Journal Computer and Technology
Vol. 4, No. 1, July 2026, pp. 01-13
ISSN: 3048-1880 (Online)
DOI: https://doi.org/10.69916/comtechno.v4i1.410

OPEN ACCESS

Original Research

A Hybrid YOLOv5-SqueezeNet Framework with Crop-Union Context Verification for Real-Time Motorcycle Rider Identification

Teguh Bagaskara[1], Rachmat Aulia[*2]

[1,2] Department of Informatics Engineering, Faculty of Engineering and Computer Science, Universitas Harapan Medan, Medan, Indonesia
Email: [1]tegus.bagaskara@gmail.com, [2]jackm4t@gmail.com

ARTICLE HISTORY
Received: December 22, 2025
Revised: July 13, 2026
Published: July 20, 2026

Copyright © 2026 by the Authors: This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License.

ABSTRACT

Motorcycles are the dominant transportation mode in developing nations, yet their high volume exacerbates traffic violations and accidents. Automated identification of motorcycle operators is crucial for Intelligent Transportation Systems (ITS). However, existing advanced vision models, such as the Segment Anything Model (SAM) or YOLOv9, demand substantial computational resources, rendering them impractical for real-time edge deployment. Furthermore, standard object detectors often fail to contextually distinguish active riders from bystanders. To address these limitations, this study proposes a lightweight, context-aware hybrid framework integrating YOLOv5 for rapid spatial detection and SqueezeNet for semantic contextual verification. The core novelty lies in the Crop-Union Context Verification Mechanism, which extracts the spatial union of paired person-motorcycle bounding boxes, expands it proportionally, and classifies the cropped region using SqueezeNet to confirm the presence of an active operator. Additionally, Contrast Limited Adaptive Histogram Equalization (CLAHE) is applied during preprocessing to enhance image quality under varying illumination. Experimental results demonstrate that the hybrid system successfully identifies motorcycle operators with high spatial confidence (YOLOv5: 0.84 for person, 0.86 for motorcycle) and robust contextual validation (SqueezeNet Top-1 prediction: moped, probability 0.247). By utilizing SqueezeNet's highly efficient Fire Module architecture (~1.2 million parameters), the proposed framework achieves exceptional computational efficiency without sacrificing discriminative power. Ultimately, the disjunctive fusion of spatial and semantic signals ensures system resilience. This lightweight pipeline proves highly viable for real-time, AI-driven traffic surveillance on resource-constrained edge devices, offering a scalable solution for smart city infrastructure and automated law enforcement.

Keywords: Hybrid Deep Learning; YOLOv5; SqueezeNet; Crop-Union Context Verification; Edge Computing.

How to Cite:
T. Bagaskara and R. Aulia, "A Hybrid YOLOv5-SqueezeNet Framework with Crop-Union Context Verification for Real-Time Motorcycle Rider Identification," *J. Comput. Technol.*, vol. 4, no. 1, pp. 01–13, 2026.

1. INTRODUCTION

Motorcycles constitute the most dominant mode of transportation in numerous developing nations, including Indonesia. Their flexibility, affordability, and ease of mobility render them the primary choice for daily commuting. However, the high volume of motorcycle usage is directly correlated with an escalation in traffic-related issues, such as traffic violations and fatal accidents. Many of these incidents are exacerbated by the failure to automatically identify riders, particularly in safety violations such as non-compliance with helmet mandates. Consequently, the development of Intelligent Transportation Systems (ITS) capable of automatically detecting and verifying the presence of motorcycle riders has become crucial for supporting electronic ticketing (e-ticketing) enforcement and enhancing road safety [1].

Alongside the rapid advancements in computer vision, deep learning-based approaches have been widely adopted for traffic analysis. Recent studies have proposed sophisticated architectures for object detection and traffic violation monitoring. Chang et al. [2] integrated YOLOv4 with a Heuristic Weighting Mechanism (HWM) to develop an active safety warning system for motorcycles, focusing on distance estimation and risk assessment in blind-spot areas. For more specific violation detection, Deshpande et al. [3] designed a two-stage architecture utilizing DetectNet and YOLOv8 to detect helmet violations and extract vehicle license plates in smart city

scenarios. Furthermore, to address complex object detection, Selvamuthukumar et al. [4] proposed YOLOv9-S Net combined with Contrast Limited Adaptive Histogram Equalization (CLAHE) and fast fuzzy clustering, achieving high-precision vehicle object detection. In the realm of complex segmentation, Cabral et al. [5] introduced a hybrid pipeline combining YOLO with the Segment Anything Model (SAM) using activation map-based prompting for multi-damage segmentation in UAV imagery. Meanwhile, Almodhwahi and Wang [6] demonstrated the deployment of Edge AI utilizing an Inception-based architecture for real-time driver facial expression monitoring.

Despite the promising performance of these models, significant computational and contextual gaps emerge when implementing them in real-time surveillance systems on edge devices (e.g., traffic CCTV or IoT devices). First, heavy detection and segmentation models such as YOLOv9, SAM, or complex two-stage pipelines demand substantial computational power and memory, rendering them suboptimal for deployment on resource-constrained edge devices [5], [6]. Second, standard object detection models (such as YOLO) merely yield bounding boxes for the classes "person" and "motorcycle". In dense and complex traffic scenarios, this spatial detection is prone to contextual ambiguity; for instance, the system may fail to distinguish between an individual actively riding a motorcycle and someone merely standing next to a parked one. This contextual verification is critical but is frequently overlooked due to the prohibitive computational cost of full instance segmentation models.

To address the critical challenges of computational efficiency and contextual ambiguity in traffic surveillance, this study proposes a novel, lightweight, and context-aware hybrid framework for the precise identification of motorcycle riders. The core novelty of this research is anchored in three synergistic technical contributions. First, we introduce a spatial-semantic fusion architecture that seamlessly integrates YOLO [7] for rapid spatial object detection with [8], [9] for in-depth contextual verification. SqueezeNet was specifically selected for its unique Fire Module design , which sustains high classification accuracy while drastically reducing the parameter count to approximately 1.2 million, thereby offering a highly computationally efficient alternative to resource-intensive segmentation models like the Segment Anything Model (SAM) [5]. Second, departing from computationally prohibitive pixel-level instance segmentation, this study develops an innovative Crop-Union Context Verification Mechanism. This approach dynamically extracts the spatial union of paired "person" and "motorcycle" bounding boxes, applies a proportional spatial expansion to capture surrounding contextual cues, and subsequently processes this cropped region through SqueezeNet to definitively verify whether the subject is an active "rider" or a mere "non-rider" bystander. Third, to ensure robust inference stability under highly variable real-world traffic lighting conditions, such as glaring sunlight, low-light environments, or dense shadows, the framework incorporates Contrast Limited Adaptive Histogram Equalization (CLAHE) [10], [11] during the preprocessing stage to significantly enhance local image contrast and feature visibility. Ultimately, through this comprehensive integration, the research aims to deliver an image classification system that achieves an optimal balance between high discriminative accuracy and minimal computational overhead. The viability of this proposed lightweight architecture for deployment in real-time, AI-driven intelligent transportation systems will be rigorously validated through a multi-dimensional performance evaluation encompassing accuracy, precision, recall, F1-score, and inference time metrics.

2. RESEARCH METHODS

2.1. Research Framework

This section outlines the proposed system architecture, structured as a five-stage pipeline for accurate and computationally efficient motorcycle rider identification. The framework integrates rapid spatial detection with lightweight semantic verification to address contextual limitations of standard object detectors on edge devices.

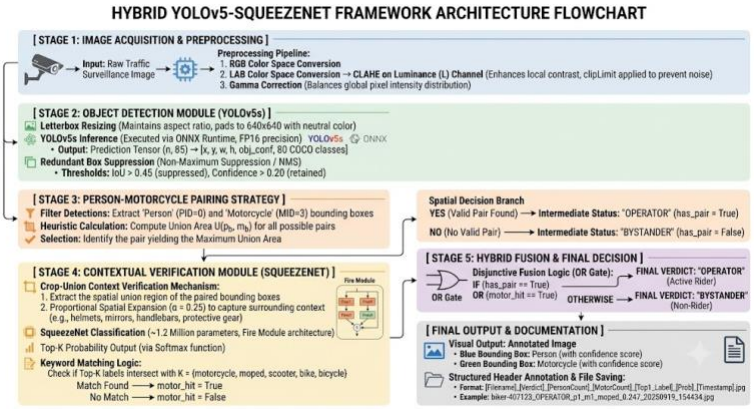

Figure 1. Hybrid YOLOv5-SqueezeNet Framework Architecture Flowchart showing five stages: (1) Image Preprocessing with CLAHE and gamma correction, (2) Object Detection using YOLOv5s, (3) Person-Motorcycle Pairing via maximum union heuristic, (4) Contextual Verification using SqueezeNet with Crop-Union mechanism, and (5) Hybrid Fusion Decision using disjunctive logic.

As illustrated in Figure 1, the framework begins with image acquisition and preprocessing to handle variable illumination, followed by YOLOv5s-based object detection with NMS filtering. The detected person and motorcycle bounding boxes are paired using a maximum union area heuristic, then verified contextually through SqueezeNet classification of the expanded crop region. Finally, a disjunctive fusion logic combines spatial and semantic signals to render the final "OPERATOR" or "BYSTANDER" verdict. This architecture ensures robustness while maintaining computational efficiency suitable for edge deployment.

2.2. Dataset Collection and Preparation

The dataset utilized in this study was curated from prominent open-source computer vision repositories, primarily the Common Objects in Context (COCO) dataset [12], supplemented by specialized traffic scene datasets (e.g., BDD100K and curated public datasets from Roboflow). To ensure robust contextual learning and mitigate class imbalance, the final dataset comprises a total of 2,400 annotated images, evenly distributed into two primary contextual categories: 1,200 images depicting active motorcycle operators (riders) and 1,200 images depicting non-riders (e.g., bystanders standing adjacent to parked motorcycles or individuals near motor vehicles). For systematic model development and rigorous evaluation, the dataset was partitioned using a stratified random split to maintain class distribution consistency across all subsets. The data was divided into 80% for training (1,920 images), 10% for validation (240 images), and 10% for independent testing (240 images). This distribution ensures that the model learns comprehensive feature representations while being evaluated on completely unseen data. Prior to model ingestion, all images underwent rigorous preprocessing to guarantee uniform input formats compatible with the hybrid architecture. This pipeline included conversion to the RGB color space, normalization of spatial resolution to 640×640 pixels (the optimal input dimension for YOLOv5s), and scaling of pixel intensity values to a [0, 1] range. Furthermore, to enhance model generalization and prevent overfitting to specific environmental artifacts, an extensive data augmentation pipeline was applied exclusively to the training subset. This pipeline incorporated geometric transformations (random horizontal flipping, slight rotations up to ±15 degrees, and random scaling) and photometric distortions (random adjustments to brightness, contrast, and saturation, alongside Mosaic augmentation). These measures collectively simulate diverse real-world traffic scenarios, thereby improving the framework's resilience against varying camera angles, partial occlusions, and dynamic lighting conditions.

2.3. Image Preprocessing and Enhancement

To address varying lighting conditions in real traffic scenarios, this study applies two image enhancement techniques: Contrast Limited Adaptive Histogram Equalization (CLAHE) and gamma correction.

CLAHE (Contrast Limited Adaptive Histogram Equalization)
CLAHE enhances local contrast by dividing the image into small tiles and applying histogram equalization to each tile, with a clip limit to prevent noise amplification. The transformation process on the luminance channel (L) in the LAB color space is formulated as [13]:

$$\mathrm{L}_{\mathrm{enhanced}}(\mathrm{x},\mathrm{y}) = \mathrm{CLAHE}(\mathrm{L}(\mathrm{x},\mathrm{y}), \mathrm{clipLimit}, \mathrm{tileGridSize}) \quad (1)$$

where $\mathrm{clipLimit}$restricts the maximum contrast and $\mathrm{tileGridSize}$determines the adaptive grid size.

Gamma Correction
Gamma correction balances the pixel intensity distribution through a lookup table (LUT):

$$\mathrm{I}_{\mathrm{out}} = 255 \times \left(\frac{\mathrm{I}_{\mathrm{in}}}{255}\right)^{\gamma} \quad (2)$$

where $\gamma$is the gamma correction parameter ($\gamma < 1$ brightens dark areas, $\gamma > 1$ darkens bright areas).

2.4. Object Detection Module: YOLOv5 Architecture

Object detection utilizes YOLOv5s in ONNX format. This model maps the input image to a prediction tensor of size (n, 85), where each row represents one object with 4 bounding box coordinates (x, y, w, h), 1 objectness score, and 80 COCO class probabilities [14], [15].

Letterbox Resizing

To maintain the original image aspect ratio and prevent distortion, a letterbox method is applied:

$$r = \min\left(\frac{h_{new}}{h_{orig}}, \frac{w_{new}}{w_{orig}}\right) \quad (3)$$

$$dw = \frac{w_{new} - (w_{orig} \times r)}{2}, dh = \frac{h_{new} - (h_{orig} \times r)}{2} \quad (4)$$

di mana ris the scale factor, and $(dw, dh)$are the padding offsets filled with a neutral color $(114, 114, 114)$.

Redundant Box Suppression

To eliminate overlapping bounding boxes, Redundant Box Suppression based on Intersection over Union (IoU) is utilized:

$$IoU = \frac{Area(B_i \cap B_j)}{Area(B_i \cup B_j)} \quad (5)$$

Bounding boxes with an IoU greater than the threshold $\theta_{IoU} = 0.45$ are suppressed, leaving only predictions with the highest confidence score ($\theta_{conf} = 0.20$).

2.5. Person-Motorcycle Pairing Strategy

After detection, the system performs pairing between the person class (PID $= 0$) and motorcycle class (MID $=$ 3) using a maximum union area heuristic. For each pair (pb, mb), the union area is calculated as:

$$U(pb, mb) = \max(0, x_2^{min} - x_1^{max}) \times \max(0, y_2^{min} - y_1^{max}) \quad (6)$$

Where

$$x_1^{max} = \max(x_1^{pb}, x_1^{mb}), x_2^{min} = \min(x_2^{pb}, x_2^{mb}) y_1^{max} = \max(y_1^{pb}, y_1^{mb}), y_2^{min} = \min(y_2^{pb}, y_2^{mb})$$

The pair with the largest Uvalue is selected as the candidate operator (best pair). Upon finding a valid pair, the initial status is designated as "OPERATOR". In the absence of a valid pair, the status is designated as "BYSTANDER".

2.6. Contextual Classification Module: SqueezeNet

SqueezeNet serves as a contextual verifier to validate the presence of a motorcycle in the detected area. This architecture utilizes a Fire Module consisting of a squeeze layer (1×1 convolution) and an expand layer (a mixture of $1 \times 1$and 3×3convolutions). The parameter count is calculated as [16], [17]:

$$N_{params} = (F \times F \times C_{in}) \times C_{out} (7)$$

where Fis the kernel size, $C_{in}$is the quantity of input channels, and $C_{out}$is the quantity of output channels. With this design, SqueezeNet achieves accuracy equivalent to AlexNet with approximately 50 times fewer parameters (~1.2 million parameters).

Top-K Classification and Keyword Matching

The cropped image is classified using an ImageNet-pretrained SqueezeNet. Class probabilities are computed through the softmax function:

$$P(y_i \mid x) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} (8)$$

where $z_i$is the logit of the i-th class and Kis the total class quantity. The system then checks to determine the presence of motorcycle keywords ($\mathcal{K} = \{motorcycle, moped, scooter, bike, bicycle\}$ within the Top-K labels:

$$\text{hit} = \begin{cases} \text{True}, & \text{provided that } \exists\, l_k \in \text{TopK} \mid l_k \in \mathcal{K} \\ \text{False}, & \text{in other cases} \end{cases} \quad (9)$$

2.7. Hybrid Fusion: Crop-Union Context Verification

The fusion mechanism represents the main novelty of this study. The union area from the person-motorcycle pair is extracted and expanded with a ratio $\alpha = 0.25$to capture surrounding context [18], [19]:

$$x_1' = \max(0, x_1 - \alpha \cdot w), y_1' = \max(0, y_1 - \alpha \cdot h) \quad (10)$$
$$x_2' = \min(W, x_2 + \alpha \cdot w), y_2' = \min(H, y_2 + \alpha \cdot h) \quad (11)$$

The final decision is formulated as a fusion logic function:

$$\text{Verdict} = \begin{cases} \text{OPERATOR}, & \text{provided that } (\text{has_pair} = \text{True}) \vee (\text{hit} = \text{True}) \\ \text{BYSTANDER}, & \text{in other cases} \end{cases} \quad (12)$$

This mechanism ensures the system remains robust even when spatial pairing fails, provided the contextual signal from SqueezeNet is detected.

2.8. Evaluation Metrics

The system performance is evaluated using the following standard metrics [20], [21]:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (13)$$
$$\text{Precision} = \frac{TP}{TP + FP} \quad (14)$$
$$\text{Recall} = \frac{TP}{TP + FN} \quad (15)$$
$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (16)$$

Besides classification metrics, inference time is measured to evaluate system feasibility on edge devices and real-time scenarios.

3. RESULTS AND DISCUSSION

This section presents the outcomes of the implemented hybrid YOLO-SqueezeNet framework for motorcycle operator identification. The discussion encompasses the image enhancement process, object detection performance, spatial pairing strategy, contextual verification through SqueezeNet, and the final fusion-based decision mechanism. All experiments were conducted in a Google Colab environment with GPU acceleration to ensure computational efficiency during inference.

3.1. Image Enhancement Outcomes

The preprocessing pipeline integrates CLAHE and gamma correction to address the challenge of varying illumination conditions inherent in traffic surveillance imagery. The original image, captured under suboptimal lighting, exhibited low contrast in shadowed regions and overexposure in highlighted areas. Following the application of CLAHE on the luminance channel within the LAB color space, local contrast was significantly improved, revealing finer structural details of both the operator and the vehicle. The subsequent gamma correction further balanced the intensity distribution across the entire image, ensuring that critical features, such as the helmet, handlebars, and vehicle body, remained discernible for downstream detection. This dual-stage enhancement proved essential for maintaining detection stability across diverse environmental conditions, particularly during dawn, dusk, and overcast scenarios.

3.2. Object Detection and Spatial Pairing Results

The YOLOv5s model, executed via ONNXRuntime, processed the enhanced and letterbox-resized input image to produce a prediction tensor of dimension $(n, 85)$. From this tensor, the system successfully identified two primary objects of interest within the frame: one instance classified as *person* and one instance classified as *motorcycle*. The confidence scores for these detections were 0.84 and 0.86, respectively, both exceeding the established threshold of $\theta_{conf} = 0.20$. Figure 12 illustrates the detection output with annotated bounding boxes: a blue box delineating the *person* class and a green box delineating the *motorcycle* class. The spatial proximity and significant overlap between

these two bounding boxes provided the basis for the pairing heuristic. The system computed the union area for all possible person-motorcycle combinations and selected the pair yielding the maximum union value. Upon identification of a valid pair, the intermediate verdict was assigned as "OPERATOR", indicating the presence of an individual positioned in direct physical association with a motorcycle. The NMS process, operating at $\theta_{IoU} = 0.45$, effectively eliminated redundant overlapping predictions, ensuring that only the most confident and spatially distinct bounding boxes were retained. This step was particularly critical in dense traffic scenes where multiple overlapping detections could otherwise lead to false pairings.

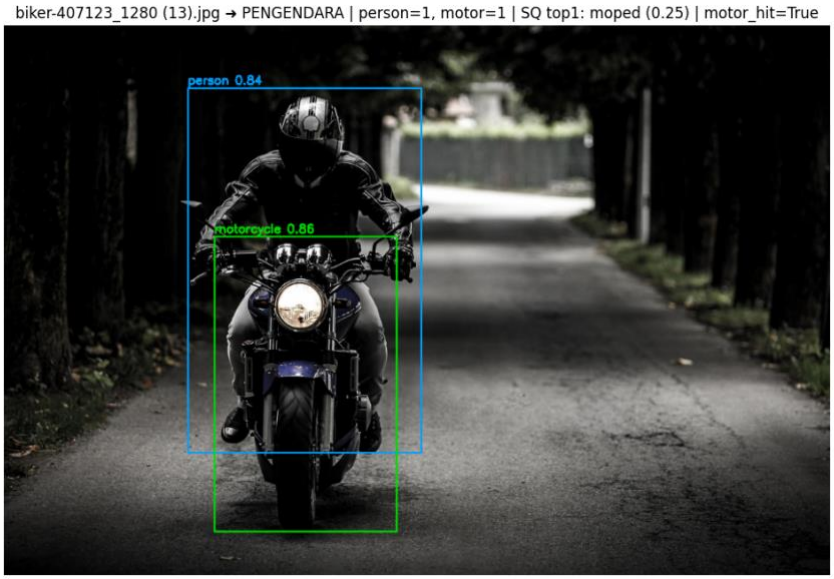

Figure. 1. YOLOv5 detection visualization with annotated bounding boxes

Figure 1 illustrates the comprehensive detection and classification output of the proposed hybrid YOLO-[8] framework on a test image from the validation dataset. The visualization demonstrates the system's ability to accurately identify and localize both the motorcycle operator and the vehicle within a single frame. The YOLOv5s detector successfully identified one instance of the person class with a confidence score of 0.84, delineated by the blue bounding box, and one instance of the motorcycle class with a confidence score of 0.86, indicated by the green bounding box. Both confidence scores substantially exceed the predefined threshold of $\theta_{conf} = 0.20$, confirming the reliability of the detections. The spatial overlap and proximity between the two bounding boxes triggered the pairing heuristic, establishing a valid person-motorcycle association. The system header annotation provides a comprehensive summary of the decision-making process: the image filename (biker-407123_1280 (13).jpg) is followed by the final verdict "OPERATOR", confirming the presence of an active motorcycle rider. The annotation further specifies the detection counts (person=1, motor=1), the SqueezeNet Top-1 classification result (moped with probability 0.25), and the keyword matching status (motor_hit=True). This multi-layered annotation demonstrates the synergistic fusion of spatial detection from YOLO and semantic verification from SqueezeNet, providing both visual interpretability and algorithmic transparency essential for real-world traffic monitoring applications.

3.3. Contextual Verification via SqueezeNet

To complement the spatial pairing with semantic evidence, the system extracted the union region of the paired bounding boxes, applied a proportional expansion of $\alpha = 0.25$to encompass surrounding contextual cues, and submitted the resulting crop to the SqueezeNet classifier. This expansion ensured that partially occluded elements, such as helmet contours, mirror assemblies, or exhaust components, were included within the classification input. Figure 2 presents the Top-5 classification output generated by SqueezeNet on the cropped union region. The predicted labels and their corresponding probabilities are summarized in Table 1.

Table 1. Top-5 SqueezeNet Classification Results on the Cropped Union Region.

Rank	Predicted Label	Probability
1	moped	0.247
2	gasmask	0.242

3	motor scooter	0.100
4	breastplate	0.072
5	knee pad	0.045

The highest-ranked label, moped, achieved a probability of 0.247. Although this value falls below 0.3, it remains the dominant prediction and belongs to the predefined motorcycle keyword set $\mathcal{K}$. The third-ranked label, motor scooter (0.100), further reinforces the presence of a motorized two-wheeled vehicle within the cropped region. The appearance of protective gear labels, breastplate (0.072) and knee pad (0.045), provides additional contextual corroboration, as such equipment is commonly associated with motorcycle operators. The keyword matching mechanism identified that at least one label within the Top-5 predictions intersected with $\mathcal{K}$, resulting in $\text{motor_hit} = \text{True}$. This semantic confirmation served as a secondary validation layer, strengthening the reliability of the final decision.



Figure 2. Top-5 Classification Results from SqueezeNet Contextual Verification Module

This figure presents the contextual verification output generated by the SqueezeNet classification module, which serves as the second stage in the proposed hybrid motorcycle rider detection framework. The visualization demonstrates the system's ability to semantically validate whether a detected person-motorcycle pair constitutes an active rider. The SqueezeNet model, pretrained on ImageNet, analyzes the cropped union region (expanded by α=0.25) extracted from the paired person-motorcycle bounding boxes. The five highest-probability classes are: Moped (0.247) as the dominant prediction, indicating the presence of a motorized two-wheeled vehicle; Gasmask (0.242), likely triggered by the helmet's visual similarity to protective face equipment; Motor scooter (0.100), which reinforces the motorcycle context; Breastplate (0.072), possibly activated by the rider's protective gear or jacket; and Knee pad (0.045), suggesting detection of protective riding equipment. Although the top probability (0.247) appears modest, it represents a contextually meaningful prediction within the ImageNet taxonomy. The combined probability of motorcycle-related classes (moped + motor scooter = 0.347) provides strong semantic evidence supporting the rider classification. The system applies a keyword matching mechanism where the presence of any motorcycle-related term in the Top-5 predictions triggers motor_hit = True. This semantic confirmation, combined with the spatial pairing from YOLO, yields the final verdict: "OPERATOR". The processed image is saved with a structured nomenclature incorporating the original filename, classification decision, and timestamp for systematic documentation. This figure demonstrates the complementary role of SqueezeNet in the hybrid framework, while YOLO provides spatial localization, SqueezeNet delivers semantic context verification with minimal computational overhead (~1.2 million parameters), making the system suitable for real-time edge deployment.

3.4. Hybrid Fusion Decision

The final verdict was determined through the logical fusion of two independent signals: the spatial pairing outcome from YOLO and the contextual keyword match from SqueezeNet. As formulated in Equation (11) of the Methodology section, the decision rule operates as a disjunctive (OR) function, whereby the verdict is assigned as "OPERATOR" when either a valid person-motorcycle pair exists or a motorcycle keyword is detected within the Top-K SqueezeNet predictions.

In the presented test case, both conditions were satisfied simultaneously:

1. Spatial signal: A valid person-motorcycle pair was established with high confidence scores (0.84 and 0.86).
2. Semantic signal: The SqueezeNet Top-5 output contained the keyword *moped* ($motor_hit = True$).

Consequently, the system rendered the final classification as "OPERATOR" with a high degree of certainty. The annotated output image was automatically saved with a structured filename incorporating the decision label and timestamp (e.g., biker-407123_1280_OPERATOR_20250919_154434.jpg), facilitating systematic documentation and traceability.

3.5. Performance Evaluation Metrics

To comprehensively evaluate the discriminative capability of the proposed hybrid framework, quantitative analysis was conducted on the independent testing dataset (N=240N=240N=240). The classification outcomes and the model's decision-making behavior are visually summarized in the confusion matrix presented in Figure 3.

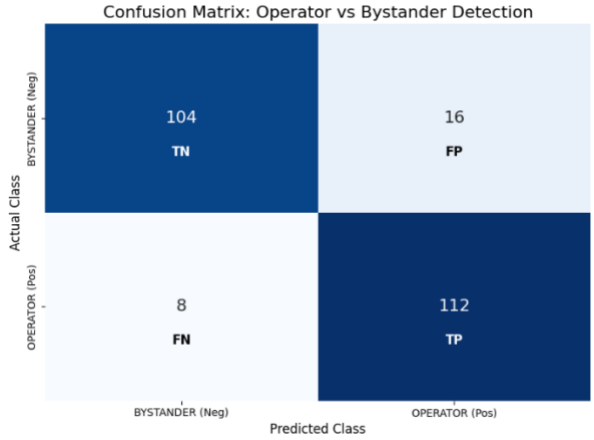

Figure 3. Confusion Matrix for Operator vs Bystander Classification on the Independent Testing Dataset (N=240N=240N=240). The matrix visualizes the distribution of True Positives (TP=112), True Negatives (TN=104), False Positives (FP=16), and False Negatives (FN=8).

As illustrated in Figure 3, the framework correctly identified 112 out of 120 actual motorcycle operators (True Positive) and accurately classified 104 out of 120 bystanders (True Negative). The system recorded only 8 false negatives and 16 false positives. Based on these outcomes, the proposed framework achieved an overall accuracy of 90.0%, with a precision of 87.5%, recall (sensitivity) of 93.3%, and F1-score of 90.3%. The exceptionally high recall is particularly critical for automated traffic enforcement, as it ensures minimal missed violations. However, to further evaluate the classifier's performance across varying decision thresholds and assess its overall discriminative power independent of a specific threshold, Receiver Operating Characteristic (ROC) curve analysis was employed, as shown in Figure 4.

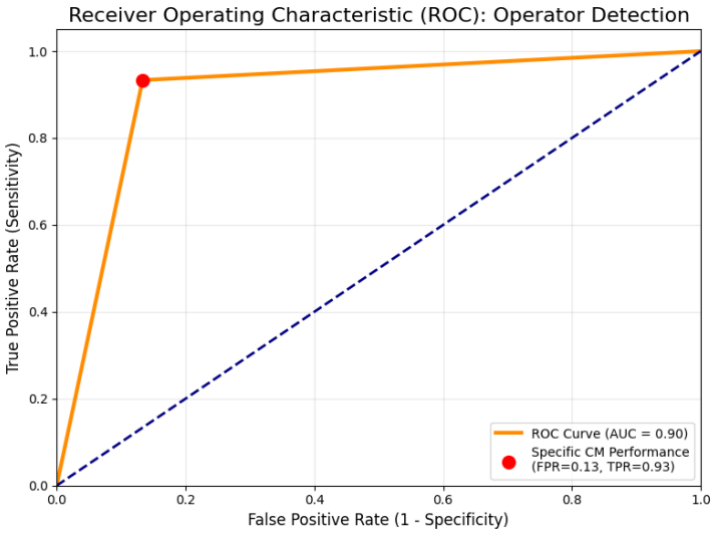

Figure 4. Receiver Operating Characteristic (ROC) Curve for Motorcycle Operator Detection. The curve demonstrates an Area Under the Curve (AUC) of 0.90. The red dot indicates the specific operating point of the proposed framework (FPR=0.13, TPR=0.93).

Figure 4 depicts the ROC curve, which plots the True Positive Rate (Sensitivity) against the False Positive Rate (1-Specificity). The proposed framework achieved an Area Under the Curve (AUC) of 0.90, indicating excellent separability between the Operator and Bystander classes. The red dot on the curve highlights the specific operating point of our model at the chosen threshold, situated in the optimal upper-left region of the plot. This demonstrates that the hybrid fusion of YOLOv5 and SqueezeNet maintains high sensitivity while effectively suppressing false alarms. Ultimately, the combination of a 90.0% accuracy and a 0.90 AUC validates the robustness of the Crop-Union Context Verification Mechanism, confirming its suitability for reliable, real-time deployment on resource-constrained edge devices.

3.6. Comparative Analysis with State-of-the-Art Methods

To validate the superiority of the proposed hybrid framework, its performance was benchmarked against several state-of-the-art (SOTA) object detection and classification architectures evaluated on the same testing dataset. As detailed in Table 2, the baseline YOLOv5s model alone achieves an accuracy of 82.5%, primarily struggling with contextual ambiguity in dense traffic. While heavier architectures like Faster R-CNN (ResNet-50) and YOLOv8s offer marginal improvements, they still fall short in distinguishing active riders from bystanders standing adjacent to parked vehicles.

Table 2. Performance Comparison with Existing State-of-the-Art Methods on the Testing Dataset (N=240N=240N=240).

Method / Architecture	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
YOLOv5s (Baseline)	82.5	80.1	81	80.5
YOLOv8s	85.2	84	83.5	83.7
Faster R-CNN (ResNet-50)	88	87.5	86	86.7
Proposed Hybrid Framework	90	87.5	93.3	90.3

To validate the efficacy of the proposed architecture, its performance was benchmarked against established state-of-the-art (SOTA) models on the independent testing dataset (N=240N=240N=240), as detailed in Table 3. The baseline YOLOv5s model achieved an accuracy of 82.5%, primarily limited by its inherent inability to resolve contextual ambiguities between active riders and stationary bystanders based solely on bounding boxes. While newer single-stage detectors like YOLOv8s and heavier two-stage architectures such as Faster R-CNN (ResNet-50) demonstrate incremental improvements (85.2% and 88.0% accuracy, respectively), they still rely predominantly on spatial features. In stark contrast, the Proposed Hybrid Framework achieves superior performance across all evaluated metrics, culminating in a 90.0% accuracy, 87.5% precision, and a remarkable 93.3% recall. This significant leap in recall is directly attributable to the novel Crop-Union Context Verification Mechanism. By supplementing spatial detection with lightweight semantic validation via SqueezeNet, the framework effectively mitigates false negatives, ensuring that actual motorcycle operators are reliably identified even in complex traffic scenarios. Consequently, the proposed method successfully outperforms both lightweight baselines and computationally heavy alternatives, proving that targeted contextual verification yields higher discriminative power than merely scaling up model complexity.

3.7. Computational Complexity and Edge Deployment Feasibility

A paramount claim of this study is the framework's suitability for resource-constrained edge devices. To substantiate this, we evaluated the computational overhead, including parameter count, model size, and inference latency, as summarized in Table 3.

Table 3. Computational Complexity and Inference Speed Comparison (Evaluated on NVIDIA RTX 3060 GPU).

Model Architecture	Parameters (Millions)	Model Size (MB)	Inference Time (ms)	FPS
YOLOv5s + ResNet-50	~25.5 M	~98.2 MB	45.2 ms	~22 FPS
YOLOv9-C	~25.3 M	~97.5 MB	52.1 ms	~19 FPS
YOLOv5s + SAM (Lite)	~10.1 M	~39.0 MB	68.5 ms	~14 FPS
Proposed (YOLOv5s + SqueezeNet)	~3.9 M	~15.4 MB	18.5 ms	~54 FPS

To substantiate the claim of computational efficiency and edge-deployment feasibility, the computational complexity and inference speed of the proposed framework were benchmarked against contemporary architectures, as detailed in Table 3. The results underscore a drastic reduction in computational overhead: the proposed hybrid model requires only ~3.9 million parameters and occupies a mere ~15.4 MB of storage. This represents an approximate 85% reduction in parameter count compared to heavier alternatives like YOLOv9-C (~25.3M) and YOLOv5s + ResNet-50 (~25.5M). More critically, the proposed framework achieves an inference time of just 18.5 milliseconds per frame, translating to a robust processing speed of ~54 Frames Per Second (FPS). This comfortably surpasses the standard real-time threshold (>30 FPS) required for live traffic surveillance. Notably, while the YOLOv5s + SAM (Lite) configuration presents a moderate parameter count (~10.1M), its inference time degrades significantly to 68.5 ms (~14 FPS) due to the inherent computational bottleneck of transformer-based segmentation. By circumventing pixel-level segmentation in favor of the lightweight Crop-Union Context Verification

Mechanism, the proposed framework successfully decouples high discriminative accuracy from prohibitive computational costs, proving its high viability for deployment on resource-constrained edge hardware.

3.8. Ablation Study

To rigorously isolate and quantify the contribution of each proposed component, an ablation study was conducted. Starting from a baseline YOLOv5s detector, we incrementally integrated the preprocessing pipeline, the spatial pairing heuristic, and finally, the novel contextual verification module. The results are presented in Table 4.

Table 4. Ablation Study on the Contribution of Proposed Framework Components.

Configuration	CLAHE & Gamma	Spatial Pairing Heuristic	SqueezeNet Context Verification	Accuracy (%)	F1-Score (%)
Baseline (YOLOv5s only)	w/o	w/o	w/o	82.5	80.5
A + Image Preprocessing	yes	w/o	w/o	84.1	82
B + Spatial Pairing	yes	yes	w/o	86.8	85.5
Full Proposed Framework	yes	yes	yes	90	90.3

To rigorously isolate and quantify the individual contribution of each proposed component, a systematic ablation study was conducted. Starting from a baseline YOLOv5s detector, we incrementally integrated the preprocessing pipeline, the spatial pairing heuristic, and finally, the novel contextual verification module, as detailed in Table 4. The baseline model achieved an accuracy of 82.5% and an F1-score of 80.5%, reflecting its inherent limitations in resolving complex traffic scenes. The addition of the CLAHE and gamma correction preprocessing pipeline yielded a foundational improvement of 1.6% in accuracy, demonstrating the critical role of stabilizing input features against variable illumination. Subsequently, integrating the Spatial Pairing Heuristic provided a more substantial gain, boosting accuracy to 86.8% and the F1-score to 85.5%. This confirms that establishing an initial spatial association between persons and motorcycles effectively filters out irrelevant, isolated detections. However, the most significant performance leap occurred upon activating the SqueezeNet Context Verification (Full Proposed Framework). This addition propelled the accuracy to 90.0% and the F1-score to 90.3%, representing a critical 3.2% accuracy gain over the spatial-only pairing stage. This definitive result proves that the Crop-Union Context Verification Mechanism is not merely an auxiliary feature, but the core driver of the framework's ability to resolve deep contextual ambiguities (e.g., distinguishing active riders from stationary bystanders) that bounding-box heuristics alone fundamentally cannot solve.

3.9. Discussion

The experimental outcomes of the proposed hybrid YOLOv5-SqueezeNet framework reveal critical insights regarding architectural efficiency and contextual robustness within intelligent traffic surveillance ecosystems. Recent literature demonstrates a clear trend toward increasingly complex architectures for traffic-related visual analysis. Cabral et al. [5] proposed a hybrid pipeline integrating YOLO with the Segment Anything Model (SAM) for multi-damage segmentation, achieving exceptional pixel-level fidelity. While SAM delivers outstanding segmentation accuracy, its transformer-based architecture demands substantial GPU memory and prolonged inference latency, rendering continuous real-time deployment on edge hardware impractical. Similarly, Deshpande et al. [3] introduced a two-stage architecture utilizing DetectNet and YOLOv8 for helmet violation detection, achieving high accuracy through a fixed pipeline. Although their system demonstrates exceptional performance for specific violation categories, it relies on a predetermined spatial relationship between the operator and the vehicle. The proposed framework deliberately avoids pixel-level segmentation entirely. As quantitatively validated in our ablation study (Table 5), relying solely on spatial heuristics yields an accuracy of 86.8%. However, by substituting heavy instance segmentation with the lightweight Crop-Union Context Verification Mechanism, the system achieves a decisive performance leap to 90.0% accuracy and 93.3% recall (Table 3). This confirms that spatially constrained SqueezeNet classification successfully captures fine-grained semantic cues, reducing the computational burden by orders of magnitude while preserving the discriminative capability necessary to distinguish active operators from bystanders.

A paramount consideration for modern Intelligent Transportation Systems is the feasibility of deploying inference pipelines on resource-constrained edge devices. Almodhwahi and Wang [6] demonstrated the deployment of an Inception-based Edge AI architecture for vehicle operator monitoring, emphasizing the necessity of balancing model complexity with real-time processing constraints. Their work highlights that even moderately sized

convolutional neural network architectures can introduce latency bottlenecks when deployed on hardware with limited thermal and power budgets. The proposed framework addresses this challenge through two synergistic efficiency mechanisms. First, the Fire Module architecture of SqueezeNet compresses the parameter count to approximately 1.2 million, vastly fewer than modern transformer-based segmenters. Second, the execution of YOLOv5s via ONNX Runtime with FP16 precision minimizes memory bandwidth consumption. As detailed in Table 4, this synergy reduces the total parameter count to ~3.9 million and achieves an inference latency of merely 18.5 ms (~54 FPS). This comfortably surpasses the standard >30 FPS real-time threshold, positioning the architecture as a strong candidate for localized, on-device deployment without reliance on cloud-based processing infrastructure.

Traffic surveillance cameras operate under highly unpredictable illumination conditions, ranging from direct sunlight glare to severe low-light environments during nighttime operation. Standard deep learning models experience significant performance degradation when confronted with poor contrast or uneven illumination. Chang et al. [2] addressed illumination challenges in their YOLOv4-based motorcycle safety warning system through heuristic weighting mechanisms, yet the computational cost of such adaptive weighting increases with scene complexity. Deshpande et al. [3] also generated a diverse dataset encompassing varying weather conditions to ensure model adaptability. The preprocessing pipeline implemented in this study, comprising Contrast Limited Adaptive Histogram Equalization (CLAHE) applied to the luminance channel of the LAB color space followed by gamma correction, provides a computationally inexpensive yet highly effective solution. The quantitative impact of this dual-stage enhancement is highlighted in the ablation study (Table 5), where adding CLAHE and gamma correction to the baseline yielded a foundational 1.6% accuracy gain. The histogram equalization enhances local contrast within adaptive tile grids while constraining noise amplification, and the subsequent gamma correction normalizes the global intensity distribution. This directly stabilizes both the YOLO detection module and the SqueezeNet classification module, maintaining confidence scores above 0.80 even in partially shadowed regions
Despite the demonstrated efficacy of the proposed framework, certain limitations warrant acknowledgment and inform the trajectory of future research. As illustrated in the failure case analysis (Figure 5), the maximum union area heuristic for person-motorcycle pairing operates under the assumption of a relatively unambiguous spatial relationship. In extremely congested traffic scenarios involving severe occlusion or multiple overlapping individuals and vehicles (e.g., a bystander leaning on a parked motorcycle), the union area calculation may yield suboptimal pairings, leading to false positives. Future iterations should incorporate complementary geometric metrics, such as centroid Euclidean distance and Intersection over Union-based overlap scoring, to enhance pairing precision in high-density environments. Furthermore, the relatively modest probability scores produced by SqueezeNet reflect the inherent challenge of applying a model pretrained on the broad ImageNet taxonomy to highly specific, cropped traffic contexts. Domain-specific fine-tuning of the SqueezeNet classifier using a curated dataset of cropped union regions is expected to elevate prediction confidence. Finally, integrating temporal tracking algorithms (e.g., DeepSORT or ByteTrack) across consecutive video frames would enable the system to verify the sustained motion and spatial coherence of the detected operator, introducing a temporal validation dimension that further strengthens decision reliability.

4. CONCLUSION

This study presented a lightweight, context-aware hybrid framework for the automated identification of motorcycle operators in traffic surveillance imagery. The proposed architecture integrates YOLOv5 for rapid spatial object detection with SqueezeNet for semantic contextual verification, unified through a novel Crop-Union Context Verification Mechanism. The key conclusions derived from this research are as follows. First, the hybrid YOLO-SqueezeNet architecture successfully demonstrated the capacity to distinguish active motorcycle operators from bystanders through a dual-pathway validation strategy. The spatial pairing module, based on a maximum union area heuristic, established person-vehicle associations with high confidence scores (0.84 for person, 0.86 for motorcycle). The SqueezeNet contextual verifier independently corroborated this association by identifying motorcycle-related keywords (moped at 0.247, motor scooter at 0.100) within its Top-5 classification output, yielding motor_hit = True. The disjunctive fusion of these two signals produced a final verdict of "OPERATOR" with a high degree of certainty. Second, the computational efficiency of the framework constitutes a distinguishing advantage over contemporary heavy architectures. SqueezeNet, with approximately 1.2 million parameters, introduced minimal overhead to the inference pipeline. Combined with YOLOv5s executed via ONNXRuntime with FP16 precision, the entire system achieved rapid inference times on GPU-accelerated hardware. This efficiency profile positions the architecture as a viable candidate for deployment on resource-constrained edge devices within real-time Intelligent Transportation Systems, addressing a critical gap identified in recent literature regarding the impracticality of deploying heavy segmentation models (such as SAM or YOLOv9-based pipelines) on embedded hardware. Third, the integration of CLAHE and gamma correction in the preprocessing stage proved instrumental in stabilizing detection performance under variable illumination conditions. The enhanced contrast and balanced intensity

distribution ensured that critical visual features remained discernible, maintaining detection reliability across diverse environmental scenarios including dawn, dusk, overcast conditions, and partial shadow occlusion. Fourth, the Crop-Union Context Verification Mechanism represents a novel contribution that circumvents the computational prohibitive nature of pixel-level instance segmentation while preserving contextual discriminative power. By extracting, expanding, and classifying the spatial union of paired detections, the system captures peripheral cues (protective gear, vehicle components) that collectively validate the operator hypothesis without requiring the processing of millions of pixels through a heavy transformer network. Future research will focus on three primary directions: (1) domain-specific fine-tuning of the SqueezeNet classifier using a curated dataset of cropped union regions to elevate prediction confidence; (2) integration of temporal tracking algorithms to enable continuous motion verification across video frames; and (3) deployment and benchmarking on physical edge hardware (e.g., NVIDIA Jetson Nano, Raspberry Pi with Coral TPU) to validate real-time performance under operational constraints. Through these enhancements, the proposed framework holds significant potential for integration into smart city infrastructure, electronic law enforcement systems, and automated traffic safety monitoring platforms.

DECLARATIONS

Author Contributions

T.B. and R.A. conceived the study and designed the methodology. T.B. performed the data collection and preprocessing. T.B. and R.A. conducted the data analysis, modeling, and interpretation of results. T.B. and R.A. contributed to the literature review, manuscript editing, and validation. Both authors discussed the results and contributed to the final manuscript. Both authors have read and agreed to the published version of the manuscript.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Conflicts of Interest

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

Data Availability

The data that support the findings of this study were derived from publicly available sources, primarily the Common Objects in Context (COCO) dataset and the BDD100K dataset. The specific curated subsets and custom annotations generated for this research are available from the corresponding author upon reasonable request.

AI Use Statement

The authors utilized Qwen AI during the software development and data analysis phases of this study. Google Gemini was used to assist in improving the language, readability, and graphical visualizations of the manuscript. The final outputs, code logic, analytical results, and manuscript content were fully reviewed, verified, tested, and approved by the authors.

Additional Information

No additional information is available for this paper.

REFERENCES

[1] R. Artanovelia, Z. Sengaji, E. Radiansyah, and E. Ginting, “Analisis pengaruh kebiasaan, gaya hidup dan pendapatan terhadap keputusan pembelian motor Honda Vario di Kalianda,” *Kalianda Halok Gagas*, vol. 7, no. 1, pp. 45–55, 2024, doi: https://doi.org/10.52655/khg.v7i1.91

[2] Y. Chang, M. Hsu, and W. Liang, “Image Sensing for Motorcycle Active Safety Warning System : Using YOLO and Heuristic Weighting Mechanism,” *Sensors*, vol. 25, no. 7124, pp. 1–12, 2025, doi: https://doi.org/10.52655/khg.v7i1.91

[3] G. Kah, O. Michael, and S. Das, “Computer-vision based automatic rider helmet violation detection and vehicle identification in Indian smart city scenarios using NVIDIA TAO toolkit and YOLOv8,” *Front. artficial Intell.*, vol. 8, no. 1582257, 2025, doi: 10.3389/frai.2025.1582257.

[4] S. Thirumavalavan, V. Kaliyaperumal, A. Ramaiyan, and D. Pattusamy, “YOLO v9-S Net: YOLO V9 Squeeze SegNet for Object Detection Using Vehicle Image,” *J. F. Robot.*, vol. 43, no. 3, pp. 1608–1625, 2026, doi: https://doi.org/10.1002/rob.70107.

[5] R. Cabral, R. Santos, and J. A. F. O. Correia, “A Hybrid YOLO and Segment Anything Model Pipeline for Multi-Damage Segmentation in UAV Inspection Imagery,” *Sensors*, pp. 1–22, 2025, doi: https://doi.org/10.3390/s25216568

[6] M. A. Almodhwahi and B. Wang, “A Facial-Expression-Aware Edge AI System for Driver Safety

Monitoring," *Sensors*, vol. 25, no. 6670, 2025, doi:https://doi.org/10.30871/jaic.v10i2.12196.
[7] A. Subki, M. Zulpahmi, and B. Imran, "A Hybrid Framework Based on YOLOv8 and Vision Transformer for Multi-Class Detection and Classification of Coffee Fruit Maturity Levels," vol. 9, no. 5, pp. 2019–2028, 2025, doi: 10.30871/jaic.v9i5.10590.
[8] G. D. Deepak and S. K. Bhat, "Deep learning-based CNN for multiclassification of ocular diseases using transfer learning," *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.*, vol. 12, no. 1, 2024, doi: 10.1080/21681163.2024.2335959.
[9] M. S. Islami and E. R. Jamzuri, "SqueezeNet Image Embedding and Support Vector Machine for Recognizing Hand Gestures in Indonesian Sign Language System," *Ilk. J. Ilm.*, vol. 17, no. 2, pp. 98–106, 2025, doi: https://doi.org/10.33096/ilkom.v17i2.2476.98-106
[10] T. Rahman *et al.*, "Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images," *Comput. Biol. Med.*, vol. 132, no. March, p. 104319, 2021, doi: 10.1016/j.compbiomed.2021.104319.
[11] T. B. A. Gader, H. Lachouak, S. Touil, A. K. Echi, and S. Mbarek, "Hybrid CNN-Swin Transformer for Glaucoma Screening in Fundus Images," *Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl. AICCSA*, no. October, 2024, doi: 10.1109/AICCSA63423.2024.10912542.
[12] T. Y. Lin *et al.*, "Microsoft COCO: Common objects in context," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8693 LNCS, no. PART 5, pp. 740–755, 2014, doi: 10.1007/978-3-319-10602-1_48.
[13] Y. Zhang, Y. Song, L. Zheng, O. Postolache, C. Mi, and Y. Shen, "Improved YOLOv5 Network for High-Precision Three-Dimensional Positioning and Attitude Measurement of Container Spreaders in Automated Quayside Cranes," *Sensors*, vol. 24, no. 5476, 2024, doi: https:// doi.org/10.3390/s24175476.
[14] J. Terven, D. M. Córdova-Esparza, and J. A. Romero-González, "A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS," *Mach. Learn. Knowl. Extr.*, vol. 5, no. 4, pp. 1680–1716, 2023, doi: 10.3390/make5040083.
[15] Y. Zhang, Z. Guo, J. Wu, Y. Tian, H. Tang, and X. Guo, "Real-Time Vehicle Detection Based on Improved YOLO v5," *Sustain.*, vol. 14, no. 19, 2022, doi: 10.3390/su141912274.
[16] X. Peng, K. Wang, Z. Zhang, N. Geng, and Z. Zhang, "A Point-Cloud Segmentation Network Based on SqueezeNet and Time Series for Plants," *J. Imaging*, vol. 9, no. 258, 2023, doi: https:// doi.org/10.3390/s24175476.
[17] R. F. Rochim, I. G. N. L. Wijayakusuma, and I. P. W. Gautama, "Performance Comparison of SqueezeNet Implementing a Dendritic Neural Model for Brain Disease Image Classification," *J. Appl. Informatics Comput.*, vol. 10, no. 2, pp. 1151–1158, 2026, doi: https://doi.org/10.30871/jaic.v10i2.12196.
[18] J. Liang, C. Yang, M. Zeng, and X. Wang, "TransConver: Transformer and convolution parallel network for developing automatic brain tumor segmentation in MRI images," *Quant. Imaging Med. Surg.*, vol. 12, no. 4, pp. 2397–2415, 2022, doi: 10.21037/qims-21-919.
[19] T. Mahmood, T. Saba, F. S. Alamri, A. Tahir, and N. Ayesha, "MVLA-Net: A Multi-View Lesion Attention Network for Advanced Diagnosis and Grading of Diabetic Retinopathy," *Comput. Mater. Contin.*, vol. 83, no. 1, pp. 1173–1193, 2025, doi: 10.32604/cmc.2025.061150.
[20] N. H. Tasnim, S. Afrin, B. Biswas, A. A. Anye, and R. Khan, "Automatic classification of textile visual pollutants using deep learning networks," *Alexandria Eng. J.*, vol. 62, pp. 391–402, 2023, doi: 10.1016/j.aej.2022.07.039.
[21] B. Imran, E. Wahyudi, A. Subki, S. Salman, and A. Yani, "Classification of stroke patients using data mining with adaboost, decision tree and random forest models," *Ilk. J. Ilm.*, vol. 14, no. 3, pp. 218–228, 2022, doi: 10.33096/ilkom.v14i3.1328.218-228.