to diverse spatial arrangements by merging four images into a single composite. After this stage, Mosaic was disabled to allow the training to focus on individual images. Additional augmentations included RandAugment, which randomly applies a variety of transformations to further increase data diversity. We also incorporated horizontal flipping with a 50% probability, HSV adjustments (hue 0.015, saturation 0.7, value 0.4), scaling (up to $\pm 50\%$), translation (up to $\pm 10\%$), random erasing of up to 40% of an image region to encourage the model to learn from partial information, and a flipped Copy-Paste strategy to introduce mirrored object segments. These augmentations collectively helped the model handle real-world variability in lighting, occlusion, scale, and positioning. An overview of the augmentation methods is shown in Figure 5.

Training Results

The fine-tuning of the YOLO11 classification model on our custom vehicle dataset demonstrated clear improvements throughout the training process. Over approximately 150 epochs, the model achieved consistent reductions in loss and significant gains in

classification accuracy (Figures 6 and 7). This steady progress reflects the model's ability to adapt to the fine-grained distinctions required for vehicle recognition, even when faced with a large number of visually similar classes. These results show the model's capacity to learn subtle differences across 625 vehicle classes.



Figure 5. Examples of data augmentation techniques applied during training (Mosaic, horizontal flip, etc.).

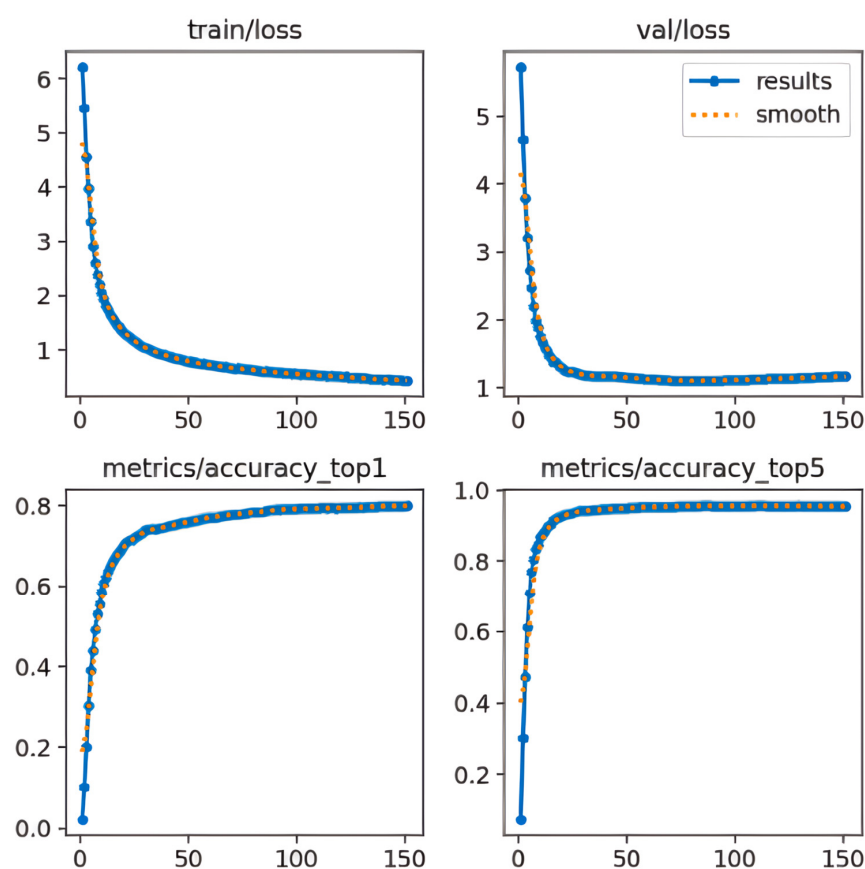


Figure 6. Classification model training performance. The plots show a steady decrease in training and validation loss, while Top-1 accuracy reached ~80% and Top-5 accuracy ~99%, indicating effective learning.

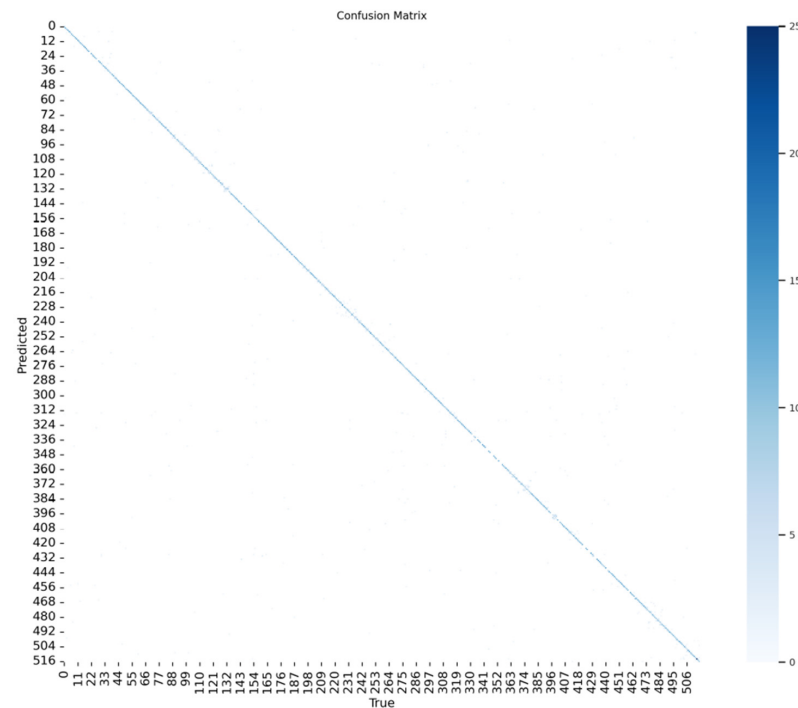


Figure 7. Confusion Matrix for Model Classification Performance.

- **Loss Metrics:**

The training loss began at a high initial value of over 6 and gradually decreased as the model optimized its weights, eventually reaching less than 0.7 by the end of training. The validation loss exhibited a similar downward trend, starting at around 5.5 and stabilizing slightly above 1.0. The close alignment between the training and validation loss curves indicates strong generalization to unseen data, while the parallel decline suggests stable learning and the effective use of data augmentation. This is particularly relevant for fine-grained classification, where subtle visual differences (such as headlight or grille design) distinguish vehicle classes.

- **Accuracy Performance:**

The model's classification accuracy improved substantially during training. Top-1 accuracy, which measures the model's ability to correctly predict the exact class with its first choice, climbed from around 5% at the beginning to approximately 80% by the final epochs. This sharp increase highlights the model's developing capability to make precise distinctions between different vehicle types as training progressed.

Top-5 accuracy reached nearly 99%, indicating that the correct vehicle class was almost always included within the model's top five predictions. This metric is particularly important in real-world applications, where a high likelihood of a correct identification among the top few predictions can still provide valuable insights for tasks like emissions estimation, fleet analysis, and traffic monitoring.

The reported metrics reflect the performance of the best model checkpoint obtained from a single optimized training run, selected based on the lowest validation loss.

The confusion matrix (Figure 7) demonstrates robust overall performance, with a prominent diagonal indicating that the model correctly identifies the vast majority of the 625 vehicle classes. A closer examination of these results reveals specific challenges inherent to fine-grained recognition. The primary source of error stems from inter-generational similarity, where misclassifications frequently occur between successive versions of the same model (e.g., distinguishing a Volkswagen Golf VI from a Golf VII). These errors

are attributed to subtle ‘facelifts’ that involve only minor changes to lighting signatures or bumpers, which are difficult to resolve. A second cluster of errors arises from badge engineering, where the model struggles to differentiate vehicles sharing identical platforms and body structures, such as the Peugeot Partner and Citroën Berlingo. In these instances, the visual distinctions are often limited to badges or slight grille variations, making these specific pairs exceptionally difficult for the model to differentiate reliably.

6. Validation

To assess the real-world performance of our vehicle recognition system, we validated it using traffic camera footage captured from European roads. This evaluation tested the entire pipeline, combining both the detection and classification stages.

6.1. Detection Performance

In terms of detection, our model successfully identified and located all passing vehicles across the test footage. Using the YOLO11 architecture, it was able to efficiently track cars, buses, trucks, and motorcycles even in busy traffic scenarios, ensuring that no vehicles were missed during the analysis.

6.2. Classification Performance

For classification, the model achieved an accuracy of 80% in correctly identifying the make, model, and generation of vehicles. This level of accuracy is consistent with the demands of fine-grained recognition tasks, which require distinguishing between highly similar models and even different generations of the same vehicle. While factors like the distance of vehicles from the camera and subtle design differences introduced challenges, the model still maintained a strong performance. Vehicles captured from rear angles (common in traffic camera placements) were still classified correctly in most cases, indicating that the learned features generalize well to real-world traffic footage (Figures 8 and 9).



Figure 8. Real-world classification examples. The model accurately identifies vehicles from challenging rear-view angles common in traffic monitoring.

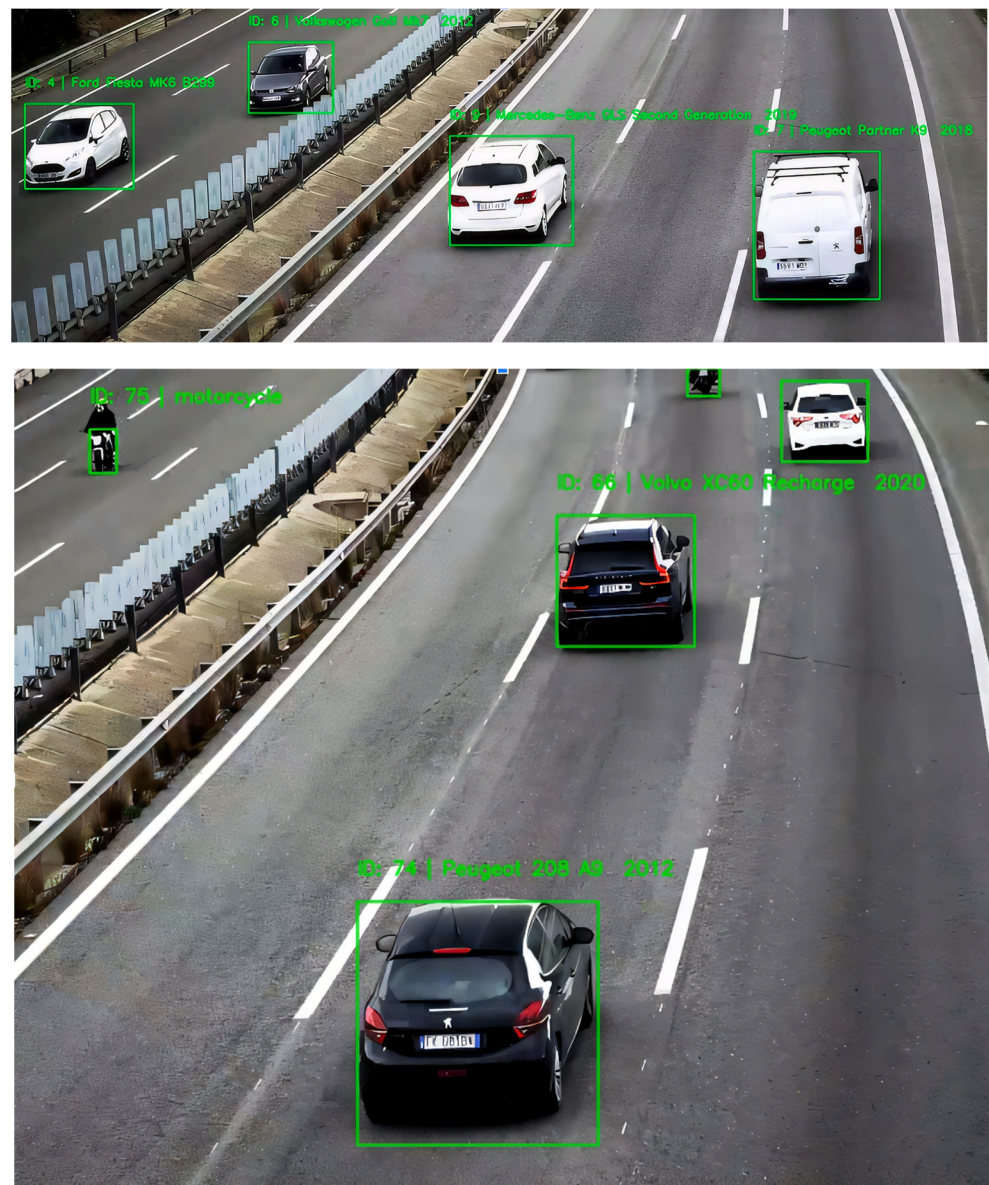


Figure 9. The complete two-stage system in action, applying successful vehicle detection (boxes) and fine-grained classification (labels) to a real traffic scene.

6.3. Comparison with the Stanford Cars Model

To further evaluate our model's effectiveness, we compared its performance with a YOLO11 classification model trained on the Stanford Cars dataset. When tested on the same European traffic footage (Table 2), the Stanford Cars model achieved significantly lower accuracy. This performance gap can be attributed to two main factors:

- **Limited Make Coverage:** Many common European makes, such as Škoda, SEAT, Peugeot, and Opel, are not well represented in the Stanford Cars dataset, which is heavily U.S.-centric. This lack of coverage led to frequent misclassifications for European models.
- **Viewpoint Limitations:** The Stanford Cars dataset primarily contains ideal side and front views, while our traffic footage often captures vehicles from the rear (especially when mounted on roadside poles or traffic lights). Our model, trained on more diverse perspectives, handled these rear-view images with much greater accuracy.

Table 2. Comparison between model trained on our dataset and Stanford dataset.

Capture Angle	Stanford	Our Dataset
Front View	54.90%	79%
Rear View	23.33%	80%

7. Conclusions

This work addresses a critical gap in existing Vehicle Make and Model Recognition (VMMR) systems: the lack of fine-grained, geographically relevant datasets that reflect the diversity of vehicles and real-world imaging conditions on European roads. To overcome these limitations, we developed a comprehensive dataset curated specifically for fine-grained recognition tasks in European contexts. The dataset includes over 625 distinct vehicle classes, with annotations for make, model, generation, and approximate production year. It incorporates varied viewpoints (front, rear, and side) and captures a wide range of lighting conditions and real-world occlusions. Complementing this dataset, we proposed a specialized two-stage recognition system. By training this system on our new dataset, we validated our approach with an 80% accuracy rate in real-world testing, significantly outperforming U.S.-centric baselines tested on European roads.

Nevertheless, opportunities for refinement remain. While the dataset introduces substantial improvements in class diversity and viewpoint representation, certain vehicle classes (particularly trucks and buses) are still slightly underrepresented, which may limit the model's ability to generalize across all traffic participants. Moreover, since the current dataset primarily consists of clear-weather footage, the system's robustness in adverse scenarios requires further validation. In real-world deployments, environmental factors such as heavy rain, fog, or low-light nighttime conditions can significantly reduce image contrast and obscure fine-grained details, leading to lower confidence scores and increased misclassification rates. Incorporating additional samples from these classes, as well as from adverse weather conditions, nighttime scenes, and less common vehicle types, could further enhance the system's robustness and applicability.

Author Contributions: Conceptualization, A.E.; Methodology, A.E.; Validation, A.E.; Investigation, A.E.; Resources, A.E.; Data curation, A.E.; Writing—original draft, A.E.; Writing—review & editing, A.J.J.; Supervision, A.J.J.; Project administration, A.J.J. All authors have read and agreed to the published version of the manuscript.

Funding: The results of this research have been funded by the Government of Catalonia and the ERDF Program of Catalonia 2021–2027 (contract EC-2024-16).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to internal privacy constraints.

Acknowledgments: This work is supported by the European Union's research and innovation programs through the SENSE project (grant agreement no. 101167948) and the Citcom.ai project (grant agreement no. 101100720). The algorithm developed in this study will be available as part of the EU LDT Toolbox under the project EDIC BUILD (grant agreement no. 101226211).

Conflicts of Interest: Authors A.E and A.J.J. were employed by the company Libelium.

References

1. Semiromizadeh, N.; Manzari, O.N.; Shokouhi, S.B.; Mirzakuchaki, S. Enhancing Vehicle Make and Model Recognition with 3D Attention Modules. *arXiv* **2025**. [\[CrossRef\]](#)
2. Yang, S.; Liu, Y.; Liu, Z.; Xu, C.; Du, X. Enhanced Vehicle Logo Detection Method based on Self-Attention Mechanism for Electric Vehicle application. *World Electr. Veh. J.* **2024**, *15*, 467. [\[CrossRef\]](#)
3. Wang, D.; Guo, J.; Zhang, C. A novel hybrid deep learning model for complex systems: A case of train delay prediction. *Adv. Civ. Eng.* **2024**, *2024*, 8163062. [\[CrossRef\]](#)
4. Liu, D. Progressive multi-task Anti-Noise learning and distilling frameworks for fine-grained vehicle recognition. *arXiv* **2024**. [\[CrossRef\]](#)
5. Fang, J.; Zhou, Y.; Yu, Y.; Du, S. Fine-Grained vehicle model recognition using a Coarse-to-Fine convolutional neural network architecture. *IEEE Trans. Intell. Transp. Syst.* **2016**, *18*, 1782–1792. [\[CrossRef\]](#)
6. Sochor, J.; Herout, A.; Havel, J. BoxCars: 3D Boxes as CNN Input for Improved Fine-Grained Vehicle Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016. [\[CrossRef\]](#)
7. Llorca, D.F.; Colas, D.; Daza, I.G.; Parra, I.; Sotelo, M.A. Vehicle model recognition using geometry and appearance of car emblems from rear view images. In Proceedings of the 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), Qingdao, China, 8–11 October 2014; pp. 3094–3099. [\[CrossRef\]](#)
8. Biglari, M.; Soleimani, A.; Hassanpour, H. Part-based recognition of vehicle make and model. *IET Image Process.* **2017**, *11*, 483–491. [\[CrossRef\]](#)
9. Bularz, M.; Przystalski, K.; Ogorzałek, M. Car make and model recognition system using rear-lamp features and convolutional neural networks. *Multimed. Tools Appl.* **2023**, *83*, 4151–4165. [\[CrossRef\]](#)
10. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**. [\[CrossRef\]](#)
11. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022. [\[CrossRef\]](#)
12. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16000–16009. [\[CrossRef\]](#)
13. He, J.; Chen, J.-N.; Liu, S.; Kortylewski, A.; Yang, C.; Bai, Y.; Wang, C. TransFG: A Transformer Architecture for Fine-Grained Recognition. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 852–860. [\[CrossRef\]](#)
14. Krause, J.; Deng, J.; Stark, M.; Li, F.-F. Collecting a Large-Scale Dataset of Fine-Grained Cars. 2013. Available online: <https://ai.stanford.edu/~jkrause/papers/fgvc13.pdf> (accessed on 20 January 2026).
15. Yang, L.; Luo, P.; Loy, C.C.; Tang, X. A large-scale car dataset for fine-grained categorization and verification. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3973–3981. [\[CrossRef\]](#)
16. 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), Las Vegas, NV, USA, 27–30 June 2016; pp. 2167–2175. [\[CrossRef\]](#)
17. Hu, Y.; Li, S.; Yan, J.; Shao, W.; Luo, X. CAR-1000: A new Large Scale Fine-Grained Visual Categorization Dataset. *arXiv* **2025**. [\[CrossRef\]](#)
18. Lyu, Y.; Schiopu, I.; Cornelis, B.; Munteanu, A. Framework for vehicle make and model Recognition—A new Large-Scale dataset and an efficient Two-Branch–Two-Stage deep learning architecture. *Sensors* **2022**, *22*, 8439. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Agrawal, S. Global License Plate Dataset. *arXiv* **2024**. [\[CrossRef\]](#)
20. Zwemer, M.H.; Brouwers, G.M.Y.E.; Wijnhoven, R.G.J.; De With, P.H.N. Semi-automatic training of a vehicle make and model recognition system. In *Image Analysis and Processing—ICIAP 2017*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2017; pp. 321–332. [\[CrossRef\]](#)
21. Manzoor, M.A.; Morgan, Y.; Bais, A. Real-Time vehicle make and model recognition system. *Mach. Learn. Knowl. Extr.* **2019**, *1*, 611–629. [\[CrossRef\]](#)
22. Maurya, T.; Kumar, S.; Rai, M.; Saxena, A.K.; Goel, N.; Gupta, G. Real time vehicle Classification using Deep Learning—Smart Traffic Management. *Eng. Rep.* **2025**, *7*, e70082. [\[CrossRef\]](#)
23. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Chu, G.; Vasudevan, V.; Zhu, Y.; Pang, R.; et al. Searching for MobileNetV3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324. [\[CrossRef\]](#)
24. Tan, M.; Le, Q.V. EfficientNetV2: Smaller models and faster training. *arXiv* **2021**. [\[CrossRef\]](#)

25. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**. [[CrossRef](#)]
26. Muñoz, A.; Thomas, N.; Vapsi, A.; Borrajo, D. Veri-Car: Towards open-world vehicle information retrieval. *arXiv* **2024**. [[CrossRef](#)]
27. Zhang, H.; Zhou, W.; Liu, G.; Wang, Z.; Qian, Z. Fine-Grained Vehicle Make and Model Recognition Framework based on magnetic fingerprint. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 8460–8472. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.