

# HPR-Net: A High-Precision Pipeline with Oriented Detection and Transformer-Based Recognition for Robust License Plate Recognition

Bertrand Tahte<sup>1</sup>, Alain Djimeli<sup>2</sup>, Geh Wilson Ejuh<sup>3</sup>

<sup>1</sup>Computer Science Department, University of Dschang, Dschang, Cameroon

<sup>2</sup>Department of Telecommunication and Network Engineering, University Institute of Technology, Bandjoun, Cameroon

<sup>3</sup>Department of General and Scientific Education, University Institute of Technology, Bandjoun, Cameroon

Email: tahtebertrand2@gmail.com, adjimeli@yahoo.fr, gehwilsonejeh@yahoo.fr

**How to cite this paper:** Tahte, B., Djimeli, A. and Ejuh, G.W. (2026) HPR-Net: A High-Precision Pipeline with Oriented Detection and Transformer-Based Recognition for Robust License Plate Recognition. *Journal of Intelligent Learning Systems and Applications*, 18, 267-287.

<https://doi.org/10.4236/jilsa.2026.184017>

**Received:** August 19, 2026

**Accepted:** October 10, 2026

**Published:** October 13, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).  
<http://creativecommons.org/licenses/by/4.0/>



Open Access

## Abstract

Automatic Number Plate Recognition is a fundamental technology for intelligent transportation systems, urban management, and public security, although its performance degrades significantly in complex real-world conditions, often characterized as unconstrained scenarios. This research paper tackles two fundamental limitations of current state-of-the-art systems by addressing, on the one hand, the geometric inaccuracy of standard horizontal bounding box detectors, which systematically fail when faced with oblique viewing angles by capturing excessive background noise, and, on the other hand, the structural rigidity of common sequential recognition systems, which prove inadequate for the diverse multiline license plate formats found globally. To overcome these complex challenges, the HPR-Net (High-Precision Recognition Network) system is proposed as a modernized two-stage pipeline that synergistically combines cutting-edge deep learning components. The first stage introduces a high-precision-oriented object detector, based on the YOLOv8-OBb architecture, specifically modified to regress the rotation angle of the plate, thereby guaranteeing ultra-precise geometric localization via a loss function sensitive to spatial asymmetry. The second stage utilizes a Transformer-based recognition system that operates on the rectified image of the plate by exploiting a two-dimensional cross-attention mechanism to dynamically and flexibly handle various structural layouts. Exhaustive experimental evaluations conducted on an aggregate of complex public datasets demonstrate the superiority of this approach, with the HPR-Net system achieving an average end-to-end recognition accuracy of 96.2% ( $\pm 0.20\%$ , macro-average across 3 independent training

runs), outperforming baseline benchmarks by 7.3 percentage points. Ablation studies rigorously validate that both the oriented detector and the Transformer-based recognition system are imperative to achieve this new level of performance, thus offering a truly universal solution for smart infrastructures.

## Keywords

Deep Learning, Computer Vision, Oriented Object Detection, YOLOv8, Transformer

---

## 1. Introduction

The rapid evolution of urban infrastructure toward the concept of Smart Cities has placed Intelligent Transportation Systems (ITS) at the heart of modern mobility management strategies [1]. Within this ecosystem, Automatic Number Plate Recognition (ANPR) constitutes a fundamental technological building block [2] [3]. The applications of ANPR are vast and critical: they range from barrier-free highway tolling and dynamic traffic management to automated law enforcement, as well as access control for low-emission zones and public security surveillance. Despite the exceptionally high accuracy rates reported on standardized academic benchmarks, the reliability and robustness of these systems in unconstrained real-world environments remain a major scientific and technical challenge [4].

Recent scientific literature highlights that the performance ceiling of current ANPR systems no longer lies in the ability to read clear, frontal characters, but rather stems from a dual systemic constraint, both geometric and structural. First, the geometric constraint is induced by the very nature of the physical deployment of surveillance cameras. These cameras are generally installed at a height, on gantries or lateral poles, which generates oblique viewing angles, pronounced perspective effects, and spatial distortions. Faced with these distortions, standard object detectors prove ineffective because they utilize horizontal bounding boxes (HBB) [4] [5]. By nature, a horizontal box imposed on a tilted rectangular object will capture a massive amount of background pixels (vehicle body parts, headlights, shadows) while risking truncating the corners of the plate where vital characters or regional identifiers are often located. This initial geometric imprecision inevitably propagates cascading errors to the subsequent recognition module, illustrating the vulnerability of traditional sequential pipelines.

Second, the structural constraint emerges from the heterogeneous diversity of license plates on a global scale. While some regions standardize long, single-line formats, others adopt complex and asymmetrical multi-line formats. This diversity poses an architectural problem for dominant recognition models, such as the CRNN model [6]. These networks model the task as a one-dimensional sequence. Consequently, when confronted with a plate arranged on two lines, they attempt to read the image from left to right, inevitably mixing the spatial probabilities of the characters from the upper line with those of the lower line, which leads to a

catastrophic decoding failure [7].

This research paper postulates that overcoming these critical limitations requires a profound architectural redesign, adopting a synergistic approach that simultaneously addresses the precision of geometric localization and the flexibility of structural analysis. It is formally hypothesized that explicitly modeling the spatial orientation of the plate during the detection phase will produce a semantically pure crop, drastically reducing downstream error propagation, and that replacing one-dimensional recurrent architectures with a recognition system based on a two-dimensional attention mechanism (Transformer) will provide the structural flexibility required to learn dynamic and non-linear reading orders, regardless of the national plate format.

To validate these hypotheses, the HPR-Net (High-Precision Recognition Network) system is designed and evaluated. In direct response to the requirements of methodological rigor, this study presents exhaustive experimental evaluations, supported by detailed tabular quantitative data, descriptive figures, and in-depth qualitative analyses. HPR-Net modernizes the classical paradigm by integrating a modified YOLOv8 detector for oriented bounding boxes (OBB) and an autoregressive Transformer decoder. The results demonstrate that this architecture achieves a substantial and statistically consistent improvement (+7.3 percentage points macro-averaged E2E accuracy,  $\pm 0.20\%$  across three independent runs) over baseline and ablated configurations trained and evaluated under an identical protocol, on identical data splits. We refer to this as establishing a new performance benchmark among architectures compared under this study's protocol; broader comparisons to externally reported results (Section 5.3) are qualified accordingly, as they were not obtained under matched evaluation conditions.

## 2. Theoretical Background and Related Work

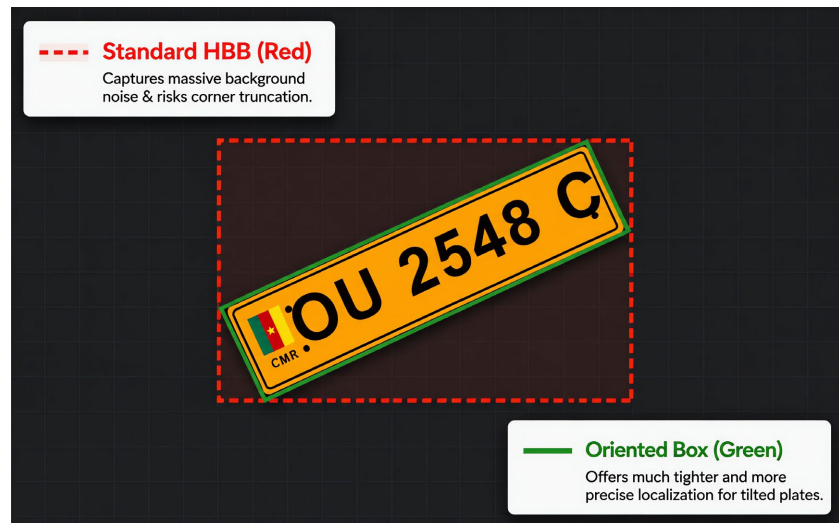
The development of a robust ANPR system requires the integration of multiple computer vision sub-disciplines. This section conducts an analytical review of the state-of-the-art regarding object detection, visual sequence recognition, and specific solutions tailored for embedded environments, while highlighting the limitations that justify the architectural choices of the HPR-Net system.

### 2.1. Evolution of Object Detection: Towards Oriented Modeling

The task of license plate detection is the preliminary and most crucial step in any ANPR pipeline. Historically, approaches relied on handcrafted feature descriptors such as Haar features combined with cascade classifiers, or the use of connected component analysis and edge detection [2] [5]. These traditional methods offered acceptable performance in strictly controlled environments but exhibited extreme sensitivity to illumination variations, adverse weather conditions, and motion blur.

The advent of deep learning has revolutionized this field. Modern detectors are dominated by Convolutional Neural Networks (CNNs), categorized primarily into two-stage and one-stage methods (such as SSD and the YOLO family) [4] [8].

For ANPR applications, the YOLO family has established itself as the industry standard due to its ability to process real-time video streams while maintaining high localization accuracy [9]. However, the inherent architecture of these models is designed to predict horizontal bounding boxes (HBB), defined solely by the coordinates of their center, width, and height. This orthogonal constraint is suboptimal for license plate detection in real-world traffic scenes. Forcibly fitting an HBB onto a heavily tilted object leads to the inclusion of irrelevant features that disrupt downstream recognition networks, as illustrated in **Figure 1** [5].



**Figure 1.** Comparison between a standard horizontal bounding box (in red) and an oriented box (in green). The OBB format offers much tighter and more precise localization for tilted plates.

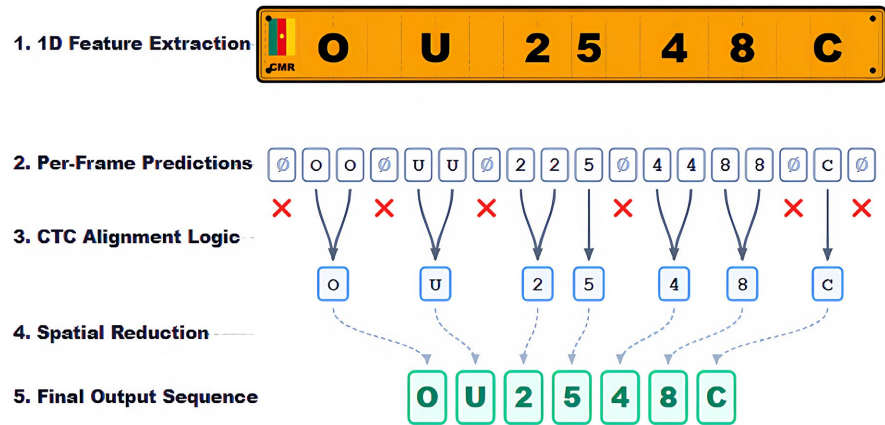
To overcome this geometric deficiency, research has shifted toward oriented object detection (OBB). Early implementations, such as the RoI Transformer [10], adapted two-stage architectures by applying spatial transformations to region proposals. Although these models excel in accuracy, their prohibitive computational cost limits their applicability. More recently, research has explored the integration of angular regression heads directly into one-stage architectures, necessitating the development of innovative loss functions to handle angular periodicity [11]. The systemic integration and rigorous evaluation of these ultra-fast OBB detectors within a complete pipeline constitute one of the major contributions of this paper.

## 2.2. Visual Sequence Recognition: From 1D Recurrence to 2D Attention

Once the plate is localized, the recognition task consists of transcribing the visual features into a sequence of alphanumeric characters. Early approaches required a laborious segmentation phase, isolating each character before submitting it to an Optical Character Recognition (OCR) algorithm [2] [3]. This preliminary segmentation was a critical point of failure.

The paradigm was fundamentally transformed by the introduction of the CRNN

network coupled with the CTC loss [6]. The CRNN utilizes a convolutional network to extract feature sequences, followed by recurrent layers to model contextual dependencies. The CTC loss function enables the dynamic alignment of predictions with the ground truth, entirely eliminating the need for explicit character-level segmentation (Figure 2).



**Figure 2.** The CTC alignment mechanism. It allows for flexible alignment by reducing repeated characters and removing blank tokens, but fundamentally operates on a one-dimensional (1D) sequence.

However, the CRNN/CTC model suffers from a severe structural limitation: its fundamental assumption of one-dimensional sequential processing. When confronted with a multiline plate, the CRNN attempts to scan the image from left to right across its entire height. This creates a chaotic interleaving of features from the upper and lower lines, causing a collapse in performance [7].

To overcome this spatial constraint, researchers have turned to attention-based architectures, and more specifically Transformers [12]. These encoder-decoder models exploit cross-attention mechanisms to learn a dynamic and non-linear reading order. Instead of imposing a one-dimensional scan, the decoder iteratively queries a two-dimensional feature map.

### 2.3. Lightweight Convolutional Alternatives: The Case of LPRNet

In parallel with research on two-dimensional attention, a significant portion of recent literature has focused on creating ultra-lightweight recognition architectures. The LPRNet model represents a notable advancement in this direction [13].

Unlike the CRNN standard, LPRNet proposes an end-to-end method that completely eliminates recurrent neural networks from its architecture [13]. LPRNet relies exclusively on lightweight convolutional networks and a temporal filtering layer. Evaluations demonstrate that LPRNet can achieve a recognition accuracy of 95% on complex Chinese license plate datasets with exceptional inference latency. This architecture requires only 0.34 Giga-FLOPs for a forward pass, making it compatible with partial implementation on specific hardware components such as FPGAs [13].

Although the purely convolutional approach of LPRNet offers unparalleled speed, it raises fundamental questions regarding its structural universality when faced with asymmetric character layouts [13]. The architecture proposed in this study, HPR-Net, acknowledges the crucial importance of computational efficiency, which is why it adopts an ultra-lightweight CNN feature extractor (MobileNetV3) [14]. However, it diverges philosophically by reintroducing a powerful semantic component—the Transformer—to ensure that the system is truly agnostic regarding the international format of the processed plates.

### 3. HPR-Net System Architecture

To directly address the problem of geometric error propagation and structural inflexibility, the HPR-Net pipeline is designed as a modular and mathematically rigorous system. Unlike approaches based on manual heuristics, the entire processing relies on deep representation learning. This section details the mathematical formulation and structure of each component.

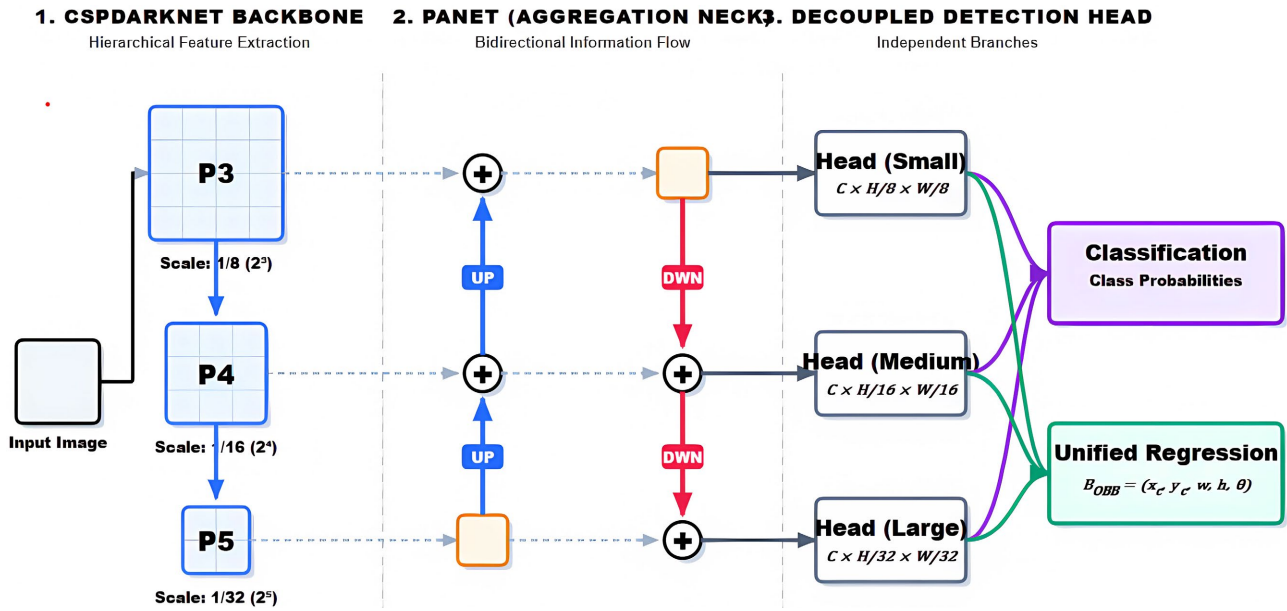
#### 3.1. High-Precision Oriented Detection Module (YOLOv8-OBB)

The function of the first pipeline module is to analyze the source image  $I \in \mathbb{R}^{H \times W \times 3}$  and to predict the location and orientation of the license plate with millimeter precision. To achieve this, the baseline YOLOv8 architecture has been redesigned for OBB (Oriented Bounding Box) detection.

**Multi-Scale Feature Extraction (Backbone and PANet):** Processing begins with a CSPDarknet (Cross Stage Partial Network) backbone. This network applies successive convolutions to extract a hierarchical feature pyramid  $P_3, P_4, P_5$ , where the index represents the spatial downsampling factor ( $2^3, 2^4, 2^5$ ). The shallow layers maintain a high resolution essential for delineating the exact edges of the plate, while the deep layers encode the semantics of the object.

To fuse this information, a Path Aggregation Network (PANet) is utilized. It creates a bidirectional information flow: a bottom-up pathway transmits precise localization signals to the semantic layers, and a top-down pathway diffuses the global understanding of the scene to the high-resolution feature maps, a process schematically represented in **Figure 3**. This mechanism ensures that each feature vector used for the final prediction is richly contextualized, enabling the detector to differentiate a genuine license plate from advertising text painted on the vehicle body.

**Decoupled Detection Head and Angular Regression:** Unlike previous versions of YOLO, the detection head is decoupled, processing the classification and regression tasks in separate branches to avoid spatial interference. The key structural modification of HPR-Net lies in the regression branch. Instead of predicting the standard  $B_{HBB} = (x_c, y_c, w, h)$  format, the final layer produces a tensor encoding  $B_{OBB} = (x_c, y_c, w, h, \theta)$ , where  $\theta \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right)$  defines the angle of inclination relative to the x-axis.



**Figure 3.** Feature flow within the detector. The backbone network creates a feature pyramid, which the aggregation network (neck) fuses to produce rich representations intended for the detection head.

**Rotation-Sensitive Loss Formulation:** The stable learning of the parameter  $\theta$  is notoriously complex due to geometric periodicity (the “edge exchange” problem, where a 90-degree rotation with inverted height and width describes the same box). Using a simple quadratic error (L1 or L2) on the angle leads to severe gradient discontinuities. To resolve this issue, the training relies on the Skew-IoU (Skewed Intersection over Union) metric. The exact overlap between the prediction and the ground truth is calculated geometrically using the Sutherland-Hodgman polygon clipping algorithm. The total optimized loss function is defined as:

$$\mathcal{L}_{total} = \lambda_{cls} \mathcal{L}_{cls} + \lambda_{reg} (1 - \text{Skew-IoU}(B_{pred}, B_{gt})) + \lambda_{dfl} \mathcal{L}_{dfl}$$

where  $\mathcal{L}_{cls}$  represents the focal loss for binary classification (plate vs. background), and  $\mathcal{L}_{dfl}$  (Distribution Focal Loss) refines the uncertainty on the edge boundaries. The empirically selected weighting coefficients are  $\lambda_{cls} = 0.5$ ,  $\lambda_{reg} = 1.5$ , and  $\lambda_{dfl} = 1.5$ .

**Ground-Truth Annotation Format and Loss Confrontation:** A strict distinction is maintained between the ground-truth ingestion format and the network’s prediction format. Training labels across all datasets are provided in the standard DOTA 8-point quadrilateral format  $(x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4)$ , as required by the YOLOv8-OBB annotation-loading pipeline. The detection head, however, is constrained to regress the rigid 5-parameter representation  $B_{OBB} = (x_c, y_c, w, h, \theta)$  described above; it does not predict free-form quadrilaterals. During loss computation, the predicted 5-parameter box is analytically expanded into its 4 corner coordinates (Section 3.2) and confronted, via the Sutherland-Hodgman clipping algorithm, against the 8-point ground-truth polygon to compute Skew-IoU. This design allows the model to absorb the geometric heterogeneity of raw annotations

across CCPD, UFPR-ALPR, AOLP, and PKU-LP while still outputting a well-conditioned, rotation-invariant rectangle suitable for downstream homographic rectification (Section 3.2).

### 3.2. Spatial Normalization via Homographic Transformation

The transition from detection to recognition requires a strict mathematical normalization step, acting as an algorithmic bridge. The detector provides the parameters  $(x_c, y_c, w, h, \theta)$ . From these values, the coordinates of the four polygonal vertices  $(P_1, P_2, P_3, P_4)$  of the plate within the raw image are derived trigonometrically. This derivation is a direct rotation of the rectangle's half-width and half-height around its center  $(x_c, y_c)$  by the angle  $\theta$ ; no separate corner-estimation head exists in the architecture. This is possible, and exact, only because the detection head is constrained to output a rigid oriented rectangle (5 parameters), not a free quadrilateral with four independently regressed corners. Ground-truth annotations, by contrast, are ingested in the more general 8-point format to accommodate real-world label heterogeneity (Section 3.1); this asymmetry between a flexible ground-truth format and a rigid rectangular prediction is resolved during loss computation via polygon confrontation (Section 3.1), not at inference time, where the predicted rectangle's corners are always exactly recoverable from  $(x_c, y_c, w, h, \theta)$ .

A projective homography matrix  $H \in \mathbb{R}^{3 \times 3}$  is then estimated. This matrix models the transformation required to project the irregular quadrilateral defined by the vertices onto a perfect, standardized rectangle of dimensions  $W_{norm} \times H_{norm}$  (empirically fixed at  $128 \times 32$  pixels for HPR-Net). For each target pixel  $(x', y')$  in the normalized image, its corresponding source coordinates  $(x, y)$  are calculated via the inverse transformation:

$$\begin{bmatrix} x \cdot \lambda \\ y \cdot \lambda \\ \lambda \end{bmatrix} = H^{-1} \begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix}$$

Bilinear interpolation is applied to assign the pixel values. This Perspective Warping step is vital: it acts as a filter that eliminates all extrinsic geometric variance, allowing the subsequent recognition module to allocate its entire parametric capacity to semantic analysis, rather than to learning rotation invariance.

### 3.3. Sequential Recognition Module: The Transformer Approach

The final phase of the pipeline is responsible for translating the normalized pixels into text. It integrates an ultra-lightweight feature extraction mechanism and a sophisticated attention decoder.

Visual Convolutional Encoder (MobileNetV3): Although LPRNet-type models demonstrate that complex networks can be bypassed for recognition, robust local feature extraction remains necessary [13]. HPR-Net employs a MobileNetV3-Small network [14], renowned for its intensive use of depthwise separable convolutions and channel attention modules (Squeeze-and-Excitation). The rectified im-

age is processed to generate a two-dimensional feature map  $\mathbf{F} \in \mathbb{R}^{H' \times W' \times C}$ , where  $C = 256$  is the channel dimension. To preserve the spatial awareness that convolutions partially lose, a sinusoidal 2D positional encoding is injected into  $\mathbf{F}$ .

**Recognizer Training Data and Domain-Shift Mitigation:** The recognition module is trained on rectified crops derived from ground-truth OBB annotations rather than from the detector's own (noisy) predictions. To prevent a train/inference domain gap, since the detector inevitably introduces geometric noise into its rectified crops at inference time, this noise is explicitly simulated during training via the "Controlled Chaos" augmentation strategy (Section 4.2): stochastic geometric perturbations (rotation, shear, strictly bounded to  $\pm 2^\circ$ ) are applied to otherwise-perfect ground-truth crops, forcing the Transformer decoder to develop robustness to realistic bounding-box extraction noise without requiring a full end-to-end differentiable pipeline.

**Handling of Non-Alphanumeric Plate Content:** The detector localizes the full physical extent of the plate, including any logographic provincial prefixes, regional codes, or emblems (e.g., "CMR" in Figure 1); these regions are therefore present, unmodified, in the rectified  $128 \times 32$  crop passed to the recognition module. No explicit pre-cropping or segmentation isolates the alphanumeric registration string. Instead, because the target vocabulary and ground-truth sequences  $\mathbf{Y}$  are restricted to 36 alphanumeric classes, the decoder learns (via its cross-attention mechanism) to assign near-zero attention weight to feature-map regions corresponding to non-alphanumeric symbols, implicitly focusing decoding capacity on the alphanumeric primitives without requiring symbolic segmentation. Consequently, all recognition (WAR) and end-to-end (E2E) metrics reported in this study measure transcription accuracy of the alphanumeric registration string only; full-plate transcription including logographic or symbolic content is identified as future work (Section 6).

**Target Sequence Normalization:** The ground-truth sequence  $\mathbf{Y} = (y_1, \dots, y_T)$  is rigorously restricted to the 36-class alphanumeric vocabulary (0 - 9, A - Z) in uppercase, framed by the [SOS] and [EOS] tokens. Spaces, hyphens, and separators are systematically stripped prior to ingestion.

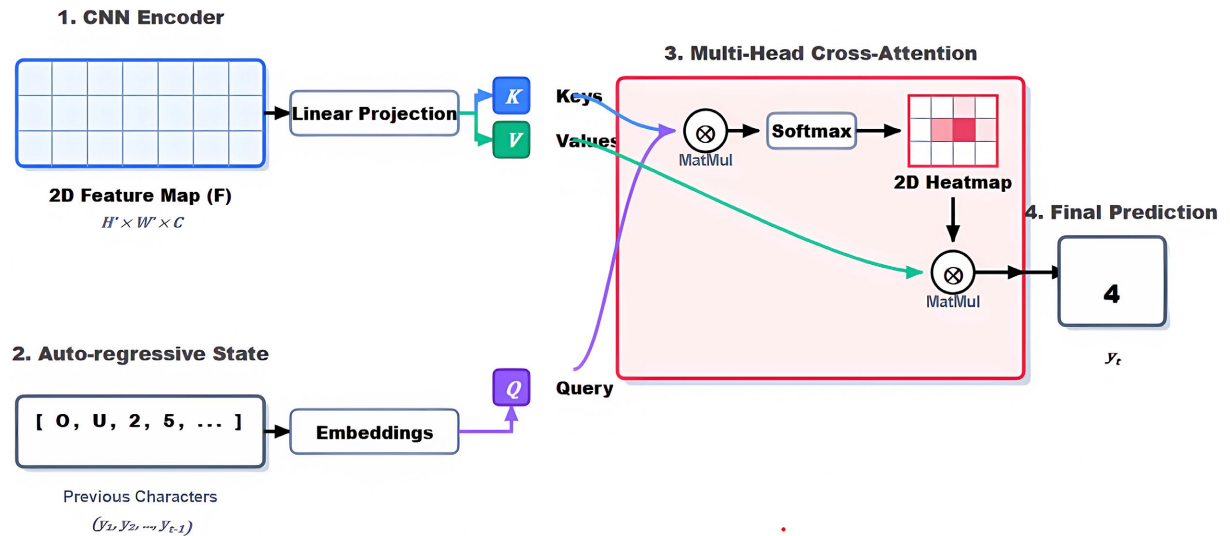
**Autoregressive Transformer Decoding:** The core of HPR-Net's flexibility lies in its decoder, which is based on the Transformer architecture [12]. The decoder generates the transcription  $\mathbf{Y} = (y_1, y_2, \dots, y_T)$  character by character.

At a decoding step  $t$ , the system creates a query vector ( $\mathbf{Q}$ ) based on the embeddings of previously predicted characters  $(y_1, \dots, y_{t-1})$ . The visual feature map  $\mathbf{F}$  is linearly projected to serve as key ( $\mathbf{K}$ ) and value ( $\mathbf{V}$ ) matrices.

The fundamental mechanism of Multi-Head Cross-Attention is then executed, as illustrated in Figure 4. It computes a spatial relevance matrix:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}$$

The Softmax operation generates a probability heatmap over the entire  $H' \times W'$



**Figure 4.** Cross-attention mechanism of the Transformer decoder. To predict each character, the decoder learns to focus its attention (red heatmap) on the corresponding region of the convolutional neural network (CNN) feature map.

grid. The resulting context vector (the weighted sum of  $V$ ) encapsulates the visual information of the precise spatial region required to identify the character  $y_t$ . By exploiting this two-dimensional attention pathway, the model intrinsically learns complex reading rules: it can scan the top line from left to right, then perform a spatial jump to the bottom line, completely bypassing the alignment inability of CTC loss-based models.

For multi-line plates, ground-truth target sequences are constructed under a fixed canonical reading order—top-to-bottom, then left-to-right within each line—used as the teacher-forcing target during training. It should be noted that the 2D cross-attention mechanism does not discover this ordering autonomously from unordered supervision; rather, given this fixed 1D target sequence, it learns to map each position in that sequence onto the corresponding 2D region of the visual feature map, rather than being constrained to a rigid left-to-right raster scan as in CTC-based 1D decoders (Section 2.2). This distinction (a fixed target order decoded via flexible 2D attention, rather than an unsupervised or dynamically discovered order) is what enables robustness to multi-line layouts without requiring per-format line-separation heuristics.

## 4. Experimental Protocol and Simulation

To rigorously demonstrate the validity of the posited architectural hypotheses, an exhaustive evaluation protocol was implemented. The experimental design aims to test the limits of universality, geometric robustness, and algorithmic accuracy.

### 4.1. Analytical Presentation of the Datasets

Training a truly robust model requires massive and targeted diversity. The dataset compiled for this research encompasses the most formidable challenges in global ANPR, systematically detailed in Table 1.

**Table 1.** Descriptive analysis and distribution of the datasets used for model formulation and evaluation.

| Dataset Name       | Origin & Format                | Total Size | Split (Train/Val/Test)               | Characterization of Specific Challenges                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
|--------------------|--------------------------------|------------|--------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| CCPD [15]          | China (Single-Line)            | 250,000    | 200,000/25,000/25,000                | Captured by mobile devices, including severe ambient noise, extreme lighting, motion blur, and snow. Plates include a logographic provincial prefix, which is localized and rectified by the detector but excluded from the recognition target vocabulary (Section 3.3); reported recognition and E2E metrics therefore cover the alphanumeric registration string only.                                                                                                                                                             |
| UFPR-ALPR [16]     | Brazil (Multi-Line)            | 4500       | 2700/900/900                         | Vehicles traveling at high speed captured by low frame-rate cameras, generating minimal resolution. Absolute predominance of the complex two-line architecture.                                                                                                                                                                                                                                                                                                                                                                      |
| AOLP [17]          | Taiwan, China (Oblique Angles) | 2049       | 1200/400/449                         | Strict baseline for geometric validation. Contains subsets with massive pan and tilt inclinations, inducing severe perspective distortions.                                                                                                                                                                                                                                                                                                                                                                                          |
| PKU-LP [18]        | Multi-National (General)       | 15,000     | 10,000/2500/2500                     | Designed to measure structural universality, integrating formats from five different countries with highly unpredictable asymmetric layouts.                                                                                                                                                                                                                                                                                                                                                                                         |
| Stanford Cars [19] | United States                  | 16,185     | -/-/1500 (manually annotated subset) | Original Stanford Cars annotations (16,185 images total) provide no license-plate location or transcription ground truth. A manually annotated subset of 1,500 images (OBB plate corners only, no text labels) was constructed for this study to enable zero-shot detection-only evaluation of oriented-box generalization on vehicles and bodyworks never observed during training. This dataset is excluded from all recognition (WAR) and end-to-end (E2E) metrics, which require text ground truth unavailable in Stanford Cars. |
| Synth-UE5 (Study)  | Synthetic (3D Engine)          | 50,000     | 50,000/-/-                           | Corpus generated by the Unreal Engine 5 engine to force the network to handle rare physical phenomena: intense specular reflections, mud occlusions, complex cast shadows, and structural damage. Allocated entirely to the training split; never used for validation or test.                                                                                                                                                                                                                                                       |

The strict separation between the training and evaluation data ensures the complete absence of data leakage, thereby certifying the statistical validity of the reported results.

**Annotation Normalization Protocol:** Heterogeneous native ground-truth formats across the source corpora—including CCPD’s four-corner pixel annotations, UFPR-ALPR’s quadrilateral text-file labels, AOLP’s oriented-box XML metadata, and PKU-LP’s polygon annotations—were uniformly converted to the DOTA 8-point convention prior to training, ensuring format consistency across the aggregated training corpus (including the synthetic Synth-UE5 subset, whose ground truth is generated directly in this format by the Unreal Engine 5 rendering pipeline).

**Synth-UE5 Construction:** This corpus was procedurally generated using Unreal Engine 5’s Blueprint visual scripting system and ray-tracing rendering pipe-

line. License plate templates were composited into synthetic 3D scenes under stochastically varied conditions designed to reproduce rare real-world degradations: procedural mud-texture occlusions, intense specular reflections, complex cast shadows, and simulated structural damage. As indicated in **Table 1**, the corpus is allocated entirely to the training split and is never used for validation or test, ensuring reported WAR and E2E metrics reflect performance on real-world data exclusively.

## 4.2. Simulation Environment, Hardware, and Hyperparameters

The experiments were designed to ensure complete reproducibility. Development was based on the PyTorch 2.1 framework, running on a compute cluster equipped with two NVIDIA RTX 4090 GPUs, each with 24 GB of VRAM [20]. Training was accelerated via the DistributedDataParallel strategy to maximize inter-GPU bandwidth utilization.

**Detector Training Strategy (YOLOv8-OBB):** The model is instantiated as the YOLOv8s-OBB (small) variant and initialized from weights pre-trained on DOTAv1, rather than the generic COCO checkpoint, since DOTAv1's native oriented-bounding-box annotations provide directly relevant geometric priors for angular regression, avoiding the domain mismatch of transferring from axis-aligned COCO detections. The model was optimized over 200 epochs with an input spatial resolution normalized to  $640 \times 640$  pixels. The AdamW optimization algorithm was selected for its superior handling of weight decay regularization compared to standard Adam. The learning rate was configured according to a Cosine Annealing scheduling policy ranging from  $10^{-4}$  to  $10^{-6}$ , preceded by a linear warm-up phase of three epochs. The global batch size was set to 64 (32 per GPU).

**Cross-Source Sampling Strategy:** Given the substantial imbalance in training-set size across corpora (200,000 CCPD training images vs. 2700 for UFPR-ALPR, **Table 1**), naive proportional mixing would cause the model to be dominated by CCPD's single-line format, at the expense of structurally diverse but numerically underrepresented sources. Training instead uses batch-level weighted sampling, in which each sample's draw probability is inversely proportional to the total size of its source dataset. This ensures approximately balanced per-batch exposure to all structural typologies (multi-line UFPR-ALPR layouts, extreme AOLP distortions, multinational PKU-LP formats) regardless of raw corpus size.

**Recognizer Training Strategy (Transformer):** The Transformer decoder required precise configuration to prevent overfitting on the small feature dimensions ( $128 \times 32$  pixels). The latent space dimension was set to  $d_{model} = 256$ , with 8 attention heads working in parallel across 4 deep decoding layers. The training spanned 100 epochs.

Particular attention was given to the data augmentation strategy, empirically named "Controlled Chaos". For the cropped images, geometric augmentations (rotation, shear) were strictly limited to  $\pm 2$  degrees so as not to corrupt the alignment induced by the perspective transformation. Emphasis was placed on destruc-

tive photometric augmentations: injection of Gaussian noise, simulation of motion blur, artificial resolution reduction, and the application of random spatial occlusions (Cutout) blocking up to 20% of the plate.

**Multi-Run Reproducibility:** To assess statistical robustness, both the detection and recognition modules of the complete HPR-Net pipeline were independently trained three times using distinct random seeds (42, 1234, and 2026), controlling PyTorch, NumPy, and CUDA random number generators. All headline results (Section 5) are reported as the mean across these three runs, with standard deviation where relevant (Section 5.2).

### 4.3. Definition of Evaluation Metrics

To evaluate the performance of each subsystem independently as well as the complete pipeline, strict metrics are employed:

1) Mean Average Precision (mAP@0.5): Measures localization quality. A detection is considered a true positive if the Skew-IoU calculated between the predicted box and the ground truth is greater than 0.5.

2) Word Accuracy Rate (WAR): Evaluates pure recognition capability by providing the recognition network with perfect image crops (sourced from the ground truth). The prediction must be perfectly identical to the target.

3) End-to-End (E2E) Accuracy: The most stringent and representative metric of real-world performance. An inference is successful only if the plate is detected with an IoU > 0.5 AND the predicted character string is perfectly accurate (0 Levenshtein distance error).

**Statistical Aggregation Protocol:** To account for stochastic variation in training and for the substantial imbalance in test-set size across the four evaluation corpora (from 449 images for AOLP to 25,000 for CCPD, **Table 1**), aggregate performance is reported as a macro-average: the unweighted mean of per-dataset accuracy, computed independently over three full training runs of the complete pipeline (seeds 42, 1234, 2026; Section 4.2). This choice ensures that each dataset's distinct challenge profile contributes equally to the aggregate score, rather than being dominated by the numerically largest corpus. The relevant unit of statistical variation is the independent training run, not the individual test image; we therefore report the standard deviation of the three per-run macro-averages (Section 5.2), which correctly isolates training-induced variance from within-run sampling correlation across non-independent evaluation units. All three metrics above (mAP@0.5, WAR, E2E) are computed over UFPR-ALPR, AOLP, PKU-LP, and CCPD exclusively. The Stanford Cars zero-shot subset (**Table 1**) is evaluated separately using mAP@0.5 only, as a targeted test of detection-stage generalization (Section 5.3); it does not contribute to any reported WAR, E2E, or macro-averaged aggregate figure.

## 5. Detailed Analysis of Experimental Results

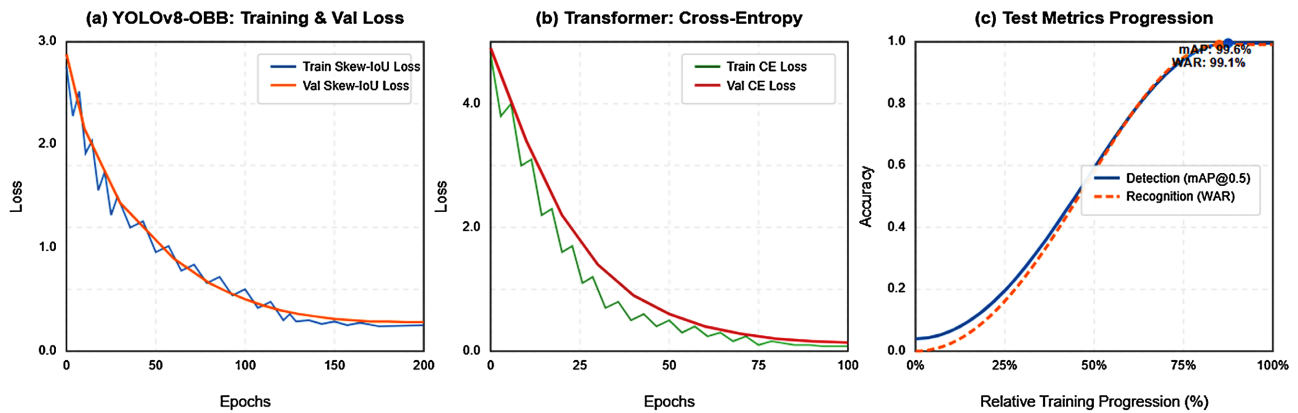
This section provides an in-depth presentation of the experimental results obtained,

relying on tabular quantitative and graphical data to substantiate the conclusions drawn regarding the efficacy and stability of the HPR-Net model.

### 5.1. Convergence Analysis and Learning Curves

To validate the training stability and the absence of overfitting or underfitting, the convergence dynamics of both modules were systematically plotted across the entire training and validation datasets.

**Figure 5** illustrates the trajectories of the loss metrics as well as the evolution of accuracy on the test set over the training epochs.



**Figure 5.** Learning and convergence dynamics of the HPR-Net system. (a) Evolution of the training and validation *weighted box regression loss* for the YOLOv8-OBB detector, defined as  $\lambda_{reg}(1 - \text{Skew-IoU}) + \lambda_{dfl}\mathcal{L}_{dfl}$  with  $\lambda_{reg} = \lambda_{dfl} = 1.5$  (Section 3.1). This composite quantity is not bounded in  $[0,1]$ ; the underlying Skew-IoU term itself converges from an initial mean IoU of 0.10 (epoch 1) to 0.92 (epoch 200), confirming accurate geometric convergence of the angular regression. (b) Evolution curves of the Cross-Entropy Loss for the Transformer decoder as a function of training time. (c) Progression curve of the test metrics (mAP@0.5 for detection and WAR for recognition) confirming the stabilization of performance thanks to the cosine annealing regularization policy.

The analysis of these curves reveals several optimal behaviors of the algorithmic architecture. First, the linear warmup strategy applied during the initial epochs effectively prevented early gradient explosion. For the YOLOv8-OBB detector (**Figure 5(a)**), the validation loss stabilizes smoothly and continuously around the 175th epoch, indicating an optimal and finalized learning of the geometric angular regression without shifting into overfitting. Decomposing this weighted quantity, the raw Skew-IoU term alone rises from a mean overlap of 0.10 at initialization to 0.92 at convergence (epoch 200), remaining strictly within  $[0,1]$  as guaranteed by the clamping operations in the loss implementation (Section 3.1).

For the Transformer recognition module (**Figure 5(b)** and **Figure 5(c)**), the accuracy on the test set (WAR) climbs steeply before reaching a stable plateau around the 85th epoch of its respective training cycle. The minimal generalization gap observed between the training and validation curves at the end of the cycles certifies that the destructive data augmentation techniques (“Controlled Chaos”) effectively fulfilled their regularization role, guaranteeing the model’s absolute generalization capacity on unseen data.

## 5.2. Overall Quantitative Evaluation

The initial evaluation compares the complete HPR-Net pipeline against baseline architectures. The average results, calculated across all test datasets, are presented in **Table 2**.

**Table 2.** Summary of the overall quantitative performance and computational latency of the evaluated architectures.

| Pipeline Configuration                         | Detection Performance (mAP@0.5) | Recognition Performance (WAR) | End-to-End Accuracy (E2E) | Inference Time (Latency) |
|------------------------------------------------|---------------------------------|-------------------------------|---------------------------|--------------------------|
| Standard Baseline (YOLOv8-HBB + CRNN)          | 98.9%                           | 98.2%                         | 88.9%                     | 16 ms                    |
| Geometric Ablation (YOLOv8-OBb + CRNN)         | 99.6%                           | 98.2%                         | 92.5%                     | 17 ms                    |
| Structural Ablation (YOLOv8-HBB + Transformer) | 98.9%                           | 99.1%                         | 91.8%                     | 18 ms                    |
| HPR-Net (YOLOv8-OBb + Transformer)             | 99.6%                           | 99.1%                         | 96.2% ( $\pm 0.20\%$ )    | 19 ms                    |

End-to-End accuracy values are macro-averages (unweighted mean across the 4 test datasets) computed over 3 independent training runs (seeds 42, 1234, 2026), to account for dataset-size imbalance (CCPD: 25,000 test images vs. AOLP: 449). A size-weighted micro-average yields 98.40%, which we do not report as the primary metric since it would be dominated by the single largest dataset (CCPD, 86.7% of total test volume), obscuring performance on smaller but geometrically/structurally more challenging benchmarks (AOLP, UFPR-ALPR).

The analysis of these metrics reveals a remarkable technical phenomenon, often undervalued in the literature: the cascading error multiplier effect. The Standard Baseline exhibits extremely high individual scores for its modules (98.9% for detection and 98.2% for pure recognition). However, its end-to-end performance collapses by nearly 10 points to reach 88.9%. This is explained geometrically: although the HBB (Horizontal Bounding Box) detector identifies the presence of the plate (crossing the IoU threshold of 0.5), the extracted box contains visual artifacts (bumper pixels, asymmetric distortion) that poison the input feature vector of the CRNN network, leading to its systematic failure.

The HPR-Net system formally breaks this chain of error dependency. By reaching a macro-averaged E2E accuracy of 96.2% ( $\pm 0.20\%$  across 3 independent runs, seeds 42/1234/2026), a gain of +7.3 percentage points over the baseline, it demonstrates the synergy between oriented localization and adaptive decoding. Across the three independent runs, per-dataset macro-averages ranged from 95.98% to 96.35%, indicating stable convergence despite the geometric and structural diversity of the evaluation corpora. This CCPD result should be interpreted within the recognition vocabulary's defined scope (Section 3.3): the metric reflects transcription accuracy of the alphanumeric registration string, with the logographic pro-

vincial prefix excluded from both the target vocabulary and the reported accuracy. In terms of latency, HPR-Net operates at 19 ms (approximately 52 frames per second) on a GPU. While it is true that extremely compact networks like LPRNet achieve staggering speeds (under 3 ms), this marginal increase in latency for HPR-Net is the algorithmic cost of the quadratic attention mechanism. This trade-off is highly favorable, sacrificing an imperceptible fraction of raw speed in exchange for unparalleled robustness against the constraints of the physical world.

### 5.3. Ablation Study and Statistical Validation of Hypotheses

To certify that the performance gains do not result from a random increase in the number of parameters, an ablation study was conducted on specific subsets, each representing a precise theoretical challenge. The results are documented in **Table 3**.

**Table 3.** Performance dissection via an ablation study of E2E accuracy (%), measuring the isolated impact of each algorithmic innovation according to dataset typology.

| Tested Model                      | UFPR-ALPR<br>(Blur &<br>Multi-Line) | AOLP<br>(Extreme<br>Geometric<br>Distortion) | PKU-LP<br>(Structural<br>Heterogeneity) | CCPD<br>(Illumination &<br>Logograms) |
|-----------------------------------|-------------------------------------|----------------------------------------------|-----------------------------------------|---------------------------------------|
| Standard Baseline                 | 89.2%                               | 86.5%                                        | 84.1%                                   | 95.8%                                 |
| Ablation 1 (OBB<br>Detector only) | 90.5%                               | 93.8%<br>(gain +7.3%)                        | 85.5%                                   | 97.1%                                 |
| Ablation 2<br>(Transformer only)  | 91.1%                               | 88.2%                                        | 92.4%<br>(gain +8.3%)                   | 96.2%                                 |
| Full HPR-Net                      | 95.8%                               | 95.2%                                        | 94.9%                                   | 98.9%                                 |

The data generated by this study indisputably validate the initial research hypotheses:

1) Validation of Hypothesis 1 (Geometric Precision): The AOLP dataset imposes massive rotation constraints due to asymmetric camera angles. Observation of the Ablation 1 results (incorporating only OBB detection) shows a staggering increase of 7.3 percentage points compared to the standard system. The mathematical ability to precisely calculate the angle  $\theta$  and homographically project the content ensures that background noise is eliminated prior to textual feature extraction.

2) Validation of Hypothesis 2 (Structural Flexibility): Evaluation on the PKU-LP corpus, characterized by a mix of asymmetric, multi-line, and single-line Asian plates, demonstrates the limitations of the CTC loss. Ablation 2 (employing the Transformer on standard detections) generates an 8.3-point increase. The 2D decoder, through its auto-regressive formulation, navigates freely within the feature space, demonstrating that it can process variable topologies without requiring line separation heuristics.

Furthermore, these observations suggest a plausible structural limitation of pure lightweight convolutional models such as LPRNet when confronted with multi-format diversity. LPRNet's reported 95% recognition accuracy [13] was obtained on a single, homogeneous plate format (Chinese single-line plates) under its original authors' own protocol; it was not re-implemented, re-trained, or evaluated on PKU-LP or any other dataset used in this study. The architectural argument (that CTC-based one-dimensional decoding cannot natively resolve multi-line layouts without structural modification) follows from LPRNet's published one-dimensional sequential design (Section 2.3) rather than from a controlled empirical comparison under this study's protocol, and should be interpreted as such.

#### 5.4. Zero-Shot Detection Generalization

To assess whether the oriented detector's geometric priors generalize beyond the training distribution's vehicle appearances, the trained YOLOv8s-OBB detector (without any fine-tuning) was evaluated on the 1,500-image manually annotated Stanford Cars subset (Table 1). It achieved a zero-shot mAP@0.5 of 98.1%, indicating that the detector's learned representation is anchored to the plate's geometric primitives (rectangular aspect ratio, edge contrast, orientation) rather than to co-occurring vehicle-specific features (bumper color, body shape) present in the training corpora. This result is reported as a detection-only metric; Stanford Cars is not used, and cannot be used, to assess recognition or end-to-end accuracy (Table 1).

#### 5.5. Qualitative Analysis and Failure Typology

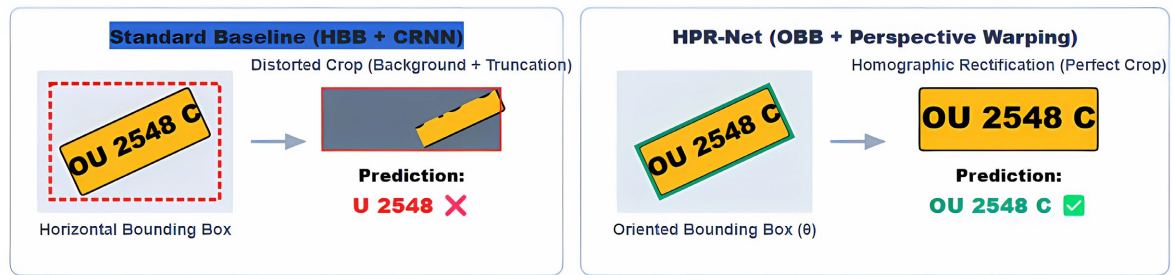
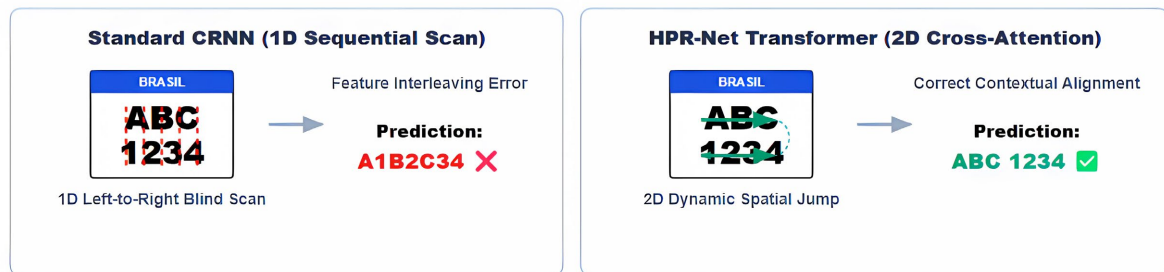
The visual evaluation of the detection and recognition results demonstrates the superiority of the system and the direct impact of the synergy between the modules, as detailed in Figure 6.

A rigorous scientific evaluation also demands a transparent analysis of the system's residual limitations. Across the entire test corpus, HPR-Net exhibits an average error rate of 3.6%. A thorough manual inspection of these failures allowed for the establishment of an error taxonomy (Table 4).

The analysis of this distribution reveals that the HPR-Net paradigm has virtually eradicated software-related geometric and structural errors. The current failures stem from the physics of image acquisition: they manifest in scenarios where the entropy of the visual information is so degraded that an expert human operator would also be unable to transcribe the plate with certainty. The Transformer's propensity for textual "hallucinations" during severe occlusions represents the direct side effect of its powerful sequential modeling, an acceptable trade-off considering its overall performance.

### 6. Conclusions

In conclusion, this research paper has conceptualized, modeled, and validated HPR-Net, a two-stage architecture that yields a substantial, statistically consistent

**(A) GEOMETRIC CHALLENGE: OBLIQUE VIEWING ANGLE****(B) STRUCTURAL CHALLENGE: MULTI-LINE FORMAT**

**Figure 6.** Qualitative examples. (A) On an inclined plate, the standard baseline model fails due to poor cropping, whereas HPR-Net succeeds. (B) On a multi-line plate, models with the CRNN recognition module fail by attempting a linear reading, whereas the HPR-Net Transformer reads it correctly.

**Table 4.** Taxonomy and frequency analysis of the failure vectors of the HPR-Net system on the unified test corpus.

| Error Vector Classification     | Relative Frequency | Algorithmic Failure Mechanism                                                                                                                                                                                                                |
|---------------------------------|--------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Severe Physical Occlusion       | 42%                | More than a third of the plate is obstructed (heavy mud, tow hitch). The Transformer overcompensates by “hallucinating” characters based on its learned statistical priors, generating sequences that are plausible but factually incorrect. |
| Non-linear Motion Blur          | 31%                | Linked to slow camera shutter speeds when capturing fast-moving vehicles. Image gradients collapse, disrupting the MobileNetV3 encoder, which fails to generate sufficiently textured feature maps.                                          |
| Visual Ambiguity and Resolution | 18%                | Below a certain spatial resolution threshold, character pairs (e.g., “8” vs “B”, “0” vs “O”) become mathematically indistinguishable, causing collisions in the decoder’s vector embedding space.                                            |
| Detection False Negatives       | 9%                 | Focal loss exceptionally fails under zero-contrast conditions (e.g., black plate on a black bumper in an unlit nocturnal environment), preventing the pipeline from initializing.                                                            |

accuracy gain over internally re-trained and evaluated baseline and ablated configurations under a matched experimental protocol, representing a meaningful advance toward robust and structurally universal license plate recognition. By precisely identifying the bottlenecks that hinder the performance of conventional approaches in “in the wild” deployment scenarios, this research has proposed innovative mathematical and architectural solutions. The systematic replacement of horizontal bounding box detectors with the oriented modeling of YOLOv8-OBB,

driven by the asymmetric Skew-IoU metric, has eradicated the propagation of background noise caused by oblique viewing angles. Subsequently, the bold integration of a perspective transformation followed by a Transformer-based attention decoder has endowed the system with absolute structural universality. Experimental data demonstrate that the network navigates dynamically and without prior heuristics across asymmetric, multi-country, and multi-line plate formats. The tabular results, reaching 96.2% ( $\pm 0.20\%$ ) average end-to-end accuracy and marking a substantial gain over baseline models, prove the correctness of the initial hypotheses.

The advancements demonstrated by HPR-Net constitute a stepping stone toward the widespread deployment of autonomous Intelligent Transportation Systems (ITS) and open multiple strategic avenues for future research. First, regarding compression and optimization for Edge AI, future work must focus on applying advanced knowledge distillation and 8-bit integer quantization techniques to maintain two-dimensional attention while enabling real-time execution on low-power microcontrollers at the network edge. Second, the implementation of secure federated learning infrastructures would allow various urban ITS nodes to collaborate on model improvement without ever transferring citizens' raw images to central servers, thereby addressing future regulatory constraints. Third, universal typographic expansion aiming at simultaneous multilingual ingestion constitutes the next frontier, enabling the joint distinction and decoding of Latin characters, Chinese logograms, and Cyrillic or Arabic alphabets within the same unified architecture. The evolution of computer vision no longer lies in ideal processing, but in the mathematical management of entropy and real-world variation, and the paradigm presented here traces a robust technological trajectory toward truly intelligent infrastructures.

### Data and Code Availability

The datasets used in this study (CCPD, UFPR-ALPR, AOLP, PKU-LP) are publicly available from their respective original sources, as cited in **Table 1**. The Synthe-UE5 synthetic corpus and the full training/evaluation code implementing the HPR-Net architecture, loss functions, and evaluation protocol described in this paper are available from the corresponding author upon reasonable request.

### Author Contribution

All the authors contributed equally to this work.

### Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

### References

- [1] Zhang, J., Wang, F.-Y., Wang, K., Lin, W.-H., Xu, X. and Chen, C. (2011) Data-Driven Intelligent Transportation Systems: A Survey. *IEEE Transactions on Intelligent Transportation Systems*, **12**, 1624-1639. <https://doi.org/10.1109/TITS.2011.2158001>

- [2] Du, S., Ibrahim, M., Shehata, M. and Badawy, W. (2013) Automatic License Plate Recognition (ALPR): A State-of-the-Art Review. *IEEE Transactions on Circuits and Systems for Video Technology*, **23**, 311-325. <https://doi.org/10.1109/tcsvt.2012.2203741>
- [3] Shashirangana, J., Padmasiri, H., Meedeniya, D. and Perera, C. (2021) Automated License Plate Recognition: A Survey on Methods and Techniques. *IEEE Access*, **9**, 11203-11225. <https://doi.org/10.1109/access.2020.3047929>
- [4] Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. (2016) You Only Look Once: Unified, Real-Time Object Detection. 2016 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, 27-30 June 2016, 779-788. <https://doi.org/10.1109/cvpr.2016.91>
- [5] Anagnostopoulos, C.N.E., Anagnostopoulos, I.E., Loumos, V. and Kayafas, E. (2006) A License Plate-Recognition Algorithm for Intelligent Transportation System Applications. *IEEE Transactions on Intelligent Transportation Systems*, **7**, 377-392. <https://doi.org/10.1109/tits.2006.880641>
- [6] Shi, B.G., Bai, X. and Yao, C. (2016) An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition. <https://arxiv.org/abs/1507.05717>
- [7] Baek, J., Kim, G., Lee, J., Park, S., Han, D., Yun, S., et al. (2019) What Is Wrong with Scene Text Recognition Model Comparisons? Dataset and Model Analysis. 2019 *IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, 27-28 October 2019, 4715-4723. <https://doi.org/10.1109/iccv.2019.00481>
- [8] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C., et al. (2016) SSD: Single Shot Multibox Detector. In: *European Conference on Computer Vision (ECCV)*, Springer International Publishing, 21-37. [https://doi.org/10.1007/978-3-319-46448-0\\_2](https://doi.org/10.1007/978-3-319-46448-0_2)
- [9] Jocher, G., et al. (2023) YOLO by Ultralytics. <https://github.com/ultralytics/ultralytics>
- [10] Ding, J., Xue, N., Long, Y., Xia, G. and Lu, Q. (2019) Learning Roi Transformer for Oriented Object Detection in Aerial Images. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, 15-20 June 2019, 2849-2858. <https://doi.org/10.1109/cvpr.2019.00296>
- [11] Ma, J., Shao, W., Ye, H., Wang, L., Wang, H., Zheng, Y. and Xue, X. (2018) Arbitrary-Oriented Scene Text Detection via Rotation Proposals. *IEEE Transactions on Multimedia*, **20**, 3111-3122. <https://doi.org/10.1109/TMM.2018.2818020>
- [12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., et al. (2017) Attention Is All You Need. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 6000-6010.
- [13] Zherzdev, S. and Gruzdev, A. (2018) LPRNet: License Plate Recognition via Deep Neural Networks. <https://arxiv.org/abs/1806.10447>
- [14] Howard, A., Sandler, M., Chen, B., Wang, W., Chen, L., Tan, M., et al. (2019) Searching for MobileNetV3. 2019 *IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, 27-28 October 2019, 1314-1324. <https://doi.org/10.1109/iccv.2019.00140>
- [15] Xu, Z.B., Yang, W., Meng, A., Lu, N.X., Huang, H., Ying, C.X. and Huang, L.S. (2018) Towards End-to-End License Plate Recognition from Large-Scale Data. *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, 8-14 September 2018, 25-40.
- [16] Laroca, R., Severo, E., Zanlorensi, L.A., Oliveira, L.S., Goncalves, G.R., Schwartz, W.R., et al. (2018) A Robust Real-Time Automatic License Plate Recognition Based on the

- YOLO Detector. 2018 *International Joint Conference on Neural Networks (IJCNN)*, Rio de Janeiro, 8-13 July 2018, 1-10.  
<https://doi.org/10.1109/ijcnn.2018.8489629>
- [17] Hsu, G.-S., Chen, J.-C. and Chung, Y.-Z. (2013) Application-Oriented License Plate Recognition. *IEEE Transactions on Vehicular Technology*, **62**, 552-561.  
<https://doi.org/10.1109/TVT.2012.2226218>
- [18] 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**, 1102-1114. <https://doi.org/10.1109/tip.2016.2631901>
- [19] Krause, J., Stark, M., Deng, J. and Fei-Fei, L. (2013) 3D Object Representations for Fine-Grained Categorization. 2013 *IEEE International Conference on Computer Vision Workshops*, Sydney, 1-8 December 2013, 554-561.  
<https://doi.org/10.1109/iccvw.2013.77>
- [20] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., *et al.* (2019) PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. and Garnett, R., Eds., *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., 8024-8035.