Rohan Chaudhary, 2026, 14:4
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Object Detection in Autonomous Vehicles

Rohan Chaudhary[1], Mayank Chauhan[2], Dr. Partap Singh[3]
Assistant Professor, Department of CSE, Quantum University, Roorkee, India

Abstract- In the context of autonomous vehicle (AV) perception, this study examines the implementation and performance assessment of the You Only Look Once version 5 (YOLOv5) model for object detection (OD). Safe and reliable navigation requires high-throughput algorithms, which require accurate and real-time detection. Roboflow's carefully selected dataset of approximately 15,000 photos was used in the study, with a focus on specific classes like cars, trucks, pedestrians, traffic lights, and bikes. PyTorch was used to train the YOLOv5 architecture for 100 epochs to enhance generalization across different lighting and angle conditions, and robust data augmentation techniques like brightness adjustment and scaling were included. The model's efficiency is attributed to its single-stage detection approach, which places emphasis on the necessary inference speed for real-time decision-making in autonomous systems. The trained model achieved a mean average precision (<unk>text<unk>mAP<unk>) of around 85%, and both precision and recall metrics exceeded 80% in terms of numbers. Critical analysis revealed that detection accuracy was significantly higher for large, frequently occurring classes (cars) than for smaller, safety-critical classes (pedestrians and traffic lights), indicating a performance gap. Further architectural improvements are required due to the structural limitation caused by class reduction and scale variations. The paper concludes with a roadmap for improved robustness, which includes migrating to sophisticated architectures like YOLOv8 or Detection Transformers (DETR), to achieve near-perfect recall rates for high-consistency objects.

Keywords: Autonomous Vehicles, Object Detection, YOLOv5, Deep Learning, Mean Average Precision, Computer Vision, Real-Time Systems.

I. INTRODUCTION

Object detection is crucial for autonomous navigation

AV technology's rapid advancement requires autonomous vehicles (AVs) to accurately perceive and understand their surroundings in real-time. Object detection (OD) is the foundational technology that provides the necessary information to identify and categorize various objects essential for safe navigation, including cars, pedestrians, cyclists, and road markings [1, 1]. To successfully implement autonomous vehicles (AVs), it is crucial to transition from traditional rule-based systems to highly reliable, deep learning-powered perception systems.

Safe and reliable autonomous operation can only be achieved by meeting two essential, often mutually exclusive, requirements for object detection systems: high accuracy and tremendous computational efficiency. Delayed detection can have significant consequences in dynamic driving environments, so algorithms must not only accurately detect objects but also predict them at a speed (frames per second, FPS) that allows for quick decisions and real-time path planning. The objective of detection in this field is to guarantee safety and technical correctness.
The choice of a one-stage detector like YOLOv5 clearly shows that inference speed is more important than the potential modest accuracy gain from a slower, two-stage architecture like Faster R-CNN for this project. The technical approach needs to be reconciled with the safety-critical limitations inherent in automotive applications, which is why this design decision is crucial.

The model's effectiveness is based on how well its accuracy (<unk> text<unk> mAP<unk>) maintains the required real-time throughput, while also being sufficient and reliable.

Overview of Real-Time Object Detection Architectures

The field of object detection has undergone rapid evolution, from hand-crafted features like HOG and SIFT to complex deep learning models. Today's methods are dominated by convolutional neural networks (CNNs), which can be broadly divided into two-stage and one-stage detectors. Although region

proposals are generated before classification, two-stage detectors provide higher localization accuracy at the cost of latency. One-stage detectors, like the YOLO series, process the image in one pass and simultaneously estimate class probabilities and bounding boxes.

YOLOv5 is an iteration that builds on previous iterations by combining optimized architectural components and loss functions to optimally balance speed, size, and detection performance. YOLOv5 is a highly viable solution for scenarios with real-time constraints, such as autonomous navigation, due to this optimization. Due to its incredible speed, the YOLO family has become the preferred choice for real-time applications.

Challenges Addressed by This Research

The challenges of autonomous vehicle object detection are tested against deep learning models. Two major challenges are addressed by this study through unique methods, the first being the importance of environmental robustness. Camera-based systems are highly sensitive to changes in lighting, angle, and poor visibility. Massive data augmentation is necessary for implementation, particularly for brightness adjustment.

Scaling, artificially exposing the model to different conditions to increase its generalization capabilities. This ensures that the model remains robust across different settings, as it mitigates the performance degradation typically seen in low light or poor viewing angles.

Second, the issue of scale and class imbalance needs to be addressed. Automotive datasets contain a wealth of information about large objects like cars and trucks, but fewer examples of smaller, safety-critical objects like pedestrians and traffic lights. This imbalance can impair the learning process. To achieve high detection fidelity for less frequent, more consequential classes, the project's implementation used the weighted loss function built into YOLO to balance detections across all object types and used data augmentation to replicate less represented classes.

Structure of the Paper and Novel Contribution

The remaining sections of this paper are organized to provide a complete technical description of the study. Section II examines the relevant literature on deep learning perception in AVs, emphasizing the trade-off between speed and accuracy. The methodology is described in detail in Section III, including preprocessing, augmentation techniques, and dataset characteristics, as well as how limitations such as annotation noise were handled. The YOLOv5 architecture and training details, including the specific loss functions and constraint management required for effective training, are covered in detail in Section IV. Quantitative results (\text{mAP}, precision, recall) are presented in Section V, which also examines performance differences between object classes in more detail and connects theoretical difficulties with empirical weaknesses. To achieve greater safety resilience, especially for vulnerable road users, advanced architectural transitions are explored in Section VI, which summarizes the results and outlines an organized research roadmap.

II. RELATED WORK

Deep Learning Approaches for Autonomous Vehicle Perception

The need for high-accuracy, real-time object detection is driving the adoption of advanced deep learning techniques in autonomous systems. Multi-stage detectors, which demonstrated high localization accuracy but had high computational costs and latency, were often investigated in early research. This trade-off was later addressed with the development of single-stage detectors, particularly the YOLO series, which achieved the high-speed inference required for quick, safe decision-making in dynamic driving environments.

The architecture's status as a powerful solution in the field of object detection for AVs has been cemented by YOLO's ongoing development, which resulted in YOLOv5. YOLOv5's effectiveness is primarily due to its optimized design, which allows it to process data in real time while maintaining a high level of accuracy.

YOLO Architecture and Improvements

In autonomous applications, YOLOv5 is often cited as a highly effective alternative to OD. Its optimized architecture, which delivers high accuracy while maintaining robust speed, is its core strength. With advanced feature fusion mechanisms such as the PANet neck and CSPDarknet53 backbone, the model has a robust architectural foundation.

A study on object detection in autonomous vehicles using YOLOv5 thoroughly tested the model's performance, confirming its superiority in both accuracy and speed for this domain. Similar scholarly research supports the model's use. Ongoing efforts to improve the core YOLOv5 algorithm, often aimed at improving robustness to small or small objects, are evident in further comparative analysis. For example, the WOG-YOLO algorithm demonstrated improved performance due to specific hyperparameter optimizations of YOLOv5.

Performance in detecting cyclists and pedestrians. In experiments, WOG-YOLO outperformed the standard YOLOv5 implementation, which achieved an 85% mAP in the same test, while achieving a 90% mAP for these critical classes. This comparison constitutes an important performance benchmark: the documented ability of the optimized YOLO variant, especially for vulnerable road users, confirms the potential for large, targeted improvements, while the approximately 85% mAP achieved in the current study is generally reliable.

Challenges of Real-World Automotive Datasets

The AV operating environment places significant constraints on perception systems, leading to three main data-centric issues discussed in the relevant literature:

Scale variation and multi – scale detection

In AV environments, objects are viewed from a great distance, so detectors must perform well at multiple scales. Even with a reduced pixel representation, distant objects need to be accurately detected. YOLOv5 addresses this by using a combined Feature Pyramid Network (FPN) and Path Aggregation Network (PANet) in its neck structure. To maintain robust detection across the entire scale range – a crucial capability for AV operation – this architecture ensures that rich semantic information – essential for classification – is combined with fine-grained details, which are required for small, distant objects.

Imbalance and Rarity

In real-world datasets, the data volume is dominated by large, frequent items, which typically exhibit a large class imbalance. This diverse representation results in fewer updates to the learning algorithm from underrepresented groups, such as cyclists or pedestrians, which may lead to reduced accuracy and recall of these essential safety items.

The literature on object detection proves that weighted loss functions, such as those that use dynamic focusing mechanisms, can mitigate this problem by making rare or difficult examples more important during gradient calculation, thereby improving detection robustness for sparse classes.

Environmental Adversity

Because AVs use visual sensors, detection is highly sensitive to ambient effects. Research shows that low light and brightness changes significantly degrade the performance of new detectors like YOLO. When combined with other factors like noise or small object size, this performance degradation is often exacerbated. True robustness against extreme conditions (fog, snow) often requires sensor fusion, combining camera data with depth- and range-sensing methods like LiDAR and radar, although data enhancement techniques like brightness adjustment can artificially increase robustness.

Future Directions: Transformer-Based Detectors

The field is rapidly moving towards more reliable, transformer-based architectures, even as CNN-based detectors like YOLOv5 offer efficiency. Using attention mechanisms that better model global dependencies within the scene, detection transformers (DETR) and real-time detection transformers (RTDETR) are starting to demonstrate state-of-the-art results, offering better solutions to difficult localization and multi-scale detection problems than conventional CNNs. These transformer-based models represent the next technological advancement for scenarios requiring

the highest accuracy, especially for solving the scale imbalance problem with small, distant objects, justifying their inclusion in future research planning.

III. EXPERIMENTAL METHODOLOGY

The research method utilized the advanced capabilities of the YOLOv5 algorithm for real-time object detection, and the process steps involved careful dataset compilation, rigorous training, and thorough evaluation.

Dataset Acquisition and Characteristics

The "Car Detection" dataset was obtained from Roboflow. Approximately 15,000 photos with 97,942 annotations across 11 classes make up this collection, which is an enhanced version of the original Udacity self-driving car dataset. In the detection task, priority was given to the car, truck, pedestrian, traffic light, and bike classes – all essential for autonomous navigation [point 3]. To guarantee reliable model evaluation, the dataset was divided into standard subsets: 70% for training, 20% for validation, and 10% for testing [point 3].

Data Preprocessing and Augmentation

To ensure compatibility with the YOLOv5 input pipeline, images were first preprocessed, including resizing and pixel standardization. Comprehensive data augmentation was an essential part of this method. Roboflow automatically added augmentations to the training data, including flipping, rotation, brightness adjustment, and scaling [point 3]. For low-light detection in AVs, these methods—especially brightness and scaling adjustments—are crucial to improving model generalization and robustness against varying lighting and angle conditions. Furthermore, mosaic data augmentation, which combines four images together to increase dataset diversity and improve detection of small objects, is naturally incorporated into the YOLOv5 framework.

Addressing Data Challenges

Ensuring consistent bounding box alignment and class label consistency were the main challenges of preprocessing [point 3]. Additionally, several methods were used to address the issue of within-class imbalance, which occurs when common objects (such as cars) outnumber safety-critical objects (such as pedestrians):

1. **Data augmentation:** This technique artificially replicates examples from underrepresented classes [point 3].
2. **Weighted Loss Function:** The YOLO loss function's built-in mechanisms, including a multi-task loss component, help balance detections across all object types by giving greater weight to rare or difficult examples.

This varied approach to preprocessing and dataset management was essential to build a strong training phase foundation and ensure the model could learn from a complex yet diverse automotive perception environment.

IV. IMPLEMENTATION DETAILS

Development Environment and Frameworks

Several key deep learning and computer vision frameworks were used in the development of the object detection system [point 4]. The YOLOv5 model was trained using PyTorch, providing effective GPU acceleration [point 4]. Essential image and video processing tasks, such as reading input streams and displaying detection results during inference, were performed using OpenCV [point 4]. Preparing the dataset, formatting annotations, and enabling initial data augmentation were all made possible by Roboflow [point 4].

YOLOv5 Training Configuration

A custom dataset was used to train the YOLOv5 model for 100 epochs [point 4]. During training, pre-processed images were fed into the network, and the model repeatedly learned to minimize the difference between the ground truth labels and the predicted bounding boxes.

The multi-component loss function that the YOLOv5 training loop optimizes includes:

1. **Bounding Box Loss:** Calculates the error in box size and localization. CIoU (Complete Intersection over Union) loss is commonly used for this regression task.

2. **Objectness Loss:** Uses Binary Cross Entropy (BCE) to calculate how likely it is that a predicted box contains an object.
3. **Classification Loss:** Uses BCE for multi-label support and quantifies the error in correctly assigning an object class (e.g., "car," "pedestrian").

Technical issues such as CUDA memory errors were addressed by reducing the batch size of the method during the optimization process and resizing the input image to fit GPU memory constraints—a common way to manage computational resources..

Anchor Box Optimization

The model's estimation of an object's size and aspect ratio is effectively guided by a set of pre-defined bounding box sizes, called anchor boxes. When anchors are properly adjusted, the model can more accurately identify objects of varying sizes and aspect ratios [point 4]. Unlike traditional techniques that simply use k-means clustering, YOLOv5 uses a genetic algorithm (GA) to incorporate an adaptive anchor box calculation mechanism. By comparing their intersection over union (IoU) fitness with the ground truth boxes, the GA iteratively optimizes the initial anchor box dimensions, making the network more likely to recognize objects of varying sizes and shapes encountered in the driving environment.

Inference and Visualization

Model During model testing, YOLOv5's built-in detection commands were used to run inference on both image and video files [point 4]. To visually validate performance, when the detection results were superimposed directly onto the input media, the bounding box, class label, and associated confidence score for each detected object were displayed [point 4]. The accuracy of real-time detection in dynamic scenes was confirmed by this visual verification, which complemented the numerical evaluation. To extend the detection capabilities to real-time object tracking and maintain a consistent understanding of vehicle movements, the DeepSORT algorithm was added after training.

This algorithm creates associations between detections in subsequent frames.

Metric	Result	Interpretation
Mean Average Precision (\text{mAP})	\approx 85\%	The model demonstrated reliable performance averaged across all object classes [Point 5].
Precision	> 80\%	Indicates that a high percentage of the model's positive predictions were correct, minimizing false positives.
Recall	> 80\%	Indicates minimal false negatives; the model successfully most of the actual object scene.

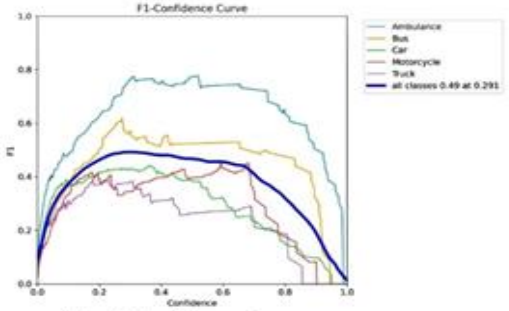


Figure 1: F1-confidence confidence curve.

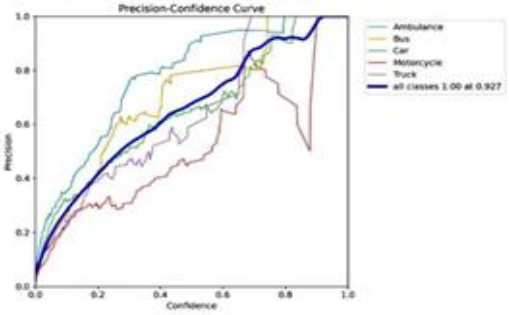


Figure 2: Precision-confidence confidence curve.

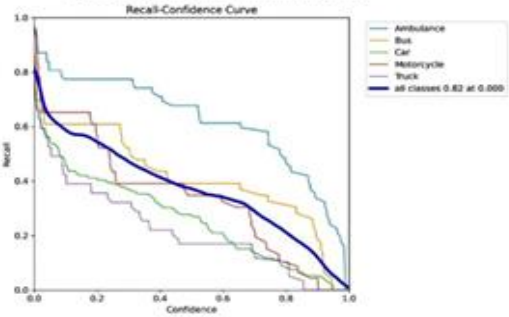


Figure 3: Recall-confidence confidence curve.

V. RESULTS AND ANALYSIS

Quantitative Performance Metrics

The performance of the trained YOLOv5 model was rigorously assessed using standard object detection

metrics: Mean Average Precision (\text{mAP}), Precision, and Recall. The final evaluation yielded strong results, affirming the model's suitability for real- time AV perception. In safety-critical applications like autonomous driving, high precision (>80%) is particularly important because it reduces false positives, which can lead to unnecessary braking or inconsiderate actions.

Additionally, high recall (>80%) validates the model's ability to recognize most important objects, reducing the likelihood of collisions due to missed events (false negatives).

Disparity in Class Performance

Analysis of the performance of each class revealed that, despite strong overall metrics, there was a noticeable gap [point 5]. Cars and trucks, in particular, which had a large number of data samples, were identified with the best accuracy and confidence. On the other hand, smaller, less common, and more safety-critical objects, such as traffic lights and pedestrians, had slightly lower detection rates [point 5].

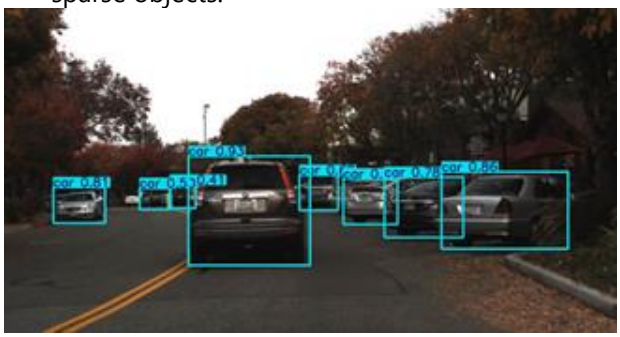
The well-known problem of scale variation and class imbalance in computer vision datasets accounts for this performance gap. Compared to larger, frequently occurring objects, pedestrians and traffic lights are actually more difficult to accurately detect and classify because they often occupy a smaller area in the image (smaller scale objects) and have fewer training examples (class imbalance).

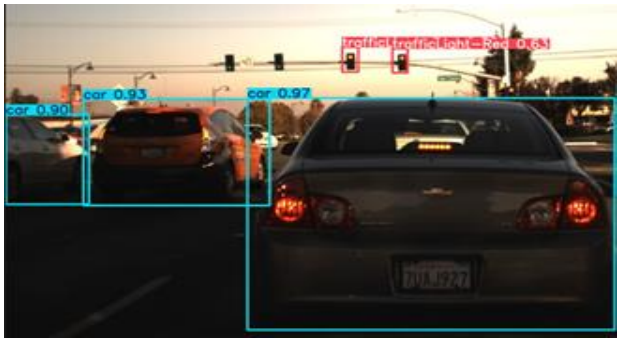
Validation and Future Enhancement Roadmap

The model's output on test videos, which successfully identified multiple objects in a difficult driving scene, was examined to validate performance numerically (using the metrics described above) and visually [Point 5].

A roadmap for future improvements is suggested, focusing specifically on improving detection accuracy for vulnerable road users and traffic signals to address the observed performance disparity and further increase safety margins. Potential improvements include:

- **Extended Training and Hyperparameter Tuning**: Adjusting training and hyperparameters, such as the learning rate, for additional approaches so that model can converge on ideal weights. [Point 5].
- **Dataset Augmentation:** Using a larger, more balanced dataset to provide the model with a wider range of examples for underrepresented classes [point 5].
- **Higher Resolution Inputs:** To improve the detection of small, distant objects like pedestrians and traffic lights, higher-resolution input images are used for inference and training [point 5].
- **Architectural Migration:** Investigating newer and more sophisticated architectures like YOLOv8 or transformer-based models (like DETR or RT-DETR) [point 5]. These models have proven to be more adept at handling multi-scale detection and complex localization, which can significantly improve \text{mAP} for small and sparse objects.





VI. CONCLUSION

This study successfully implemented and evaluated the YOLOv5 object detection model for real-time perception in autonomous driving systems. When fully utilized, the model demonstrated strong quantitative performance with a mean average precision (\text{mAP}) of approximately 85% using a pre-annotated dataset from Roboflow and advanced data augmentation techniques, confirming its potential as a highly effective, real-time solution for vehicle perception [point 5].

The study showed that the one-stage YOLOv5 architecture meets the critical requirement of fast predictions needed for quick decision-making in autonomous vehicles. However, the analysis also revealed a major technical issue: compared to larger vehicles, the model demonstrated lower detection confidence for smaller, safety-critical objects like pedestrians and traffic lights [point 5]. This performance gap, which is believed to be due to scale variation and class imbalance, highlights the need for continued improvement in AV perception systems. Future research should prioritize increasing recall for vulnerable road users and system robustness in difficult conditions [Point 5].

Improving data balance, using high-resolution inputs to capture finer details, and testing next-generation models like YOLOv8 or DETR/RT-DETR to achieve near-perfect detection fidelity for all object classes are some of the suggested improvements. Achieving this increased robustness and reliability is crucial for the safe and widespread adoption of fully autonomous transportation systems.

REFERENCES

1. H. Zhu, X. Yan, H. Tang, Y. Chang, B. Li, and X. Yuan, "Moving Object Detection With Deep CNNs," IEEE Access, vol. 8, pp. 29729–29741, 2020.
2. H. Zhu, H. Wei, B. Li, X. Yuan, and N. Kehtarnavaz, "Real-Time Moving Object Detection in High-Resolution Video Sensing," Sensors, vol. 20, no. 12, p. 3591, Jun. 2020.
3. S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A Survey of Deep Learning Techniques for Autonomous Driving," Journal of Field Robotics, vol. 37, no. 3, pp. 362–386, Apr. 2020.
4. A. Balasubramaniam and S. Pasricha, "Object Detection in Autonomous Vehicles: Status and Open Challenges."
5. R. Novickis, A. Levinskis, R. Kadikis, V. Fescenko, and K. Ozols, "Functional Architecture for Autonomous Driving and its Implementation," in Proc. 2020 17th Biennial Baltic Electronics Conference (BEC), Tallinn, Estonia, Oct. 2020, pp. 1–6.
6. S. Kuutti, R. Bowden, Y. Jin, P. Barber, and S. Fallah, "A Survey of Deep Learning Applications to Autonomous Vehicle Control," arXiv, Dec. 23, 2019.