This is an author-prepared version of a paper accepted for publication in the *Journal of the Brazilian Computer Society*.
The final version will be available at the SBC-OpenLib Library.

Toward Unified Fine-Grained Vehicle Classification and Automatic License Plate Recognition

Gabriel E. Lima* [**Federal University of Paraná** | ***gelima@inf.ufpr.br***]
Valfride Nascimento [**Federal University of Paraná** | ***vwnascimento@inf.ufpr.br***]
Eduardo Santos [**Paraná Military Police, Federal University of Paraná** | ***ed.santos@pm.pr.gov.br***]
Eduil Nascimento Jr. [**Paraná Military Police** | ***eduiljunior@pm.pr.gov.br***]
Rayson Laroca [**Pontifical Catholic University of Paraná, Federal University of Paraná** | ***rayson@ppgia.pucpr.br***]
David Menotti [**Federal University of Paraná** | ***menotti@inf.ufpr.br***]

* *Department of Informatics, Federal University of Paraná, R. Evaristo F. Ferreira da Costa 391, Jardim das Américas, Curitiba, PR, 81530-090, Brazil.*

Abstract. Extracting vehicle information from surveillance images is essential for intelligent transportation systems, enabling applications such as traffic monitoring and criminal investigations. While Automatic License Plate Recognition (ALPR) is widely used, Fine-Grained Vehicle Classification (FGVC) offers a complementary approach by identifying vehicles based on attributes such as color, make, model, and type. Although there have been advances in this field, existing studies often assume well-controlled conditions, explore limited attributes, and overlook FGVC integration with ALPR. To address these gaps, we introduce UFPR-VeSV, a dataset comprising 24,945 images of 16,297 unique vehicles with annotations for 13 colors, 26 makes, 136 models, and 14 types. Collected from the Military Police of Paraná (Brazil) surveillance system, the dataset captures diverse real-world conditions, including partial occlusions, nighttime infrared imaging, and varying lighting. All FGVC annotations were validated using license plate information, with text and corner annotations also being provided. A qualitative and quantitative comparison with established datasets confirmed the challenging nature of our dataset. A benchmark using five deep learning models further validated this, revealing specific challenges such as handling multicolored vehicles, infrared images, and distinguishing between vehicle models that share a common platform. Additionally, we apply two optical character recognition models to license plate recognition and explore the joint use of FGVC and ALPR. The results highlight the potential of integrating these complementary tasks for real-world applications. The UFPR-VeSV dataset is publicly available at: `https://github.com/Lima001/UFPR-VeSV-Dataset`.

Keywords: Intelligent Transportation Systems, Fine-Grained Vehicle Classification, Automatic License Plate Recognition, Surveillance.

1 Introduction

Extracting vehicle information from surveillance images is a crucial aspect of Intelligent Transportation Systems (ITS), enabling applications such as traffic monitoring and criminal investigations [Yang *et al*., 2015; He *et al*., 2024a; Laroca *et al*., 2025]. Traditional vehicle identification methods primarily rely on Automatic License Plate Recognition (ALPR). However, ALPR techniques are susceptible to partial occlusions, viewpoint variations, and poor image quality, all of which can degrade recognition performance [Fan and Zhao, 2022; Nascimento *et al*., 2024; Wojcik *et al*., 2025].

Fine-Grained Vehicle Classification (FGVC) presents a complementary solution by categorizing vehicles based on attributes such as color, make, model, type, and year. Unlike License Plates (LPs), these attributes are often more resilient to occlusions and viewpoint changes. Additionally, although distinguishing visually similar vehicles is an inherently complex task, recent research has achieved remarkable results [Wang *et al*., 2020; Lu *et al*., 2023].

However, a review of the literature reveals that FGVC research assumes at least one of the following controlled conditions: fixed viewpoints, sufficient lighting, or high-quality images. These assumptions fail to fully capture the challenges of real-world surveillance scenarios. Moreover, existing research frequently addresses make and model recognition separately from color and type recognition [Amirkhani and Barshooi, 2023; Hu *et al*., 2023], hindering a comprehensive exploration of vehicle attributes.

Another unexplored aspect in existing research is the integration of FGVC and ALPR, which could offer significant benefits [Oliveira *et al*., 2021]. By cross-referencing FGVC attributes with LP records, ALPR system errors can be identified, reducing false positives and enhancing information retrieval. Furthermore, FGVC can help refine the search space for ALPR, particularly in forensic scenarios where LPs are low-quality or occluded [Nascimento *et al*., 2025].

To bridge these gaps, we present the publicly available UFPR Vehicle Surveillance (UFPR-VeSV) dataset, consisting of 24,945 images of 16,297 unique vehicles. It includes annotations across 13 color classes, 26 makes, 136 models, and 14 vehicle types. Sourced from the Military Police of Paraná (Brazil) surveillance system, the dataset reflects diverse real-world conditions, including frontal and rear views, partial occlusions, varying lighting, and nighttime infrared imaging. LPs were manually annotated and used to retrieve vehicle information for validating the FGVC annotations. The dataset also includes annotations for both LP characters and corner coordinates.

To establish the value of the UFPR-VeSV dataset, this paper makes the following contributions. First, we conduct a comprehensive comparison with existing datasets to highlight the dataset's novelty. Second, we empirically validate its challenging nature by showing that a simple transfer-learning approach, effective on related datasets, fails to achieve adequate results on ours. Third, we establish performance baselines by benchmarking five deep learning models for FGVC tasks and evaluating two optical character recognition models for License Plate Recognition (LPR). Finally, as a step toward a unified solution, we explore the integration of FGVC and ALPR and analyse its effectiveness for enhancing vehicle information retrieval.

The dataset's design, which prioritized high-quality FGVC annotations, resulted in the exclusion of truly unconstrained ALPR scenarios (see Section 3.2). This limitation likely explains the high LPR performance, which surpassed individual FGVC tasks. Despite this observation, the experiments confirm that integrating ALPR and FGVC is a promising strategy for enhancing vehicle information retrieval. This joint analysis represents a significant step toward a unified approach, and the challenges identified herein provide a foundation for future research.

A preliminary version of this research was published at the 2024 Conference on Graphics, Patterns, and Images (SIBGRAPI) [Lima *et al*., 2024]. This work expands upon that study in several key aspects. First, we broaden the scope beyond vehicle color recognition to include the recognition of vehicle make, model, and type. We also introduce a larger and more diverse dataset that more accurately reflects real-world conditions. To highlight its contributions, we compare the proposed dataset with prominent existing ones in ALPR and FGVC. Additionally, the methodology has been enhanced through the integration of an additional deep-learning model, along with improvements in training and evaluation procedures. Lastly, we present LPR results on the new dataset and introduce a novel contribution by integrating our best-performing LPR and FGVC methods.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 introduces the UFPR-VeSV dataset. Section 4 provides a comparative analysis with existing datasets. Section 5 presents the experimental evaluation of FGVC tasks. Section 6 details the experimental evaluation of LPR and its subsequent integration with FGVC. Finally, Section 7 concludes the paper and discusses future research directions.

2 Related Work

This section provides an overview of recent studies and datasets in the fields of FGVC and ALPR. Section 2.1 focuses on vehicle color recognition, while Section 2.2 covers vehicle make and model recognition, and Section 2.3 is related to vehicle type recognition. Section 2.4 reviews ALPR-related research. A differentiation of existing studies and this work is presented separately in Section 2.5.

2.1 Vehicle Color Recognition

Vehicle color recognition plays a significant role in vehicle identification, as color information is visually distinctive, less affected by occlusions, and remains stable across viewpoints [Chen *et al*., 2014]. Early studies used small datasets captured in controlled environments, relying on handcrafted feature extraction methods combined with machine learning classifiers for color prediction [Baek *et al*., 2007; Son *et al*., 2007; Dule *et al*., 2010]. These works laid the groundwork for more robust and advanced methodologies.

The first large-scale, publicly available dataset for this task was introduced by Chen *et al*. [2014], comprising 15,601 frontal-view images categorized into eight colors. Their initial approach employed a region-of-interest selector and support vector machines. Due to its diverse conditions – including variations in lighting, haze, and overexposure – the dataset became a popular benchmark for subsequent research using Convolutional Neural Network (CNN) models [Fu *et al*., 2020; Zhang *et al*., 2018; Hu *et al*., 2015].

Following that, researchers have introduced new datasets with new scenarios to be explored. Wang *et al*. [2021] proposed a dataset of 32,220 rear-view images, classified into 11 colors and 75 subcategories, and explored a hybrid CNN-Vision Transformer (ViT) model for recognition. Hu *et al*. [2023] introduced the Vehicle Color-24 dataset, consisting of 31,232 frontal-view images distributed across 24 color classes, and developed a CNN with multi-scale feature fusion and a specialized loss function to address class imbalance, reaching promising results.

More recently, in [Lima *et al*., 2024], we introduced UFPR-VCR, a dataset designed to capture real-world challenges such as partial occlusions and nighttime conditions. It consists of 10,039 images taken under varying conditions, including both frontal and rear views. By evaluating four deep learning architectures, we achieved a peak accuracy of 66.2%, highlighting the complexity of color recognition in unconstrained environments and underscoring the importance of investigating such scenarios.

2.2 Vehicle Make and Model Recognition

Vehicle make and model recognition is a challenging task in FGVC due to its high intra-class variability and low inter-class differences [Wang *et al*., 2020; Oliveira *et al*., 2021]. Early research in this area began with the Stanford Cars-196 dataset [Krause *et al*., 2013a,b], which comprises 16,185 web-sourced images from 196 vehicle models. The initial benchmark, achieved using the BubbleBank algorithm [Deng *et al*., 2013], was later surpassed by deep learning-based approaches [Yu *et al*., 2022; Lu *et al*., 2023], which have since become the standard in the field.

To better reflect real-world conditions, Yang *et al*. [2015] introduced CompCars, a dataset containing both web-sourced and surveillance images. Its surveillance subset (CompCars-SV) includes 44,481 frontal-view images of 281 vehicle models annotated with multiple attributes. Similarly, Sochor *et al*. [2016] developed the BoxCars dataset, consisting of 63,750 surveillance images with multi-viewpoint data and 3D

bounding box annotations, further challenging the recognition of vehicle make and model.

Wang *et al*. [2020] introduced MPF-Cars, a large-scale dataset with 335,011 images from 2,019 models and 180 manufacturers. Their proposed three-branch CNN model leveraged full-vehicle, front, and logo views to improve identification performance. Likewise, Kuhn and Moreira [2021] presented BRCars, a dataset of 300,325 images covering 427 car models from online advertisements. Differently from related works, the dataset includes both exterior and interior view images, with recognition performed without distinguishing between the two.

Recent research has focused on advancing recognition methods and scenarios. Amirkhani and Barshooi [2023] introduced the DeepCar 5.0 dataset, utilizing CNNs to analyze vehicle headlights, grilles, and bumpers for enhanced recognition. Additionally, Wolf *et al*. [2024] explored an open-set recognition scenario and proposed a knowledge-distillation-based label smoothing approach, improving both closed-set and open-set recognition on the CompCars-SV dataset.

2.3 Vehicle Type Recognition

Vehicle type recognition provides a coarse-grained classification level when compared to vehicle make and model recognition, distinguishing between categories such as cars, trucks, and motorcycles. Early studies used handcrafted feature extraction methods applied to small and limited datasets [Ferryman *et al*., 1995; Jolly *et al*., 1996; Lai *et al*., 2001; Wu *et al*., 2001; Ma and Grimson, 2005].

Dong *et al*. [2015] pioneered the use of deep learning in recognizing vehicle type and introduced BIT-Vehicle, a dataset of 9,850 high-resolution frontal-view images from six types. Hu *et al*. [2017] later proposed a multi-task CNN for joint vehicle localization and type recognition, and presented the SYSU-Vehicle dataset with 5,000 web-sourced images and five classes. Later, Shvai *et al*. [2020] expanded the field by compiling a dataset of 73,638 toll booth images and integrating CNN features with optical sensor data.

Recent studies have explored unconventional scenarios to improve recognition under challenging conditions. For example, Basak and Suresh [2024] employed residual dense networks to generate super-resolved images, enhancing accuracy in low-resolution images. In a different approach, Luo *et al*. [2024] developed a method for satellite imagery, combining DenseNet [Huang *et al*., 2017] and Transformer-in-Transformer [Han *et al*., 2021] layers to extract fine-grained spatial features.

2.4 Automatic License Plate Recognition

ALPR systems typically comprise two main components: License Plate Detection (LPD) and LPR [Laroca *et al*., 2023b, 2025]. LPD identifies the license plate region within an image, while LPR extracts and interprets the characters. One of the earliest widely adopted datasets within the area is AOLP [Hsu *et al*., 2013], which includes 2,049 images across three subsets, each tailored to different ALPR applications.

Further research aimed to improve evaluation scenarios proposing datasets such as PKU [Yuan *et al*., 2017], SSIG-ALPR [Gonçalves *et al*., 2018], and UFPR-ALPR [Laroca *et al*., 2018]. Nevertheless, Xu *et al*. [2018] identified limitations within those datasets in either scale (containing fewer than 10,000 images) or diversity. This led to the creation of CCPD, a large-scale dataset with 250,000 images captured by roadside parking management personnel using handheld cameras.

Further expanding dataset diversity, Laroca *et al*. [2022] introduced RodoSol-ALPR, a dataset of 20,000 toll booth images collected under varying conditions, including different times of day, weather scenarios, and camera distances. It was the first publicly available dataset within the field to include Mercosur license plates. Additionally, it remains the largest dataset with annotated motorcycle images.

Recent research has focused on improving feature extraction to better handle generalization in real-world scenarios. Rao *et al*. [2024] and Liu *et al*. [2024] utilized spatial attention mechanisms to enhance both LP detection and recognition. Moreover, Liu *et al*. [2024] explored the use of synthetically generated data to increase dataset diversity, demonstrating its effectiveness in enhancing recognition performance.

Lastly, it is worth pointing out the work from Oliveira *et al*. [2021], which introduced the Vehicle-Rear dataset. It contains 3,000 rear-view images with extra annotations for vehicle color, make, and model. The authors proposed a dual-stream CNN network and integrated vehicle appearance features with LPR to perform vehicle identification. Despite including annotations suitable for both ALPR and FGVC tasks, the dataset was not explored for the latter.

2.5 Key Differentiators and Contributions

This work's primary contribution is the introduction of the UFPR-VeSV dataset, designed to capture vehicles under challenging surveillance conditions often absent in existing benchmarks. Beyond its capture conditions, the dataset introduces novel complexities for FGVC tasks. For color recognition, it includes scenarios rarely addressed in the literature, such as infrared images and multicolored vehicles. For type recognition, a finer class granularity is employed to distinguish between challenging categories (e.g., motorcycles versus scooters). For model recognition, the dataset includes vehicles that are visually similar due to shared manufacturing platforms, even across different vehicle types. The dataset also incorporates rear-view images of trucks with obstructed bodies and motorcycles.

Finally, our research methodology sets this work apart. First, we not only evaluate FGVC tasks in isolation but also assess the methods for vehicle color, make, model, and type recognition jointly, revealing specific challenges in that combined context. Second, we benchmark LPR methods on our proposed dataset. Finally, we conduct a specific analysis integrating the outputs of both LPR and FGVC models. This unified approach allows us to investigate recognition challenges under adverse conditions and showcase how these complementary systems can be combined.

3 UFPR-VeSV Dataset

UFPR-VeSV comprises 24,945 images of 16,297 unique vehicles, specifically designed to support FGVC research in real-world surveillance scenarios. It encompasses vehicles classified into 13 colors, 26 makes, 136 models, and 14 types. The dataset exhibits a highly unbalanced distribution across these attributes (see Figure 1), reflecting the characteristics of Brazilian traffic [Ministério dos Transportes, 2024; Farias and Croquer, 2023; Celestino, 2021]. This poses a challenge for recognition models, which often struggle with underrepresented classes [Huang *et al.*, 2016; Ochal *et al.*, 2023].

The images were sourced from the Military Police of Paraná's surveillance system within a single municipality, captured by distinct cameras positioned on highways, urban streets, and rural roads. The cameras operate under diverse conditions, capturing vehicles from varying angles and distances, with environmental factors such as lighting, weather, and motion blur influencing image quality. Additionally, approximately 3% of the images were manually captured by police officers during monitoring operations, introducing further data variety.

Although the dataset is designed to focus on a single vehicle per image, additional vehicles may appear due to camera positioning and perspective, as shown in Figure 2. In some cases, multiple LPs are visible but partially occluded. These scenarios are retained to reflect real-world conditions and the challenges they pose for ALPR and FGVC-based systems.

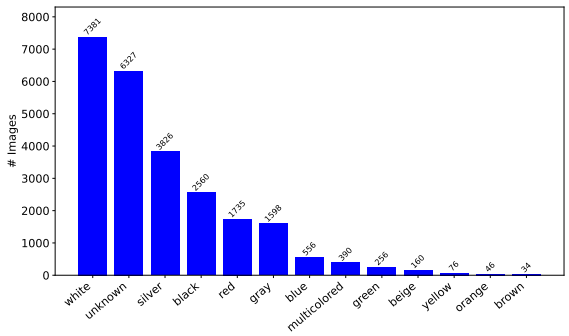
Figure 2. Examples of images featuring multiple vehicles due to camera perspective. The background vehicle is highlighted with a green border, while the main vehicle is shadowed to enhance contrast.

The dataset spans a wide temporal range, including both daytime and nighttime conditions. While timestamps are not available, images are categorized by the camera's capture mode. Nighttime images, primarily captured in infrared mode, account for 5,372 images (21.5%). While infrared imaging improves visibility in low light, it also presents challenges, such as reduced contrast in fine details and potential overexposure from vehicle headlights (see Figure 3).

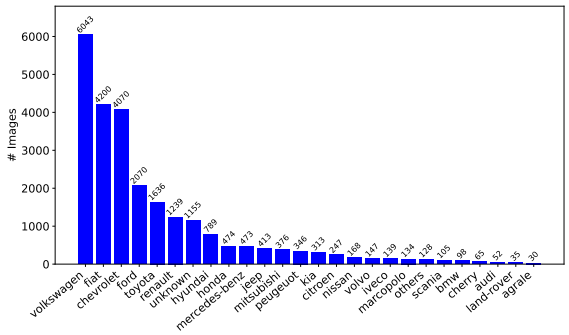
(a) Adequate **(b)** Reduced contrast **(c)** Light overexposure

Figure 3. Examples of infrared images under varying conditions: (a) optimal visibility with enhanced perception, (b) reduced contrast affecting detail clarity, and (c) overexposure caused by vehicle headlights.

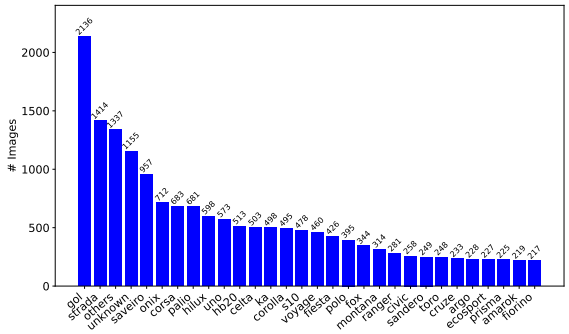
Another factor contributing to the diversity of the dataset is the viewpoint of the vehicles. UFPR-VeSV captures multiple viewpoints, including frontal, rear, three-quarter, and high-angle perspectives. To standardize annotations, each image is



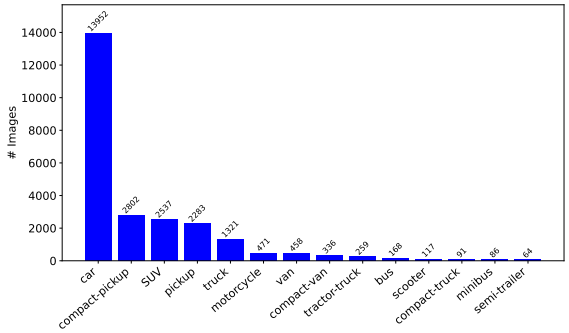
(a) Distribution of vehicle colors in the UFPR-VeSV dataset.



(b) Distribution of vehicle makes in the UFPR-VeSV dataset.



(c) Distribution of the 30 most common vehicle models in the UFPR-VeSV dataset.



(d) Distribution of vehicle types in the UFPR-VeSV dataset.

Figure 1. Distribution of vehicles across the attributes of color (a), make (b), model (c) and type (d) in the UFPR-VeSV dataset. For better visualization, only the 30 most common vehicle models are displayed in (c), representing 63.7% of the total images.

categorized as either front or rear view based on the visibility of the LP. As a result, the dataset contains 13,842 rear-view and 11,103 frontal-view images.

The dataset contains 16,297 LPs with two distinct layouts: Brazilian (5,171 LPs) and Mercosur (11,126 LPs). These LPs are captured under diverse conditions, including varying

angles, resolutions, and levels of noise (see Figure 4). We remark that 0.2% of the LPs contain illegible characters. These images were retained due to their minimal impact on overall performance, their representation of real-world obstructions and degradations, and their non-interference with FGVC tasks. Importantly, accurate LP information was successfully retrieved even in cases involving occlusion (see Section 3.2 for annotation details).

Figure 4. Example of LPs cropped from images captured under diverse conditions, showcasing variations in resolution, perspective, and image quality. The corresponding annotated LP text is shown below each image.

Regarding ALPR, the UFPR-VeSV dataset includes annotations for both LP characters and corner coordinates. Figure 5 shows the character distribution across the seven LP positions in the proposed dataset. While digits are relatively evenly distributed, letters exhibit significant imbalances: LPs are more likely to start with certain letters, such as “A” and “B”. Character distribution in ALPR datasets is inherently imbalanced due to region-specific LP allocation policies [Gonçalves *et al.*, 2018; Laroca *et al.*, 2022].

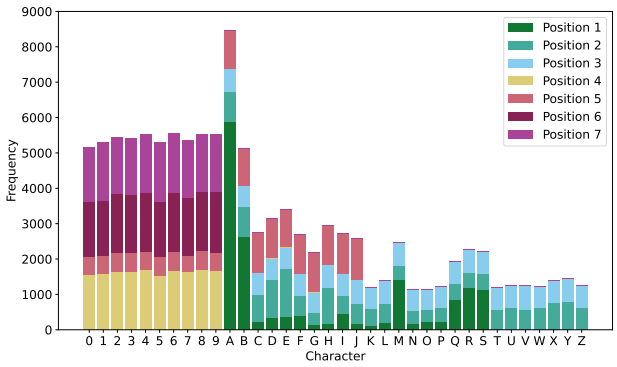


Figure 5. Distribution of LP character classes in the UFPR-VeSV dataset.

With the dataset’s key characteristics established, the following sections detail its creation process. Section 3.1 elaborates on the image selection and preprocessing methods. Section 3.2 describes the annotation methodology, including LP text labeling and FGVC attribute verification. Section 3.3 defines the dataset splitting protocol to ensure reproducibility. Finally, Section 3.4 discusses the privacy safeguards implemented to address ethical considerations.

3.1 Image Collection and Preprocessing

Initially, 30,240 images were collected from the Military Police of Paraná’s surveillance system. Each image was manually inspected to evaluate its suitability for the study, followed by a filtering process to eliminate samples that did not meet the research criteria. A total of 1,253 images were discarded due to factors such as extreme LP occlusions, severe image degradation, and poor vehicle framing. Furthermore, the images were grouped based on LP information, and highly similar samples – such as those with similar viewpoints and lighting conditions – were removed, resulting in the exclusion of an additional 4,042 images.

As the police system collects images from various surveillance sources, the dataset exhibited inconsistencies in format. Some images were already cropped around vehicles, while others included significant background content. To standardize the dataset, the YOLOv11 model [Ultralytics, 2025] was employed for vehicle detection and precise cropping. This model was selected due to its strong detection performance and its widespread adoption in both academic and industrial applications [Nyi Myo *et al.*, 2025; He *et al.*, 2024b; Khanam and Hussain, 2024].

A manual review was carried out to ensure proper vehicle cropping, especially in images containing multiple detected vehicles. Manual intervention was also necessary in cases where the model failed to detect a vehicle or produced inaccurate results. This was particularly important in challenging scenarios, such as motorcycles captured in low-light conditions or vehicles appearing at a distance.

Additional standardization was performed to address the presence of green borders found in some pre-cropped images. To ensure visual consistency, a 5-pixel border was removed from all sides of each image. A manual review was then conducted to verify that no LPs were significantly affected or occluded by this adjustment. Figure 6 illustrates an example of this issue and the resulting image after processing.

(a) Original **(b)** Processed

Figure 6. Example of border standardization. (a) original image with a green border; (b) result after removing a 5-pixel margin from all sides.

Finally, it is important to highlight that the images were not resized to uniform dimensions. As a result, the dataset retains a variety of image sizes, with widths ranging from 89 to 2,110 pixels and heights from 135 to 1,408 pixels. This approach prevents distortions that could affect recognition tasks, allowing resizing to be handled as needed for specific models and applications.

3.2 Annotation Process

LP text annotations were initially performed manually and subsequently used to automate the retrieval of FGVC attributes via the Brazilian National Traffic Secretariat (SENATRAN) database. The retrieved vehicle information was then manually verified against the vehicle’s visual appearance to ensure consistency. In cases of discrepancies, the LP text was re-annotated manually, followed by another round of automated retrieval and verification of the FGVC attributes.

The use of manually annotated LPs during the FGVC annotation process carries an important implication: the LPs in the dataset are human-recognizable. While the dataset captures an unconstrained surveillance scenario for FGVC

tasks, the ALPR context had to be restricted to ensure LP readability. This introduces a limitation in fully replicating real-world conditions, as it substantially reduces the presence of challenging cases typically faced by ALPR systems.

However, this limitation represents a necessary step toward advancing FGVC research and its integration with ALPR. To date, no existing study has jointly addressed vehicle color, make, model, and type recognition, nor explored their combined use with LP recognition (as detailed in Section 2). The primary goal of UFPR-VeSV is to establish a foundation for future research in this direction. We anticipate that future datasets will build upon and enhance the scenarios and discussions presented here, ultimately contributing to the progress of both FGVC and ALPR fields.

To better support FGVC-related tasks, the color, make, and model annotations in the UFPR-VeSV dataset were refined to maximize its utility across all attributes. These adjustments aimed to include previously overlooked scenarios, enabling a more comprehensive evaluation of their impact on recognition performance and better reflecting the challenges of real-world ITS applications.

Infrared images were grouped into a dedicated color class due to their inherent lack of color data. Vehicles officially categorized or visually similar to multicolored – a non-literal translation of *fantasia* in Portuguese, used when no predominant color is identifiable – were also retained in their own color class. Motorcycles and scooters had color, make, and model attributes reassigned to an "unknown" class due to their underexplored nature in FGVC literature and the limited visual information from these images (Figure 7a). The same strategy was applied to rear-view truck-like vehicles, where cargo compartments frequently obstruct the main body, making accurate FGVC impossible (Figures 7b and 7c).

(a) (b) (c)

Figure 7. Examples of vehicles assigned to the "unknown" class for color, make, and model annotations: (a) a red Yamaha motorcycle; (b) a white Volkswagen truck; (c) a red Mercedes-Benz truck. In all cases, the main body of the vehicle is obstructed, rendering the identification of color, make, and model impossible.

Classes with fewer than 25 samples were adjusted to reduce extreme class imbalance. Underrepresented colors were merged into visually similar classes (e.g., purple into blue, garnet into red, gold into beige) while underrepresented make/model were consolidated into a generic "others" class. Additionally, vehicle models from different production years were grouped into a single class, in contrast to related datasets [Yang *et al.*, 2015; Kuhn and Moreira, 2021]. This decision was made to avoid excessive fragmentation, ensuring both class balance and experimental reliability.

Corner annotations were obtained for each LP using a two-stage approach. We use a YOLOv11 model [Ultralytics, 2025] to detect the LP regions within the input image, and then CDCC-NET [Laroca *et al.*, 2021] regressed the corner coordinate values. Both models were fine-tuned on popular public datasets collected from multiple regions, such as AOLP [Hsu *et al.*, 2013], UFPR-ALPR [Laroca *et al.*, 2018], CLPD [Zhang *et al.*, 2021], among others. A matching algorithm was used to handle multiple detections within an image, assigning a match if all LP corners fell within the vehicle's bounding box. In cases where multiple plates were visible (as in Figure 2), manual identification was performed. Manual corrections were applied when the YOLOv11 model misidentified non-LP textual regions.

Additional annotations – such as vehicle viewpoint and camera capture mode – were manually labeled and carefully verified to improve the overall quality of the dataset.

3.3 Splitting Protocol

To ensure robust and unbiased model evaluation on the UFPR-VeSV dataset, we developed a custom 5-fold evaluation protocol. This methodology is specifically designed to prevent data leakage and mitigate partitioning bias, thereby providing a more reliable estimate of a model's true generalization performance.

The protocol begins by partitioning the entire dataset into five non-overlapping folds, a process governed by two critical constraints. The first constraint is the prevention of data leakage, an issue that can lead to overly optimistic performance estimates [Laroca *et al.*, 2023a]; this is achieved by confining all images of the same vehicle (i.e. with the same LP) to a single fold. The second constraint is multi-attribute stratification, where partitioning is performed simultaneously across all four vehicle attributes – color, make, model, and type – to ensure each fold preserves the class distribution of the original dataset as close as possible.

Following that, we generated ten distinct 3:1:1 train-validation-test splits from the five folds. Our protocol is designed to deterministically use each of the $C(5,3) = 10$ unique combinations of three folds as the training set exactly once. To achieve this, each of the five folds (indexed 0 to 4) is used as the test set exactly twice. The two validation sets for that test set are then assigned using the rules $val_1 = (test + 1) \pmod 5$, and $val_2 = (test + 2) \pmod 5$. The training set for each split is subsequently formed by the three remaining folds. The specific files defining all train-validation-test splits are publicly available to ensure reproducibility.

3.4 Privacy concerns

To comply with ethical and legal guidelines, privacy-sensitive elements were addressed. While LPs do not constitute personal data in Brazil – since they cannot be directly linked to vehicle owners – some images contain identifiable faces of drivers or pedestrians. To mitigate this, the RetinaFace model [Deng *et al.*, 2020] was employed to detect facial regions for further blurring. After automated processing, a manual verification step was conducted, allowing for minor corrections to ensure the quality and effectiveness of the anonymization process.

4 Comparative Analysis

This section provides a two-stage comparative analysis to highlight the contributions of the UFPR-VeSV dataset. First, a qualitative analysis (Section 4.1) reveals that preceding benchmarks are often limited and fail to capture real-world diversity, rendering them less complex. Subsequently, a quantitative analysis (Section 4.2) provides empirical evidence for this claim.

4.1 Qualitative Analysis

Table 1 provides a structured comparison of related datasets. The total number of images and unique vehicle identities are also included. When details were not provided in the original studies, they are marked as unknown (unk.). The number of classes is omitted to prevent inconsistencies, as different datasets may define class labels in varying ways. For instance, some datasets merge make and model information into a single class, while others treat them hierarchically.

Table 1. Comparison between UFPR-VeSV (proposed in this work) and related datasets.

Dataset	Images	Vehicles	Source	Viewpoint	Annotations			
					Color	Make/Model	Type	ALPR
Ferryman *et al.* [1995]	176	unk.	Field	Frontal/Rear	-	-	✓	-
Jolly *et al.* [1996]	393	unk.	Field	n/a	-	-	✓	-
Lai *et al.* [2001]	unk.	unk.	Field	Rear	-	-	✓	-
Wu *et al.* [2001]	800	unk.	Field	Frontal	-	-	✓	-
Ma and Grimson [2005]	unk.	unk.	Field	Frontal	-	-	✓	-
Baek *et al.* [2007]	500	unk.	Field	Frontal	✓	-	-	-
Dule *et al.* [2010]	1,960	unk.	Field	Frontal	✓	-	-	-
AOLP [Hsu *et al.*, 2013]	2,049	1,286	Field	Frontal/Rear	-	-	-	✓
Stanford Cars-196 [Krause *et al.*, 2013a]	16,185	unk.	Web	Frontal/Rear	-	✓	-	-
Chen *et al.* [2014]	15,601	unk.	Field	Frontal	✓	-	-	-
BIT-Vehicle [Dong *et al.*, 2015]	9,850	unk.	Field	Frontal	-	-	✓	-
CompCars-SV [Yang *et al.*, 2015]	44,481	unk.	Field	Frontal	✓	✓	-	-
BoxCars [Sochor *et al.*, 2016]	63,750	21,250	Field	Frontal/Rear	-	✓	-	-
SSIG-ALPR [Gonçalves *et al.*, 2018]	2,000	815	Field	Frontal/Rear	-	-	-	✓
PKU [Yuan *et al.*, 2017]	3,977	1,933	Field	Frontal	-	-	-	✓
SYSU-Vehicle [Hu *et al.*, 2017]	5,000	unk.	Web	Frontal/Rear	-	-	✓	-
CCPD [Xu *et al.*, 2018]	250,000	unk.	Field	Frontal/Rear	-	-	-	✓
UFPR-ALPR [Laroca *et al.*, 2018]	4,500	150	Field	Frontal/Rear	-	-	-	✓
Shvai *et al.* [2020]	73,638	unk.	Field	Frontal	-	-	✓	-
MPF-Cars [Wang *et al.*, 2020]	335,011	71,305	Field	Frontal	-	✓	-	-
BR-Cars [Kuhn and Moreira, 2021]	300,325	52,000	Web	Frontal/Rear	-	✓	-	-
Vehicle-Rear [Oliveira *et al.*, 2021]	26,160*	2,966	Field	Rear	✓	✓	-	✓
Wang *et al.* [2021]	32,220	unk.	Field	Rear	✓	-	-	-
RodoSol-ALPR [Laroca *et al.*, 2022]	20,000	12,785	Field	Frontal/Rear†	-	-	-	✓
DeepCar 5.0 [Amirkhani and Barshooi, 2023]	40,185	unk.	Web	Frontal	-	✓	-	-
Vehicle Color-24 Hu *et al.* [2023]	31,232	unk.	Field	Frontal	✓	-	-	-
Basak and Suresh [2024]	7,242	unk.	Field	Frontal/Rear	-	-	✓	-
Luo *et al.* [2024]	57,984	unk.	Field	n/a	-	-	✓	-
UFPR-VCR [Lima *et al.*, 2024]	10,039	9,502	Field	Frontal/Rear	✓	-	-	-
UFPR-VeSV (ours)	24,945	16,297	Field	Frontal/Rear	✓	✓	✓	✓

*Counting reported for Vehicle-Rear vehicle frames.
†Only motorcycle images are rear-view.

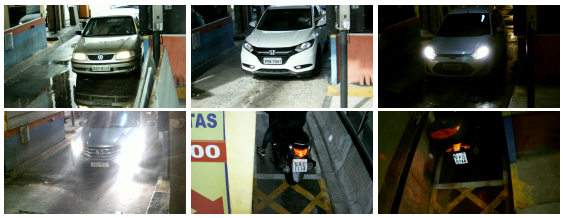
Following the FGVC literature, we also categorize datasets based on their image sources. Field-sourced datasets are derived from surveillance or traffic monitoring systems, capturing vehicles in their natural operational conditions. In contrast, web-sourced datasets comprise images gathered online from advertisements and curated photographic scenes. Although web-sourced datasets contribute to FGVC research, they often fail to reflect real-world surveillance conditions.

Datasets are further categorized based on the images' viewpoint. Frontal and rear viewpoints indicate which plate is in view. Other perspectives are labeled as not applicable (n/a) due to the absence of a visible plate, which are needed for our ALPR experiments and its integration with FGVC.

As Table 1 illustrates, most datasets specialize in either FGVC or ALPR, and FGVC benchmarks typically treat attributes like color, type, and make/model independently. The only exceptions that annotate all these attributes are Vehicle-Rear and UFPR-VeSV. However, Vehicle-Rear is limited to 3,000 unique vehicles from a single viewpoint and was not used for FGVC in its original study. In contrast, UFPR-VeSV provides a more diverse collection of over 16,000 distinct vehicles captured from multiple viewpoints and under varied real-world conditions.

(a) Images from the dataset proposed by Chen *et al.* [2014].

(b) Images from CompCars-SV, proposed by Yang *et al.* [2015].



(c) Images from RodoSol-ALPR, proposed by Laroca *et al.* [2022].

(d) Images from UFPR-VeSV, proposed in this work.

Figure 8. Representative images from three public datasets and UFPR-VeSV. Our dataset features significantly more challenging scenarios, with vehicles captured from diverse viewpoints, environments, lighting conditions, image quality levels, and nighttime infrared imaging.

A common limitation among existing datasets is the lack of scenario diversity (illustrated in Figure 8). Datasets for vehicle color recognition are collected under controlled conditions, creating optimistic conditions that artificially boost recognition accuracy [Lima *et al.*, 2024]. Similarly, while other established datasets (e.g., CompCars-SV and RodoSol-ALPR) incorporate variations in weather and time, they still lack diversity in viewpoints and capture locations. In contrast, UFPR-VeSV introduces more diverse and challenging scenarios, capturing vehicles under varying illumination, partial occlusions, and complex real-world conditions.

4.2 Quantitative Analysis

This section presents an empirical evaluation designed to highlight the limited diversity and lack of challenging scenarios in related datasets. To this end, we benchmarked five deep learning classifiers on two widely used datasets – Chen *et al*. [2014] dataset for color recognition and CompCars-SV [Yang *et al*., 2015] for model recognition – and contrasted their performance against the UFPR-VeSV dataset.

The evaluation considered five architectures: EfficientNet-V2 [Tan and Le, 2021], MobileNet-V3 [Howard *et al*., 2019], ResNet-50 [He *et al*., 2016], Swin Transformer-V2 [Liu *et al*., 2022], and ViT-b16 [Dosovitskiy *et al*., 2021]. They were chosen due to their strong performance in computer vision tasks, broad support across deep learning frameworks, and adoption in related work [Hassan *et al*., 2021; Kuhn and Moreira, 2021; Lima *et al*., 2024].

To isolate the effect of dataset complexity, a controlled transfer-learning strategy was employed, inspired by our prior work [Lima *et al*., 2024]. This approach establishes a baseline for comparison. Each classifier was initialized with ImageNet-pretrained weights, and only the final classification layer was trained to adapt to the target classes.

All classifiers were trained for up to 500 epochs using the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$), Cross-entropy loss, a batch size of 128, and a weight decay of 10^{-5}. We used an initial learning rate of 10^{-3}, reduced by a factor of 10 after ten epochs of stagnant validation performance; an early-stopping condition was triggered after 15 epochs without improvement. All images were resized to a 224×224 input, preserving the aspect ratio by scaling the longest side and padding the shorter side. The training set was subjected to an additional data augmentation pipeline (random rotation, scaling, shearing, brightness/contrast adjustments, motion blur, and random masking[1]). As the final step for all images, normalization was applied using the standard ImageNet mean and standard deviation.

The evaluation was conducted using specific protocols for each dataset: our custom splitting protocol for UFPR-VeSV (see Section 3.3), the methodology from our prior work [Lima *et al*., 2024] for Chen *et al*. [2014] dataset, and the original protocol for CompCars [Yang *et al*., 2015]. Classification performance was measured using macro accuracy (Ma-Acc), micro accuracy (Mi-Acc), and F1-score (F1). Micro accuracy reflects the overall performance, while macro accuracy provides a balanced measure that gives equal weight to both frequent and rare classes.

The results in Table 2 reveal a clear performance gap. The simple transfer-learning approach was effective on the reference benchmarks; ViT-b16 achieved over 92% micro-accuracy, a result close to literature reports [Hu *et al*., 2023; Amirkhani and Barshooi, 2023]. In contrast, classifiers performed worse on UFPR-VeSV dataset due to its challenging nature and more realistic evaluation setting. This highlights that benchmarks with limited diversity risk producing optimistic evaluations that do not reflect practical performance.

[1]A file specifying all parameters of the data augmentation pipeline will be included with the dataset release.

Table 2. Performance metrics (%) achieved by all classifiers using the transfer-learning approach on the Chen *et al*. [2014] dataset (vehicle color recognition), the CompCars dataset [Yang *et al*., 2015] (vehicle model recognition), and the proposed UFPR-VeSV dataset (vehicle color, make, model, and type recognition). Results on the proposed dataset are averaged over ten runs, with standard deviations shown in parentheses.

Results for color [Chen *et al*., 2014] and model [Yang *et al*., 2015] recognition benchmarks.

Classifier	*Chen* et al. *[2014]*			*CompCars [Yang* et al*., 2015]*		
	Mi-Acc	Ma-Acc	F1	Mi-Acc	Ma-Acc	F1
EfficientNet-V2	85.5	85.6	86.8	73.5	70.5	72.2
MobileNet-V3	88.1	90.2	90.4	82.0	79.2	81.2
ResNet-50	92.7	93.4	93.8	87.4	82.5	85.0
Swin Transformer-V2	91.8	93.0	93.6	92.4	89.0	90.1
ViT-b16	**92.9**	**93.5**	**94.0**	**94.6**	**92.9**	**93.9**

Results on the UFPR-VeSV dataset.

Classifier	*UFPR-VeSV Color*			*UFPR-VeSV Type*		
	Mi-Acc	Ma-Acc	F1	Mi-Acc	Ma-Acc	F1
EfficientNet-V2	77.8 (0.6)	46.1 (1.7)	47.0 (1.8)	75.2 (0.5)	55.6 (2.7)	59.7 (2.7)
MobileNet-V3	81.0 (0.5)	50.9 (1.4)	53.2 (1.6)	76.9 (0.3)	56.4 (2.0)	61.4 (2.6)
ResNet-50	83.2 (0.5)	54.2 (0.8)	56.6 (1.1)	81.7 (0.4)	65.4 (1.7)	70.4 (1.4)
Swin Transformer-V2	**87.2 (0.4)**	56.5 (1.5)	60.2 (1.7)	84.4 (0.7)	**72.8(1.8)**	**76.8(1.9)**
ViT-b16	**87.5 (0.4)**	**60.2 (1.8)**	**63.3 (1.6)**	**86.3 (0.7)**	**75.0 (2.7)**	**77.1 (2.3)**

Classifier	*UFPR-VeSV Make*			*UFPR-VeSV Model*		
	Mi-Acc	Ma-Acc	F1	Mi-Acc	Ma-Acc	F1
EfficientNet-V2	42.6 (0.6)	23.7 (0.9)	24.1 (0.9)	35.3 (0.7)	21.5 (0.8)	22.6 (0.8)
MobileNet-V3	45.7 (1.0)	26.3 (1.3)	28.4 (1.4)	40.8 (0.7)	25.2 (1.0)	27.7 (1.2)
ResNet-50	51.0 (1.3)	30.9 (1.3)	33.4 (1.3)	46.8 (0.9)	28.1 (0.7)	30.4 (0.9)
Swin Transformer-V2	51.6 (1.0)	32.0 (1.2)	32.7 (1.4)	49.8 (0.7)	32.8 (1.1)	34.8 (1.0)
ViT-b16	**58.1 (0.5)**	**39.1 (1.1)**	**42.0 (1.0)**	**57.4 (1.3)**	**40.8 (1.1)**	**43.7 (1.4)**

5 Fine-grained Vehicle Classification

This section establishes the FGVC performance baseline on the UFPR-VeSV dataset. The analysis is split into two parts. In Section 5.1, we benchmark five deep learning classifiers for color, make, model, and type recognition and analyze the results for each task individually. In Section 5.2, we use the predictions from the best-performing classifier to analyze the results when these tasks are considered in combination.

5.1 Single-Task Analysis

This section establishes the FGVC baseline on the UFPR-VeSV dataset. Unlike the transfer-learning strategy in Section 4.2, we now employ an end-to-end fine-tuning strategy for five deep learning architectures: EfficientNet-V2, MobileNet-V3, ResNet-50, Swin Transformer-V2, and ViT-b16. The splitting protocol, loss function, and evaluation metrics remain consistent with the prior qualitative analysis.

All classifiers were initialized with ImageNet pre-trained weights, and all layers were set as trainable. Training was conducted for up to 500 epochs, with early stopping halting the process after 30 epochs without validation improvement. We used the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with a 5×10^{-4} weight decay and a 64 batch size. The initial learning rate was 10^{-2}, reduced by a factor of 10 after 20 epochs of stagnant validation loss.

A comprehensive data augmentation pipeline was applied during training to enhance generalization. This included random resized cropping (224×224 pixels), random horizontal flipping, and RandAugment [Cubuk *et al*., 2020] (two transformations, intensity 9). For inference, images were also resized to 224×224 while preserving their aspect ratio. As a final preprocessing step, all images were normalized using the standard ImageNet mean and standard deviation.

Table 3 presents the performance of each classifier across

the FGVC tasks. For the attention-based models (ViT and Swin Transformer), the results shown are from the transfer-learning approach (Section 4.2), as this strategy yielded superior performance. We attribute this is because the large data requirements of transformers can limit generalization when all layers are fine-tuned.

Table 3. Performance metrics (%) achieved by all classifiers using the end-to-end fine-tuning approach on vehicle color, make, model, and type recognition (averaged over 10 runs). Standard deviations are shown in parentheses.

Classifier	*UFPR-VeSV Color*			*UFPR-VeSV Type*		
	Mi-Acc	Ma-Acc	F1	Mi-Acc	Ma-Acc	F1
EfficientNet-V2	**93.5 (0.6)**	**71.5 (2.3)**	**73.8 (2.3)**	**96.1 (0.7)**	**89.0 (2.0)**	**90.2 (1.6)**
MobileNet-V3	**93.2 (0.7)**	**71.2 (2.9)**	**73.2 (2.5)**	95.2 (0.7)	85.6 (1.9)	87.4 (1.6)
ResNet-50	93.1 (0.4)	69.6 (2.6)	71.8 (2.3)	95.5 (0.6)	86.8 (2.3)	88.4 (1.9)
Swin Transformer-V2	87.2 (0.4)	56.5 (1.5)	60.2 (1.7)	84.4 (0.67	72.8 (1.8)	76.8 (1.9)
ViT-b16	87.5 (0.4)	60.2 (1.8)	63.3 (1.6)	86.3 (0.7)	75.0 (2.7)	77.1 (2.3)

Classifier	*UFPR-VeSV Make*			*UFPR-VeSV Model*		
	Mi-Acc	Ma-Acc	F1	Mi-Acc	Ma-Acc	F1
EfficientNet-V2	**94.4 (0.6)**	**85.0 (1.6)**	**86.4 (1.4)**	**90.9 (0.6)**	**86.2 (1.1)**	**87.3 (0.8)**
MobileNet-V3	91.3 (0.7)	78.0 1.7)	80.2 1.4)	86.5 (0.8)	79.2 (1.2)	80.9 (1.1)
ResNet-50	93.6 (0.5)	83.5 (1.7)	85.0 (1.1)	89.9 (0.9)	84.6 (1.2)	85.7 (0.8)
Swin Transformer-V2	51.6 (1.0)	32.0 (1.2)	32.7 (1.4)	49.8 (0.7)	32.8 (1.1)	34.8 (1.0)
ViT-b16	58.1 (0.5)	39.1 (1.1)	42.0 (1.0)	57.4 (1.3)	40.8 (1.1)	43.7 (1.4)

A general trend across all classifiers was the significant gap between micro-accuracy and the macro metrics (macro-accuracy and F1-score). This discrepancy is an expected result of the dataset's severe class imbalance, which causes classifiers to perform well on frequent classes but struggle with underrepresented ones. With this in mind, we selected EfficientNet-V2 for a detailed analysis to identify the specific challenges for each task, as it achieved the best performance across the experiments.

In color recognition, the classifier performed worst for beige, brown, blue, green, and gray. This is likely due to lighting variations and close shade similarities, such as darker shades appearing black or lighter beige appearing silver. The "multicolored" class was also challenging, with the classifier often predicting one of the vehicle's colors. This error is attributed to image perspective and illumination conditions that can make one color appear dominant (see Figure 9).

(a) Red **(b)** Black

Figure 9. Examples of misclassified multicolored vehicle from different viewpoints. The predicted color is shown below each image. Depending on the camera angle and illumination, a single color can appear dominant, which causes the classifier to classify the vehicle based on that one color.

For make recognition, the "others" class had the lowest performance ($\approx$ 30%). This suggests that a superclass for less common makes is ineffective, highlighting the need for out-of-distribution methods. Another common source of error was inter-manufacturer confusion for similar vehicle types; for example, Land-Rover which primarily sells SUVs in Brazil was often misclassified by more representative SUV manufacturers.

In vehicle model recognition, three primary sources of error were identified. First, a single vehicle platform is often sold in multiple body-style variants (e.g., sedan, hatch, compact pickup) that are visually similar, especially from the front. Second, some models are sold under different names by different manufacturers in Brazil (see Figure 10). Finally, a consistent design language across different models from the same make also contributed to errors.

(a) Peugeot Boxer **(b)** Fiat Ducato

Figure 10. Examples of misclassified vehicle models from different manufacturers. Original make and model are displayed below each image. In (a), a rear-view Boxer was misclassified as a Ducato, while in (b) a rear-view Ducato was misclassified as a Boxer. Aside from minor brand-specific markings, the vehicles have a very similar structure, making accurate differentiation challenging.

In type recognition, misclassifications were frequent between related types, such as scooter versus motorcycle and truck classes (i.e., compact-truck, tractor-truck, and truck). Another common error was the confusion of semi-trailers with trucks, as their images typically include the towing truck. Finally, compact-pickups and cars were also confused when derived from the same base vehicle platform.

The preceding analysis highlighted the difficulty of distinguishing similar vehicles under challenging real-world FGVC conditions. However, relying on a single prediction is often insufficient for practical applications, especially in public security (e.g., criminal investigation or forensics), where a limited set of high-probability candidates is more useful than a single classification. Therefore, the performance of EfficientNet-V2 was further analyzed using top-1, top-2, and top-3 predictions, as detailed in Table 4.

Table 4. Top-1, top-2, and top-3 performance metrics (%) for EfficientNet-V2 on vehicle color, make, model, and type recognition (UFPR-VeSV). Results are reported as the mean over 10 runs. Standard deviations are in parentheses.

Task	Top-1			Top-2			Top-3		
	Mi-Acc	Ma-Acc	F1	Mi-Acc	Ma-Acc	F1	Mi-Acc	Ma-Acc	F1
Color	93.5 (0.6)	71.5 (2.3)	73.8 (2.3)	98.0 (0.3)	87.0 (1.7)	89.4 (1.5)	99.1 (0.1)	93.5 (1.4)	95.0 (1.0)
Make	94.4 (0.6)	85.0 (1.6)	86.4 (1.4)	96.8 (0.4)	90.8 (1.3)	91.6 (1.3)	97.8 (0.3)	92.9 (1.3)	93.7 (1.2)
Model	90.9 (0.6)	86.2 (1.1)	87.3 (0.8)	95.3 (0.5)	91.5 (1.1)	92.3 (0.9)	96.7 (0.5)	93.8 (0.9)	94.5 (0.7)
Type	96.1 (0.7)	89.0 (2.0)	90.2 (1.6)	99.2 (0.2)	97.2 (1.2)	97.7 (0.8)	99.7 (0.1)	98.9 (0.8)	99.0 (0.6)

The improvement in top-k metrics demonstrates that the correct classification is frequently included within the top three candidates. This suggests the classifier is partially robust to the dataset's inherent ambiguities, a finding with practical implications for the deployment of real-world systems. However, substantial room for improvement remains, particularly in the macro-accuracy metrics for color, make, and model recognition.

A broader analysis of classification errors revealed that nighttime infrared images were a principal source of misclassification. Despite representing only 21.5% of the dataset, these images contributed to 53.4%, 42.9%, and 39.4% of total top-1 errors for make, model, and type recognition, respectively. The issue persisted even in the top-3 analysis, where this condition accounted for $\approx$ 60% of the remaining errors for

the same tasks.

In contrast to the clear negative impact of infrared conditions, viewpoint had a mixed effect on performance. The frontal view improved make recognition, likely due to the visibility of the manufacturer's badge. Conversely, this view was less effective for model and type recognition, as it offers fewer distinguishing features.

5.2 Joint-Task Analysis

The previous section established baselines by evaluating the color, make, model, and type recognition tasks in isolation. However, this information is semantically connected: a hierarchical relationship exists between make, model, and type, while color is an orthogonal property. Moreover, in practical surveillance applications, queries can range from a single attribute (e.g., "blue cars") to a complex, multi-attribute conjunction (e.g., "blue Ford models").

Therefore, this work also evaluates simultaneous accuracy to measure performance on these conjunctive tasks. Using the predictions from the previously trained EfficientNet-V2 classifiers (the best performing method), this metric is defined as the percentage of images for which all attributes within a specified set are correctly predicted. Table 5 illustrates how this metric changes as the set of required attributes expands. The expansion follows the logical "type-make-model" hierarchy, with the independent "color" attribute included as the final component.

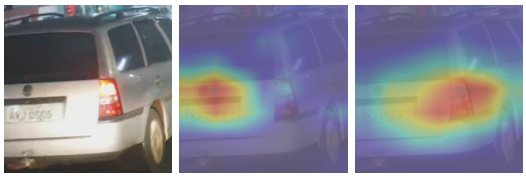
Table 5. EfficientNet-V2 simultaneous accuracy (%) for conjunctive attribute recognition, detailing the performance as the set of required attributes expands. Results are reported as mean (standard deviation in parentheses) over 10 runs.

Attribute set	Accuracy
Type	96.1 (0.7)
Type & Make	91.5 (1.0)
Type & Make & Model	85.5 (1.1)
Type & Make & Model & Color	80.2 (1.2)

As shown in Table 5, simultaneous accuracy predictably degrades as more attributes are required, a drop attributed to the compounding of individual error rates. This highlights a significant gap: while single-task accuracy can exceed 90% (Table 3, Section 5.1), the ability to produce a simultaneously correct vehicle description remains a challenge, underscoring the need for improved joint-task recognition.

Beyond this performance degradation, the isolated training/evaluation approach introduces a more critical issue: logically inconsistent predictions. The vehicle attributes are semantically bound; for example, the "Fiat Ducato" model inherently belongs to the "Fiat" make. Because the classifiers were trained independently, they are unaware of these dependencies. This leads to nonsensical outputs, such as 2.9% of predictions where the model was correctly identified, but its corresponding make was not.

Such inconsistencies represent a critical failure in a practical system. This issue is explained by the classifiers learning different, independent features for each task. We confirm this divergence by Grad-CAM [Selvaraju *et al*., 2017] attention map analysis, which shows that the make recognition focuses primarily on the manufacturer's badge, while the model recognition relies on other features, such as headlights (see an example in Figure 11). Thus, make recognition could fail if the badge is hidden or blurry, while model recognition can still succeed if its required features remain visible.



(a) Original **(b)** Make **(c)** Model

Figure 11. Grad-CAM [Selvaraju *et al*., 2017] attention maps for a Volkswagen Parati. (a) Original image. (b) The make recognition map, focusing on the manufacturer's badge. (c) The model recognition map, relying on other features, such as the headlights.

This analysis reveals a key limitation for practical FGVC: isolated models are insufficient, as they can produce logically inconsistent results. Methods must, therefore, enforce hierarchical consistency. Multi-task learning is a promising solution, as it would require the classifier to learn shared representations that explicitly encompass this hierarchy, which would reduce or eliminate such contradictions.

6 ALPR and FGVC Integration

This section evaluates ALPR and its integration with FGVC. These two tasks are crucial for vehicle identification but are often studied independently, missing their combined potential. First, Section 6.1 establishes an LPR baseline by assessing the performance of two state-of-the-art optical character recognition methods on the UFPR-VeSV dataset. Following this, Section 6.2 analyzes the potential of a joint system, quantifying how FGVC can serve as a support mechanism to enhance ALPR robustness.

6.1 License Plate Recognition

This experiment assessed the performance of two optical character recognition methods on the LPs extracted from the UFPR-VeSV images: GP-ALPR [Liu *et al*., 2024] and ParSeq-Tiny [Bautista and Atienza, 2022]. GP-ALPR is specifically designed to handle irregular LPs using deformable spatial attention and global perception modules. ParSeq-Tiny is a general scene text recognition method that combines context-free non-autoregressive and context-aware autoregressive inference through permutation language modeling.

These models were selected for their state-of-the-art performance, widespread adoption in prior research [Nascimento *et al*., 2024; Du *et al*., 2025], and public availability, which facilitates reproducibility. Both models were trained from scratch in accordance with the original methodologies proposed by their respective authors. Their performance was evaluated using two metrics: LP-level accuracy, which reflects the percentage of LPs correctly recognized in their entirety, and character-level accuracy, which measures the proportion of individual characters accurately identified.

Table 6 compares the LPR results for GP-ALPR and ParSeq-Tiny. The latter achieved the highest performance, with an

average LP-level accuracy of 98.0%. This high accuracy was an expected outcome; FGVC annotations were primarily based on official vehicle data retrieved using the LPs, thus ensuring that all LPs are human-recognizable in some form.

Table 6. Comparison of LP-level and character-level accuracy (%) achieved by the GP-ALPR and ParSeq-Tiny models on the UFPR-VeSV dataset for the LPR task. Results represent the average performance over 10 runs using different dataset splits, with standard deviations reported in parentheses.

Model	LP-level accuracy	Char-level accuracy
GP ALPR	93.8 (0.3)	98.7 (0.1)
ParSeq-Tiny	**98.0** (**0.4**)	**99.7** (**0.1**)

Despite the high accuracy, ParSeq-Tiny – the best-performing method – still failed in specific scenarios, as shown in Figure 12. The errors highlight persistent challenges, including highly degraded or low-contrast characters, illumination obstructions, excessive blurring, and physically deformed LPs. Low image quality led to confusion between structurally similar characters, such as "H" for "M" and "M" for "N". Finally, infrared images were a significant source of failure, accounting for 46.2% of all misrecognitions.

Figure 12. Example of misrecognized LPs. For each image, the ground-truth label is shown above the model's prediction, with incorrectly recognized characters highlighted in red. The failure cases include severely degraded characters (a, h, i), illumination-induced obstructions (b), low-contrast characters (c, j), blurring (d, g), and physically deformed LPs (e, f).

6.2 Joint System Analysis

The previous section showed that baseline LPR methods can produce errors that compromise vehicle identification. To address this, this section analyzes the potential for an FGVC system to function as a support mechanism to enhance ALPR robustness. This analysis establishes a clear path toward unified systems by showing their benefits and challenges.

To evaluate the joint system's performance, we established a clear methodology. First, we used the predictions from the best-performing methods identified in the previous sections: ParSeq-Tiny for LPR and EfficientNet-V2 for FGVC tasks. Second, we defined a "correct FGVC set." This term is used for metric computation and means that all attributes within a specific combination were predicted correctly for a given image. Based on this, we defined three metrics that reflect a practical application scenario.

- **Validation Rate** P(FGVC set correct | LPR correct): Measures how often the FGVC support system agrees with a correct LPR prediction.
- **Conflict Rate** P(FGVC set incorrect | LPR correct): Quantifies how often the FGVC support system predicts at least one attribute incorrectly, conflicting with the correct LPR.
- **Recovery Rate** P(FGVC set correct | LPR incorrect): Measures how often the FGVC support system correctly identifies all vehicle attributes, given the LPR fails.

The results in Table 7 reveal an expected trade-off. As more FGVC attributes are required, the Validation and Recovery Rates decrease while the Conflict Rate increases. This is a direct consequence of the cumulative error challenge identified in Section 5.2, amplified by our strict condition that a single attribute error fails the entire FGVC set. This methodology explains why the most demanding five-attribute combination yields the worst rates.

Table 7. Performance metrics of the integrated ALPR-FGVC system. All values are percentages (%). Validation and Conflict Rates are conditional on a correct LPR; Recovery Rate is conditional on an incorrect LPR. Results are reported as mean (standard deviation in parentheses) over 10 runs.

Tasks	Validation Rate ↑	Conflict Rate ↓	Recovery Rate ↑
LPR & Color	93.6 (0.5)	6.5 (0.5)	**92.7** (**0.7**)
LPR & Make	94.6 (0.5)	5.4 (0.5)	85.7 (0.8)
LPR & Model	91.1 (0.6)	8.9 (0.6)	82.1 (1.4)
LPR & Type	**96.2** (**0.6**)	**3.8** (**0.6**)	92.1 (0.9)
LPR & Color & Make	88.7 (0.7)	11.3 (0.7)	80.1 (0.9)
LPR & Color & Model	85.4 (0.8)	14.6 (0.8)	76.8 (1.4)
LPR & Color & Type	90.2 (0.7)	9.8 (0.7)	85.5 (0.9)
LPR & Make & Model	88.2 (0.8)	11.8 (0.8)	75.7 (1.3)
LPR & Make & Type	91.8 (0.9)	8.2 (0.9)	80.1 (1.2)
LPR & Model & Type	88.9 (1.0)	11.1 (1.0)	78.7 (1.5)
LPR & Color & Make & Model	82.8 (0.9)	17.2 (0.9)	71.1 (1.2)
LPR & Color & Make & Type	86.1 (1.0)	13.9 (1.0)	75.0 (1.2)
LPR & Color & Model & Type	83.4 (1.0)	16.6 (1.0)	73.5 (1.4)
LPR & Make & Model & Type	86.4 (1.1)	13.6 (1.1)	72.9 (1.3)
LPR & Color & Make & Model & Type	81.1 (1.1)	18.9 (1.1)	68.4 (0.7)

Despite the trade-off, the results demonstrate the system's powerful capability as a fail-safe. The Recovery Rate shows that even in the strictest case, the FGVC system correctly identified a complete vehicle description in 68.4% of LPR failures. This information is invaluable for practical use, allowing for cross-checking official records or flagging an LPR output as erroneous.

The results also highlight a trade-off between reliability and robustness. Relaxing the FGVC set to a single attribute yields high reliability; for instance, the "LPR + Type" configuration had the highest Validation Rate (96.2%) and the lowest Conflict Rate (3.8%). However, this approach lacks robustness. A single attribute like "Type" can be misleading, as a misrecognized LP might coincidentally match a database record for a different vehicle sharing the same type. In contrast, a multi-attribute set provides a more helpful descriptor (e.g., "Car, Fiat, Uno, White") that reduces the likelihood of such a false match. Therefore, while this multi-attribute check is currently less reliable, it represents the desirable goal for a unified system.

Furthermore, FGVC can also function as a fail-safe in scenarios where an ALPR-only system would fail completely. In real-world surveillance, factors like headlight glare or occlusion can render an LP illegible. In these cases, the FGVC system can still be used to analyze the vehicle image and provide valuable information. Figure 13 shows illustrative examples of such scenarios.

(a) Color unk. Fiat Palio **(b)** White Chevrolet Celta

Figure 13. Examples of ALPR failures where the LP is illegible due to (a) light glare and (b) occlusion. Despite these failures, our FGVC classifier acts as a fail-safe, correctly identifying the type, make, and model for both vehicles. Color recognition is correct for (b), while the infrared image (a) is classified as "unknown" due to the absence of color information.

Finally, we acknowledge this initial analysis has some limitations. First, our methodology relied on two simplifications: our "correct FGVC set" definition was a strict condition, and we used ground truth to identify errors. A practical system should be able to leverage partially correct information and must employ a conflict arbitration method (like confidence scores) to decide whether to trust the LPR or the FGVC outcomes. Second, our analysis is dataset-limited and we could not assess the "false match risk" – where an incorrect LPR might coincidentally retrieve a vehicle record that matches the FGVC output. Evaluating this risk requires access to an entire official vehicle database (including LP and attribute records) and remains a key direction for future research.

7 Conclusions

In this work, we introduce UFPR-VeSV, a public dataset for FGVC and ALPR research in surveillance scenarios. It contains annotations for 13 vehicle colors, 26 manufacturers, 136 models, and 14 types (all validated against official records), plus LP text labels and corner annotations. A quantitative and qualitative analysis confirms the dataset captures challenging real-world conditions. As the first of its kind, UFPR-VeSV supports both independent and combined research in FGVC and ALPR.

Benchmark experiments for vehicle color, make, model, and type recognition were conducted, with EfficientNet-V2 achieving the best overall performance. All recognition tasks achieved micro-accuracy scores above 90%. Despite this performance, an error analysis revealed remaining challenges that warrant further attention, including infrared images, multicolored vehicles, and distinguishing similar body-style model variants.

Furthermore, our analysis of joint FGVC tasks revealed that combining attributes reduces simultaneous recognition rates and leads to inconsistent predictions, such as a correct model with an incorrect make. A promising solution is to replace isolated classifiers with a multi-task learning [Caruana, 1997] framework. This approach can enforce hierarchical consistency and leverage natural correlations to enhance generalization across tasks.

We also benchmarked two methods for the LPR task. ParSeq-Tiny achieved the highest recognition rate, exceeding 98%. Nonetheless, error analysis revealed areas for improvement, particularly in handling distorted or physically deformed LPs. Building on this, we explored the combined use of FGVC and LPR, demonstrating that FGVC functions as fail-safe. Our results showed that even in the strictest case, the FGVC system provided a complete, correct vehicle description for $\approx 68\%$ of LPR failures.

Note that the UFPR-VeSV dataset's focus on high quality FGVC annotations – relying on official records retrieved from LP information – means it does not fully capture the challenges of unconstrained ALPR. In such scenarios, systems face significant legibility issues, as over 25% of images are inadequate for recognition [Wojcik *et al*., 2025]. This further highlights the importance of our proposed ALPR-FGVC integration, which can help validate correct recognitions or reject unreliable outputs. Both actions are crucial for reducing false positives.

Future research is needed to develop methods for weighing predictions from different recognition systems and assessing their reliability. A promising solution is selective prediction [Geifman and El-Yaniv, 2017], but this necessitates classifier calibration, as standard confidence scores are often overconfident [Guo *et al*., 2017; Minderer *et al*., 2021; Laroca *et al*., 2023b]. Extending these ideas to top-k settings also raises open questions, such as how to handle illogical Make–Model combinations in joint FGVC tasks or how to define top-k predictions in ALPR (e.g., at the character level or for entire sequences). A thorough exploration of these issues remains an important direction for future work.

Looking ahead, we plan to release a large-scale dataset inspired by UFPR-VeSV, comprising over a million surveillance images with greater diversity and more unconstrained conditions – especially for advancing ALPR research in such scenarios. This new dataset will likely feature semi-automatically generated annotations, highlighting the need to explore strategies for effectively training models with noisy or coarse labels [Lucio *et al*., 2019], while striking a balance between annotation efficiency and recognition performance.

Acknowledgments

This study was partially funded by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES)* – Finance Code 001, and by the *Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)* (# 315409/2023-1 and # 312565/2023-2). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Quadro RTX 8000 GPU, which was used in part of this research. We also extend our sincere thanks to students Leonardo Carpwiski, Rafael Marques, and Thiago Assunção for their assistance in labeling the LP texts in the UFPR-VeSV dataset.

Declarations

Authors' Contributions

Gabriel E. Lima is the primary contributor and writer of the manuscript. Valfride Nascimento assisted with experiment validation and participated in the manuscript review. Eduardo Santos contributed to data selection and was involved in the manuscript review process. Eduil Nascimento Jr. provided the image resources and contributed to the manuscript review. Rayson Laroca co-advised the project, supporting its conceptualization and methodology, and

contributed to the review and editing of the manuscript. David Menotti supervised the project, provided resources, and contributed to the review and editing of the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The resources generated and analyzed during this study are available at `https://github.com/Lima001/UFPR-VeSV-Dataset`.

References

Amirkhani, A. and Barshooi, A. H. (2023). Deepcar 5.0: Vehicle make and model recognition under challenging conditions. *IEEE Transactions on Intelligent Transportation Systems*, 24(1):541–553. doi:10.1109/TITS.2022.3212921.

Baek, N., Park, S.-M., Kim, K.-J., and Park, S.-B. (2007). Vehicle color classification based on the support vector machine method. In *International Conference on Intelligent Computing*, pages 1133–1139. doi:10.1007/978-3-540-74282-1_127.

Basak, S. and Suresh, S. (2024). Vehicle detection and type classification in low resolution congested traffic scenes using image super resolution. *Multimedia Tools and Applications*, 83(8):21825–21847. doi:10.1007/s11042-023-16337-2.

Bautista, D. and Atienza, R. (2022). Scene text recognition with permuted autoregressive sequence models. In *European Conference on Computer Vision (ECCV)*, pages 178–196. doi:10.1007/978-3-031-19815-1_11.

Caruana, R. (1997). Multitask learning. *Machine learning*, 28:41–75. doi:10.1023/A:1007379606734.

Celestino, M. (2021). 10 marcas que mais venderam carros na década. https://www.webmotors.com.br/wm1/noticias/10-marcas-que-mais-venderam-carros-na-decada. Accessed: 2025-02-19.

Chen, P., Bai, X., and Liu, W. (2014). Vehicle color recognition on urban road by feature context. *IEEE Transactions on Intelligent Transportation Systems*, 15(5):2340–2346. doi:10.1109/TITS.2014.2308897.

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. (2020). Randaugment: Practical automated data augmentation with a reduced search space. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3008–3017. doi:10.1109/CVPRW50498.2020.00359.

Deng, J., Guo, J., Ververas, E., Kotsia, I., and Zafeiriou, S. (2020). RetinaFace: Single-shot multi-level face localisation in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5202–5211. doi:10.1109/CVPR42600.2020.00525.

Deng, J., Krause, J., and Fei-Fei, L. (2013). Fine-grained crowdsourcing for fine-grained recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. doi:10.1109/CVPR.2013.81.

Dong, Z., Wu, Y., Pei, M., and Jia, Y. (2015). Vehicle type classification using a semisupervised convolutional neural network. *IEEE Transactions on Intelligent Transportation Systems*, 16(4):2247–2256. doi:10.1109/TITS.2015.2402438.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, pages 1–22.

Du, Y., Chen, Z., Su, Y., Jia, C., and Jiang, Y.-G. (2025). Instruction-guided scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–16. doi:10.1109/TPAMI.2025.3525526.

Dule, E., Gökmen, M., and Beratoğlu, M. S. (2010). A convenient feature vector construction for vehicle color recognition. In *WSEAS International Conference on Neural Networks, Evolutionary Computing and Fuzzy systems*, page 250–255. doi:10.5555/1863431.1863473.

Fan, X. and Zhao, W. (2022). Improving robustness of license plates automatic recognition in natural scenes. *IEEE Transactions on Intelligent Transportation Systems*, 23(10):18845–18854. doi:10.1109/TITS.2022.3151475.

Farias, V. and Croquer, G. (2023). Por que o carro colorido sumiu? 67% dos veículos no Brasil são brancos, pretos ou cinzas. https://g1.globo.com/economia/noticia/2023/08/20/por-que-o-carro-colorido-sumiu-67percent-dos-veiculos-no-brasil-sao-brancos-pretos-ou-cinzas.ghtml. Accessed: 2025-02-19.

Ferryman, J. M., Worrall, A. D., Sullivan, G. D., and Baker, K. D. (1995). A generic deformable model for vehicle recognition. In *British Machine Vision Conference (BMVC)*, page 127–136. doi:10.5555/236190.236202.

Fu, H., Ma, H., Wang, G., Zhang, X., and Zhang, Y. (2020). MCFF-CNN: Multiscale comprehensive feature fusion convolutional neural network for vehicle color recognition based on residual learning. *Neurocomputing*, 395:178–187. doi:10.1016/j.neucom.2018.02.111.

Geifman, Y. and El-Yaniv, R. (2017). Selective classification for deep neural networks. In *International Conference on Neural Information Processing Systems (NeurIPS)*, page 4885–4894. doi:10.5555/3295222.3295241.

Gonçalves, G. R., Diniz, M. A., Laroca, R., Menotti, D., and Schwartz, W. R. (2018). Real-time automatic license plate recognition through deep multi-task networks. In *Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 110–117. doi:10.1109/SIBGRAPI.2018.00021.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR. doi:10.5555/3305381.3305518.

Han, K., Xiao, A., Wu, E., Guo, J., XU, C., and Wang, Y. (2021). Transformer in transformer. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Sys-*

tems, volume 34, pages 15908–15919. Curran Associates, Inc. doi:10.5555/3540261.3541478.

Hassan, A., Ali, M., Durrani, N. M., and Tahir, M. A. (2021). An empirical analysis of deep learning architectures for vehicle make and model recognition. *IEEE Access*, 9:91487–91499. doi:10.1109/ACCESS.2021.3090766.

He, C., Wang, D., Cai, Z., Zeng, J., and Fu, F. (2024a). A vehicle matching algorithm by maximizing travel time probability based on automatic license plate recognition data. *IEEE Transactions on Intelligent Transportation Systems*, 25(8):9103–9114. doi:10.1109/TITS.2024.3358625.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. doi:10.1109/CVPR.2016.90.

He, L., Zhou, Y., Liu, L., and Ma, J. (2024b). Research and application of YOLOv11-based object segmentation in intelligent recognition at construction sites. *Buildings*, 14(12). doi:10.3390/buildings14123777.

Howard, A., Sandler, M., Chen, B., Wang, W., Chen, L.-C., Tan, M., Chu, G., Vasudevan, V., Zhu, Y., Pang, R., Adam, H., and Le, Q. (2019). Searching for MobileNetV3. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1314–1324. doi:10.1109/ICCV.2019.00140.

Hsu, G.-S., Chen, J.-C., and Chung, Y.-Z. (2013). Application-oriented license plate recognition. *IEEE Transactions on Vehicular Technology*, 62(2):552–561. doi:10.1109/TVT.2012.2226218.

Hu, B., Lai, J.-H., and Guo, C.-C. (2017). Location-aware fine-grained vehicle type recognition using multi-task deep networks. *Neurocomputing*, 243:60–68. doi:10.1016/j.neucom.2017.02.085.

Hu, C., Bai, X., Qi, L., Chen, P., Xue, G., and Mei, L. (2015). Vehicle color recognition with spatial pyramid deep learning. *IEEE Transactions on Intelligent Transportation Systems*, 16(5):2925–2934. doi:10.1109/TITS.2015.2430892.

Hu, M., Bai, L., Fan, J., Zhao, S., and Chen, E. (2023). Vehicle color recognition based on smooth modulation neural network with multi-scale feature fusion. *Frontiers of Computer Science*, 17(3):173321. doi:10.1007/s11704-022-1389-x.

Huang, C., Li, Y., Loy, C. C., and Tang, X. (2016). Learning deep representation for imbalanced classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5375–5384. doi:10.1109/CVPR.2016.580.

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. doi:10.1109/CVPR.2017.243.

Jolly, M.-P., Lakshmanan, S., and Jain, A. (1996). Vehicle segmentation and classification using deformable templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(3):293–308. doi:10.1109/34.485557.

Khanam, R. and Hussain, M. (2024). YOLOv11: An overview of the key architectural enhancements. *arXiv preprint*. doi:10.48550/arXiv.2410.17725.

Krause, J., Deng, J., Stark, M., and Fei-Fei, L. (2013a). Collecting a large-scale dataset of fine-grained cars. In *Second Workshop on Fine-Grained Visual Categorisation (FGVC), in conjunction with CVPR*. available at `https://ai.stanford.edu/~jkrause/papers/fgvc13.pdf`.

Krause, J., Stark, M., Deng, J., and Fei-Fei, L. (2013b). 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 554–561. doi:10.1109/ICCVW.2013.77.

Kuhn, D. M. and Moreira, V. P. (2021). BRCars: a dataset for fine-grained classification of car images. In *2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 231–238. doi:10.1109/SIBGRAPI54419.2021.00039.

Lai, A., Fung, G., and Yung, N. (2001). Vehicle type classification from visual-based dimension estimation. In *IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 201–206. doi:10.1109/ITSC.2001.948656.

Laroca, R., Araujo, A. B., Zanlorensi, L. A., De Almeida, E. C., and Menotti, D. (2021). Towards image-based automatic meter reading in unconstrained scenarios: A robust and efficient approach. *IEEE Access*, 9:67569–67584. doi:10.1109/ACCESS.2021.3077415.

Laroca, R., Cardoso, E. V., Lucio, D. R., Estevam, V., and Menotti, D. (2022). On the cross-dataset generalization in license plate recognition. In *International Conference on Computer Vision Theory and Applications (VISAPP)*, pages 166–178. doi:10.5220/0010846800003124.

Laroca, R., Estevam, V., Britto Jr., A. S., Minetto, R., and Menotti, D. (2023a). Do we train on test data? The impact of near-duplicates on license plate recognition. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. doi:10.1109/IJCNN54540.2023.10191584.

Laroca, R., Estevam, V., Moreira, G. J. P., Minetto, R., and Menotti, D. (2025). Advancing multinational license plate recognition through synthetic and real data fusion: A comprehensive evaluation. *IET Intelligent Transport Systems*, 19(1):e70086. doi:10.1049/itr2.70086.

Laroca, R., Severo, E., Zanlorensi, L. A., Oliveira, L. S., Gonçalves, G. R., Schwartz, W. R., and Menotti, D. (2018). A robust real-time automatic license plate recognition based on the YOLO detector. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–10. doi:10.1109/IJCNN.2018.8489629.

Laroca, R., Zanlorensi, L. A., Estevam, V., Minetto, R., and Menotti, D. (2023b). Leveraging model fusion for improved license plate recognition. In *Iberoamerican Congress on Pattern Recognition (CIARP)*, pages 60–75. doi:10.1007/978-3-031-49249-5_5.

Lima, G. E., Laroca, R., Santos, E., Nascimento Jr., E., and Menotti, D. (2024). Toward enhancing vehicle color recognition in adverse conditions: A dataset and benchmark. In *Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 1–6. doi:10.1109/SIBGRAPI62404.2024.10716307.

Liu, Q., Chen, S.-L., Chen, Y.-X., and Yin, X.-C. (2024). Improving license plate recognition via diverse stylistic plate generation. *Pattern Recognition Letters*, 183:117–124. doi:10.1016/j.patrec.2024.05.005.

Liu, Y.-Y., Liu, Q., Chen, S.-L., Chen, F., and Yin, X.-C. (2024). Irregular license plate recognition via global information integration. In *International Conference on Multimedia Modeling*, pages 325–339. doi:10.1007/978-3-031-53308-2_24.

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., Wei, F., and Guo, B. (2022). Swin transformer v2: Scaling up capacity and resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11999–12009. doi:10.1109/CVPR52688.2022.01170.

Lu, L., Cai, Y., Huang, H., and Wang, P. (2023). An efficient fine-grained vehicle recognition method based on part-level feature optimization. *Neurocomputing*, 536:40–49. doi:10.1016/j.neucom.2023.03.035.

Lucio, D. R., Laroca, R., Zanlorensi, L. A., Moreira, G., and Menotti, D. (2019). Simultaneous iris and periocular region detection using coarse annotations. In *Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 178–185. doi:10.1109/SIBGRAPI.2019.00032.

Luo, R., Song, Y., Ye, L., and Su, R. (2024). Dense-tnt: Efficient vehicle type classification neural network using satellite imagery. *Sensors*, 24(23). doi:10.3390/s24237662.

Ma, X. and Grimson, W. (2005). Edge-based rich representation for vehicle classification. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1185–1192. doi:10.1109/ICCV.2005.80.

Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., and Lucic, M. (2021). Revisiting the calibration of modern neural networks. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 15682–15694. Curran Associates, Inc. doi:10.5555/3540261.3541461.

Ministério dos Transportes (2024). Frota nacional (junho de 2024). https://www.gov.br/transportes/pt-br/assuntos/transito/conteudo-Senatran/frota-de-veiculos-2024. Accessed: 2025-02-19.

Nascimento, V., Laroca, R., Ribeiro, R. O., Schwartz, W. R., and Menotti, D. (2024). Enhancing license plate super-resolution: A layout-aware and character-driven approach. *Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 1–6. doi:10.1109/SIBGRAPI62404.2024.10716303.

Nascimento, V., Lima, G. E., Ribeiro, R. O., Schwartz, W. R., Laroca, R., and Menotti, D. (2025). Toward advancing license plate super-resolution in real-world scenarios: A dataset and benchmark. *Journal of the Brazilian Computer Society*, 1(31):435–449. doi:10.5753/jbcs.2025.5159.

Nyi Myo, N., Boonkong, A., Khampitak, K., and Hormdee, D. (2025). A two-point association tracking system incorporated with YOLOv11 for real-time visual tracking of laparoscopic surgical instruments. *IEEE Access*, 13:12225–12238. doi:10.1109/ACCESS.2025.3529710.

Ochal, M., Patacchiola, M., Vazquez, J., Storkey, A., and Wang, S. (2023). Few-shot learning with class imbalance. *IEEE Transactions on Artificial Intelligence*, 4(5):1348–1358. doi:10.1109/TAI.2023.3298303.

Oliveira, I. O., Laroca, R., Menotti, D., Fonseca, K. V. O., and Minetto, R. (2021). Vehicle-Rear: A new dataset to explore feature fusion for vehicle identification using convolutional neural networks. *IEEE Access*, 9:101065–101077. doi:10.1109/ACCESS.2021.3097964.

Rao, Z., Yang, D., Chen, N., and Liu, J. (2024). License plate recognition system in unconstrained scenes via a new image correction scheme and improved CRNN. *Expert Systems with Applications*, 243:122878. doi:10.1016/j.eswa.2023.122878.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626. doi:10.1109/ICCV.2017.74.

Shvai, N., Hasnat, A., Meicler, A., and Nakib, A. (2020). Accurate classification for automatic vehicle-type recognition based on ensemble classifiers. *IEEE Transactions on Intelligent Transportation Systems*, 21(3):1288–1297. doi:10.1109/TITS.2019.2906821.

Sochor, J., Herout, A., and Havel, J. (2016). BoxCars: 3D boxes as CNN input for improved fine-grained vehicle recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3006–3015. doi:10.1109/CVPR.2016.328.

Son, J.-W., Park, S.-B., and Kim, K.-J. (2007). A convolution kernel method for color recognition. In *International Conference on Advanced Language Processing and Web Information Technology*, pages 242–247. doi:10.1109/ALPIT.2007.28.

Tan, M. and Le, Q. (2021). EfficientNetV2: Smaller models and faster training. In *International Conf. on Machine Learning*, pages 10096–10106.

Ultralytics (2025). YOLOv11. https://docs.ultralytics.com/models/yolo11/. Accessed: 2025-03-04.

Wang, H., Peng, J., Zhao, Y., and Fu, X. (2020). Multi-path deep CNNs for fine-grained car recognition. *IEEE Transactions on Vehicular Technology*, 69(10):10484–10493. doi:10.1109/TVT.2020.3009162.

Wang, Y., Wang, C., Zheng, Y., Fu, H., and Ma, H. (2021). Transformer based neural network for fine-grained classification of vehicle color. In *International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 118–124. doi:10.1109/MIPR51284.2021.00025.

Wojcik, L., Lima, G. E., Nascimento, V., Nascimento Jr., E., Laroca, R., and Menotti, D. (2025). LPLC: A dataset for license plate legibility classification. *Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 1–6. doi:10.1109/SIBGRAPI67909.2025.11223367.

Wolf, S., Loran, D., and Beyerer, J. (2024). Knowledge-distillation-based label smoothing for fine-grained open-set vehicle recognition. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 330–340. doi:10.1109/WACVW60836.2024.00041.

Wu, W., QiSen, Z., and Mingjun, W. (2001). A method of vehicle classification using models and neural networks. In *IEEE Vehicular Technology Conference*, pages 3022–3026. doi:10.1109/VETECS.2001.944158.

Xu, Z., Yang, W., Meng, A., Lu, N., Huang, H., Ying, C., and Huang, L. (2018). Towards end-to-end license plate detection and recognition: A large dataset and baseline. In *European Conference on Computer Vision (ECCV)*. doi:10.1007/978-3-030-01261-8_16.

Yang, L., Luo, P., Loy, C. C., and Tang, X. (2015). A large-scale car dataset for fine-grained categorization and verification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3973–3981.

doi:10.1109/CVPR.2015.7299023.

Yu, Y., Liu, H., Fu, Y., Jia, W., Yu, J., and Yan, Z. (2022). Embedding pose information for multiview vehicle model recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8):5467–5480. doi:10.1109/TCSVT.2022.3151116.

Yuan, Y., Zou, W., Zhao, Y., Wang, X., Hu, X., and Komodakis, N. (2017). A robust and efficient approach to license plate detection. *IEEE Transactions on Image Processing*, 26(3):1102–1114. doi:10.1109/TIP.2016.2631901.

Zhang, L., Wang, P., Li, H., Li, Z., Shen, C., and Zhang, Y. (2021). A robust attentional framework for license plate recognition in the wild. *IEEE Transactions on Intelligent Transportation Systems*, 22(11):6967–6976. doi:10.1109/TITS.2020.3000072.

Zhang, Q., Zhuo, L., Li, J., Zhang, J., Zhang, H., and Li, X. (2018). Vehicle color recognition using multiple-layer feature representations of lightweight convolutional neural network. *Signal Processing*, 147:146–153. doi:10.1016/j.sigpro.2018.01.021.