Lightweight Dual-YOLOv8 Instance-Aware Semantic Segmentation for Real-Time Autonomous Driving on Edge ARM/GPU Platforms

Safa Teboulbi[1], Seifeddine Messaoud[2], Mohamed Ali Hajjaji[3], Mohamed Atri[4], Abdellatif Mtibaa[5]
Laboratory of SI2E-University of Sfax-ISIMM of Monastir, University of Monastir, Monastir 5000, Tunisia[1]
Laboratory of Condensed Matter and Nanoscience (LR11ES40)-Faculty of Sciences of Monastir,
University of Monastir, Monastir 5019, Tunisia[2]
Laboratory of RLANTIS-FSM, University of Monastir, Tunisia[3]
ISSAT of Sousse, University of Sousse, Sousse 4003, Tunisia[3]
Computer Engineering Department-College of Computer Science, King Khalid University, Abha 61421, Saudi Arabia[4]
Laboratory of SI2E-National Engineering School of Sfax, University of Sfax, Sfax 3038, Tunisia[5]

***Abstract*—Semantic segmentation is a fundamental component of autonomous driving systems, enabling accurate scene understanding and object-level perception. However, achieving precise instance-level delineation while maintaining real-time performance on resource-constrained platforms remains a significant challenge, particularly for edge deployment scenarios. This paper proposes a lightweight dual-YOLOv8 fusion framework for instance-aware semantic segmentation in autonomous driving applications. The proposed approach integrates YOLOv8n-seg and YOLOv8s-seg through a multi-scale fusion strategy that exploits their complementary feature representations to improve the segmentation of road-relevant objects, including cars, buses, trucks, and motorcycles. The framework is evaluated on the Reetiquetado de Vehiculos dataset using standard instance-level segmentation metrics. Experimental results demonstrate strong performance, achieving an overall mAP@0.5 of 92.9% and mAP@0.5:0.95 of 80.8%, while maintaining real-time inference with an average processing time of 7.9 ms per image (126 FPS) on an NVIDIA RTX 3050 GPU. Class-wise and confidence-based analyses confirm consistent segmentation accuracy across vehicle categories, highlighting the robustness of the proposed fusion strategy in handling scale variation, occlusions, and object diversity. In addition, an embedded deployment analysis provides insight into the feasibility and practical constraints of deploying the proposed framework on representative edge platforms. Overall, the proposed dual-YOLOv8 fusion framework achieves an effective balance between segmentation accuracy and computational efficiency, making it suitable for real-time autonomous driving perception on edge ARM/GPU platforms and Advanced Driver Assistance Systems (ADAS).**

***Keywords*—Autonomous driving; instance-aware semantic segmentation; real-time instance segmentation; YOLOv8; dual-model fusion; edge deployment**

I. Introduction

Autonomous driving systems rely heavily on accurate and real-time perception to ensure safe navigation in complex and dynamic environments. Among perception tasks, semantic and instance-aware segmentation play a crucial role by enabling fine-grained scene understanding and precise localization of surrounding objects, such as vehicles, pedestrians, and road elements. Unlike object detection, which provides coarse bounding boxes, instance-aware semantic segmentation delivers pixel-level object delineation while preserving individual object identities, making it particularly valuable for downstream tasks including tracking, motion prediction, and decision making in autonomous driving systems [1], [2].

Recent advances in deep learning have significantly improved segmentation performance in autonomous driving scenarios. Early approaches focused on semantic segmentation using convolutional neural networks, while more recent methods emphasize instance-aware segmentation to better handle crowded scenes and overlapping objects [3], [4]. However, achieving high segmentation accuracy often comes at the cost of increased computational complexity, which limits real-time deployment on resource-constrained platforms commonly used in autonomous vehicles and Advanced Driver Assistance Systems (ADAS).

To address this challenge, lightweight and real-time instance segmentation models have gained increasing attention. YOLO-based architectures have emerged as a promising solution due to their unified design and high inference speed. Methods such as YOLACT [3] and Insta-YOLO [5] demonstrated that real-time instance segmentation is feasible, while subsequent works further improved efficiency and robustness through architectural refinements and task-specific optimizations [6], [7]. More recently, YOLOv8-based segmentation models have shown strong performance across various applications, motivating their adoption for real-time perception tasks [8], [9].

Despite these advances, existing instance-aware semantic segmentation approaches typically rely on a single segmentation model and tend to optimize either segmentation accuracy or inference speed in isolation. As a result, such designs often struggle to simultaneously maintain high accuracy and low latency across diverse object scales, frequent occlusions, and dense urban traffic scenes, particularly when deployed on resource-constrained edge ARM/GPU platforms. Although multimodal fusion strategies combining heterogeneous sensors such as cameras and LiDAR have been explored, model-level fusion within a single visual modality remains largely underinvestigated for real-time autonomous driving. Consequently, there exists a clear gap for lightweight fusion strategies that can

exploit complementary representations from multiple compact models to improve segmentation robustness without sacrificing real-time performance or deployability.

In this work, we address the above limitations by proposing a lightweight dual-YOLOv8 fusion framework for instance-aware semantic segmentation in autonomous driving, specifically designed to balance accuracy, real-time performance, and edge deployability.

The main contributions of this paper are summarized as follows:

- A lightweight dual-YOLOv8 fusion framework for instance-aware semantic segmentation that exploits complementary multi-scale representations from YOLOv8n-seg and YOLOv8s-seg while preserving real-time performance.
- A comprehensive experimental analysis of accuracy–latency trade-offs for instance-level segmentation in autonomous driving scenarios, including class-wise, confidence-based, and qualitative evaluations.
- An embedded deployment-oriented study that analyzes the feasibility and limitations of deploying dual-model segmentation pipelines on edge ARM CPU and low-power GPU platforms.
- Empirical insights into model-level fusion strategies for real-time perception, providing reusable design guidelines for other resource-constrained segmentation tasks.

The remainder of this paper is organized as follows. Section II reviews related work on instance-aware semantic segmentation and real-time perception in autonomous driving. Section III details the proposed dual-YOLOv8 fusion framework. Section IV describes the experimental setup, including the training platform and implementation details, training configuration, dataset description, and evaluation metrics. Section V presents the experimental results and provides an in-depth analysis and discussion, encompassing dataset characteristics, training dynamics, class-wise performance, quantitative and qualitative segmentation results, embedded deployment considerations, and a comparative study with state-of-the-art methods. Finally, Section VI concludes the paper and outlines directions for future research.

II. Related Work

A. Instance-Aware Semantic Segmentation

Instance-aware semantic segmentation has been widely studied in the context of autonomous driving due to its ability to provide detailed scene understanding. Early works explored clustering-based and embedding-based approaches to separate object instances within semantic classes [1]. Later, end-to-end deep learning architectures capable of jointly predicting object instances and semantic labels significantly improved segmentation quality in complex driving scenes [3], [4].

Several studies addressed challenges such as occlusions, class imbalance, and missed detections in autonomous driving environments. CompleteInst [10] proposed an efficient network for handling missed detections, while joint detection and segmentation frameworks improved consistency between localization and segmentation outputs [11]. Recent surveys further highlight the importance of instance-aware segmentation and summarize advances in architectures, evaluation metrics, and applications [2], [12].

B. YOLO-Based Real-Time Instance Segmentation

YOLO-based models have gained popularity for real-time instance segmentation due to their unified detection and segmentation pipelines and high computational efficiency. YOLACT [3] and YOLACT++ demonstrated competitive real-time performance by decoupling mask generation from detection. Insta-YOLO [5] further explored lightweight designs for real-time instance segmentation.

More recent works extended YOLO architectures to improve segmentation accuracy through contour regression, multi-task learning, and architectural refinements [6], [7]. With the introduction of YOLOv8, several studies proposed enhanced segmentation variants for real-time applications, including construction site monitoring and debris detection [8], [9]. These studies demonstrate the flexibility and effectiveness of YOLOv8-based segmentation models.

C. Fusion Strategies and Advanced Segmentation Models

Beyond single-model approaches, fusion strategies have been explored to enhance segmentation robustness. While many works focus on multimodal fusion using camera and LiDAR data [13], model-level fusion within a single modality remains less explored for real-time instance-aware segmentation in autonomous driving. Transformer-based segmentation models such as Mask2Former [14] and foundation models like AD-SAM [15] have shown strong performance but remain computationally expensive for real-time deployment.

D. Positioning of the Proposed Approach

In contrast to existing YOLO-based instance segmentation methods that typically rely on a single network architecture, this work introduces a fusion-based dual-YOLOv8 framework that integrates two lightweight segmentation models operating at complementary scales. While prior approaches primarily focus on architectural refinements within a single model to improve either accuracy or speed, such designs may struggle to maintain robustness under significant object scale variation, occlusions, and complex urban traffic scenes.

By jointly leveraging YOLOv8n-seg for efficient and low-latency inference and YOLOv8s-seg for richer multi-scale feature representation, the proposed framework exploits model-level complementarity to enhance instance-aware semantic segmentation performance without incurring substantial computational overhead. Unlike multimodal fusion strategies that combine heterogeneous sensor inputs such as camera and LiDAR data, the proposed approach operates within a single visual modality, making it particularly suitable for real-time and edge deployment in autonomous driving and ADAS applications.

Unlike conventional ensemble approaches that independently aggregate predictions from multiple models, the proposed framework introduces a deployment-oriented dual-network fusion strategy explicitly designed for real-time

instance-aware segmentation. Rather than modifying individual YOLO architectures, this work demonstrates that fusing two carefully selected lightweight segmentation models enables cross-model scale complementarity that cannot be achieved by a single network without significantly increasing complexity. This design choice distinguishes the proposed approach from prior YOLO-based enhancements focused on single-model architectural refinement and provides a practical and generalizable fusion paradigm for real-time perception systems under strict latency and resource constraints.

As a result, the proposed dual-YOLOv8 fusion strategy addresses a key limitation of current real-time instance segmentation methods by improving segmentation robustness and accuracy while preserving practical real-time performance on resource-constrained platforms.

III. Proposed Work

To address the growing demand for accurate, real-time, and edge-deployable perception in autonomous driving systems, this work proposes a lightweight dual-network fusion framework for instance-aware semantic segmentation of vehicles. The primary objective of the proposed approach is to achieve precise pixel-level object delineation while maintaining computational efficiency suitable for real-time deployment on resource-constrained platforms. An overview of the proposed framework is illustrated in Fig. 1, which integrates dataset preparation, a dual-network segmentation architecture, and representative qualitative outputs within a unified perception pipeline.

The framework is built upon the Reetiquetado de Vehiculos dataset (version 2), hosted on Roboflow Universe, which provides instance-level segmentation annotations for five vehicle categories: *car*, *bus*, *motorcycle*, *truck*, and *truck3*. As depicted in the upper-left part of Fig. 1, the dataset includes images captured under diverse real-world traffic conditions, encompassing variations in illumination, viewpoint, vehicle scale, and background complexity. Such diversity is essential for training segmentation models capable of robust generalization across complex urban environments. The dataset is divided into training, validation, and test subsets to ensure a consistent and reproducible evaluation protocol.

At the core of the proposed framework lies a dual-YOLOv8 segmentation architecture that integrates the complementary strengths of the lightweight YOLOv8n-seg and the more expressive YOLOv8s-seg models through a model-level fusion strategy. Unlike conventional single-model approaches, the proposed design exploits the heterogeneity of lightweight and moderately scaled networks to improve robustness across varying object sizes and scene complexities. YOLOv8n-seg emphasizes fast inference and low computational cost, while YOLOv8s-seg provides richer multi-scale feature representations that enhance segmentation accuracy. Their integration enables the framework to capture both fine-grained local details and higher-level contextual information.

Each YOLOv8 segmentation model consists of a CSP-Darknet backbone for efficient feature extraction, a Path Aggregation Feature Pyramid Network (PAFPN) for enhanced multi-scale feature aggregation, and a decoupled segmentation head that jointly optimizes object localization, classification, and instance-level mask prediction. This architectural design facilitates accurate separation of individual vehicle instances while preserving semantic class information, thereby fulfilling the requirements of instance-aware semantic segmentation. The proposed dual-network fusion strategy leverages complementary feature representations from both models, improving robustness to object scale variation, partial occlusions, and visually cluttered backgrounds.

Finally, the bottom-right section of Fig. 1 presents qualitative segmentation results across multiple driving scenarios. The visual examples demonstrate the ability of the proposed framework to generate accurate and consistent instance-level segmentations for different vehicle categories, even under challenging conditions such as partial occlusions and complex urban backgrounds. These qualitative observations confirm that the proposed dual-YOLOv8 fusion framework achieves a favorable balance between segmentation accuracy and computational efficiency.

In summary, the proposed framework provides an efficient, scalable, and edge-friendly solution for instance-aware semantic segmentation in autonomous driving, combining complementary YOLOv8 segmentation models and high-quality instance annotations to address key perception challenges in real-world traffic environments.

IV. Experimental Setup

A. Training Platform and Implementation Details

All experiments were conducted on a workstation running Ubuntu 24.04, equipped with an 11th Gen Intel(R) Core(TM) i7-11800H CPU operating at 2.30 GHz, 16 GB of RAM, and an NVIDIA GeForce RTX 3050 GPU. The proposed framework was implemented using the Python programming language and trained using the PyTorch deep learning framework with CUDA acceleration. Automatic Mixed Precision (AMP) was enabled to improve training efficiency and reduce memory usage.

The instance-aware semantic segmentation models, namely YOLOv8n-seg and YOLOv8s-seg, were implemented based on the official YOLOv8 repositories. The use of official implementations ensures correctness, reproducibility, and practical relevance of the experimental results.

B. Training Configuration

The proposed framework was trained for a total of 100 epochs using a batch size of 8 and an input resolution of 640×640 pixels. Optimization was performed using the Adam optimizer with an initial learning rate of 0.01. A cosine learning-rate scheduling strategy was employed to progressively reduce the learning rate during training, promoting stable convergence. An early-stopping mechanism with a patience of 20 epochs was applied to prevent overfitting by terminating training when no improvement in validation performance was observed.

Automatic Mixed Precision (AMP) was enabled to accelerate training and reduce memory consumption. Pretrained weights were utilized to initialize the models, facilitating faster convergence and improved generalization. Data augmentation

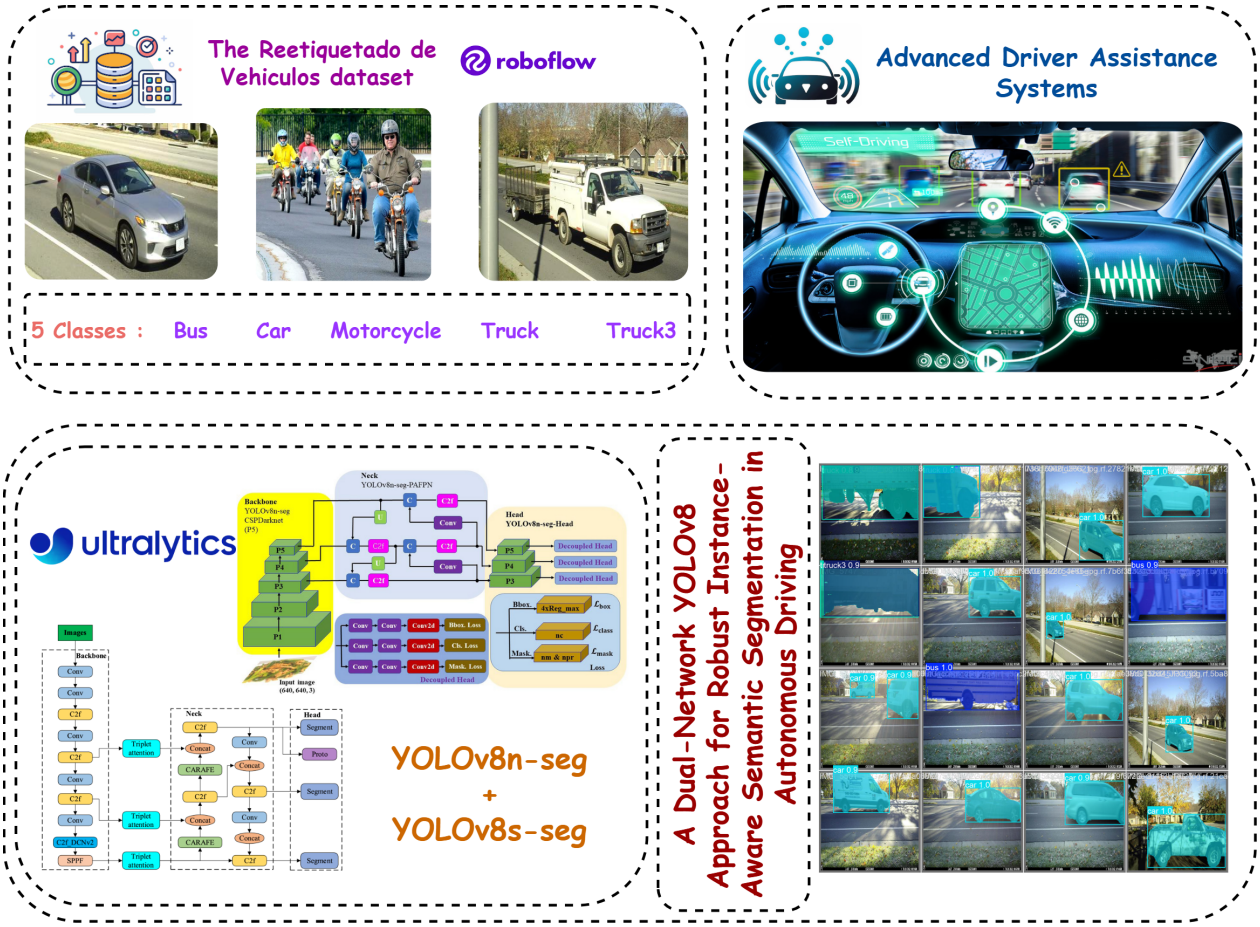

Fig. 1. Overview of the proposed dual-YOLOv8 fusion framework for instance-aware semantic segmentation.

techniques, including blurring, grayscale conversion, contrast-limited adaptive histogram equalization (CLAHE), mosaic augmentation, mixup, and random erasing, were applied during training to enhance robustness against variations in scale, illumination, and appearance.

C. Dataset Description

The experiments were conducted using the Reetiquetado de Vehiculos dataset (version 2), a publicly available benchmark hosted on Roboflow Universe [16]. The dataset contains annotated images representing five vehicle categories: *bus*, *car*, *motorcycle*, *truck*, and *truck3*. It is organized into three subsets, namely training, validation, and test splits, enabling a consistent and reproducible evaluation protocol. The Reetiquetado de Vehiculos dataset was selected due to its vehicle-centric focus, instance-level annotations, and suitability for evaluating real-time segmentation performance in autonomous driving scenarios.

The dataset includes complete instance-aware semantic segmentation annotations for all vehicle classes, allowing pixel-level delineation of individual object instances. The diversity of vehicle types and balanced class distribution make it suitable for evaluating segmentation performance on both frequently occurring and less-represented vehicle categories. No additional re-annotation was performed, and the dataset was used as provided, ensuring fair and reproducible experimental conditions.

D. Evaluation Metrics and Performance Measures

The proposed dual-YOLOv8 fusion framework is evaluated using standard metrics for instance-aware semantic segmentation. Performance is assessed at the instance level using precision, recall, F1-score, and mean Average Precision (mAP), all computed on the predicted segmentation masks.

Precision (P) measures the proportion of correctly predicted instances among all predicted instances, while recall (R) reflects the proportion of ground-truth instances that are correctly detected. These metrics are defined as [17], [18]:

$$P = \frac{TP}{TP + FP}, \tag{1}$$

$$R = \frac{TP}{TP + FN}, \tag{2}$$

where, TP, FP, and FN denote the numbers of true positives, false positives, and false negatives, respectively.

To provide a balanced evaluation of segmentation performance, the F1-score is computed as the harmonic mean of precision and recall:

$$F1 = \frac{2 \cdot P \cdot R}{P + R}. \tag{3}$$

Segmentation accuracy is further quantified using the Intersection over Union (IoU), which measures the overlap between a predicted mask M_p and its corresponding ground-truth mask M_{gt}:

$$IoU = \frac{|M_p \cap M_{gt}|}{|M_p \cup M_{gt}|}. \tag{4}$$

Based on the IoU criterion, the mean Average Precision (mAP) is employed to summarize overall segmentation performance. The Average Precision (AP) is computed as the area under the Precision–Recall curve for each class, and the mAP is obtained by averaging AP values across all C object classes:

$$mAP = \frac{1}{C} \sum_{c=1}^{C} AP_c. \tag{5}$$

Following the standard COCO evaluation protocol, segmentation performance is reported at an IoU threshold of 0.5 (mAP@0.5) as well as averaged over multiple IoU thresholds ranging from 0.5 to 0.95 in steps of 0.05 (mAP@0.5:0.95), thereby capturing both localization accuracy and mask quality.

In addition, confidence-based performance curves, including Precision–Confidence, Recall–Confidence, Precision–Recall, and F1–Confidence curves, are analyzed to examine the behavior of the segmentation model under varying confidence thresholds. This analysis provides insights into the trade-off between detection reliability and instance coverage, which is particularly important for real-time autonomous driving applications.

V. Results Analysis and Discussion

This section presents a comprehensive analysis and discussion of the experimental results obtained with the proposed dual-YOLOv8 fusion framework, highlighting its performance in terms of segmentation accuracy, robustness, and computational efficiency.

A. Statistical and Spatial Characteristics of the Dataset

Fig. 2 illustrates the distribution and spatial characteristics of the annotated instances in the Reetiquetado de Vehiculos dataset. The class distribution, shown in the upper-left plot, reveals a moderate class imbalance, with *car* instances being the most frequent, followed by *truck* and *truck3*, while *bus* and *motorcycle* appear less frequently. This imbalance reflects realistic traffic conditions and motivates the need for robust segmentation models capable of handling both dominant and under-represented classes.

The upper-right visualization overlays normalized bounding boxes of all instances, highlighting the diversity in object scales and aspect ratios present in the dataset. The lower-left plot shows the spatial distribution of object centers, indicating that vehicles are predominantly located near the central and lower regions of the images, which is consistent with typical on-road camera viewpoints. Finally, the lower-right plot depicts the relationship between object width and height, demonstrating a wide variation in vehicle sizes and shapes. Together, these statistics confirm that the dataset captures substantial variability in object position, scale, and class frequency, presenting a challenging and realistic benchmark for instance-aware semantic segmentation in autonomous driving scenarios.

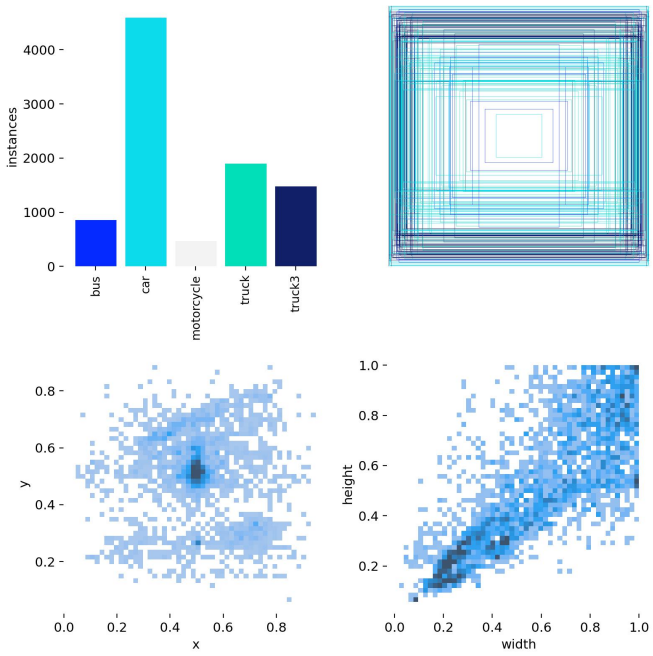

Fig. 2. Statistical analysis of the Reetiquetado de Vehiculos dataset, including class distribution, normalized bounding-box overlap, spatial distribution of object centers, and width–height relationships of annotated vehicle instances.

B. Correlation Analysis of Bounding-Box Attributes

Fig. 3 presents a correlogram illustrating the relationships among normalized bounding-box attributes in the Reetiquetado de Vehiculos dataset, including object center coordinates (x, y), width, and height. The diagonal histograms reveal the marginal distributions of each attribute, showing that object centers are predominantly concentrated near the central regions of the images, while width and height values span a broad range. Off-diagonal plots highlight the correlations between variables, most notably a strong positive relationship between bounding-box width and height, indicating consistent aspect ratios across vehicle types. The joint distributions involving spatial coordinates further confirm that larger objects tend to appear closer to the camera viewpoint, while smaller instances are more dispersed. Overall, this analysis demonstrates the structural diversity and realistic spatial correlations present in the dataset, reinforcing its suitability as a challenging benchmark for instance-aware semantic segmentation in autonomous driving scenarios.

C. Training Dynamics and Convergence Analysis

The learning curves presented in Fig. 4 demonstrate a stable and well-behaved convergence of the YOLOv8-seg training process over 100 epochs. All training loss components,

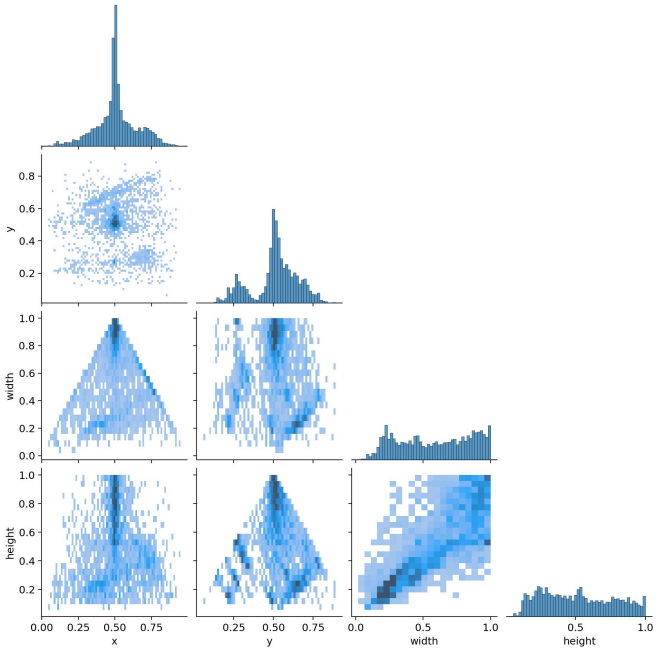

Fig. 3. Correlogram of normalized bounding-box attributes showing marginal distributions and pairwise correlations between object center coordinates, width, and height.

including bounding box regression loss, segmentation loss, classification loss, and distribution focal loss (DFL), exhibit a smooth and monotonic decrease throughout training. This behavior indicates effective optimization of both localization accuracy and pixel-level mask prediction, without oscillations or sudden instabilities that would suggest optimization difficulties. A similar decreasing trend is observed for the validation losses, which closely follow their corresponding training curves. The absence of sustained divergence between training and validation losses suggests that the model generalizes well to unseen data and does not suffer from significant overfitting. Minor fluctuations in the validation curves, particularly for classification and segmentation losses, can be attributed to the inherent complexity of UAV urban scenes and class imbalance, but these variations remain controlled and diminish as training progresses. From a performance perspective, both detection and segmentation metrics improve rapidly during the early training epochs, reflecting fast learning of low-level and mid-level visual features. Precision and recall for bounding boxes increase steadily before reaching a plateau, indicating that the model achieves a stable balance between false positives and false negatives. A similar trend is observed for mask precision and recall, confirming that improvements in object localization directly translate into better instance-level segmentation quality. The mAP@0.5 and mAP@0.5:0.95 curves for both bounding boxes and masks show consistent growth followed by gradual saturation after approximately 70–80 epochs. This saturation behavior indicates that the model approaches its optimal representational capacity under the given architecture and training configuration, with diminishing returns beyond this point. Notably, the close alignment between box-level and mask-level mAP curves highlights the strong coupling between detection and segmentation heads in the proposed framework.

Overall, these training dynamics confirm that the adopted optimization strategy, loss formulation, and architectural design lead to reliable convergence, stable validation performance, and balanced improvement across detection and instance segmentation tasks. This stability is particularly important for UAV-based applications, where robustness and generalization under challenging imaging conditions are critical.

D. Class-wise Performance Analysis

Fig. 5 presents the normalized confusion matrix of the proposed instance-aware semantic segmentation framework across the five vehicle categories. The strong diagonal dominance indicates high classification accuracy for all classes, with particularly strong performance for *car* (0.93), *bus* (0.89), *motorcycle* (0.87), *truck3* (0.85), and *truck* (0.81). These results confirm that the framework effectively distinguishes between different vehicle categories despite variations in appearance and scale.

Most misclassifications arise between visually similar categories, particularly *truck* and *truck3*, which exhibit closely related geometric structures and appearance patterns. A limited degree of confusion with the background class is also observed, mainly for large-scale vehicles, and can be attributed to partial occlusions, truncated instances, or ambiguous object boundaries in complex traffic scenes. Despite these challenges, the predominantly low off-diagonal values in the confusion matrix indicate strong class separability and consistent semantic discrimination. Overall, this analysis confirms the robustness of the proposed dual-YOLOv8 fusion framework in effectively managing inter-class similarity while preserving accurate instance-level segmentation performance.

E. Mask Precision-Recall Analysis

Fig. 6 shows the mask precision–recall curves for each vehicle category and for all classes combined. The curves exhibit strong performance across a wide range of recall values, indicating that the proposed framework maintains high precision while successfully retrieving most object instances. The aggregated curve achieves an overall mAP@0.5 of 92.9%, confirming the effectiveness of the proposed dual-YOLOv8 fusion approach for instance-aware semantic segmentation. Class-wise results demonstrate particularly strong performance for *car* and *bus*, while slightly lower precision is observed for visually challenging categories such as *truck*. Overall, the curves indicate a favorable balance between precision and recall, highlighting the robustness of the segmentation framework across different vehicle types.

F. Mask Precision-Confidence Analysis

Fig. 7 illustrates the mask precision–confidence curves for each vehicle class and for all classes combined. Precision increases steadily as the confidence threshold rises, indicating a progressive reduction in false-positive segmentation predictions. The aggregated curve shows that the proposed framework achieves near-perfect precision at high confidence levels, reaching a precision of 1.00 at a confidence threshold of approximately 0.98. Class-wise curves exhibit consistent behavior, with minor variations reflecting differences in object appearance and class frequency. These results demonstrate the

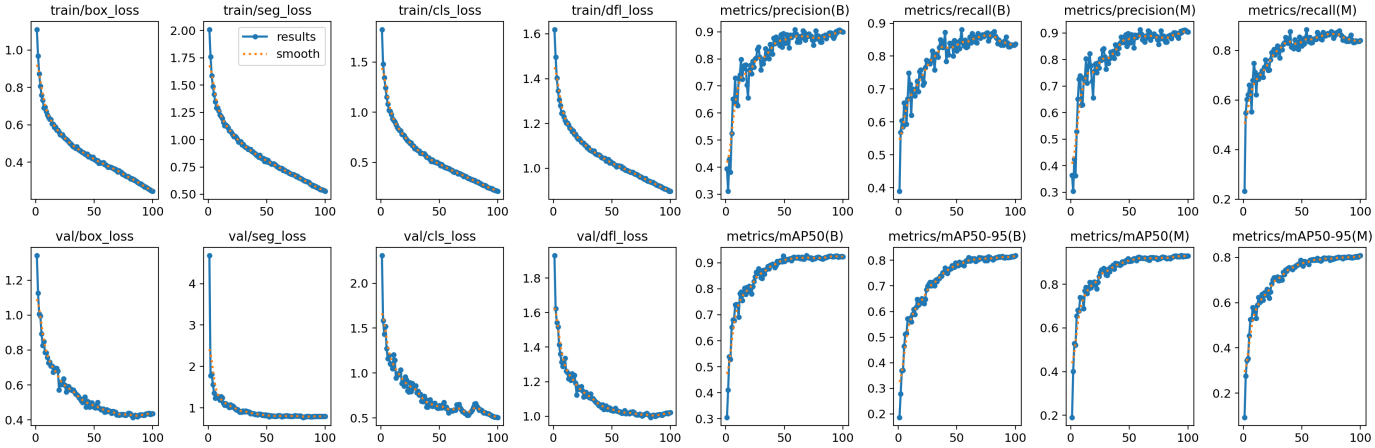

Fig. 4. Training and validation curves of YOLOv8-seg, over 100 epochs, for Bounding Box Prediction (B) and Segmentation Outputs (M).

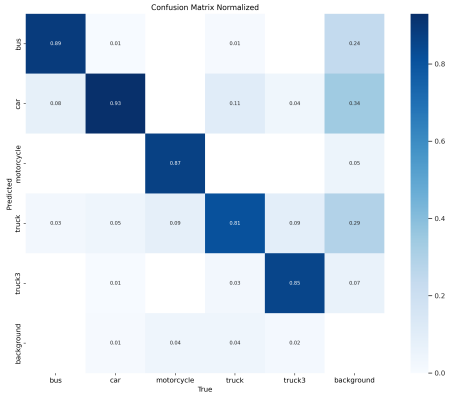

Fig. 5. Normalized confusion matrix analysis.

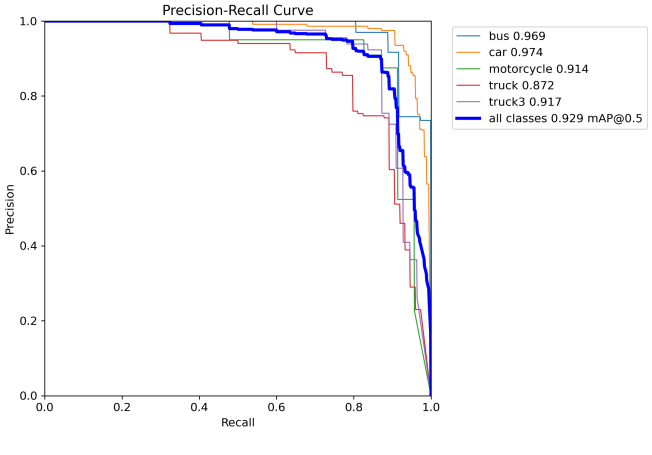

Fig. 6. Mask precision-recall curves analysis.

reliability of the proposed instance-aware semantic segmentation framework in producing highly accurate segmentation outputs when appropriate confidence thresholds are applied.

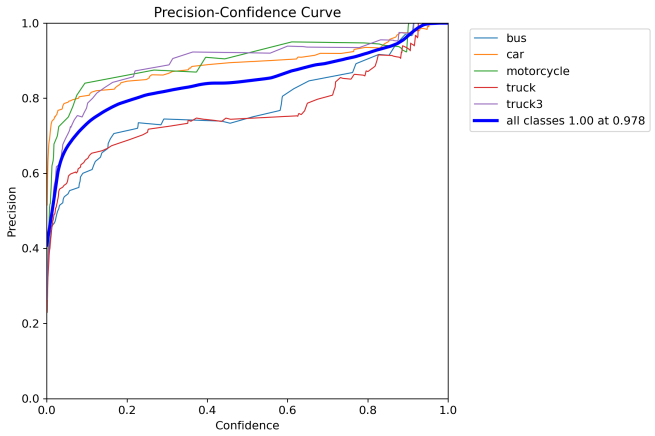

Fig. 7. Mask precision-confidence analysis.

G. Mask Recall-Confidence Analysis

Fig. 8 presents the mask recall–confidence curves for individual vehicle classes and for all classes combined. High recall values are observed at low confidence thresholds, indicating that the proposed framework successfully detects and segments most object instances. As the confidence threshold increases, recall gradually decreases due to the stricter acceptance of predictions. The aggregated curve shows an overall recall of approximately 0.96 at zero confidence and maintains stable performance across a wide confidence range before dropping sharply at very high thresholds. Class-wise trends remain consistent, with slight variations reflecting differences in object size, visibility, and class frequency. These results highlight the trade-off between recall and confidence threshold selection in instance-aware semantic segmentation.

H. Mask F1-Confidence Analysis

Fig. 9 presents the mask F1–confidence curves for individual vehicle classes and for all classes combined. The F1 score increases rapidly at low confidence thresholds and remains stable over a broad confidence range, indicating a balanced trade-off between precision and recall. The aggregated

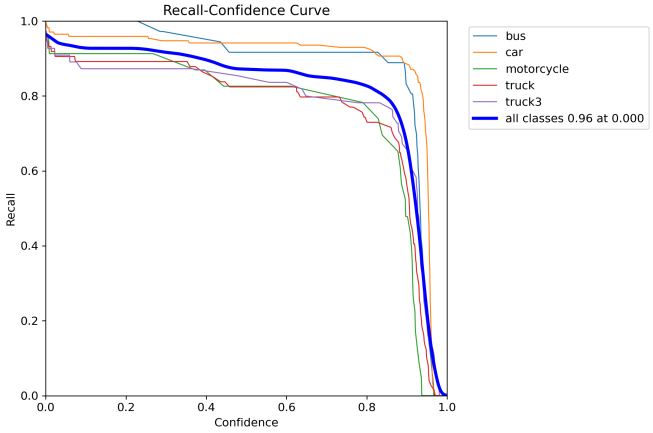

Fig. 8. Mask recall-confidence curves analysis.

curve reaches a maximum F1 score of approximately 0.87 at a confidence threshold of 0.77, which can be considered an optimal operating point for the proposed instance-aware semantic segmentation framework. Class-wise curves exhibit similar trends, with minor variations reflecting differences in object appearance and class distribution. These results provide a practical guideline for selecting confidence thresholds that balance segmentation accuracy and completeness in real-time applications.

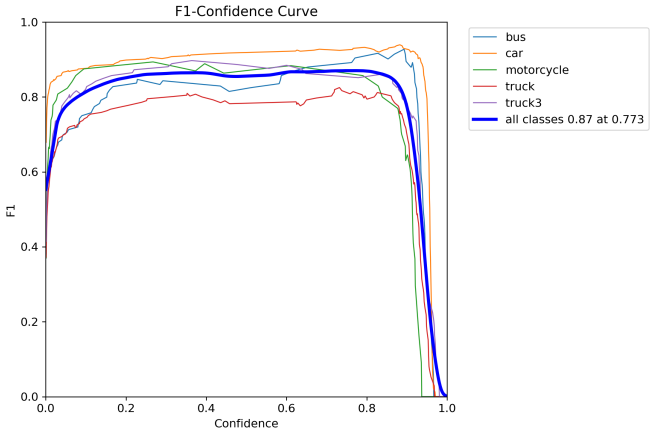

Fig. 9. Mask F1-confidence analysis.

I. Embedded Deployment: ARM vs. GPU

Although the proposed Dual-YOLOv8 fusion framework achieves strong segmentation accuracy and real-time inference on a desktop GPU, embedded deployment introduces stricter constraints in compute throughput, memory bandwidth, and runtime overhead. On the workstation platform used in this study (Intel i7-11800H + NVIDIA RTX 3050), the proposed method attains an overall mask $\text{mAP}_{0.5}$ of 92.9 ($\text{mAP}_{0.5:0.95}$ of 80.8) with an average processing time of approximately 7.9 ms per image at 640×640, corresponding to about 126 FPS.

To contextualize deployment on resource-constrained devices, we consider two representative edge targets: 1) an ARM CPU-only platform (Raspberry Pi 5) and 2) a low-power GPU platform (NVIDIA Jetson Nano). Raspberry Pi 5 is built around the Broadcom BCM2712 SoC featuring a quad-core Arm Cortex-A76 CPU operating at 2.4 GHz. Jetson Nano integrates a 128-core NVIDIA Maxwell GPU and a quad-core Arm Cortex-A57 CPU, and is advertised at 472 GFLOPS AI performance. For a fair real-time characterization on embedded systems, end-to-end latency is defined as:

$$T_{\text{e2e}} = T_{\text{pre}} + T_{\text{net}} + T_{\text{post}}, \tag{6}$$

where, T_{pre} includes resize/letterbox and normalization, T_{net} is the model forward pass, and T_{post} includes confidence filtering, NMS, and mask decoding. Unless explicitly stated, visualization/overlay rendering and camera/video I/O are excluded, since they depend heavily on the application pipeline rather than the model itself.

Table I reports the desktop reference (measured in this paper) and *estimated* end-to-end timing on Raspberry Pi 5 and Jetson Nano for batch=1 and 640×640. The embedded results are estimates (not directly measured in this study) and are provided to guide the expected order of magnitude when selecting an appropriate deployment target and optimization strategy.

The GPU-accelerated Jetson Nano is expected to provide substantially lower end-to-end latency than CPU-only execution due to CUDA/TensorRT acceleration, making it the more suitable target for running the full dual-model fusion pipeline at 640×640 when near real-time constraints are required. However, Jetson Nano lacks comprehensive support for INT8 across all layers in TensorRT workflows on this class of hardware, so FP16 (or FP32) is typically the practical deployment choice.

In contrast, Raspberry Pi 5 inference is dominated by T_{net} on CPU, and thus benefits most from 1) quantization and 2) reducing input size and/or simplifying the model. Quantized inference on Raspberry Pi-class devices is commonly performed using INT8 arithmetic to reduce model size and improve runtime performance, but it may introduce a small accuracy drop, motivating either post-training quantization with calibration data (PTQ) or quantization-aware training (QAT).

Finally, these embedded estimates reinforce the main conclusion drawn from the workstation experiments: while the proposed fusion strategy offers an excellent accuracy–efficiency balance on a desktop GPU (92.9 $\text{mAP}_{0.5}$ at 7.9 ms/image), deployment on CPU-only embedded devices typically requires additional compression (quantization/pruning) and/or architectural simplification to meet strict real-time constraints.

J. Qualitative Segmentation Results

In addition to quantitative metrics, qualitative results on unseen validation images further demonstrate the robustness and visual consistency of the proposed framework.

Fig. 10 presents qualitative instance-aware semantic segmentation results obtained on the validation set using the

TABLE I. END-TO-END LATENCY BREAKDOWN

Platform	Precision	T_{pre} (ms)	T_{net} (ms)	T_{post} (ms)	T_{e2e} (ms)	FPS
RTX 3050 (reference)	FP16/AMP	–	–	–	7.9	126
Jetson Nano (GPU)	FP16 (TensorRT)	4–7	35–60	8–15	50–82	12–20
	FP32	4–7	55–90	10–18	69–115	9–14
Raspberry Pi 5 (CPU)	FP32	25–45	600–1000	70–120	695–1165	0.9–1.4
	INT8 (PTQ/QAT)	25–45	280–520	60–100	365–665	1.5–2.7

proposed fusion-based dual-YOLOv8 framework. The examples demonstrate accurate segmentation of multiple vehicle categories, including cars, buses, trucks, and motorcycles, across diverse traffic scenes. The model successfully delineates vehicle boundaries at the pixel level while preserving correct semantic labels and high confidence scores. Notably, the framework handles variations in object scale, partial occlusions, and complex backgrounds, such as roadside vegetation and shadows, with consistent segmentation quality. The results also show reliable separation of adjacent objects and stable predictions for both near-field and distant vehicles. These qualitative observations corroborate the quantitative evaluation, confirming the robustness and visual consistency of the proposed instance-aware semantic segmentation approach in real-world driving scenarios.

Fig. 10. Qualitative instance-aware semantic segmentation results on the validation set produced by the proposed fusion-based dual-YOLOv8 framework.

K. Comparative Study with State-of-the-Art

Table II presents an indicative comparison between the proposed dual-YOLOv8 fusion framework and representative real-time instance segmentation methods reported in the literature. Early one-stage approaches such as YOLACT and YOLACT++ achieve moderate segmentation accuracy on the COCO benchmark, with Mask AP values of 29.8% and 34.1%, respectively, while operating at approximately 33–34 FPS. Subsequent YOLO-based methods, including Insta-YOLO and YOLO-Core, improve inference speed and segmentation performance through architectural refinements, with YOLO-Core reaching 38.6% COCO Mask AP at 45 FPS. More recent lightweight designs tailored for traffic and urban perception, such as PSC-YOLO and UDS-YOLO, emphasize efficiency and robustness in complex road scenes. PSC-YOLO reports a relative improvement of 2.0% in mask average precision over YOLOv8n-seg while maintaining real-time performance at approximately 91 FPS, whereas UDS-YOLO achieves 39.5% mAP@0.5 on the Cityscapes dataset with a processing speed of 106.6 FPS. In contrast, the proposed dual-YOLOv8 fusion framework attains a substantially higher segmentation accuracy, achieving 92.9% mAP@0.5 while sustaining real-time inference at 126 FPS on the evaluated dataset. Although the reported results are obtained under different datasets and evaluation protocols, the comparison highlights the favorable accuracy–efficiency trade-off of the proposed approach and its suitability for real-time autonomous driving applications.

Overall, the experimental findings and comparative analysis confirm that the proposed approach achieves a favorable balance between segmentation accuracy and real-time performance, reinforcing its suitability for practical deployment in autonomous driving and ADAS applications.

VI. CONCLUSION AND FUTURE WORK

In this paper, a lightweight dual-YOLOv8 fusion framework for instance-aware semantic segmentation in autonomous driving scenarios has been presented. By combining the complementary strengths of YOLOv8n-seg and YOLOv8s-seg, the proposed approach enhances robustness to object scale variation and occlusions while maintaining real-time inference performance. Experimental results on a vehicle-centric dataset demonstrate the effectiveness of the framework, achieving an overall mAP@0.5 of 92.9% with an average inference time of 7.9 ms per image (126 FPS) on an NVIDIA RTX 3050 GPU. Both quantitative and qualitative evaluations confirm accurate and consistent instance-level segmentation in challenging traffic scenes.

While the proposed framework demonstrates strong performance, several limitations should be acknowledged. The experimental evaluation is limited to vehicle categories and does not yet include vulnerable road users such as pedestrians or cyclists. In addition, the embedded deployment analysis on ARM and low-power GPU platforms is indicative, with latency figures partially based on estimated performance rather than exhaustive hardware benchmarking. The current fusion

TABLE II. INDICATIVE COMPARISON WITH REAL-TIME INSTANCE SEGMENTATION METHODS

Method	Backbone	Reported Metric	Speed (FPS)
YOLACT [3]	ResNet-101	29.8 (COCO Mask AP)	33
YOLACT++ [19]	ResNet-101	34.1 (COCO Mask AP)	34
Insta-YOLO [5]	YOLOv3	–	30
YOLO-Core [6]	CSP-based	38.6 (COCO Mask AP)	45
PSC-YOLO [20]	YOLOv8n-based	+2.0% Mask AP vs YOLOv8n-seg	∼91
UDS-YOLO [21]	YOLOv8-seg	39.5 / 24.0 (Cityscapes)	106.6
YOLOv8-seg [9]	YOLOv8s	Dataset-specific	120
Proposed Method	**Dual-YOLOv8 Fusion**	**92.9 (mAP@0.5)**	**126**

strategy also employs fixed model combinations and does not dynamically adapt to scene complexity.

Beyond quantitative performance, this study highlights the potential of model-level fusion as a practical design strategy for real-time perception systems in autonomous driving and ADAS. By demonstrating that complementary lightweight networks can outperform single-model designs under strict latency constraints, the proposed framework provides insights for future perception architectures targeting resource-constrained and safety-critical environments.

Future work will explore extensions to multimodal perception by incorporating additional sensors such as LiDAR or radar, as well as adaptive fusion mechanisms with confidence-aware weighting to dynamically balance model contributions across varying scene complexities.

REFERENCES

[1] B. De Brabandere, D. Neven, and L. Van Gool, "Semantic instance segmentation for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, pp. 7–9, 2017.

[2] J. M. Molina, J. P. Llerena, L. Usero, and M. A. Patricio, "Advances in instance segmentation: Technologies, metrics and applications in computer vision," *Neurocomputing*, Art. no. 129584, 2025.

[3] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT: Real-time instance segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pp. 9157–9166, 2019.

[4] S. Mohanapriya, P. Natesan, M. S. Saranya, P. Indhumathi, S. T. P. Mohananpriya, and R. Monisha, "Instance segmentation for autonomous vehicle," *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 9, pp. 565–570, 2021.

[5] E. Mohamed, A. Shaker, A. El-Sallab, and M. Hadhoud, "Insta-YOLO: Real-time instance segmentation," *arXiv preprint arXiv:2102.06777*, 2021.

[6] H. Liu, W. Xiong, and Y. Zhang, "YOLO-Core: Contour regression for efficient instance segmentation," *Mach. Intell. Res.*, vol. 20, no. 5, pp. 716–728, 2023.

[7] X. Chang, H. Pan, W. Sun, and H. Gao, "YolTrack: Multitask learning based real-time multi-object tracking and segmentation for autonomous vehicles," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 12, pp. 5323–5333, 2021.

[8] R. Bai *et al.*, "Automated construction site monitoring based on improved YOLOv8-seg instance segmentation algorithm," *IEEE Access*, vol. 11, pp. 139082–139096, 2023.

[9] A. A. Alsuwaylimi, "Enhanced YOLOv8-seg instance segmentation for real-time submerged debris detection," *IEEE Access*, vol. 12, pp. 117833–117848, 2024.

[10] H. Wang, S. Zhu, L. Chen, Y. Li, and T. Luo, "CompleteInst: An efficient instance segmentation network for missed detection scene of autonomous driving," *Sensors*, vol. 23, no. 22, p. 9102, 2023.

[11] S. Abdigapporov, S. Miraliev, V. Kakani, and H. Kim, "Joint multiclass object detection and semantic segmentation for autonomous driving," *IEEE Access*, vol. 11, pp. 37637–37649, 2023.

[12] M. A. Elhassan *et al.*, "Real-time semantic segmentation for autonomous driving: A review of CNNs, transformers, and beyond," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 36, no. 10, p. 102226, 2024.

[13] K. Geng, G. Dong, G. Yin, *et al.*, "Deep dual-modal traffic objects instance segmentation method using camera and LiDAR data for autonomous driving," *Remote Sensing*, vol. 12, no. 20, p. 3274, 2020.

[14] B. Cheng *et al.*, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 1290–1299, 2022.

[15] M. Camarena *et al.*, "AD-SAM: Fine-tuning the segment anything vision foundation model for autonomous driving perception," *arXiv preprint arXiv:2510.27047*, 2025.

[16] Jebnouni, "Reetiquetado de Vehiculos dataset," Roboflow Universe, 2025. [Online]. Available: https://universe.roboflow.com/jebnouni/reetiquetado-de-vehiculos-degab

[17] S. Teboulbi, S. Messaoud, M. A. Hajjaji, *et al.*, "Real-time implementation of AI-based face mask detection and social distancing measuring system for COVID-19 prevention," *Scientific Programming*, vol. 2021, no. 1, Art. no. 8340779, 2021.

[18] S. Teboulbi, S. Messaoud, M. A. Hajjaji, *et al.*, "Fine-tuned YOLO V-10X: Real-time object detection for autonomous vehicles under multiple weather conditions," in *Proc. IEEE 22nd Int. Multi-Conf. Syst., Signals & Devices (SSD)*, pp. 1021–1026, 2025.

[19] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT++: Better real-time instance segmentation," *arXiv preprint arXiv:1912.06218*, 2020.

[20] X. Gu and G. Zhang, "PSC-YOLO: A lightweight model for urban road instance segmentation," *Journal of Real-Time Image Processing*, vol. 22, no. 2, pp. 1–13, 2025.

[21] Q. Qiang, M. Zhang, B. Zhao, *et al.*, "UDS-YOLO: An improved instance segmentation network for traffic scenarios," *The Journal of Supercomputing*, vol. 81, no. 14, pp. 1292–1307, 2025.