Revisiting Vehicle Color Recognition in Long-Tailed Surveillance Scenarios

Vinícius Orrú[1,2], Bruno H. Foggiatto[1], Gabriel E. Lima[3], David Menotti[3], and Rayson Laroca[1,3,*]

[1] Pontifical Catholic University of Paraná, Curitiba, Brazil
[2] National High Court of Brazil, Brasília, Brazil
[3] Federal University of Paraná, Curitiba, Brazil

* rayson@ppgia.pucpr.br

Abstract. Vehicle color recognition is an important cue for vehicle identification in surveillance systems, especially when license plates are illegible due to low resolution, occlusion, motion blur, or poor illumination. However, real-world vehicle color distributions are highly imbalanced, making overall accuracy insufficient to assess performance on rare but operationally relevant colors. This paper presents a comprehensive study of vehicle color recognition under severe class imbalance using UFPR-VeSV, a challenging real-world surveillance dataset. We investigate synthetic minority-class augmentation through two off-the-shelf generative strategies: text-conditioned image generation with RunDiffusion/Juggernaut-XL and image-conditioned color editing with Gemini 2.0 Flash. The curated synthetic data are combined with modern visual representations, loss reweighting, learning-rate scheduling, color-safe augmentation, foreground-aware preprocessing, and ensemble fusion. The best-performing approach achieves 94.6% micro accuracy and 79.7% macro accuracy, improving macro accuracy by 8.2 percentage points over recent literature. A manual error analysis further shows that many remaining failures are visually ambiguous even for human annotators, highlighting the practical limits of color-based vehicle identification in unconstrained surveillance imagery. The generated images and source code are publicly available at https://github.com/viniciusorru/vcr-synthetic.

Keywords: Vehicle color recognition · Synthetic data · Class imbalance.

1 Introduction

Vehicle identification in camera networks is a central task in intelligent transportation systems, supporting traffic monitoring, vehicle search, and forensic investigation [9, 25, 43]. Although vehicle identification is often associated with Automatic License Plate Recognition (ALPR) [18], license plates may be illegible due to occlusion, motion blur, poor illumination, sensor degradation, or low resolution [23, 32, 42]. Complementary visual attributes extracted from the vehicle body are therefore important in practical scenarios [26, 28, 33]. Among them,

vehicle color is particularly useful because it is intuitive for human operators, covers a larger image region than the license plate, and can support searches even when the license plate cannot be recognized [4, 13, 27, 28].

Early Vehicle Color Recognition (VCR) methods relied mainly on handcrafted features, color-space analysis, and explicit vehicle-region detection [2, 4, 11]. More recently, deep learning methods have achieved strong results on controlled benchmarks [7, 12, 41]. However, real-world surveillance imagery remains challenging due to heterogeneous viewpoints, adverse weather, changing illumination, glare, motion blur, sensor noise, nighttime acquisition, and partial occlusions. These factors directly affect the visual evidence available for color recognition, motivating recent datasets and methods designed to evaluate VCR under more realistic conditions [13, 14, 27, 28].

A further difficulty is the severe class imbalance naturally found in vehicle color distributions. In real traffic, neutral colors such as white, black, silver, and gray are far more frequent than colors such as yellow, orange, green, beige, and brown [17,31]. As a consequence, a model may achieve high overall accuracy while still performing poorly on rare colors, since micro-level performance is dominated by frequent classes. This limitation is particularly relevant in surveillance and forensic applications, where uncommon colors can be highly informative for reducing the search space. Therefore, VCR should not be assessed only by overall accuracy, but also by macro-level metrics that give equal importance to each class and better expose failures on underrepresented colors.

Prior VCR studies have addressed class imbalance mainly through reweighted losses, balanced sampling, and conventional data augmentation [13,27]. Although these strategies are useful, they remain constrained by the diversity of the available real images. For minority classes, the training samples may cover only a limited range of vehicle types, viewpoints, illumination conditions, backgrounds, and acquisition artifacts. This motivates the use of generative models as a complementary way to increase visual support for rare colors. Recent studies have shown that diffusion-based synthetic images can improve image classification when generated and filtered appropriately [1, 39]. Nevertheless, in practical VCR, synthetic data should be evaluated as part of a broader recognition pipeline, together with imbalance-aware training, robust representations, suitable augmentation, and careful error analysis.

In this work, we study VCR on UFPR-VeSV [28], a recently introduced and highly diverse real-world vehicle surveillance dataset. Rather than presenting synthetic data generation as the main contribution by itself, we provide a comprehensive study of strategies for improving minority-class recognition in VCR. We investigate two complementary off-the-shelf generation strategies: text-conditioned generation with RunDiffusion/Juggernaut-XL [36] and image-conditioned editing with Gemini 2.0 Flash [19]. The generated images are manually filtered through an inter-rater quality assessment protocol and combined with several recognition strategies, including DINOv3 features [37], loss reweighting, color-safe augmentation, background segmentation, and ensemble fusion.

Our contributions are threefold. First, we provide an extensive evaluation of modern recognition strategies for VCR under severe class imbalance, focusing on macro-level performance rather than only on overall accuracy. Second, we build and evaluate two curated synthetic image sets designed to increase the visual support for minority color classes. Third, we conduct a detailed error analysis of the final predictions, distinguishing cases that appear correctable by the model from those that remain ambiguous even for human annotators. To the best of our knowledge, this is the first VCR study to analyze the remaining errors at this level of detail, providing a clearer view of both the practical value and the limitations of vehicle color as a surveillance cue.

The remainder of this paper is organized as follows. Section 2 presents UFPR-VeSV and the proposed synthetic minority-class augmentation process. Section 3 describes the experimental setup. Section 4 reports the results and error analysis. Finally, Section 5 concludes the paper and outlines future research directions.

2 UFPR-VeSV and Synthetic Augmentation

2.1 Reference Dataset and Class Imbalance

We use UFPR-VeSV [28], a real-world surveillance dataset for unified fine-grained vehicle classification and ALPR. It contains operational-camera images acquired under heterogeneous viewpoints, distances, illumination, weather, motion blur, nighttime infrared capture, and partial occlusions, making it suitable for studying VCR beyond near-ideal imagery.

The dataset comprises 24,945 images of 16,297 unique vehicles collected from the Military Police of Paraná, Brazil. Images are annotated with 13 color classes and other attributes not explored here. As shown in Fig. 1, the distribution is highly long-tailed: white contains 7,381 images, whereas brown contains only 34. The "unknown" class mainly corresponds to infrared nighttime images and motorcycle samples for which color cannot be reliably determined; it is kept in the recognition task but is not used as a target for synthetic augmentation.

Lima et al. [28] reported strong overall results on UFPR-VeSV, but the best color-recognition configuration reached only 71.5% macro accuracy. This gap motivates our analysis of whether synthetic images, imbalance-aware training, and modern visual representations can improve minority-class prediction.

2.2 Synthetic Image Generation Strategy

Our synthetic data are intended to increase the number and diversity of training samples for underrepresented colors, not to replace real surveillance images. We consider two complementary strategies: text-to-image generation, which can create diverse vehicles, scenes, viewpoints, and lighting conditions; and image-conditioned editing, which changes vehicle color while preserving much of the original structure and scene context. RunDiffusion/Juggernaut-XL follows the first strategy, using a photorealistic model derived from Stable Diffusion

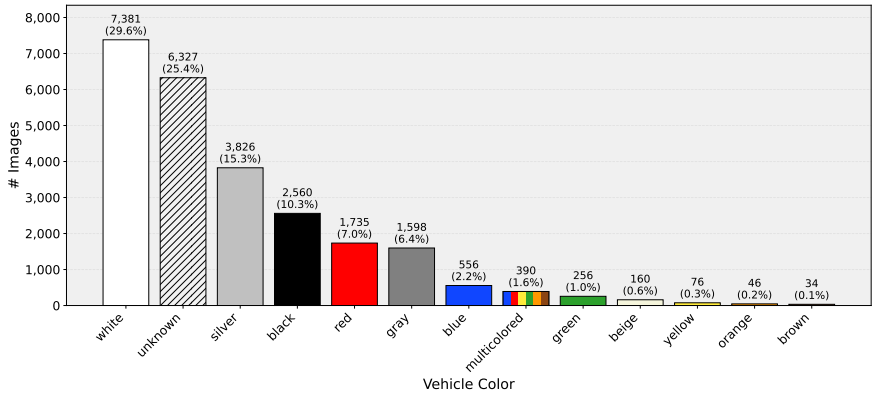

Fig. 1. Distribution of vehicle colors in UFPR-VeSV [28]. The strong imbalance motivates macro-level evaluation and synthetic data focused on minority classes.

XL [34, 36], while Gemini 2.0 Flash image generation supports multimodal image editing from image and text inputs [19]. Both models are used off the shelf, without fine-tuning.

2.3 Text-to-Image Generation with RunDiffusion/Juggernaut-XL

The RunDiffusion subset was generated with the Juggernaut-XL checkpoint [36]. For each image, the script sampled a target color by favoring underrepresented classes in the original training distribution and reducing the probability of classes already frequent among the generated candidates. It also selected a common Brazilian vehicle model with inverse-frequency weighting. Prompts combined vehicle make/model, target color, location, environmental condition, viewpoint, and distance, including Brazilian roads and urban scenes under sunny, cloudy, rainy, nighttime, sunset, and foggy conditions. For the multicolored class, the prompt explicitly required vehicles with multiple colors.

After generation, YOLOv11 (small) [40] was used to detect the main vehicle. Only the largest vehicle-related bounding box was stored with the sample metadata. Fig. 2 shows representative examples, illustrating the intended diversity of vehicles, viewpoints, and scenes for minority colors.

In total, 19,144 RunDiffusion images were generated. After quality control (see Section 2.5) and duplicate removal, 12,951 images were retained.

2.4 Image-Conditioned Color Editing with Gemini 2.0 Flash

Gemini receives an existing image from UFPR-VeSV along with a textual instruction, generating a modified version that changes vehicle body color while preserving the camera geometry, vehicle pose, background, and image quality. For each selected source image and target color, the prompt asked Gemini to change the entire vehicle body to the target color, avoid vibrant tones, keep realistic vehicle colors, and preserve a visible front or rear view. Due to API rate

Fig. 2. Representative images generated with RunDiffusion/Juggernaut-XL [36]. Samples are arranged from left to right and top to bottom as blue, black, orange, silver, red, gray, green, white, beige, yellow, brown, and multicolored.

limits, 14,040 Gemini images were generated, of which 6,028 were accepted after quality control. Fig. 3 illustrates the image-conditioned nature of this strategy.

Fig. 3. Representative images generated with Gemini. Each pair shows the original image from UFPR-VeSV [28] on the left and the generated color variant on the right.

2.5 Manual Quality Control and Inter-Rater Agreement

All synthetic images were evaluated by two independent annotators before training. The criteria were visual realism, correctness of the target color, and consistency with the vehicle surveillance domain. For RunDiffusion, annotators accepted realistic, non-deformed vehicles with the intended color; for Gemini, edited samples also had to preserve plausible vehicle appearance without visible editing artifacts. Only images accepted by both annotators were retained.

Table 1 reports the observed agreement (P_o), i.e., the proportion of cases where both annotators agreed on the same label, alongside Cohen's kappa (κ) [5],

which adjusts for the agreement expected by chance alone. The resulting κ values fall in the moderate range of the Landis-Koch scale [21], which is consistent with the inherent ambiguity of borderline color categories such as gray $\leftrightarrow$ silver, beige $\leftrightarrow$ brown, and colors captured under low-light conditions.

Table 1. Inter-rater agreement for synthetic image evaluation.

Source	N	P_o	κ	Accepted
RunDiffusion/Juggernaut-XL	19,144	0.80	0.43	12,951 (67.7%)
Gemini 2.0 Flash	14,040	0.71	0.41	6,028 (42.9%)

2.6 Final Augmented Dataset Composition

Table 2 reports the number of original and accepted synthetic images per color class. The synthetic sets mainly target minority colors, especially blue, green, yellow, orange, and brown. The final distribution also reflects the quality-control stage: some minority classes, particularly multicolored and beige, received fewer accepted samples than expected. Multicolored vehicles were often generated with a single dominant color or patterns that were difficult to label consistently, while beige frequently drifted toward visually similar classes such as white, silver, gray, or brown. These borderline cases likely reflect low visual separability, biases in the generative models' training data, and increased annotator disagreement. Majority classes are augmented less aggressively, and the "unknown" class is not augmented because its label indicates unreliable color evidence.

Table 2. Color class distribution in the original UFPR-VeSV dataset and in the accepted synthetic subsets. The "unknown" class corresponds mainly to infrared nighttime images and motorcycle samples for which color cannot be reliably determined.

Color	UFPR-VeSV	RunDiffusion	Gemini	Total	
White	7,381	173	581	8,135	(+10.2%)
Unknown	6,327	–	–	6,327	(+ 0.0%)
Silver	3,826	159	267	4,252	(+11.1%)
Black	2,560	170	488	3,218	(+25.7%)
Red	1,735	428	886	3,049	(+75.7%)
Gray	1,598	640	163	2,401	(+50.3%)
Blue	556	2,001	879	3,436	(+518.0%)
Multicolored	390	640	438	1,468	(+276.4%)
Green	256	2,135	718	3,109	(+1,114.5%)
Beige	160	365	123	648	(+305.0%)
Yellow	76	1,946	565	2,587	(+3,303.9%)
Orange	46	2,129	683	2,858	(+6,113.0%)
Brown	34	2,165	237	2,436	(+7,064.7%)
Total	**24,945**	**12,951**	**6,028**	**43,924**	**(+76.1%)**

3 Experimental Setup

3.1 Evaluation Protocol

We follow the stratified 3:1:1 train–validation–test protocol of Lima et al. [28], using ten independent splits generated with the script released by the authors. This preserves the evaluation setup and enables direct comparison with the original UFPR-VeSV benchmark. All results are reported exclusively on real test images; synthetic images are reserved for training only. RunDiffusion/Juggernaut-XL images are appended directly to each training split, while Gemini-generated images are included only when their source image belongs to the corresponding training partition, thereby preventing source-level data leakage [22, 24].

The experiments consider four training-set configurations: the original UFPR-VeSV training data, UFPR-VeSV + RunDiffusion, UFPR-VeSV + Gemini, and UFPR-VeSV + Gemini + RunDiffusion. Following prior work on VCR [27, 28], we report micro accuracy (Mi-Acc), macro accuracy (Ma-Acc), and macro F1-score (F1), averaged over the ten splits. Among these, we focus primarily on macro accuracy, as it directly reflects per-class performance and is straightforward to interpret in the context of long-tailed distributions.

3.2 Recognition Models and Experimental Design

To avoid tailoring conclusions to a single architecture, we evaluate four supervised baselines used in recent VCR studies [27, 28]: EfficientNet-V2 [38], ResNet-50 [10], Swin Transformer Tiny [29], and ViT-B/16 [6], all initialized with ImageNet-pretrained weights.

We also evaluate DINOv3 [37] as a frozen self-supervised feature extractor, motivated by its transferable representations under limited supervision. Following prior evidence that frozen DINOv2 features can be effective when the target domain remains close to natural imagery [16], we test the Small (ViT-S/16, 21M parameters), Base (ViT-B/16, 86M), and Large (ViT-L/16, 300M) variants. For each model, the frozen `[CLS]` token feeds a two-layer Multi-Layer Perceptron (MLP) head with a 512-dimensional projection, batch normalization, ReLU, dropout ($p = 0.4$), and a final classifier.

3.3 Optimization and Class-Imbalance Handling

All models are optimized with Adam [20] and trained for up to 100 epochs with early stopping. We compare standard Cross-Entropy (CE) with Weighted Cross-Entropy (WCE), a common strategy for imbalanced recognition [15, 44]. We also evaluate three learning-rate schedules: Cosine Decay (CD) [30], Linear Warmup with Cosine Decay (LWCD), and ReduceLROnPlateau. For LWCD, the learning rate increases from $\eta_{\min} = 10^{-6}$ to $\eta_{\max} = 10^{-4}$ during the first five epochs and then follows CD [8]. DINOv3 experiments use batch size 32, learning rate 10^{-3}, and weight decay 10^{-5} due to GPU memory constraints. All experiments were conducted on an NVIDIA RTX 5090 GPU with 32 GB of memory.

3.4 Color-Safe Augmentation and Foreground Preprocessing

Data augmentation must be especially conservative when color is the target attribute. We apply an Albumentations [3] pipeline with probability 0.8, randomly sampling three transformations from affine transformations ($\pm 15^\circ$ rotation and scale 0.9–1.1), horizontal flip, Gaussian noise, blur, brightness/contrast variation ($\pm 20\%$), and conservative HSV perturbations. To avoid corrupting the color signal, hue, saturation, and value shifts are limited to ± 5 levels ($\approx 2.8\%$ of the OpenCV range), ± 20 levels ($\approx 7.8\%$), and ± 10 levels ($\approx 3.9\%$ of the 8-bit range), respectively. We also carefully inspected augmented samples throughout this process to ensure that the perceived color was preserved.

We also evaluate a foreground-aware preprocessing strategy based on SAM 2 [35]. In this setting, the vehicle region is preserved while the background is blurred, as illustrated in Fig. 4. This experiment assesses whether emphasizing vehicle paint helps reduce the model's dependence on contextual cues such as the scene background, shadows, and road surface.

Fig. 4. Examples of foreground-aware preprocessing with SAM 2 [35].

3.5 Model Fusion

Finally, we evaluate whether complementary models improve robustness on minority classes. For each split, the top three candidates are selected based on validation macro accuracy and then combined by hard voting. When the three models disagree, the prediction with the highest class confidence is used only as a tiebreaker. Importantly, we also evaluated a confidence-based fusion rule, which directly selects the prediction with the highest class confidence, but it performed slightly below hard voting and the difference was not statistically significant. Therefore, we focus on hard voting in the remaining analysis.

4 Results and Discussion

4.1 Overall Performance

Table 3 compares the performance of the previous UFPR-VeSV benchmark [28] against our baselines and the final proposed approach. Among the single models trained exclusively on the original data, DINOv3-Large achieves the most favorable balance between overall and class-specific performance, reaching 94.0%

micro-accuracy and 72.5% macro-accuracy. While this represents a 1.0 percentage point (pp) improvement in macro-accuracy over the results reported by Lima et al. [28], this specific gain is not statistically significant ($p > 0.05$).

Table 3. Baseline results on the original UFPR-VeSV dataset and performance of the proposed approach. The baselines use CE loss, CD, and no data augmentation. The proposed approach combines Gemini-augmented training data, WCE, LWCD, color-safe augmentation, foreground-aware preprocessing, and hard-voting ensemble fusion.

Model	Mi-Acc (%)	Ma-Acc (%)	F1 (%)
EfficientNet-V2 (from [28])	93.5 ± 0.6	71.5 ± 2.3	73.8 ± 2.3
ResNet-50	91.2 ± 0.7	64.1 ± 2.3	66.4 ± 2.2
EfficientNet-V2	90.9 ± 1.0	66.7 ± 2.5	68.2 ± 1.9
ViT-B/16	92.8 ± 0.6	69.9 ± 3.4	72.5 ± 2.8
Swin-T	93.2 ± 0.6	72.5 ± 2.9	74.0 ± 2.8
DINOv3-Small	92.5 ± 0.6	69.1 ± 3.1	71.2 ± 2.2
DINOv3-Base	93.0 ± 0.4	69.6 ± 3.4	72.3 ± 3.1
DINOv3-Large	94.0 ± 0.6	72.5 ± 3.3	74.4 ± 2.5
Proposed approach	**94.6 ± 0.4**	**79.7 ± 4.0**	**78.8 ± 3.1**

In contrast, our full proposed pipeline reaches 79.7% macro-accuracy—a significant improvement of 7.2 pp over our DINOv3-Large baseline and 8.2 pp over the previous benchmark ($p < 0.05$). As detailed in Section 4.3, these gains are cumulative, resulting from the integration of Gemini-augmented data, loss reweighting, LWCD, color-safe augmentation, foreground-aware preprocessing, and hard-voting ensemble fusion. Notably, while the increase in micro-accuracy is more modest (from 93.5% to 94.6%), it remains statistically significant ($p < 0.05$). This suggests that our methodology primarily enhances performance on minority classes and visually challenging colors, improvements that are more effectively captured by macro-level metrics.

4.2 Effect of Synthetic Data and Loss Reweighting

Table 4 isolates the effect of training-set composition under CE and WCE, using Cosine Decay (CD) without data augmentation. To keep the analysis concise, we report the best result obtained for each configuration; the same trend holds across model architectures.

Under CE, the differences are modest: the original dataset remains competitive, while Gemini achieves the highest macro accuracy, at 74.9%. This suggests that synthetic data alone brings limited gains when the objective is dominated by frequent classes. The effect is clearer under WCE. On the original long-tailed data, reweighting is unstable and reduces macro accuracy to 59.3%. With synthetic images, WCE becomes effective again, reaching 71.5% with RunDiffusion, 75.7% with Gemini, and 71.9% with the combined set. Gemini is strongest despite adding fewer images than RunDiffusion, suggesting that domain alignment

Table 4. Test results for each training-set composition under fixed CE and WCE settings, using CD, no data augmentation, and the 10-fold average. Rows within each loss setting are ordered by macro accuracy (Ma-Acc).

Training configuration	Mi-Acc (%)	Ma-Acc (%)	F1 (%)
Cross-Entropy (CE)			
UFPR-VeSV	94.0 ± 0.5	72.5 ± 3.2	74.4 ± 2.5
UFPR-VeSV + RunDiffusion	93.5 ± 0.5	72.5 ± 4.6	74.6 ± 4.4
UFPR-VeSV + Gemini	**93.8 ± 0.3**	**74.9 ± 4.7**	**76.4 ± 3.5**
UFPR-VeSV + Gemini + RunDiffusion	93.1 ± 0.7	72.8 ± 3.8	74.1 ± 2.8
Weighted Cross-Entropy (WCE)			
UFPR-VeSV	36.4 ± 0.3	59.3 ± 3.0	55.8 ± 3.3
UFPR-VeSV + RunDiffusion	91.5 ± 0.4	71.5 ± 5.6	72.7 ± 5.4
UFPR-VeSV + Gemini	**92.7 ± 0.9**	**75.7 ± 3.0**	**75.0 ± 2.7**
UFPR-VeSV + Gemini + RunDiffusion	92.6 ± 0.1	71.9 ± 4.4	73.5 ± 3.5

is more important than volume alone. The lower result of the combined set also indicates that adding more synthetic samples is not automatically beneficial when part of the data shifts the training distribution.

4.3 Cumulative Ablation

Table 5 summarizes the best-performing path from the original DINOv3-Large baseline to the final ensemble. The rows are a practical progression rather than isolated factorial effects, since the best model may change after each modification. Still, each added component improves macro accuracy.

Table 5. Best-performing path from the original baseline to the final ensemble. The last column reports the macro-accuracy gain relative to the previous row.

Approach	Configuration	Best model	Mi-Acc (%)	Ma-Acc (%)	F1 (%)	ΔMa
Lima et al. [28]	Original	EfficientNet-V2	93.5 ± 0.6	71.5 ± 2.3	73.8 ± 2.3	–
Our baseline	Original + CD	DINOv3-Large	94.0 ± 0.5	72.5 ± 3.2	74.4 ± 2.5	–
+ Gemini data	Gemini + CD	DINOv3-Large	93.8 ± 0.3	74.9 ± 4.7	76.4 ± 3.5	+2.4
+ Loss reweighting	Gemini + WCE + CD	Swin-T	92.7 ± 0.9	75.7 ± 3.0	75.0 ± 2.7	+0.8
+ Scheduler	Gemini + WCE + LWCD	DINOv3-Large	93.2 ± 0.8	76.7 ± 4.3	75.3 ± 3.0	+1.0
+ Color-safe aug.	+ augmentation	DINOv3-Large	92.9 ± 0.5	77.3 ± 4.0	74.7 ± 2.6	+0.6
+ Segmented bg.	+ background blur	DINOv3-Large	93.0 ± 0.8	78.1 ± 4.6	76.0 ± 3.0	+0.8
+ Ensemble	hard voting	Ensemble	**94.6** ± 0.4	**79.7** ± 4.0	**78.8** ± 3.1	**+1.6**

The largest single-model gain comes from Gemini-augmented data, which raises macro accuracy from 72.5% to 74.9%. Loss reweighting and LWCD then increase the result to 76.7%, showing that synthetic data and imbalance-aware optimization are complementary. Color-safe augmentation and foreground-aware preprocessing provide smaller gains, leading to the best single model at 78.1% macro accuracy. Finally, hard voting among the three strongest DINOv3-Large variants yields 79.7% macro accuracy and 78.8% macro F1.

The selected ensemble combines three DINOv3-Large variants trained with Gemini data, WCE, and LWCD: one without augmentation, one with color-safe augmentation, and one with both color-safe augmentation and segmented-background training. We also evaluated more heterogeneous ensembles combining different architectures, such as Swin-T and ViT-B/16, but these alternatives led to lower macro accuracy. Therefore, the final composition is deliberately conservative: its members share the same backbone and synthetic source, but differ in visual invariance and foreground focus.

4.4 Error Analysis

To better understand the remaining failures, we manually analyzed all errors produced by the final hard-voting ensemble on three representative folds, selected according to the worst, median, and best macro accuracy. Each error was independently labeled by two annotators as a *model error*, when the correct color could be inferred from the image, or an *inherently ambiguous* case, when visual evidence was insufficient for reliable color assignment. Across these folds, the ensemble produced 785 errors. Considering only annotator consensus labels, 459 errors (58.5%) were considered ambiguous and 326 (41.5%) were considered model errors, with observed agreement $P_o = 73.63\%$ and Cohen's $\kappa = 0.35$.

These categories help contextualize the remaining gap after the proposed approach reaches nearly 80% macro accuracy in a highly long-tailed surveillance setting. Model errors suggest room for improvement, whereas ambiguous errors expose a more fundamental limitation of surveillance VCR. Such cases include infrared nighttime capture, severe underexposure, reflections, occlusion, and borderline hues such as gray $\leftrightarrow$ silver, silver $\leftrightarrow$ white, or beige $\leftrightarrow$ brown. Remarkably, under this criterion, the estimated upper bound is about 96.8% micro accuracy and 88.5% macro accuracy, indicating that the proposed result already closes a substantial part of the achievable gap and that many remaining failures are tied to the visual limits of the task.

Correctable model errors are concentrated in gray $\rightarrow$ black and silver $\rightarrow$ white predictions, often associated with low contrast, shadows, and highlights. Ambiguous errors are dominated by multicolored $\rightarrow$ white and gray $\leftrightarrow$ silver cases. Importantly, the ground-truth labels (except for the unknown class, which was assigned by the authors of [28] to samples whose color could not be reliably determined) come from official vehicle documentation rather than from visual annotation [28], so these cases should not be interpreted as annotation errors. Instead, they often reveal a mismatch between the documented color and the color evidence available in the surveillance image. For instance, multicolored $\rightarrow$ white errors commonly occur when only a small vehicle region supports the documented multicolored label, while gray $\leftrightarrow$ silver cases reflect a weak visual boundary between neutral colors under changing illumination. Fig. 5 illustrates these ambiguities, showing that real-world VCR is limited not only by model capacity but also by the reliability of color evidence in the image.

Fig. 5. Examples of ambiguous misclassifications. In several cases, the ground-truth label (GT) is difficult to confirm visually because of low illumination, reflections, partial views, or color evidence restricted to a small vehicle region.

5 Conclusions

This paper presented a comprehensive study of vehicle color recognition under severe class imbalance in real-world surveillance imagery. Using UFPR-VeSV [28] as a challenging benchmark, we evaluated synthetic minority-class augmentation together with modern visual representations, imbalance-aware training, learning-rate scheduling, color-safe augmentation, foreground-aware preprocessing, and ensemble fusion. The results show that synthetic data is most effective when it is combined with an appropriate training objective: while standard cross-entropy yielded only limited gains, the addition of curated synthetic images made weighted cross-entropy substantially more stable and effective.

Among the evaluated synthetic sources, Gemini 2.0 Flash yielded the greatest performance gains despite contributing fewer images than RunDiffusion/Juggernaut-XL, suggesting that domain alignment and preservation of the surveillance context matter more than synthetic data volume alone. By combining Gemini-augmented training, DINOv3-Large representations, weighted loss, linear warmup with cosine decay, color-safe augmentation, foreground-aware preprocessing, and hard-voting ensemble fusion, the proposed system achieves a macro accuracy of 79.70%, outperforming the previous state-of-the-art by 8.20 percentage points [28].

The manual error analysis further highlights an important limitation of real-world VCR: a large portion of the remaining mistakes are not simply model failures, but visually ambiguous cases that are difficult even for human annotators. Such errors are often caused by low illumination, infrared nighttime capture, reflections, occlusions, and weak boundaries between visually similar colors such as gray, silver, white, beige, and brown. Therefore, although the proposed pipeline substantially improves class-balanced recognition, the results also show that the practical limits of VCR depend not only on model capacity, but also on the amount and reliability of color evidence available in the image.

Future work will investigate synthetic augmentation in multi-task settings that jointly exploit color, type, make, model, and other vehicle attributes. We also plan to explore uncertainty-aware predictions and human-in-the-loop protocols, which may be especially useful when the visual evidence is insufficient for assigning a single reliable color label.

Ethics and Privacy Considerations The synthetic images used in this study were generated from textual prompts or by editing images from UFPR-VeSV [28]. Gemini-based samples may preserve scene elements, including visible license plates and people. Therefore, these images will be distributed and used under the same access, distribution, and research-purpose restrictions applied to the original dataset. We have already obtained authorization from the authors of UFPR-VeSV to distribute and use the synthetic images under these conditions.

Acknowledgments This study was supported in part by *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brasil*, under *Programa de Excelência Acadêmica (PROEX)* – Finance Code 001; by *Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)* (# 315409/2023-1); and by *Fundação Araucária* (# 078/2026).

References

1. Azizi, S., Kornblith, S., Saharia, C., Norouzi, M., Fleet, D.J.: Synthetic data from diffusion models improves ImageNet classification. Transactions on Machine Learning Research (2023), https://openreview.net/forum?id=DlRsoxjyPm
2. Baek, N., Park, S.M., Kim, K.J., Park, S.B.: Vehicle color classification based on the support vector machine method. In: International Conference on Intelligent Computing. pp. 1133–1139 (2007)
3. Buslaev, A., et al.: Albumentations: Fast and flexible image augmentations. Information **11**(2), 125 (2020)
4. Chen, P., Bai, X., Liu, W.: Vehicle color recognition on urban road by feature context. IEEE Transactions on Intelligent Transportation Systems **15**(5), 2340–2346 (2014)
5. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement **20**(1), 37–46 (1960)
6. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR). pp. 1–22 (2021)
7. Fu, H., Ma, H., Wang, G., Zhang, X., Zhang, Y.: MCFF-CNN: Multiscale comprehensive feature fusion convolutional neural network for vehicle color recognition based on residual learning. Neurocomputing **395**, 178–187 (2020)
8. Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., He, K.: Accurate, large minibatch SGD: Training ImageNet in 1 hour. arXiv preprint arXiv:1706.02677 (2017)
9. He, C., et al.: 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 (2024)

10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
11. Hsieh, J.W., Chen, L.C., Chen, S.Y., Chen, D.Y., Alghyaline, S., Chiang, H.F.: Vehicle color classification under different lighting conditions through color correction. IEEE Sensors Journal **15**(2), 971–983 (2015)
12. Hu, C., Bai, X., Qi, L., Chen, P., Xue, G., Mei, L.: Vehicle color recognition with spatial pyramid deep learning. IEEE Transactions on Intelligent Transportation Systems **16**(5), 2925–2934 (2015)
13. Hu, M., Bai, L., Fan, J., Zhao, S., Chen, E.: Vehicle color recognition based on smooth modulation neural network with multi-scale feature fusion. Frontiers of Computer Science **17**(3), 173321 (2023)
14. Hu, M., Wang, C., Yang, J., Wu, Y., Fan, J., Jing, B.: Rain rendering and construction of rain Vehicle Color-24 dataset. Mathematics **10**(17) (2022)
15. Huang, C., Li, Y., Loy, C.C., Tang, X.: Learning deep representation for imbalanced classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5375–5384 (2016)
16. Huang, Y., et al.: Comparative analysis of imagenet pre-trained deep learning models and dinov2 in medical imaging classification. In: IEEE Annual Computers, Software, and Applications Conference (COMPSAC). pp. 297–305 (2024)
17. iSeeCars: Cars are half as colorful as they were 20 years ago. https://www.iseecars.com/most-popular-car-colors-study (2024), accessed: 2026-04-21
18. Ismail, A., Mehri, M., Sahbani, A., Essoukri Ben Amara, N.: Automatic license plate recognition in in-the-wild scenarios: A comprehensive review, open issues, and future directions. IEEE Access **13**, 145387–145415 (2025)
19. Kampf, K., Brichtova, N.: Experiment with Gemini 2.0 Flash native image generation. Google Developers Blog (2025), `https://developers.googleblog.com/experiment-with-gemini-20-flash-native-image-generation/`, accessed: 2026-04-26
20. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint (2015), accessed: 2025-03-11
21. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159–174 (1977)
22. Laroca, R., Estevam, V., Britto Jr., A.S., Minetto, R., Menotti, D.: Do we train on test data? The impact of near-duplicates on license plate recognition. In: International Joint Conference on Neural Networks (IJCNN). pp. 1–8 (June 2023)
23. Laroca, R., Nascimento, V., Kim, D., Chung, S., Bae, S., Seo, U., Oh, S., Phung, C.M., Vo, M.G., Ye, X., Du, Y., Su, Y., Chen, Z., Heo, S., Lee, H., Na, K., Vu Nguyen, K.V., Pham, S.T., Phung, D.N.N., Le, T.P., Vo Tran, V.N., Menotti, D.: ICPR 2026 Competition on low-resolution license plate recognition. In: International Conference on Pattern Recognition (ICPR). pp. 1–20 (Aug 2026)
24. Laroca, R., Santos, M., Estevam, V., Luz, E., Menotti, D.: A first look at dataset bias in license plate recognition. In: Conference on Graphics, Patterns and Images (SIBGRAPI). pp. 234–239 (Oct 2022)
25. Laroca, R., Estevam, V., Moreira, G.J.P., Minetto, R., Menotti, D.: Advancing multinational license plate recognition through synthetic and real data fusion: A comprehensive evaluation. IET Intelligent Transport Systems **19**(1), e70086 (2025)
26. Li, L., Zang, H., Fan, X., Cheng, H., Dong, Y.: Weakly supervised bilinear convolutional neural network for fine-grained vehicle classification. IEEE Transactions on Intelligent Transportation Systems **26**(12), 22470–22481 (2025)

27. Lima, G.E., Laroca, R., Santos, E., Nascimento Jr., E., Menotti, D.: Toward enhancing vehicle color recognition in adverse conditions: A dataset and benchmark. In: Conference on Graphics, Patterns and Images. pp. 1–6 (Sept 2024)
28. Lima, G.E., Nascimento, V., Santos, E., Nascimento Jr., E., Laroca, R., Menotti, D.: Toward unified fine-grained vehicle classification and automatic license plate recognition. Journal of the Brazilian Computer Society **32**(1), 783–799 (2026)
29. Liu, Z., et al.: Swin transformer V2: Scaling up capacity and resolution. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11999–12009 (2022)
30. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts (2017), https://arxiv.org/abs/1608.03983
31. Ministeri, D.: Shades of success: Europe's favourite car colours revealed. https://www.jato.com/resources/news-and-insights/shades-success-europes-favourite-car-colours-revealed (2025), accessed: 2026-04-21
32. Nascimento, V., Laroca, R., Ribeiro, R.O., Schwartz, W.R., Menotti, D.: Enhancing license plate super-resolution: A layout-aware and character-driven approach. Conference on Graphics, Patterns and Images (SIBGRAPI) pp. 1–6 (2024)
33. Ni, X., Huttunen, H.: Vehicle attribute recognition by appearance: Computer vision methods for vehicle type, make and model classification. Journal of Signal Processing Systems **93**(4), 357–368 (2021)
34. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: Improving latent diffusion models for high-resolution image synthesis. In: International Conference on Learning Representations (2024), https://openreview.net/forum?id=di52zR8xgf
35. Ravi, N., et al.: SAM 2: Segment anything in images and videos. In: International Conference on Learning Representations (ICLR). pp. 1–42 (2025)
36. RunDiffusion: Juggernaut-XL. https://huggingface.co/RunDiffusion/Juggernaut-XL, accessed April 8, 2026.
37. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Darcet, T., Misra, I., Bojanowski, P.: DINOv3. arXiv preprint arXiv:2508.10104 (2025)
38. Tan, M., Le, Q.: EfficientNetV2: Smaller models and faster training. In: International Conference on Machine Learning (ICML). pp. 10096–10106 (2021)
39. Trabucco, B., Doherty, K., Gurinas, M.A., Salakhutdinov, R.: Effective data augmentation with diffusion models. In: International Conference on Learning Representations (ICLR). pp. 1–23 (2024)
40. Ultralytics: YOLOv11. https://docs.ultralytics.com/models/yolo11/ (2026), accessed: 2026-04-30
41. Wang, Y., Wang, C., Zheng, Y., Fu, H., Ma, H.: Transformer based neural network for fine-grained classification of vehicle color. In: International Conference on Multimedia Information Processing and Retrieval (MIPR). pp. 118–124 (2021)
42. Wojcik, L., Machoski, E.A.F., Nascimento Jr., E., Laroca, R., Menotti, D.: LPLCv2: An expanded dataset for fine-grained license plate legibility classification. International Joint Conference on Neural Networks (IJCNN) pp. 1–7 (2026)
43. Yi, X., Wang, Q., Liu, Q., Rui, Y., Ran, B.: Advances in vehicle re-identification techniques: A survey. Neurocomputing **614**, 128745 (2025)
44. Zhang, Y., Kang, B., Hooi, B., Yan, S., Feng, J.: Deep long-tailed learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(9), 10795–10816 (2023)