LPLCv2: An Expanded Dataset for Fine-Grained License Plate Legibility Classification

Lucas Wojcik*, Eduardo A. F. Machoski*, Eduil Nascimento Jr.†, Rayson Laroca‡,*, and David Menotti*

*Federal University of Paraná, Curitiba, Brazil
†Department of Technological Development and Quality, Paraná Military Police, Curitiba, Brazil
‡Pontifical Catholic University of Paraná, Curitiba, Brazil
*{lmlwojcik, eafm23, menotti}@inf.ufpr.br †eduiljunior@pm.pr.gov.br ‡rayson@ppgia.pucpr.br

***Abstract*—Modern Automatic License Plate Recognition (ALPR) systems achieve outstanding performance in controlled, well-defined scenarios. However, large-scale real-world usage remains challenging due to low-quality imaging devices, compression artifacts, and suboptimal camera installation. Identifying illegible license plates (LPs) has recently become feasible through a dedicated benchmark; however, its impact has been limited by its small size and annotation errors. In this work, we expand the original benchmark to over three times the size with two extra capture days, revise its annotations and introduce novel labels. LP-level annotations include bounding boxes, text, and legibility level, while vehicle-level annotations comprise make, model, type, and color. Image-level annotations feature camera identity, capture conditions (e.g., rain and faulty cameras), acquisition time, and day ID. We present a novel training procedure featuring an Exponential Moving Average-based loss function and a refined learning rate scheduler, addressing common mistakes in testing. These improvements enable a baseline model to achieve an 89.5% F1-score on the test set, considerably surpassing the previous state of the art. We further introduce a novel protocol to explicitly addresses camera contamination between training and evaluation splits, where results show a small impact. Dataset and code are publicly available at https://github.com/lmlwojcik/LPLCv2-Dataset.**

***Index Terms*—intelligent transportation systems, license plate recognition, license plate legibility classification**

I. Introduction

Automatic License Plate Recognition (ALPR) systems have become widely adopted in recent years, supporting applications such as road surveillance, law enforcement, and parking management [1]. Advances in deep learning have enabled ALPR systems to achieve remarkably high performance in core tasks, including license plate (LP) detection and LP text recognition [2]–[4], particularly under controlled conditions.

Several datasets commonly used in ALPR research suffer from inherent limitations, including a small number of capture devices [5], often producing high-quality images not representative of real-world surveillance conditions, limited diversity in acquisition scenarios [6], and relatively small sample sizes [7]. When deployed under challenging conditions such as rain, fog, low-light environments, or in the presence of low-resolution imagery, suboptimal camera placement, inadequate acquisition equipment, and compression artifacts, the performance of state-of-the-art ALPR systems degrades significantly [8], [9].

Fig. 1. LPs grouped by legibility class, following the definition of [10].

These shortcomings underscore the need for more diverse, realistic, and challenging ALPR datasets.

At the same time, real-world ALPR applications demand fast and reliable license plate tracking and recognition under unconstrained conditions. Critical systems, such as those used for road surveillance, depend on low false positive rates to ensure efficient resource allocation and operational trustworthiness. Incorrect predictions can lead to unnecessary use of resources and raise ethical concerns. As these issues are particularly pronounced in scenarios with low LP legibility, several strategies have been proposed to mitigate their impact.

Super-resolution (SR) is one such example. Despite recent advances in SR techniques aimed at recovering characters from severely degraded LPs [11], [12], this line of research remains incipient and is not yet reliable enough for deployment in real-world, safety-critical systems. A key limitation is the limited cross-dataset generalization of existing methods. As reported in [10], current state-of-the-art license plate SR approaches not only struggle to generalize across datasets but can also harm high-quality inputs, reducing LP legibility and ultimately degrading Optical Character Recognition (OCR) performance.

Another example is LP legibility classification, a recently explored task that aims to distinguish between legible and illegible LPs to handle each case appropriately. In large-scale real-world deployments, ALPR systems may process millions of images per day, making it computationally infeasible to apply end-to-end processing pipelines, such as SR followed by OCR, to every single image[1]. Instead, legibility classification enables early decision-making, allowing systems to discard

[1]This large-scale setting, involving millions of images per day, corresponds exactly to the operational scenario of one of our ongoing research projects.

Accepted for presentation at IJCNN 2026. The final published version will be available on IEEE Xplore.

unusable samples, selectively apply SR to severely degraded LPs, or directly perform OCR on high-quality inputs. Despite its practical relevance, prior work has reported accuracies of only around 76% in the most general evaluation setting, highlighting substantial room for improvement [10].

This work addresses key limitations of previous studies on legibility classification. We revisit an existing benchmark [10] by correcting annotation errors, introducing new and more detailed labels, and extending it with many more images that cover a broader range of real-world scenarios. The LP legibility is annotated using four classes (see Fig. 1). In addition, we propose an improved training procedure tailored to the most common error patterns observed in baseline experiments. Our main contributions are summarized as follows:

- A comprehensive revision of a previous benchmark, including the correction of mis-annotated legibility labels and its extension into a new dataset that is over three times larger than the original, with samples collected from more than 700 cameras and covering scenarios underrepresented in the literature, such as faulty cameras, low-resolution imagery, and adverse weather conditions;
- The introduction of novel, fine-grained annotations at both the image and LPs levels, including camera identifiers and capture conditions for images, as well as vehicle-related data for the LPs;
- An improved training procedure specifically designed to mitigate common errors and better handle hard samples present in the dataset;
- A cross-camera evaluation to investigate potential camera contamination between training and testing partitions.

The remainder of this work is organized as follows. Section II provides an overview of related methods and datasets. Section III describes the proposed dataset and the revisions made to the previous benchmark. Section IV details the experimental protocol and setup. Section V presents and discusses the results. Finally, Section VI concludes the paper.

II. Related Work

Recent advances in ALPR have led to very high detection and recognition rates on most public benchmarks [2], [13], [14]. Modern pipelines are able to handle a wide range of conditions within these datasets, achieving strong performance on mainstream ALPR evaluation protocols.

A key challenge for cross-dataset generalization stems from the diversity of LP layouts across countries. To address this issue, recent works have explored synthetic data generation as a means to increase layout coverage and robustness [2]. This diversity has also driven a shift from character-based OCR approaches [15], [16] toward global recognition models [17]–[19]. Global models typically rely on implicit character localization, often through attention mechanisms, enabling improved recognition across different LP layouts.

Despite these advances, most publicly available datasets remain biased toward high-quality, fully legible LPs. For Brazilian and Mercosur layouts, commonly used datasets include SSIG-SegPlate [5], UFPR-ALPR [15], and RodoSol-ALPR [20]. RodoSol-ALPR is collected in relatively controlled toll plaza environments along a single highway, while SSIG-SegPlate and UFPR-ALPR rely on a very limited number of cameras (one and three, respectively) and predominantly feature high-resolution, clearly legible LPs.

Chinese LPs are primarily represented by CCPD [21] and CLPD [7]. Although CCPD is widely adopted, it has undergone multiple revisions and expansions, leading to inconsistencies in dataset size and evaluation protocols across studies [4], [22]. Differences in test splits across versions hinder direct comparison of reported results [2], and the presence of near-duplicate samples, corresponding to images of the same vehicle/LP, across training and test sets further compromises fair evaluation [23]. CLPD introduces challenges such as skewed LPs, varying illumination, and adverse weather conditions, as well as diverse capture angles and devices. However, image quality variability remains limited, with characters remaining clearly distinguishable despite geometric distortions.

Additional datasets include AOLP [24], CD-HARD [25], and OpenALPR-EU [26]. AOLP contains images captured with both handheld and static cameras, but is restricted to close-range, high-quality views. CD-HARD and OpenALPR-EU are comparatively small, with just over 100 samples each, and also rely on closeup imagery acquired with high-quality cameras. While CD-HARD introduces severe viewpoint distortions, these datasets do not capture other common real-world degradations such as low resolution and severe blur.

Overall, existing datasets span multiple region formats but remain constrained by controlled capture conditions and high-resolution imagery. In unconstrained surveillance environments, ALPR systems must operate across diverse cameras, resolutions, and adverse conditions. As a result, the practical applicability of current state-of-the-art methods remains limited in many real-world deployments.

To address these limitations, we introduce the LPLCv2 dataset, described in detail in Section III. This benchmark incorporates a legibility measure that explicitly quantifies OCR difficulty, enabling systematic evaluation under challenging conditions. The data was collected from over 700 cameras spanning a wide range of resolutions, viewpoints, and capture distances, resulting in LPs with varying levels of degradation. In addition, we provide fine-grained annotations that allow LPLCv2 to serve as a challenging benchmark for conventional and robust ALPR tasks.

III. The proposed Dataset

The proposed dataset[2], LPLCv2, is an expanded and enriched version of LPLCv1 [10]. The original benchmark was designed primarily for legibility classification, improving ALPR pipelines by enabling different processing strategies according to LP legibility. In particular, distinguishing between legible and illegible LPs allows unsuitable detections to be filtered out prior to OCR. Building upon this foundation,

[2] LPLCv2 is available at https://github.com/lmlwojcik/LPLCv2-Dataset

Fig. 2. Representative samples from the proposed LPLCv2 dataset.

LPLCv2 substantially broadens the scope of the benchmark by incorporating additional annotations relevant to practical scenarios involving legibility assessment, while also significantly increasing the dataset size.

LPLCv2 features new images collected from the same real-world domain as LPLCv1. These were captured over three different days by traffic radar systems deployed across the Brazilian state of Paraná and were pre-processed prior to release to remove embedded metadata and anonymize capture locations. Each image contains at least one LP, and is annotated with respect to (i) the time of day (night, morning, afternoon and evening, each a period of 6 hours and starting at midnight), (ii) the day and (iii) camera IDs (with some omitted when unique camera identification was not possible), and two capture-condition flags indicating (iv) rain and (v) faulty camera operation (some cameras were observed to produce images with a corrupted red channel, resulting in a characteristic magenta color cast across the entire image).

For each annotated LP, the dataset provides annotations for (i) the enclosing rectangular bounding box (x, y, w, h), (ii) the LP text, (iii) the legibility level, (iv) vehicle attributes (make, model, type, and color), and (v) vehicle occlusion (boolean). These annotations were initially generated automatically using task-specific models, namely a YOLOv11 detector [27] for LP localization, a PARSeq-tiny model [28] for OCR, and a ResNet-50 model [29] for legibility estimation, and were subsequently manually reviewed to ensure consistency and accuracy. Vehicle attributes were retrieved from the Brazilian National Traffic Secretariat (SENATRAN) database based on the recognized LP text. Not all LPs include complete annotations: vehicle attributes are omitted when unavailable for a given LP text, while LP characters are excluded when the LP is too degraded to support reliable human validation. In some cases, however, otherwise illegible LPs could still be validated using the zoom feature of some of the cameras in the dataset.

Fig. 2 presents representative samples from LPLCv2, highlighting its diversity and practical relevance. The bottom-left image illustrates a capture from a faulty camera, a scenario that, to the best of our knowledge, is not represented in existing public datasets. The top-right image depicts rainy conditions, while the top-left image shows an LP that would otherwise be illegible, but whose characters could be validated due to the camera zoom. Together, these examples further highlight the variability present in LPLCv2, including extreme lighting, different camera distances, heterogeneous capture devices, and diverse vehicle types. As a result, our dataset offers a faithful and challenging representation of real-world ALPR operating conditions.

Table I presents the image statistics of LPLCv2, including all instances inherited from LPLCv1. In total, 26,889 new images containing 28,800 annotated LPs were added to the original 10,210 images and 12,687 LPs from LPLCv1. As previously stated, the dataset spans three distinct capture days. All images from LPLCv1 originate from a single day, with new images comprising the two other additional days: 15,284 images from one day and 11,605 from the other. Images from the *evening* and *night* correspond to night-time instances (6 p.m. to midnight for evening and midnight to 6 a.m. for night), while *morning* and *afternoon* images correspond to daytime (6 a.m. to noon for morning and noon to 6 p.m. for afternoon).

TABLE I
LPLCV2 IMAGE STATISTICS.

Images by Time of Day		Images by Attributes	
Class	Number	Attribute	Number
Morning	10,998	Faulty Camera	3,690
Afternoon	12,799	Raining	770
Evening	9,157	Has Camera ID	29,965
Night	4,145	→ Total Images	37,099

Table II reports analogous statistics for the annotated LPs in LPLCv2. All annotations were manually reviewed, and OCR outputs were cross-validated using vehicle data. Partially occluded LPs were removed from the dataset, alongside the corresponding LP-level occlusion labels. Vehicle-level occlusion is present when vehicle attributes cannot be reliably identified from the full image.

Legibility classes follow the same criteria as in LPLCv1. Perfect LPs exhibit clear characters with no visible distortion. Good LPs contain minor, non-disruptive distortions while remaining fully legible. Poor LPs display more severe distortions but still allow unambiguous character identification. Illegible LPs present substantial degradation that prevents confident

character validation, except in cases where camera zoom enables reliable interpretation.

TABLE II
LPLCv2 LP STATISTICS.

LPs by Legibility		Other Attributes	
Class	Number	Attribute	Number
Perfect	18,425	Vehicle Visible	38,089
Good	10,180	Vehicle Available	25,506
Poor	7,520	LP Text Available	36,414
Illegible	5,362	→ Total LPs	41,487

In addition to extending the dataset, we conducted a comprehensive revision of the legibility annotations inherited from LPLCv1. Several labeling errors, revealed through network mispredictions, were corrected by reassigning affected LPs to adjacent classes. The revision took into account the annotator's assessment together with the predictions from the models used in [10]. In total, 1,012 LPs were relabeled: 514 poor instances were reclassified as good, 112 good instances were reclassified as poor, 275 perfect instances were reclassified as good, and 111 good instances were reclassified as perfect. Furthermore, 787 LPs originally labeled as illegible were reclassified as poor after successful OCR validation. Representative examples of these corrections are illustrated in Fig. 3.

Fig. 3. LP legibility annotation errors observed in LPLCv1 (left) and their corresponding corrections in LPLCv2 (right).

IV. MODELING AND EXPERIMENTS

This section presents the experimental setup and modeling decisions adopted to ensure a fair, reproducible evaluation of LP legibility classification across multiple scenarios.

A. Legibility Classification

Following [10], we adopt a 5-bin, 10-fold cross-validation protocol with 40-20-40 splits for training, validation, and testing. All reported results are the average across the 10 test folds. To construct the folds, all LPs are uniformly divided into five ordered bins, each containing approximately 20% of the dataset. For each fold, two bins are assigned to training, one to validation, and the remaining two to testing. The validation bin is circularly rotated across folds (e.g., the third bin is used for validation in the first fold, the fourth in the second, and so on), producing five distinct splits. Five additional folds are obtained by swapping the training and testing partitions of each split, resulting in a total of 10 folds. All folds are released alongside LPLCv2 to ensure reproducibility.

When training on LPLCv1 [10], we use the original splits provided with the dataset for a fair comparison with prior results. For LPLCv2, newly added instances are incorporated exclusively into the training partitions of each fold derived from LPLCv1. As a result, the training set size increases from approximately 5k to over 33k LPs. Naturally, our new experiments utilize the corrected labels for samples from LPLCv1.

Legibility classification is evaluated under four scenarios. These are classification tasks over cropped LPs. The first three are reproduced from [10]. In the "Baseline" scenario, the four original legibility classes are used as prediction targets. In the second scenario, denoted "Legibility Recognition", the *perfect* and *good* classes are merged into a single label, referred to as *suitable*, and samples labeled as *illegible* are excluded. This label mapping is maintained in the third scenario, "Full Recognition", where the *illegible* class is reintroduced.

The notion of *suitable* is grounded on the reliability of LPs for OCR. Prior results [10] showed that *poor* LPs achieved a whole-plate accuracy of only 72% even with the best-performing model, whereas *good* and *perfect* samples exceeded 90%. Motivated by this performance gap, we introduce a fourth scenario, termed "Quality Filter", targeting practical OCR pre-processing. It is formulated as a binary classification task, where *suitable* comprises the *perfect* and *good* classes, and *unsuitable* comprises the *poor* and *illegible* classes. Table III summarizes the mappings for each evaluation scenario.

TABLE III
MAPPING OF ORIGINAL LEGIBILITY CLASSES TO LABELS USED IN EACH EVALUATION SCENARIO.

Scenario	**Classes**			
Baseline	Perfect	Good	Poor	Illegible
Legibility Recognition	Suitable		Poor	–
Full Recognition	Suitable		Poor	Illegible
Quality Filter	Suitable		Unsuitable	

In addition, we also introduce two new 10-fold cross-validation splits constructed exclusively from LPs extracted from images with an associated camera identifier. These comprise 33,977 LPs and define the *intra-camera* and *cross-camera* evaluation protocols. The goal is to investigate potential biases in legibility classification arising from testing on data captured by the same devices observed during training.

The inclusion of camera identifiers for samples originating from LPLCv1 reveals that the previously adopted 10-fold splits exhibited camera-level overlap between training and test sets. This form of contamination can lead to overly optimistic performance estimates [30]. We therefore release these camera-aware splits as a new component of the benchmark.

To construct the camera-based splits, the cameras are sorted in descending order according to the number of available images. Two assignment strategies are then applied in parallel. In the first strategy, images from each camera are uniformly distributed across one set of five bins. In the second strategy, all images from a given camera are assigned to the bin that currently contains the fewest samples within a separate set

of five bins. Both strategies aim to balance the total number of images per bin. Each set of bins is subsequently used to generate a distinct 10-fold split, corresponding to the intra-camera and cross-camera protocols, respectively.

B. Model and Parameters

In this work, we use the pre-trained ResNet-50 model [29] available in the PyTorch `torchvision` library. Although this model was not the top-performing architecture in [10], its performance improves substantially after the proposed training refinements and the inclusion of additional training data.

To better handle hard examples, we adopt an Exponential Moving Average (EMA) loss [31], implemented as a dynamically weighted cross-entropy loss for multi-class classification. The class weights are computed at each mini-batch based on the relative number of misclassified samples per class. The class weight tensor for batch b is defined in Eq. (1), where $mispred$ corresponds to the error count in the smoothed confusion matrix, as defined in Eq. (2) and Eq. (3). This calculation ensures a higher weight for classes with higher error counts, and the loss is stabilized by using the smoothed confusion matrix for weight calculation. This modeling is suitable for LPLCv2, which features class imbalance and critical differences between the least represented classes.

$$W(b) = min(\frac{max(mispred(b))}{mispred(b) + \epsilon}, M) \quad (1)$$

$$smooth(b) = \alpha * confusion(b) + (1-\alpha) * confusion(b) \quad (2)$$

$$mispred(b) = sum(smooth(b)) - diag(smooth(b)) \quad (3)$$

Our implementation uses $\alpha = 0.8$, $\epsilon = 1e{-}6$ and $M = 1.2$. Optimization is performed using Adam [32] with an initial learning rate of $1e{-}5$. A linear decay factor of 0.75 is applied with a patience of 5 epochs, down to a minimum learning rate of $5e{-}6$. Models are trained independently for each fold using early stopping with a patience of 20 epochs based on validation accuracy, for a maximum of $1{,}000$ epochs. The batch size is set to 64, and all layers of the network are fine-tuned rather than freezing the backbone. All experiments are conducted on an NVIDIA RTX 6000 GPU with 48 GB of RAM.

V. Results

Table IV reports the results for the *Baseline* scenario. All results are reported as F1-scores computed on the test sets and averaged across the 10 folds. The *Overall* column corresponds to the micro-averaged F1-score across all classes. The first row reproduces the results reported in [10] (LPLCv1). Subsequent rows show the cumulative impact of the improvements introduced in this work, namely the correction of legibility labels, the addition of newly annotated images to the training sets, and the usage of the EMA loss. The test sets are unchanged, and the model is trained using the same hyperparameters as in the original benchmark unless stated otherwise.

TABLE IV
Baseline results using ResNet-50, reported as F1-score (%).

Partition	Class				Overall
	Perfect	Good	Poor	Illegible	
LPLCv1	84.5%	68.0%	56.7%	73.0%	74.5%
+ Revised Labels	92.1%	83.1%	83.8%	88.2%	84.4%
+ More Images	92.8%	86.0%	86.0%	92.9%	88.7%
+ EMA	92.7%	86.2%	86.3%	93.0%	89.8%

The results show that the label correction substantially improves dataset consistency and, consequently, model performance, with the overall F1-score increasing from 74.51% to 84.44%. Expanding the training data with new samples leads to further improvement, raising performance to 88.67%, corresponding to a reduction of approximately 27% in the remaining classification errors. The introduction of the EMA loss produces an additional gain in overall performance. In particular, it the average F1-score for the three minority classes (good, poor and illegible) are slightly improved, with a marginal reduction for the majority class (perfect). Upon aggregation, the trade-off results in a higher overall F1-score.

For the EMA loss experiment, Fig. 4 presents the confusion matrix for one representative (median F1-score) fold. Most errors occur between adjacent legibility classes, indicating that the model rarely produces extreme misclassifications. For example, *good* LPs are primarily confused with *perfect* or *poor* samples, and analogous patterns are observed for other classes. This behavior suggests that the learned decision boundaries largely respect the ordinal nature of the legibility labels.

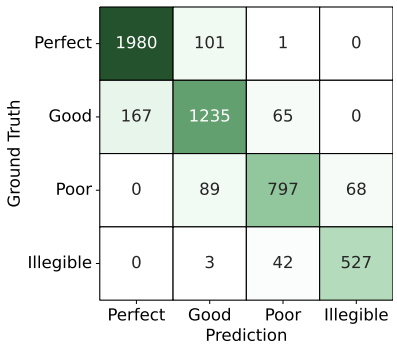

Fig. 4. Confusion matrix from one of the EMA experiment folds.

Fig. 5 provides qualitative insight into these misclassifications. Many errors arise from visually ambiguous cases in which the appearance of an LP lies close to the boundary between classes. Some *illegible* LPs retain a small number of recognizable but severely degraded characters, while some *poor* LPs exhibit strong blur patterns similar to those observed in fully illegible instances. Likewise, certain *good* and *perfect* LPs share comparable noise characteristics, reducing the visual contrast between these categories. These examples highlight that the remaining errors are largely driven by intrinsic ambiguity in the data rather than systematic model failures.

We further analyze the edge cases in Fig. 6, which corre-

Fig. 5. Examples of misclassified LPs illustrating borderline legibility cases.

spond to misclassifications where an LP is assigned to a non-adjacent legibility class. Three *illegible* LPs are incorrectly classified as *good*. As the current model operates on the entire LP patch rather than on individual characters, we conjecture that background patterns in these samples resemble those of certain *good* instances observed during training. Conversely, the single *perfect* LP misclassified as *poor* appears to be affected by low-contrast imaging conditions, which are also characteristic of some *poor* samples. Unlike genuinely *poor* LPs, however, the characters in this instance do not exhibit the obtrusive degradation associated with that class.

(a) *Illegible* LPs classified as *good*.

(b) *Perfect* LP classified as *poor*.

Fig. 6. LPs misclassified into non-adjacent legibility classes.

We report the results for the remaining scenarios in Table V (see Table III for class mapping). Performance is measured using the mean micro-averaged F1-score on the test sets, corresponding to the *Overall* metric reported in Table IV. For the *Quality Filter* scenario, not present in LPLCv1 [10], we reproduce the same setup using the original labels and parameters for a fair comparison with the proposed contributions.

TABLE V
RESNET-50 PERFORMANCE (F1-SCORE) IN THE REMAINING SCENARIOS: LEGIBILITY RECOGNITION (SUITABLE AND POOR); FULL RECOGNITION (SUITABLE, POOR AND ILLEGIBLE); AND QUALITY FILTER (SUITABLE AND UNSUITABLE).

Partition	Legibility Recognition	Full Recognition	Quality Filter
LPLCv1	92.6%	87.2%	93.3%
+ Revised Labels	95.3%	93.7%	96.4%
+ More Images	95.9%	95.2%	96.9%
+ EMA	96.5%	94.8%	96.7%

The results for the corrected labels and expanded dataset follow the same trend observed in the *Baseline* scenario, yielding consistent improvements in the test performance. In contrast, the effect of the EMA loss is less stable in scenarios with fewer target classes. While performance improves in the *Legibility Recognition* scenario, it decreases slightly in the other two scenarios. Most importantly, in the *Quality Filter* scenario, the variation is minimal and not statistically significant.

Finally, we perform the camera contamination experiments using the proposed novel protocol. In Table VI we present the results of the intra-camera and cross-camera experiments conducted under the *Baseline* scenario. Both experiments use the same hyperparameters (with EMA), differing only in the adopted 10-fold splits. Unlike the previous experiments, where newly annotated samples are added exclusively to the training sets, the camera-based splits are generated over the entire revised benchmark. Consequently, the training partitions are smaller, while the validation and test sets are larger. Even under these more restrictive conditions, the observed performance degradation is limited, indicating strong generalization across camera devices and reinforcing the robustness of the proposed benchmark.

TABLE VI
BASELINE SCENARIO RESULTS UNDER INTRA-CAMERA AND CROSS-CAMERA PROTOCOLS (F1-SCORE).

Partition	Class				Overall
	Perfect	Good	Poor	Illegible	
Intra-Camera	93.4%	82.6%	84.8%	91.6%	89.2%
Cross-Camera	93.5%	81.8%	85.7%	91.1%	88.2%

As shown by the results, separating images captured by the same camera across different dataset partitions leads to a measurable but moderate decrease in performance. This indicates that the proposed legibility classification approach is largely robust to the camera variability present in the revised benchmark. The same trend is observed across the remaining scenarios, as reported in Table VII.

TABLE VII
RESNET-50 PERFORMANCE ACROSS THE OTHER EVALUATION SCENARIOS UNDER INTRA-CAMERA AND CROSS-CAMERA SPLITS (F1-SCORE).

Partition	Legibility Recognition	Full Recognition	Quality Filter
Intra-Camera	96.2%	94.5%	96.6%
Cross-Camera	95.6%	93.8%	96.2%

VI. CONCLUSIONS

This paper advances the study of LP legibility classification by revising a previously proposed benchmark, introducing a substantially expanded dataset, and improving recognition performance through a refined training strategy. We show that correcting annotation errors in the original benchmark improves dataset consistency and leads to significant performance gains. Additional improvements are achieved by expanding the training data and adopting a tailored loss function designed to better handle difficult and ambiguous samples.

Beyond performance gains, we enhance the dataset in both diversity and descriptive power. The proposed dataset incorporates new capture conditions, including rainy environments and

imagery from faulty cameras. To the best of our knowledge, the latter represents the first public contribution of its kind within the broader ALPR literature. We also introduce richer annotations, such as camera identifiers for most images and vehicle-related metadata for most LPs. These additions enable a wider range of experimental protocols and facilitate novel studies to be conducted.

A viable direction for future work is to integrate legibility classification into a complete ALPR pipeline by leveraging the dataset's detailed annotations. This will likely require new evaluation protocols for fair model comparison and may also enable studies on vehicle identification, such as LP super-resolution [9] and fine-grained vehicle classification [33]. Another direction for further research regards the cross-model generalization of the proposed contributions, as well as cross-dataset evaluation for LPs of different layouts.

ACKNOWLEDGMENTS

This study was supported in part by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brasil*, under the *Programa de Excelência Acadêmica (PROEX)* – Finance Code 001; in part by the *Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)* (# 315409/2023-1); and in part by the *Fundação Araucária* (# 078/2025).

REFERENCES

[1] A. Ismail, M. Mehri, A. Sahbani, and N. Essoukri Ben Amara, "Automatic license plate recognition in in-the-wild scenarios: A comprehensive review, open issues, and future directions," *IEEE Access*, vol. 13, pp. 145 387–145 415, 2025.

[2] R. Laroca, V. Estevam, G. J. P. Moreira, R. Minetto, and D. Menotti, "Advancing multinational license plate recognition through synthetic and real data fusion: A comprehensive evaluation," *IET Intelligent Transport Systems*, vol. 19, no. 1, p. e70086, 2025.

[3] X. Ke, G. Zeng, and W. Guo, "An ultra-fast automatic license plate recognition approach for unconstrained scenarios," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, pp. 5172–5185, 2023.

[4] H. Ding, J. Gao, Y. Yuan, and Q. Wang, "An end-to-end contrastive license plate detector," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 1, pp. 503–516, 2024.

[5] G. R. Gonçalves, S. P. G. da Silva, D. Menotti, and W. R. Schwartz, "Benchmark for license plate character segmentation," *Journal of Electronic Imaging*, vol. 25, no. 5, p. 053034, 2016.

[6] V. Srebrić, "EnglishLP Database," 2003. [Online]. Available: https://www.zemris.fer.hr/projects/LicensePlates/english/baza_slika.zip

[7] L. Zhang *et al.*, "A robust attentional framework for license plate recognition in the wild," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 11, pp. 6967–6976, 2021.

[8] S. Wahyu *et al.*, "Fog and rain augmentation for license plate recognition in tropical country environment," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 13, no. 4, p. 3951, Dec. 2024.

[9] V. Nascimento, G. E. Lima, R. O. Ribeiro, W. R. Schwartz, R. Laroca, and D. Menotti, "Toward advancing license plate super-resolution in real-world scenarios: A dataset and benchmark," *Journal of the Brazilian Computer Society*, vol. 1, no. 31, pp. 435–449, 2025.

[10] L. Wojcik, G. E. Lima, V. Nascimento, E. Nascimento Jr., R. Laroca, and D. Menotti, "LPLC: A dataset for license plate legibility classification," *Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2025.

[11] V. Nascimento, R. Laroca, R. O. Ribeiro, W. R. Schwartz, and D. Menotti, "Enhancing license plate super-resolution: A layout-aware and character-driven approach," *Conference on Graphics, Patterns and Images (SIBGRAPI)*, pp. 1–6, 2024.

[12] Y. Pan, J. Tang, and T. Tjahjadi, "LPSRGAN: Generative adversarial networks for super-resolution of license plate image," *Neurocomputing*, vol. 580, p. 127426, 2024.

[13] R. Laroca, L. A. Zanlorensi, V. Estevam, R. Minetto, and D. Menotti, "Leveraging model fusion for improved license plate recognition," in *Iberoamerican Congress on Pattern Recognition*, Nov 2023, pp. 60–75.

[14] Q. Liu *et al.*, "Improving multi-type license plate recognition via learning globally and contrastively," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 9, pp. 11 092–11 102, 2024.

[15] R. Laroca, E. Severo, L. A. Zanlorensi, L. S. Oliveira, G. R. Gonçalves, W. R. Schwartz, and D. Menotti, "A robust real-time automatic license plate recognition based on the YOLO detector," in *International Joint Conference on Neural Networks (IJCNN)*, July 2018, pp. 1–10.

[16] S. M. Silva and C. R. Jung, "Real-time license plate detection and recognition using deep convolutional neural networks," *Journal of Visual Communication and Image Representation*, vol. 71, p. 102773, 2020.

[17] A. Alzahrani, "License plate recognition using a hybrid class attention-inception network," in *International Conference on Modelling Strategies in Mathematics*, vol. 3306, 2025, p. 060023.

[18] T.-M. Seo and D.-J. Kang, "A robust layout-independent license plate detection and recognition model based on attention method," *IEEE Access*, vol. 10, pp. 57 427–57 436, 2022.

[19] Y.-Y. Liu, Q. Liu, S.-L. Chen, F. Chen, and X.-C. Yin, "Irregular license plate recognition via global information integration," in *International Conference on Multimedia Modeling*, 2024, pp. 325–339.

[20] R. Laroca, E. V. Cardoso, D. R. Lucio, V. Estevam, and D. Menotti, "On the cross-dataset generalization in license plate recognition," in *International Conference on Computer Vision Theory and Applications (VISAPP)*, Feb 2022, pp. 166–178.

[21] Z. Xu, W. Yang, A. Meng, N. Lu, H. Huang, C. Ying, and L. Huang, "Towards end-to-end license plate detection and recognition: A large dataset and baseline," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 261–277.

[22] Y. Gao, H. Lu, S. Mu, and S. Xu, "GroupPlate: Toward multi-category license plate recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 5, pp. 5586–5599, 2023.

[23] R. Laroca, V. Estevam, A. S. Britto Jr., R. Minetto, and D. Menotti, "Do we train on test data? The impact of near-duplicates on license plate recognition," in *International Joint Conference on Neural Networks (IJCNN)*, June 2023, pp. 1–8.

[24] G. S. Hsu, J. C. Chen, and Y. Z. Chung, "Application-oriented license plate recognition," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 2, pp. 552–561, Feb 2013.

[25] S. M. Silva and C. R. Jung, "License plate detection and recognition in unconstrained scenarios," in *European Conference on Computer Vision (ECCV)*, Sept 2018, pp. 593–609.

[26] OpenALPR, "OpenALPR-EU Dataset," 2016. [Online]. Available: https://github.com/openalpr/benchmarks/tree/master/endtoend/eu

[27] Ultralytics, "YOLOv11," 2025, accessed: 2026-01-28. [Online]. Available: https://docs.ultralytics.com/models/yolo11/

[28] D. Bautista and R. Atienza, "Scene text recognition with permuted autoregressive sequence models," in *European Conference on Computer Vision (ECCV)*, 2022, pp. 178–196.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[30] R. Laroca, M. Santos, V. Estevam, E. Luz, and D. Menotti, "A first look at dataset bias in license plate recognition," in *Conference on Graphics, Patterns and Images (SIBGRAPI)*, Oct 2022, pp. 234–239.

[31] D. Morales Brotons, T. Vogels, and H. Hendrikx, "Exponential Moving Average of Weights in Deep Learning: Dynamics and Benefits," *Transactions on Machine Learning Research Journal*, pp. 1–27, Apr. 2024.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[33] G. E. Lima *et al.*, "Toward enhancing vehicle color recognition in adverse conditions: A dataset and benchmark," in *Conference on Graphics, Patterns and Images (SIBGRAPI)*, Sept 2024, pp. 1–6.