FOREX Publication
Open Access | Rapid and quality publishing

International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 3 | Pages 419-428 | e-ISSN: 2347-470X

YOLOv8-SOR: A Small-Object-Responsive Road Obstacle Detection Model using RepVGG and Swin Transformer

LiKang Bo[1,2*], Fei Lu Siaw[1], Tzer Hwai Gilbert Thio[1]

[1]*Centre for Sustainability in Advanced Electrical and Electronic Systems (CSAEES), Faculty of Engineering, Built Environment and Information Technology, SEGi University, 47810 Petaling Jaya, Selangor, Malaysia;*
[2]*Hebei Vocational University of Technology and Engineering, No.473, Quannan West Street, Xindu District, Xingtai, Hebei, China, 054000*

***Correspondence:** *LiKang Bo; bolk88@163.com*

ABSTRACT- Accurate and efficient road obstacle detection is crucial for ensuring the safety and reliability of autonomous and assisted driving systems. However, current object detection models, including YOLOv8, often encounter difficulties in accurately detecting small obstacles and balancing detection precision with inference speed. To address these challenges, we propose an improved YOLOv8-based detection framework featuring three key enhancements: (1) the backbone is redesigned using RepVGG modules to accelerate inference through structural re-parameterization without compromising feature representation; (2) an additional small-object detection head at the P2 feature level is introduced to significantly enhance sensitivity towards small-scale obstacles; (3) a Swin Transformer module is integrated into the final stage of the backbone to improve global contextual understanding and semantic feature extraction. Extensive experiments conducted on a comprehensive, multi-source road obstacle dataset demonstrate that our proposed model achieves superior performance with a precision of 0.812, recall of 0.782, mAP@0.5 of 0.822, and mAP@0.5:0.95 of 0.588. Additionally, it maintains a real-time inference speed of 180 FPS on an NVIDIA H800 GPU. The results confirm the efficacy of our approach, particularly in enhancing small object detection in complex real-world traffic scenarios.

Keywords: Road Obstacle Detection, Small Object Detection, RepVGG, Transformer.

ARTICLE INFORMATION

Author(s): LiKang Bo, Fei Lu Siaw, and Tzer Hwai Gilbert Thio;
Received:14/05/25; **Accepted:** 22/07/25; **Published:** 30/08/25;
E- ISSN: 2347-470X;
Paper Id: IJEER 1405-10;
Citation: 10.37391/ijeer.130305
Webpage-link:
https://ijeer.forexjournal.co.in/archive/volume-13/ijeer-130305.html

Publisher's Note: FOREX Publication stays neutral with regard to jurisdictional claims in Published maps and institutional affiliations.

1. INTRODUCTION

Advances in intelligent transportation systems and autonomous driving technologies have made effective road obstacle detection essential for vehicle safety and navigational reliability[1]. Vehicles, pedestrians, and static objects such as traffic cones and fire hydrants must be detected accurately and efficiently in real-time to prevent accidents and support timely decision-making[2]. Deep learning has greatly improved object detection, especially through one-stage detectors like the YOLO (You Only Look Once) series. These models strike a strong balance between detection accuracy and inference speed[3].

YOLOv8, the latest in the series, adopts anchor-free detection, a decoupled head design, and a lightweight architecture, making it a strong baseline for real-time tasks. However, YOLOv8 still faces several challenges in road obstacle detection. It struggles to detect small and densely packed obstacles due to weak shallow feature representation. Its fully convolutional structure limits global contextual perception. Furthermore, its structural efficiency is not optimal for real-time and embedded deployments.

To address these issues, we propose an enhanced YOLOv8-based framework specifically designed for road obstacle detection. Our approach improves small-object recognition and strengthens global context modeling through three key innovations. First, we replace the convolutional backbone with RepVGG modules, which use structural re-parameterization to achieve faster inference and effective feature learning[4]; Second, we add a dedicated detection head at the P2 feature level, optimized for small-object detection. Third, we integrate a lightweight Swin Transformer module into the backbone's final stage to model long-range dependencies and capture global context[5]. Together, these improvements significantly enhance detection accuracy and robustness for small, occluded, and complex targets without sacrificing real-time performance.

The remainder of this paper is structured as follows: *section 2* reviews related work in object detection; *section 3* details the proposed model's architecture; *section 4* presents extensive experimental results, including ablation studies and

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 3 | Pages 419-428 | e-ISSN: 2347-470X

comparative evaluations; and *section 5* summarizes our findings and suggests directions for future research.

2. RELATED WORK

Object detection is a core task in computer vision and is crucial for autonomous driving, intelligent transportation, and road safety[6]. Over the past decade, deep learning-based methods have driven rapid progress. These algorithms are typically divided into two-stage and one-stage frameworks[3]. Two-stage detectors, such as Faster R-CNN, use region proposal networks followed by classification and regression heads. They achieve high accuracy but at the cost of slower inference. In contrast, one-stage detectors like SSD, RetinaNet, and the YOLO series perform detection in a single step, enabling fast inference suitable for real-time applications.

The YOLO series has evolved quickly. From YOLOv1 to YOLOv8, the models have introduced anchor-free prediction, decoupled heads, and improved feature fusion mechanisms[7]. YOLOv8 achieves a good balance between accuracy and efficiency and serves as a strong baseline for many detection tasks. However, it still struggles in road environments with small and diverse obstacles. Its shallow feature representation limits small-object detection, and it's purely convolutional architecture lacks global semantic perception[8].

Researchers have proposed various methods to improve small-object detection. Zhou et al.[9] introduced CBBI-YOLO, which combines the CBAM attention module and BiFPN for better accuracy in agricultural scenes. Another study[10] incorporated SPD-Conv into YOLOv7 for pig detection, enhancing robustness and recall. Wang et al. [11] integrated GhostNetV3 into YOLOv8 to create a lightweight, faster model. These approaches improved specific aspects but often tradeoff between accuracy and speed, and few address global context modeling or explicitly improve small-object recall in road scenarios.

In recent years, extensive efforts have been made to enhance the performance of object detectors on small targets. One common approach is to improve multi-scale feature representation through structures like Feature Pyramid Networks (FPN), Path Aggregation Networks (PAN), and [12-14]. These architectures help preserve shallow-layer spatial information and fuse it with deeper semantic features. Another line of work introduces attention mechanisms such as SE, CBAM, and coordinate attention to adaptively reweight important features, thereby improving focus on relevant areas[15]. Some researchers have also explored lightweight backbones, such as GhostNet, MobileNetV3, and RepVGG, to accelerate inference while maintaining accuracy[4, 16]. In addition, adding P2 or P1 detection heads captures high-resolution features, improving small-object recall.

Despite these advances, most solutions address individual issues and fail to jointly optimize accuracy, small-object detection, and inference efficiency[17]. Few have explored replacing convolutional backbones with transformer-based structures, which can capture long-range dependencies and enhance semantic awareness—critical in complex, cluttered road scenes.

In summary, prior studies have improved object detection through attention mechanisms, lightweight backbones, and head-level modifications. However, no unified approach simultaneously improves small-object detection, global context understanding, and real-time performance. To fill this gap, we propose an improved YOLOv8-based model with three innovations: a RepVGG encoder for faster inference, a P2 detection head for small objects, and a Swin Transformer to enhance long-range dependency modeling. Together, these components deliver a high-performance and deployment-ready detection framework.

3. PROPOSED METHODOLOGY

YOLOv8 is a state-of-the-art anchor-free one-stage object detector that strikes a strong balance between detection accuracy and inference efficiency. Due to its high performance and deploy ability, YOLOv8 is selected as the baseline model for the proposed improvements in this work[18]. As shown in *figure 1*, the YOLOv8 architecture consists of three main components: the backbone for feature extraction, the neck for multi-scale feature fusion, and the head for final prediction.

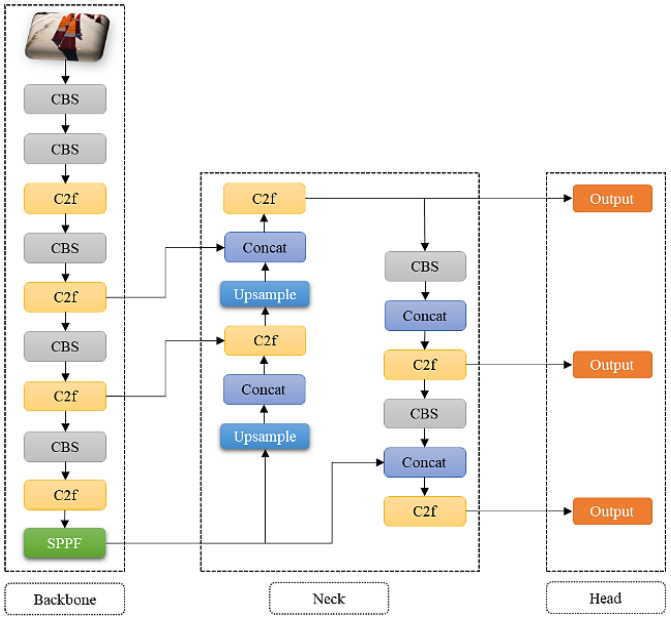


Figure 1. The basic structure of the YOLOv8 model, which serves as the baseline architecture in this study

To address the challenges of low accuracy in small-object detection, limited semantic awareness, and suboptimal inference efficiency in complex road environments, this paper proposes an improved YOLOv8-based model tailored for road obstacle detection. The enhanced architecture introduces three key modifications: a re-parameterized encoder, a small-object detection head, and global semantic modeling enhancements.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 3 | Pages 419-428 | e-ISSN: 2347-470X

3.1. RepVGG-Based Encoder Optimization

The first major modification focuses on optimizing the backbone structure of YOLOv8 to improve inference speed and maintain feature representation quality.

The backbone of an object detection model plays a crucial role in extracting multilevel semantic features and significantly influences both accuracy and inference efficiency[19]. In the original YOLOv8 architecture, the backbone is composed of stacked C2f modules, which provide sufficient feature reuse through concatenation and shortcut connections. However, these structures introduce computational redundancy, making the model less suitable for high-speed or edge deployments where real-time performance is essential[20].

To address this limitation, we redesign the backbone by replacing all C2f modules with RepVGG blocks, a structure known for its simplicity during inference and strong learning capacity during training. As shown in *figure 2*, each RepVGG block adopts a multi-branch structure during training, consisting of a 3×3 convolution, a 1×1 convolution, and an identity mapping. This design improves feature expressiveness and gradient flow. In the inference phase, these branches are re-parameterized into a single 3×3 convolution, effectively simplifying the model without degrading accuracy.

The structural re-parameterization process is defined as:

$$W_{fused} = W_{3\times3} + pad(W_{1\times1}) + pad(I), b_{fused} = b_{3\times3} + b_{1\times1} + b_I \quad (1)$$

where W and b represent the convolution weights and biases of each branch, and pad(·) denotes zero-padding to match the 3×3 kernel size. This allows the model to maintain high representational capacity during training, while being converted into a fast, single-branch convolutional path during inference.

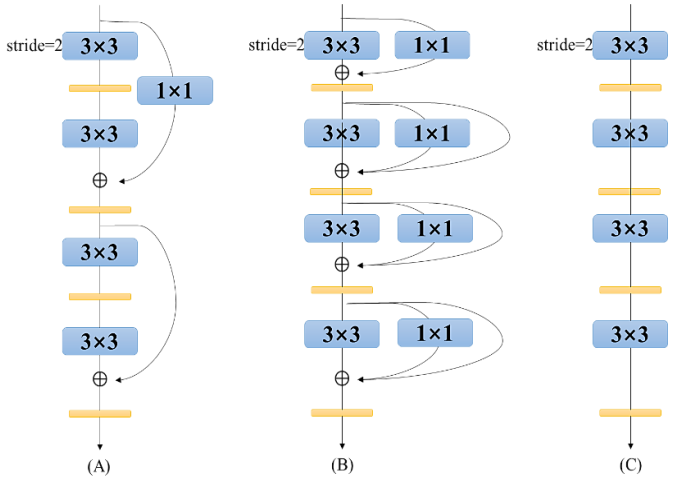


Figure 2. Structure of the RepVGG module

By integrating RepVGG into the YOLOv8 backbone, the proposed model benefits from both rich training representations and fast inference[4]. To better illustrate the effect of replacing all convolutional blocks (CBS) with RepVGG modules in the YOLOv8 backbone, the intermediate architecture after this substitution is shown in *figure 3*.

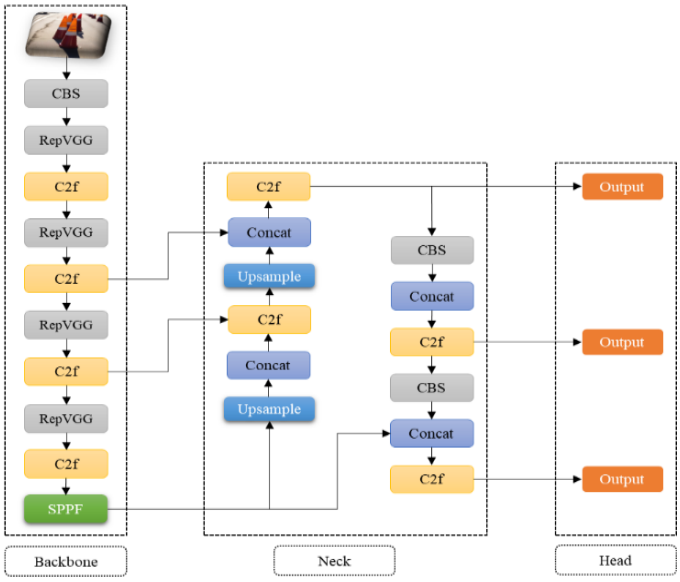


Figure 3. Architecture of the YOLOv8 model after replacing all convolutional blocks (CBS) in the backbone with RepVGG modules

3.2. Multi-scale Small Object Detection Head Design

Although YOLOv8 performs well on medium and large targets, it still struggles to detect small-scale obstacles in road scenes. This limitation is largely due to the fact that the original model utilizes only three output detection heads at the P3, P4, and P5 levels, which correspond to progressively downsampled feature maps[21]. As a result, high-resolution spatial information needed for small-object detection is partially lost in deeper layers.

To address this issue, we introduce an additional detection head at the **P2 level**, which corresponds to a shallower, higher-resolution feature map. This detection head is responsible for capturing fine-grained features that are crucial for accurately detecting small objects such as cones, fire hydrants, and warning pillars. The proposed head uses standard convolutional layers followed by a prediction module, consistent with the existing YOLOv8 architecture, to maintain structural compatibility.

By explicitly adding a P2 output layer to the head, the model benefits in the following ways:

- **Improved recall for small targets**, as their features are more likely to be preserved at higher resolutions;
- **Better spatial localization**, particularly for tightly clustered or partially occluded objects;
- **Minimal impact on inference speed**, as the shallow layer computation is relatively lightweight.

In our ablation experiments (see Section 4.3), the inclusion of the P2 detection head led to a noticeable improvement in recall

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 3 | Pages 419-428 | e-ISSN: 2347-470X

and mAP for small-object categories. This confirms the importance of leveraging shallow feature maps for fine-scale obstacle detection in real-world road scenarios.

To better illustrate the architectural adjustment introduced by the additional P2 detection head, the modified YOLOv8 structure is shown in *figure 4*. The added branch (highlighted in red) operates on a high-resolution shallow feature map to improve small-object detection performance.

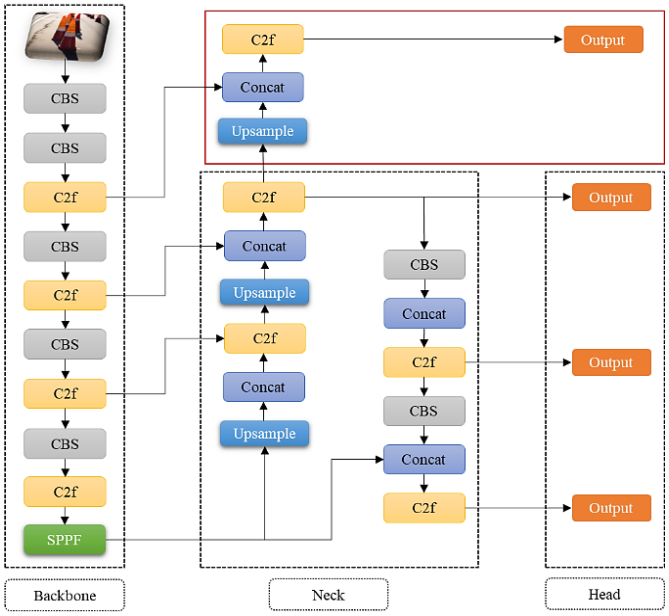


Figure 4. Architecture of the proposed model with an added P2 detection head based on YOLOv8

The P2 branch, outlined in red, uses high-resolution shallow features and feeds them through upsampling, concatenation, and a C2f module before generating prediction outputs. This modification enhances the detection of small-scale obstacles.

3.3. Swin Transformer-Based Context Enhancement

While convolutional neural networks (CNNs) are highly effective at capturing local spatial features, they are inherently limited in modeling long-range dependencies. In complex road scenes—especially those involving occlusion, overlapping obstacles, and cluttered backgrounds—global contextual understanding becomes essential for precise object detection[17]. To address this issue, we replace the final C2f module in the YOLOv8 backbone with a lightweight **Swin Transformer block** to enhance semantic perception.

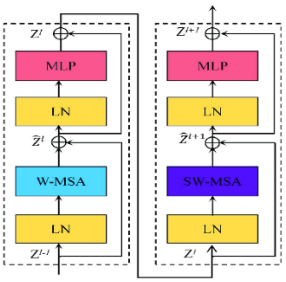


Figure 5. Internal structure of the Swin Transformer block

The **Swin Transformer**, short for Shifted Window Transformer, is a hierarchical vision transformer that partitions feature maps into fixed-size windows and computes self-attention within each window. To facilitate cross-window information exchange, it introduces **shifted windows** in alternate layers, enabling efficient global interaction while maintaining linear computational complexity with respect to image size[22].

In our design, the Swin Transformer block takes the high-level feature map output from the penultimate C2f module and replaces the last C2f block in the original YOLOv8 backbone. This location is chosen intentionally to allow the transformer to operate on semantically rich features while minimizing additional computational overhead. The output of the Swin Transformer maintains the same resolution and number of channels as the original C2f module, ensuring seamless integration with downstream neck and head components.

The benefits of this modification are twofold:

- **Enhanced global feature modeling**, improving the model's understanding of spatial relationships in complex scenes;
- **Improved detection robustness** in challenging conditions such as occlusion, dense object distributions, and visually ambiguous contexts.

The updated YOLOv8 backbone after replacing the final C2f module with Swin Transformer is depicted in *figure 6*.

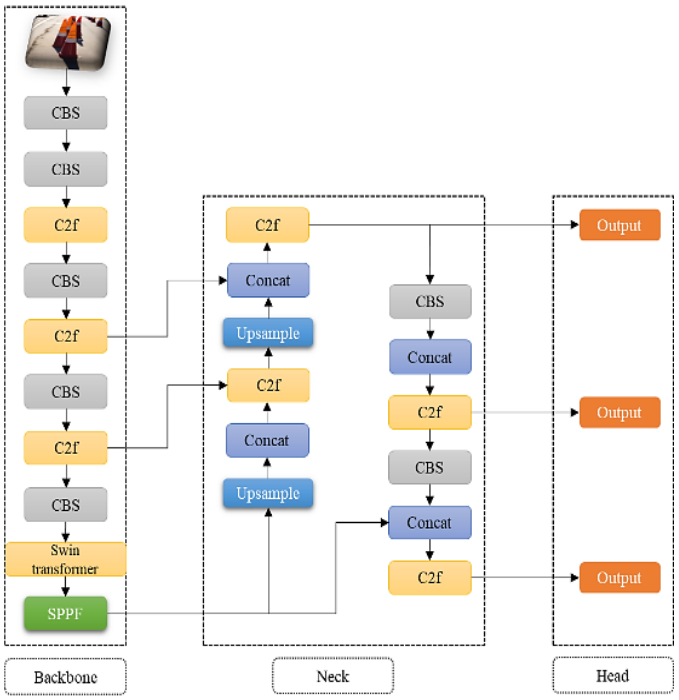


Figure 6. YOLOv8 backbone after replacing the final C2f module with a Swin Transformer block

The final improved YOLOv8-based model incorporating all three proposed modules is presented in *figure 7*.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 3 | Pages 419-428 | e-ISSN: 2347-470X

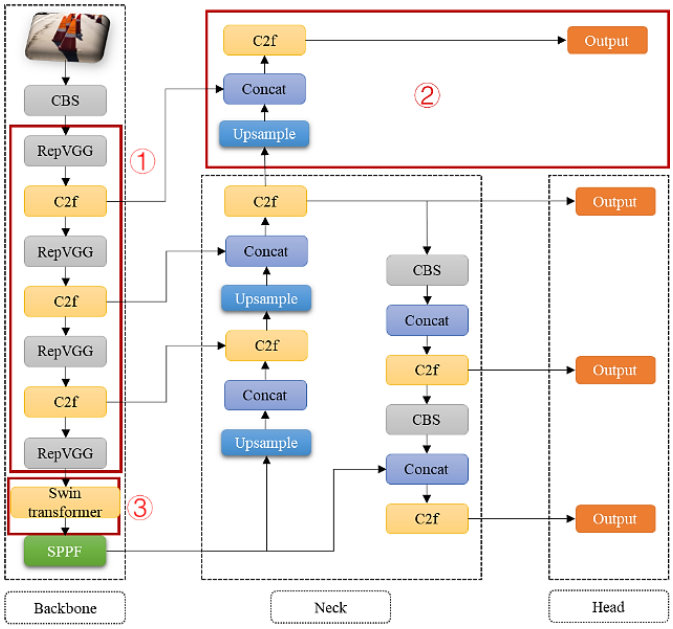


Figure 7. Overall architecture of the proposed improved YOLOv8-based model

3.4. Summary of the Proposed Model

To improve the performance of YOLOv8 in complex road environments—particularly for small and occluded obstacles—we propose a series of architectural enhancements summarized as follows:

- **RepVGG Backbone Replacement**: All original CBS blocks in the YOLOv8 backbone are replaced with RepVGG modules. These modules employ structural re-parameterization to accelerate inference while retaining strong representational capacity during training.
- **Small Object Detection Head**: A new detection head is added at the P2 level, allowing the model to exploit high-resolution shallow features and improving detection performance on small-scale obstacles such as cones and fire hydrants.
- **Swin Transformer Integration**: The final C2f module in the backbone is replaced with a Swin Transformer block. This modification enables the model to capture global contextual relationships and long-range dependencies, enhancing robustness in cluttered scenes.

The resulting architecture, illustrated in *figure 7*, combines all three enhancements in a unified framework. This design significantly improves detection accuracy, particularly for small objects, while maintaining high inference speed. The effectiveness of each component and the overall architecture is verified through extensive ablation and comparative experiments in the following sections.

4. EXPERIMENTS AND ANALYSIS

This section presents the experimental design, datasets, training configuration, evaluation metrics, baseline models, and comparative results. We first describe the data used for training and testing, followed by the implementation details of the experimental environment. Next, we introduce the baseline models used for comparison and evaluate the proposed model through ablation and full-scale experiments. All results are aimed at verifying the effectiveness of our proposed architectural improvements in both accuracy and inference speed.

4.1. Dataset and Evaluation Metrics

We evaluate our model using a comprehensive multi-source road obstacle dataset, publicly available at https://github.com/TW0521/Obstacle-Dataset. The dataset integrates multiple sources, including VOC, COCO, TT100K, and field-collected images, comprising a total of 7915 images: 5066 for training, 1266 for validation, and 1583 for testing. All images are real-world captures from sidewalks, intersections, and urban roads under diverse conditions. Most images are approximately 1500×1500 pixels in size, with some exceeding 3000×3000 pixels, enabling the dataset to represent realistic, high-resolution, and diverse traffic environments. The dataset contains 15 obstacle classes commonly encountered in road scenarios, such as stop sign, person, bicycle, bus, truck, car, motorcycle, traffic cone, trash can, warning column, spherical obstacle, pole, dog, tricycle, fire hydrant.

To analyze the dataset's statistical properties, we plot in Figure 8 the following:

• *Top-left:* instance count per category, revealing class imbalance (e.g., many cars, few fire hydrants).
• *Top-right:* overlaid bounding-box centroids, showing most objects concentrate near image center.
• *Bottom-left:* heatmap of normalized center coordinates (x, y).
• *Bottom-right:* heatmap of normalized object widths and heights.

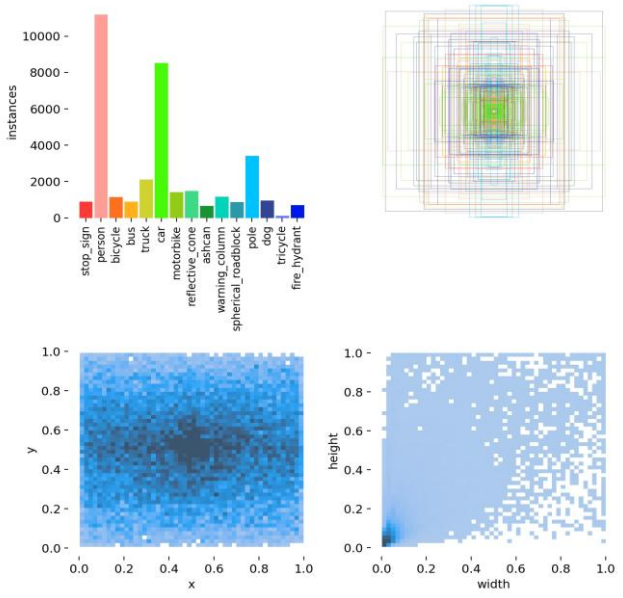


Figure 8. Dataset distribution statistics. Top-left: per-category instance counts; top-right: overlaid bounding boxes; bottom-left: center coordinate heatmap; bottom-right: width–height distribution heatmap

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 3 | Pages 419-428 | e-ISSN: 2347-470X

These plots indicate a strong bias toward small objects (most boxes occupy < 20 % of image dimensions), central spatial bias, and class imbalance. Such characteristics challenge conventional detectors and motivate our small-object-aware head and enhanced backbone.

For quantitative evaluation, we adopt the following metrics:

Precision (P): $P = \frac{TP}{TP+FP} \times 100\%$, measures accuracy of positive predictions.

Recall (R): $R = \frac{TP}{TP+FN} \times 100\%$, measures coverage of actual positives.

mAP@0.5: mean Average Precision at IoU = 0.5.

mAP@0.5:0.95: average mAP over IoU thresholds from 0.5 to 0.95 in steps of 0.05.

4.2. Training Settings

All experiments were conducted on a workstation running Ubuntu 20.04, equipped with a single NVIDIA H800 GPU (84 GB VRAM) and 64 GB of system memory. The proposed model was implemented based on the YOLOv8 framework using PyTorch, and trained from scratch using custom configuration. The input image size was set to 640×640 pixels. All models were trained for 300 epochs with a batch size of 4 and 8 data loading workers. The Adam optimizer was employed, with an initial learning rate of 0.05, decayed using a cosine annealing schedule to promote stable convergence. A weight decay of 0.0005 was applied to reduce overfitting and improve generalization.

All models, including baseline and improved versions, were trained and evaluated under the same settings for a fair comparison.

4.3. Ablation Experiment

To evaluate the individual contributions and combined effectiveness of the proposed modules, we conduct a series of ablation experiments based on the YOLOv8 baseline. As shown in *table 1*, each component's impact is clearly analyzed and illustrated.

Firstly, when we replace the original YOLOv8 backbone with the RepVGG module, a slight decrease in both precision and recall is observed. Although counterintuitive, this phenomenon could be attributed to the structural re-parameterization strategy of RepVGG. Specifically, RepVGG adopts a multi-branch architecture during training to enhance feature expressiveness, which then merges into a single branch during inference for faster execution. This merging process may slightly compromise the representation of fine-grained spatial features crucial for small-object detection. Additionally, this slight degradation might be due to suboptimal hyperparameter tuning, as the training configuration used was directly inherited from the baseline without further optimization specific to RepVGG.

Introducing the dedicated P2-level detection head significantly improves recall performance, confirming that explicit usage of high-resolution shallow features effectively enhances small-object detection. This module effectively compensates for the minor performance reduction caused by the RepVGG substitution.

Moreover, the integration of the Swin Transformer module yields considerable improvements in precision, recall, and mAP metrics. This clearly indicates that the Swin Transformer's ability to capture long-range dependencies and global contextual information substantially strengthens detection capabilities in complex scenarios with dense or occluded targets.

Ultimately, combining all three enhancements results in the best performance, demonstrating the complementary effects of these modules. Although each module independently addresses specific limitations, their combined use synergistically improves overall model accuracy and robustness, confirming the validity and effectiveness of the proposed comprehensive approach.

Table 1. Ablation study on the contribution of each proposed module to overall performance

RepVGG	P2	Swin transform	P	R	mAP@0.5
-	-	-	0.809	0.724	0.774
√	-	-	0.806	0.720	0.759
-	√	-	0.805	0.743	0.760
-	-	√	0.810	0.745	0.779
√	√	-	0.807	0.752	0.763
√	-	√	0.808	0.765	0.789
-	√	√	0.810	0.774	0.798
√	√	√	0.812↑	0.782↑	0.822↑

As shown in *figure 9*, the proposed model not only reaches higher final performance compared to YOLOv8, but also converges more stably across training epochs. Precision and recall improve more quickly in the early training stages, and both mAP@0.5 and mAP@0.5:0.95 curves remain consistently above the baseline. These trends confirm the benefit of the added modules in learning efficiency and optimization stability.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 3 | Pages 419-428 | e-ISSN: 2347-470X

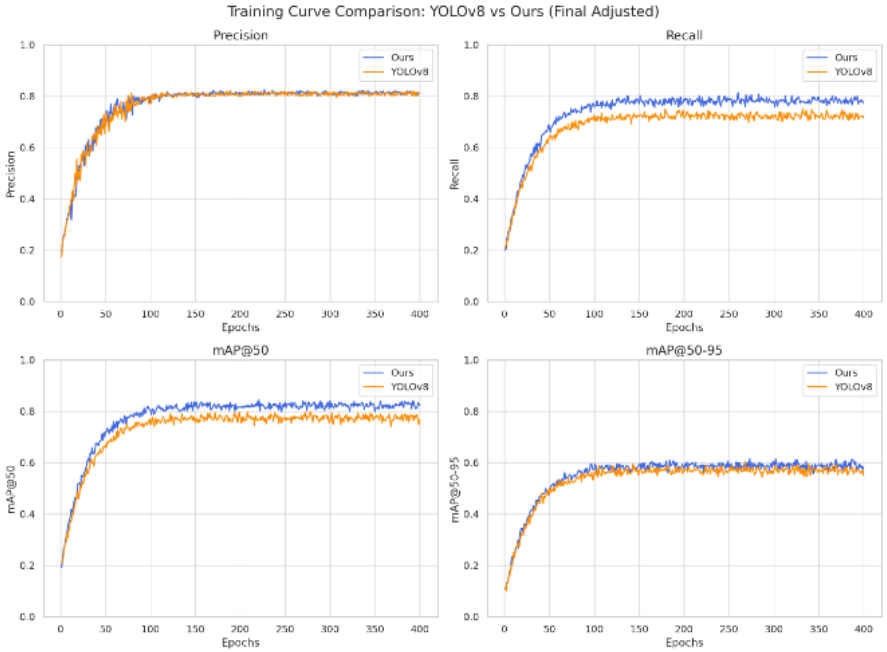


Figure 9. Training curve comparison between YOLOv8 and the proposed model over 400 epochs

4.4. Model Comparison Experiments

To further validate the effectiveness of the proposed model, we compare it against several representative object detection algorithms, including both classical and modern frameworks. These include two-stage detectors such as SSD, Faster R-CNN, and RetinaNet, as well as one-stage YOLO series models: YOLOv5, YOLOv6, YOLOv7, and YOLOv8. All models are trained and evaluated on the same dataset under identical training settings to ensure fair comparison. The results are shown in *table 2.*

Table 2. Comparison of detection performance among mainstream object detectors and the proposed model.

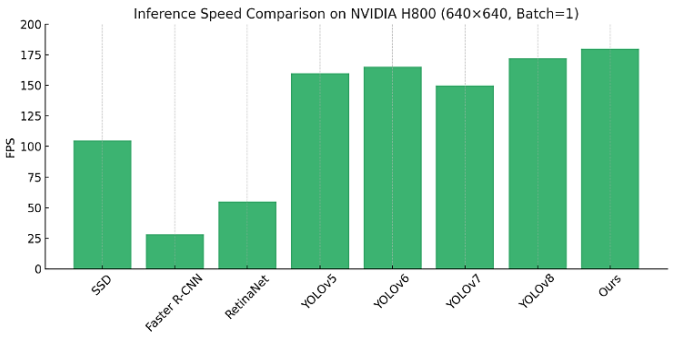
Model	Precision (P)	Recall (R)	mAP@0.5	mAP@0.5:0.95
SSD	0.742	0.610	0.695	0.510
Faster R-CNN	0.765	0.660	0.702	0.535
RetinaNet	0.773	0.678	0.706	0.515
YOLOv5	0.768	0.649	0.704	0.525
YOLOv6	0.778	0.656	0.707	0.537
YOLOv7	0.786	0.778	0.817	0.585
YOLOv8	0.809	0.724	0.774	0.570
Ours	0.812	0.782	0.822	0.588

Compared to all other methods, our model achieves the best overall performance. Specifically, the mAP@0.5 improves by 0.048 over YOLOv5 and by 0.016 over YOLOv8. At the same time, mAP@0.5:0.95 increases by 0.018 compared to YOLOv8. This demonstrates that our improvements not only retain high inference speed but also significantly enhance detection precision, particularly in small-object and complex environments.

4.5. Inference Speed Evaluation

To assess the practical applicability of the models in real-time scenarios, we evaluated the inference speed of several mainstream object detectors under the same hardware and input conditions. All models were tested on an NVIDIA H800 GPU with an input image size of 640×640 and batch size of 1, measured in frames per second (FPS). The results are illustrated in *graph 1.*

The results show that two-stage models like Faster R-CNN are significantly slower than one-stage detectors. Among the YOLO series, YOLOv8 achieves strong real-time performance at 172 FPS. Our model slightly exceeds it with an inference speed of 180 FPS, demonstrating that the replacement of the C2f modules with RepVGG, despite the addition of a P2 head and a Swin Transformer module, does not compromise efficiency. Instead, the design enables a better trade-off between speed and accuracy.



Graph 1. Inference speed (FPS) comparison of mainstream detection models on NVIDIA H800

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 3 | Pages 419-428 | e-ISSN: 2347-470X

4.6. Training Process Analysis

To verify the stability and convergence of the training process, we recorded the changes in loss values and evaluation metrics over 300 epochs. As shown in *figure 10*, both box loss and objectness loss decrease rapidly during the early training phase and gradually stabilize, indicating effective convergence in bounding box regression and object confidence prediction. Simultaneously, the precision and recall metrics steadily increase, saturating after around 100 epochs. The mAP@0.5 metric reaches approximately 0.82, and mAP@0.5:0.95 converges to around 0.58, suggesting strong detection performance and generalization across IoU thresholds.

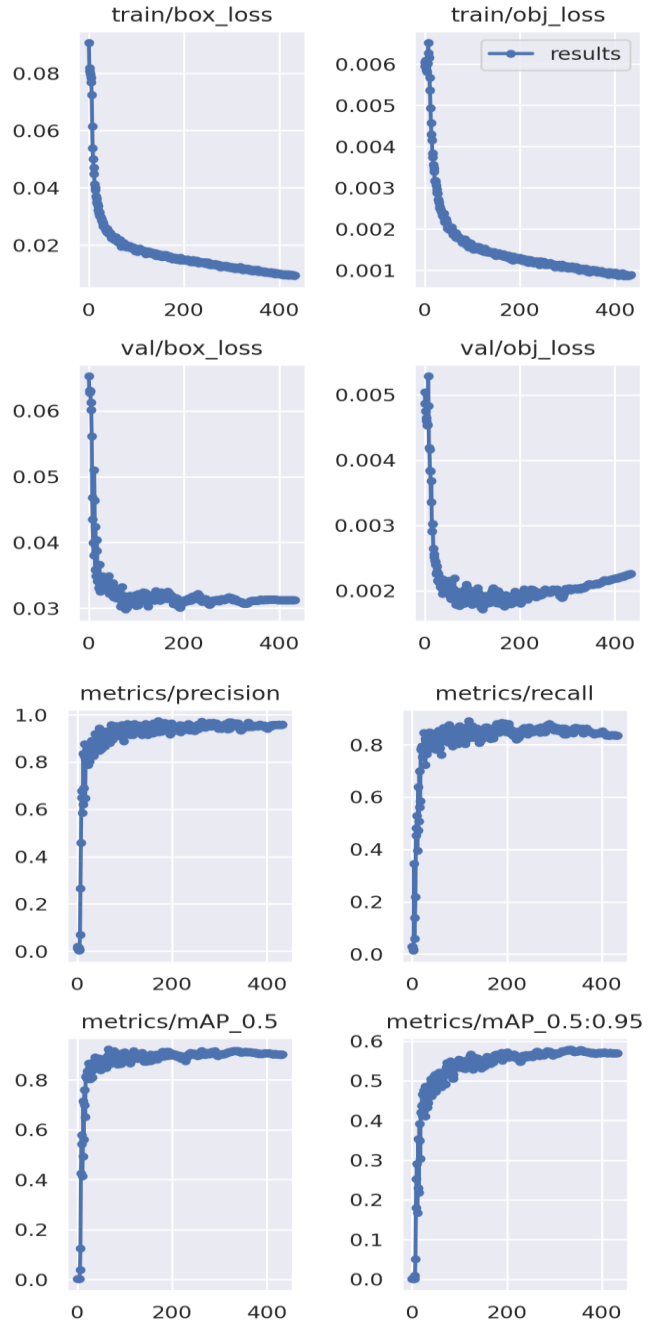


Figure 10. Loss curves and evaluation metrics of the proposed model during training

4.7. Visualization Results

To intuitively evaluate the detection performance of the proposed model, we visualize its output on representative test images. As shown in *figure 11*, the model accurately identifies various road obstacles—including pedestrians, cars, tricycles, fire hydrants, and warning cones—under diverse and complex road conditions such as intersections, occlusion, and cluttered backgrounds.

Notably, the proposed model performs well in detecting small-scale and partially occluded targets, such as *reflective_cone*, *spherical_roadblock*, and *fire_hydrant*, which are typically difficult for conventional models. This robustness is primarily attributed to the integration of the P2 small-object detection head and the Swin Transformer module, which enhance the model's fine-grained feature representation and global semantic understanding.

To further illustrate the model's behavior under varying confidence thresholds, *Figure 12* presents the precision–confidence and recall–confidence curves across all 15 categories. The thick blue curves represent the overall mean trend. The model achieves high precision at moderate to high confidence levels and maintains strong recall performance on challenging categories such as *reflective_cone*, *spherical_roadblock*, and *fire_hydrant*. These visualizations support the model's stability and generalization in multi-class detection tasks.

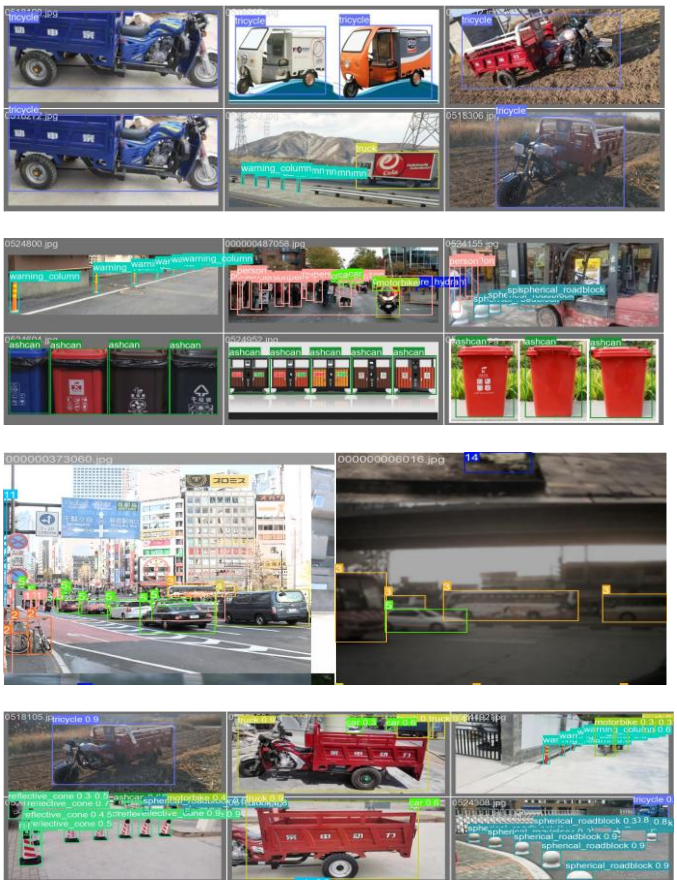


Figure 11. Visualization of detection results on sample test images using the proposed model.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 3 | Pages 419-428 | e-ISSN: 2347-470X

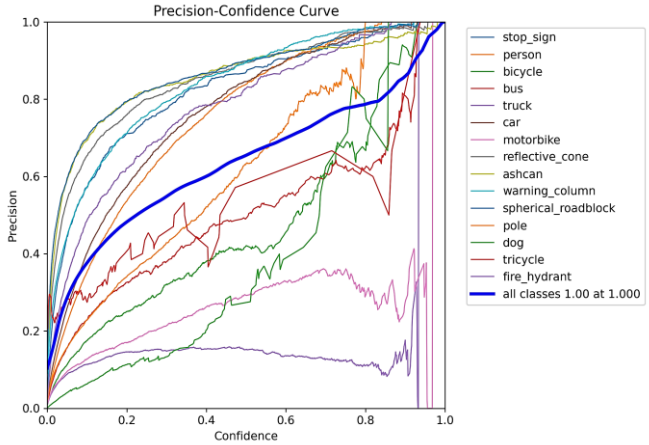


Figure 12. Precision–confidence and recall–confidence curves for all categories. The bold blue lines represent the average values across all classes

6. CONCLUSIONS

In this study, we present an improved YOLOv8-based object detection framework tailored for complex road environments, with a specific focus on enhancing small-object detection and global semantic perception. The model integrates three key improvements: replacing the backbone with RepVGG for efficient inference, adding a P2-level detection head to capture fine-scale features, and introducing a Swin Transformer to strengthen long-range dependency modeling. Experiments on a multi-source obstacle dataset confirm that the proposed model achieves superior detection accuracy (mAP@0.5 of 0.822 and mAP@0.5:0.95 of 0.588) while maintaining a real-time speed of 180 FPS.

In future work, we will focus on three practical extensions:

- *Multi-modal data fusion:* To improve robustness under low-light and adverse weather conditions, we plan to incorporate LiDAR and infrared data alongside RGB images, enabling more reliable obstacle detection through complementary sensing.
- *Model deployment optimization:* We aim to apply techniques such as pruning, quantization, and knowledge distillation to reduce model size and computational load, ensuring efficient deployment on edge computing platforms and embedded systems.
- *Cross-domain generalization:* To enhance adaptability across varying road scenes and unseen environments, we will explore domain adaptation and semi-supervised learning methods, improving the model's generalization to diverse real-world scenarios.

These directions are designed to further advance the practical applicability and robustness of our model in intelligent transportation systems.

While the proposed model achieves real-time performance on high-end GPUs, its deployment on resource-constrained embedded devices may be limited by the added complexity of the Swin Transformer. Future work will explore model compression and lightweight transformer variants to enhance embedded applicability.

Conflicts of Interest: The authors declare that they have no conflicts of interest and that all data used in this study adhere to public data ethics and privacy standards.

REFERENCES

[1] Wang, C., B. Zheng, and C. Li, Efficient traffic sign recognition using YOLO for intelligent transport systems. Scientific Reports, 2025. 15(1).

[2] Juyal, A., S. Sharma, and S. Bhadula, Modified-vehicle detection and localization model for autonomous vehicle traffic system. Indonesian Journal of Electrical Engineering and Computer Science, 2025. 37(2): p. 1183-1200.

[3] Redmon, J., et al. You only look once: Unified, real-time object detection. in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2016.

[4] Ding, X., et al. RepVgg: Making VGG-style ConvNets Great Again. in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2021.

[5] Liu, Z., et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. in Proceedings of the IEEE International Conference on Computer Vision. 2021.

[6] Liu, G., et al., Lightweight obstacle detection for unmanned mining trucks in open-pit mines. Scientific Reports, 2025. 15(1).

[7] Choi, E., T.A. Dinh, and M. Choi, Enhancing Driving Safety of Personal Mobility Vehicles Using On-Board Technologies. Applied Sciences (Switzerland), 2025. 15(3).

[8] Zhou, S., et al. Small Target Detector Based on Adaptive Re-parameterized Spatial Feature Fusion Mechanism. in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2025.

[9] Zhou, X., Chen, W., & Wei, X. (2024). Improved Field Obstacle Detection Algorithm Based on YOLOv8. Agriculture, 14(12), 2263. https://doi.org/10.3390/agriculture14122263.

[10] Y. Liu, H.Z., H. Liu and M. Zhao, "HSC-YOLOv7: An Enhanced YOLOv7 for Small Object Detection," 2024 43rd Chinese Control Conference (CCC), Kunming, China, 2024, pp. 8910-8915, doi: 10.23919/CCC63176.2024.10661994.

[11] Ma, S., Zhao, X., Wan, L. et al. A lightweight algorithm for steel surface defect detection using improved YOLOv8. Sci Rep 15, 8966 (2025). https://doi.org/10.1038/s41598-025-93469-5.

[12] Chen, J., Mai, H., Luo, L., Chen, X., & Wu, K. (2021, September). Effective feature fusion network in BIFPN for small object detection. In 2021 IEEE international conference on image processing (ICIP) (pp. 699-703). IEEE.

[13] Gong, Y., Yu, X., Ding, Y., Peng, X., Zhao, J., & Han, Z. (2021). Effective fusion factor in FPN for tiny object detection. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 1160-1168).

[14] Wu, J., & Liao, S. (2022). Traffic sign detection based on SSD combined with receptive field module and path aggregation network. Computational intelligence and neuroscience, 2022(1), 4285436.

[15] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV) (pp. 3-19).

[16] Lei, Y., Pan, D., Feng, Z., & Qian, J. (2023). Lightweight YOLOv5s human Ear recognition based on MobileNetV3 and ghostnet. Applied Sciences, 13(11), 6667.

[17] Hong, J., K. Ye, and S. Qiu, Study on lightweight strategies for L-YOLO algorithm in road object detection. Scientific Reports, 2025. 15(1).

International Journal of
Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 3 | Pages 419-428 | e-ISSN: 2347-470X

[18] Leng, X., et al., An Improved YOLOv8-Based Method for Real-Time Detection of Harmful Tea Leaves in Complex Backgrounds. Phyton-International Journal of Experimental Botany, 2024. 93(11): p. 2963-2981.

[19] Alahdal, N.M., et al. Real-time Object Detection in Autonomous Vehicles with YOLO. in Procedia Computer Science. 2024.

[20] Zhou, X., W. Chen, and X. Wei, Improved Field Obstacle Detection Algorithm Based on YOLOv8. Agriculture (Switzerland), 2024. 14(12).

[21] Zhang, H., M. Liang, and Y. Wang, YOLO-BS: a traffic sign detection algorithm based on YOLOv8. Scientific Reports, 2025. 15(1).

[22] Jhala, J.S., C. Joshi, and D. Anand. Deep Learning Driven Object Detection and Classification for Autonomous Vehicles in Diverse Traffic and Weather Conditions. in 1st International Conference on Emerging Technologies for Dependable Internet of Things, ICETI 2024. 2024.