Building an Adaptive Vietnamese License Plate Recognition and Retrieval System using Multi-Task Deep Learning

Phuoc Minh Hieu PHAM[1], Sy Sieu CAO[1], Le Phu Trung HUYNH[1*]
[1]University of Management and Technology HCMC
*Corresponding author: trung.huynhlephu@umt.edu.vn

Abstract

Automatic License Plate Recognition (ALPR) is an essential component of intelligent transportation, yet its performance is often significantly degraded by real-world image distortions and regional plate format complexities. This research addresses these challenges by proposing a highly adaptive, multi-task deep learning framework specifically designed for the Vietnamese license plate context. The system targets the unique diversity of Vietnamese plates while robustly handling low-quality image inputs.
The proposed framework operates as a multi-stage, conditional pipeline. First, a real-time object detection model is employed to localize all license plate instances. The core component of the system is a lightweight Quality Assessment Module (QAM), which acts as an intelligent router, analyzing and classifying each detected plate into one of three distinct categories: “clear”, “restorable”, or “unrestorable”. The system’s adaptive nature is demonstrated in the subsequent multi-branch routing: “restorable” images are selectively forwarded to a specialized restoration neural network. Conversely, “clear” images bypass this resource-intensive step. Finally, “unrestorable” images are rejected entirely, preventing erroneous processing and optimizing overall system throughput. A robust Optical Character Recognition (OCR) model is then used to transcribe the character string from both “clear” and successfully “restored” plates. Finally, the recognized string is used as a query key for retrieving vehicle information from an associated database. This multi-task approach—integrating detection, quality assessment, conditional restoration, and recognition—demonstrates significant accuracy improvements under challenging real-world conditions compared to traditional, non-adaptive pipelines. The system provides a robust and efficient solution for practical ALPR and information retrieval applications within the specific context of Vietnam.

Keywords— Vietnamese License plate recognition, multi-task learning, computer vision.

I. INTRODUCTION

Automatic License Plate Recognition (ALPR) serves as a cornerstone technology within modern Intelligent Transportation Systems (ITS). Its applications are extensive and critical, underpinning systems for automated toll collection, traffic law enforcement, smart parking management, and vehicle access control. The efficacy of these applications hinges on the system’s ability to provide accurate and real-time vehicle identification.

However, the performance of conventional ALPR systems degrades significantly when deployed in unconstrained real-world environments. These systems often fail when faced with a wide spectrum of image quality issues, including motion blur from fast-moving vehicles, severe glare from headlights or sunlight, low-light noise, and geometric distortions from oblique camera angles.

This research specifically addresses the challenges within the Vietnamese license plate context. This environment presents unique regional complexities, including a high diversity of plate formats, various background colors (e.g., white, blue, yellow), differing character layouts. A "one-size-fits-all" ALPR solution is often insufficient to handle this variability, leading to high error rates.

II. RELATED WORK

The proposed system comprises four sequential yet adaptive stages: (1) license plate detection, (2) image quality assessment and routing, (3) conditional image restoration, and (4) character recognition and retrieval.

2.1 *License plate detection*

Despite these advancements, existing detectors assume relatively clean input images and do not adapt to severe degradations prevalent in Vietnamese traffic surveillance systems, including extreme viewpoint angles, specular glare, motion blur, and low-resolution sensors [1]. For instance, while segmentation-based methods on edge devices [2] and OpenALPR adaptations [3] have been proposed for local deployment, they remain vulnerable to real-world distortions—particularly under extreme illumination variations such as daytime specular reflections from sunlight and nighttime low-contrast imaging due to poor lighting or headlight glare. Moreover, none incorporate pre-processing quality assessment to filter or route degraded inputs prior to recognition. This critical limitation motivates our proposed adaptive architecture. Our framework first employs a robust detector (YOLOv8-nano [4]...) and then introduces a novel Quality Assessment Module (QAM) as a distinct downstream processing step to route inputs.

2.2 *Image quality assessment and routing*

Assessing the quality of detected license plate images is crucial for robust ALPR under real-world degradations. No-reference image quality assessment (NR-IQA) methods have gained prominence due to their applicability without pristine references. Early deep learning approaches leveraged CNNs to predict perceptual quality scores [6, 7], while recent works incorporate geometric priors [8] or lightweight architectures for efficiency [9]. These methods, however, are primarily designed for general image aesthetics or distortion classification, not for task-specific routing in ALPR pipelines.

Illumination extremes pose a particularly acute challenge: daytime glare saturates plate regions, while nighttime captures suffer from noise and low signal-to-noise ratio (SNR). Although specialized enhancement models like U-Net-based day/night preprocessing have been explored [10], they are applied uniformly and cannot distinguish between recoverable and irreparable degradation—leading to unnecessary computation or persistent OCR failures.

Although multi-angle detection model have been proposed [1], none integrate real-time quality evaluation with conditional, illumination-aware restoration. Our multi-branch design addresses this gap by optimizing both accuracy and computational efficiency across diurnal cycles.

2.3 *Conditional image restoration*

Restoring degraded license plate images is essential for improving OCR accuracy in low-quality inputs. Classical techniques, such as noise modeling and blur estimation [14], rely on hand-crafted priors and perform poorly under complex, compound distortions. Traditional machine learning approaches [15] offer marginal improvements but lack generalization.

The advent of deep learning has introduced more powerful paradigms: convolutional neural networks (CNNs) [16, 15] capture local patterns effectively but suffer from limited receptive fields; Transformer-based models [17, 18, 19, 15] excel at modeling long-range dependencies through self-attention, though vanilla Vision Transformers (ViT) incur prohibitive computational cost on high-resolution inputs. Advanced generative methods, including GANs and diffusion models [15], achieve state-of-the-art perceptual quality but risk introducing artifacts (e.g., hallucinated characters) and demand significant resources—unsuitable for real-time ALPR.

Despite extensive progress in general image restoration [15], no prior ALPR system integrates task-aware quality routing with conditional, text-preserving restoration. This gap underscores the novelty of our adaptive pipeline.

2.4 *Character recognition and retrieval*

Optical Character Recognition (OCR) is the final critical stage of ALPR. Early end-to-end frameworks, exemplified by CRNN [20], established an effective baseline by combining convolutional features with recurrent sequence modeling.

Recent advances, however, have shifted towards Transformer-based architectures. Models like TrOCR [25] treat text recognition as an image-to-sequence problem, while others like SwinTextSpotter [21] synergize detection and recognition under a Swin Transformer backbone. Among these, PARSeq [22] has shown high robustness by reframing the task as a permuted autoregressive sequence problem, making it effective against occlusion.

Despite their strong performance, these state-of-the-art models are typically optimized for general scene text. They do not account for the unique typographic and linguistic constraints of specific domains, such as the fixed-length strings, mixed alphanumeric formats, and regional variations found in Vietnamese license plates. This research gap motivates the need for an architecture that is not only robust but also specifically adapted to the ALPR domain.

III. METHODOLOGY

Building upon the limitations identified in the related work, this section presents the adaptive multi-task ALPR framework specifically engineered for Vietnamese license plates under real-world imaging constraints. The system dynamically adjusts its processing path based on input quality, eliminating redundant operations while maximizing recognition accuracy.

3.1 *Overall System Architecture*

The proposed pipeline comprises four interdependent modules executed in a conditional, multi-branch manner, as depicted in Fig. 1. Unlike conventional rigid ALPR systems that apply uniform processing regardless of input quality, our framework incorporates an intelligent routing mechanism—the *Quality Assessment Module (QAM)*—to classify each detected plate and selectively activate resource-intensive restoration only when necessary.

The processing flow is defined as follows:

1. **License Plate Detection**: A real-time *YOLOv8-nano* detector localizes all plate instances in the input frame [1, 2].
2. **Image Quality Assessment and Routing**: The *QAM*, built on *MobileNetV3-Small*, classifies each cropped plate into one of three categories: "*clear*", "*restorable*", or "*unrestorable*" [8, 9, 12, 27].
3. **Conditional Image Restoration**: Only "restorable" plates are processed by a fine-tuned *Swin2SR* model to recover textual fidelity [19].
4. **Character Recognition and Retrieval**: A *CRNN* with CTC decoding transcribes the (restored or original) plate, followed by structured database lookup [24].

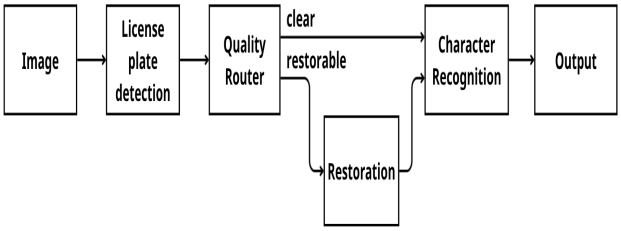

Figure 1: Block diagram of the proposed adaptive ALPR pipeline. The *QAM* enables conditional routing: clear plates bypass restoration, restorable plates are enhanced via *Swin2SR*, and unrestorable plates are rejected to prevent error propagation.

3.2 *License Plate Detection (Module 1)*

The primary objective of this module is to accurately localize all license plate instances within a full-frame input image, enabling robust downstream processing under real-world traffic conditions.

- **Model Selection**: We adopt *YOLOv8-nano* [4] as the detection backbone, building upon the foundational real-time object detection paradigm introduced in YOLO [5]. This variant is selected for its optimal trade-off between inference speed and mean Average Precision (mAP) on small, densely packed objects—critical for multi-vehicle scenes in Vietnamese urban environments. Fine-tuning on a diverse dataset of annotated Vietnam traffic images (including day/night, rain, and motion blur) further enhances robustness to regional plate variations and imaging distortions [2, 1].

- **Input and Output**: The module accepts a full-resolution input image $I \in \mathbb{R}^{H\times W\times 3}$ and outputs a set of bounding boxes $B = \{(x_{1,i}, y_{1,i}, x_{2,i}, y_{2,i}, c_i)\}_{i=1}^{N}$, where N is the number of detected plates and c_i denotes confidence score. Non-maximum suppression (NMS) with IoU threshold 0.4 eliminates redundant detections.

- **Data Flow**: For each bounding box B_i, the corresponding plate region $P_i = \text{crop}(I, B_i)$ is extracted and resized to a fixed dimension (128×64) while preserving aspect ratio via padding. These cropped patches serve as direct input to Module 2: Image Quality Assessment and Routing (QAM), enabling adaptive processing based on degradation severity.

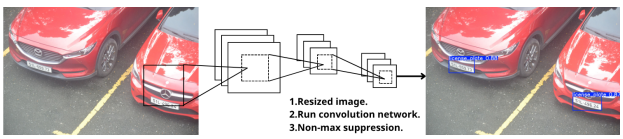

Figure 2: The YOLO detection pipeline [5]. The input image is resized, processed through a convolutional backbone, and refined via non-maximum suppression (NMS) to produce confident license plate bounding boxes. This mechanism forms the core of *Module 1*, enabling real-time localization in multi-vehicle scenes.

3.3 *Image Quality Assessment and Routing (Module 2)*

This module represents the core component of the proposed adaptive framework. Instead of processing every detected plate uniformly, this module functions as an intelligent router. Its primary task is to perform Blind Image Quality Assessment (BIQA), a concept explored in various deep learning contexts [6, 7], to determine the optimal subsequent processing path. This adaptive routing is critical for both maximizing accuracy and optimizing computational throughput.

3.3.1 Model Selection for High-Throughput Routing

The QAM must be exceptionally fast, as it analyzes every detected plate. A heavy or slow model at this stage would create a significant bottleneck for the entire system. Therefore, a lightweight Convolutional Neural Network (CNN) is required, echoing recent trends in efficient IQA model design [9, 27].

To enable this adaptive processing, we evaluated several candidate backbones: ResNet50, MobileNetV2 [11], EfficientNet-B0 [12], and MobileNetV3 [13]. We selected MobileNetV3-Small due to its superior efficiency on edge devices, achieving 3.2× fewer parameters and 2.8× lower FLOPs than EfficientNet-B0 while maintaining comparable classification accuracy.

This model is specifically engineered for high performance in resource-constrained environments [13]. It achieves its efficiency through novel architectural components such as the inverted residual linear bottleneck, a concept introduced in MobileNetV2 [11] and refined with h-swish activation functions. This architecture, detailed in Figure 3, provides a superior trade-off between accuracy and latency, making it an ideal choice for our rapid classification task.

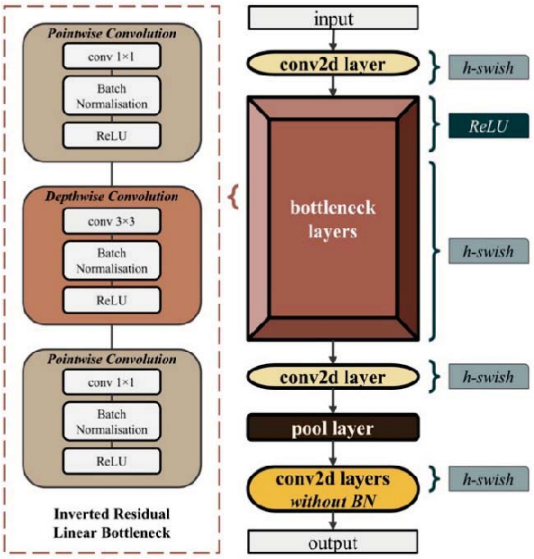

Figure 3: The general architecture of MobileNetV3, illustrating the inverted residual linear bottleneck blocks that enable its computational efficiency [13]

3.3.2 Classification Task and Routing Logic

The QAM is trained as a 3-class classifier, receiving the cropped license plate image from Module 1 and assigning it to one of three distinct categories:

1. **"Clear":** Images with high resolution, sharp focus, and no significant geometric distortion.
2. **"Restorable":** Images with moderate degradation (e.g., motion blur, glare, high skew angles) but still containing sufficient underlying information for successful restoration.
3. **"Unrestorable":** Images that are severely degraded (e.g., extreme low resolution, heavy occlusion, or complete motion blur) where recovery is deemed impossible.

This classification directly dictates the "adaptive" routing logic of the pipeline, as illustrated in Figure-4.

- **"Clear"** images bypass the restoration module entirely and are sent directly to Module 4 (Character Recognition), saving computational resources.
- **"Restorable"** images are forwarded to Module 3 (Conditional Image Restoration) for enhancement.
- **"Unrestorable"** images are rejected (Removed) from the pipeline, preventing the system from processing "garbage" data and producing a highly confident but incorrect result.

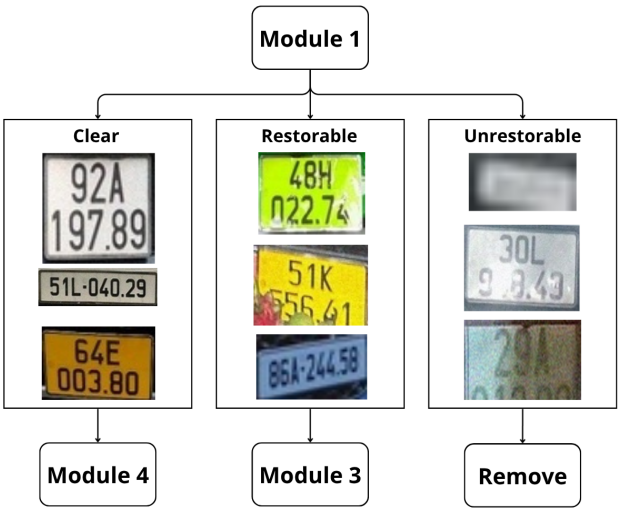

Figure 4: The 3-branch conditional routing logic of the QAM. Based on the classification, an image is either sent for recognition, restoration, or removed

3.4 *Conditional Image Restoration (Module 3)*

This module is only activated when QAM (Module 2) classifies an image as "Restorable". Its main goal is not to produce an aesthetically pleasing image, but to produce a text-legible image, in particular preserving textual integrity and avoiding the creation of hallucinated artifacts.

3.4.1 Model Selection

As analyzed in Section 2.3 (Related Work), GAN and Diffusion models, although powerful, have the risk of fabricating details, an unacceptable risk for OCR. Therefore, we need an efficient Transformer-based model.

We chose Swin2SR [19] as the backbone architecture for this module. Swin2SR combines the hierarchical Swin Transformer V2 architecture with compression super-resolution targets, making it ideal for our constraints:

1. **Efficiency**: It is designed for efficient processing, suitable for the requirements of a multi-stage pipeline.

3.5 *Character Recognition and Retrieval (Module 4)*

The final module is responsible for transcribing the character string from the processed plate images (either "clear" originals or "restored" outputs from Module 3).

As discussed in Section-2.4, while traditional CRNN [20] offers efficiency, its sequential RNN component can be less robust to the non-linear distortions and noise that may still be present even after restoration.

Therefore, to maximize accuracy, we adopt a state-of-the-art, Transformer-based backbone: PARSeq [22]. By leveraging the parallel and context-aware nature of Transformer attention mechanisms, PARSeq is inherently more robust to the partial occlusions, geometric skew, and artifacts that our restoration module may not have perfectly corrected. This choice aligns with the state-of-the-art in text recognition, which favors Transformer architectures for their superior performance [26].

To address the "general-purpose" limitation identified in our Related Work (Section-2.4), we do not use the pre-trained PARSeq model directly. Instead, the model is extensively fine-tuned on a mixed dataset of real and synthetically degraded Vietnamese plates. This critical step adapts the powerful general architecture to the specific typographic and contextual rules of the Vietnamese ALPR domain.

IV. CONCLUSION

V. REFERENCES

[1] D. Tran-Anh, K. L. Tran, and H.-N. Vu, “License plate recognition based on multi-angle view model,” *ArXiv*, vol. abs/2309.12972, 2023.

[2] D. D. Nguyen, T. S. Vo, M. H. Le, and M.-S. Nguyen, “An integration of segmentation technique on edge devices for license plate recognition,” *Ministry of Science and Technology, Vietnam*, 2025.

[3] H. Tran, G. Ma, T. Nguyen, and T. Cao, “Building vietnam’s license plate recognition system based on openalpr,” *International Journal of Multidisciplinary Research and Publications*, 2023.

[4] M. Yaseen, “What is yolov8: An in-depth exploration of the internal features of the next-generation object detector,” *ArXiv*, vol. abs/2408.15857, 2024.

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[6] S. Bianco, L. Celona, P. Napoletano, and R. Schettini, "On the use of deep learning for blind image quality assessment," *arXiv preprint arXiv:1602.05531*, April 2017.

[7] H. Talebi and P. Milanfar, "Nima: Neural image assessment," *arXiv preprint arXiv:1709.05424*, April 2018.

[8] N.-H. Shin, S.-H. Lee, and C.-S. Kim, "Blind image quality assessment based on geometric order learning," pp. 12799–12808, 2024.

[9] N. B. Nasim Jamshidi Avanaki, Abhijay Ghildyal and S. Zadtootaghaj, "Lar-iqa: A lightweight, accurate, and robust no-reference image quality assessment model," *arXiv preprint arXiv:2408.17057v2*, September 2024.

[10] P. N. Chowdhury, P. Shivakumara, R. Raghavendra, U. Pal, T. Lu, and M. Blumenstein, "A new u-net based license plate enhancement model in night and day images," in *Asian Conference on Pattern Recognition*, 2019.

[11] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, 2018.

[12] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *ArXiv*, vol. abs/1905.11946, 2019.

[13] S. Qian, C. Ning, and Y. Hu, "Mobilenetv3 for image classification," *2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, pp. 490–497, 2021.

[14] M. Maru and M. C. Parikh, "Image restoration techniques: A survey," February 2017.

[15] P. K. H. S. A. T. Nikita Singhal, Anup Kadam and Pranay, "Study of recent image restoration techniques: A comprehensive survey," June 2025.

[16] Y. L. E. B. M. D. Peng Liu, Xiaoxiao Zhou and R. Fang, "Image restoration using deep regulated convolutional networks," *arXiv preprint arXiv:1910.08853v2*, June 2024.

[17] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," *arXiv preprint arXiv:2108.10257*, August 2021.

[18] L. Agnolucci, L. Galteri, M. Bertini, and A. D. Bimbo, "Restoration of analog videos using swin-unet," in *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2022.

[19] M. V. Conde, U.-J. Choi, M. Burchi, and R. Timofte, "Swin2SR: Swinv2 transformer for compressed image super-resolution and restoration," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2022.

[20] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 2298–2304, 2015.

[21] M. Huang, Y. Liu, Z. Peng, C. Liu, D. Lin, S. Zhu, N. J. Yuan, K. Ding, and L. Jin, "Swintextspotter: Scene text spotting via better synergy between text detection and text recognition," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4583–4593, 2022.

[22] D. Bautista and R. Atienza, "Scene text recognition with permuted autoregressive sequence models," in *European Conference on Computer Vision*, 2022.

[23] R. Baena, S. Kalleli, and M. Aubry, "General detection-based text line recognition," *ArXiv*, vol. abs/2409.17095, 2024.

[24] W. Yu, M. Ibrayim, and A. Hamdulla, "Scene text recognition based on improved crnn," *Inf.*, vol. 14, p. 369, 2023.

[25] M. Li, T. Lv, L. Cui, Y. Lu, D. A. F. Florêncio, C. Zhang, Z. Li, and F. Wei, "Trocr: Transformer-based optical character recognition with pre-trained models," *ArXiv*, vol. abs/2109.10282, 2021.

[26] S. Fang, H. Xie, Y. Wang, Z. Mao, and Y. Zhang, "Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition," *2021 IEEE/CVF*

Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7094–7103, 2021.

[27] J. H. Joloudari, B. Mesbahzadeh, O. Zare, E. Arslan, R. Alizadehsani, and H. Moosaei, "No-reference image contrast assessment with customized efficientnet-b0," *arXiv preprint arXiv:2509.21967*, 2025.