A hybrid object detection Approach for Autonomous vehicles using YOLOv8n model and CNN Transformer module

Praveenchandar J [1], Shish Ahmad[2]

[1] Lincoln University College, Malaysia; [2] Integral University, Lucknow, India.

praveenjpc@gmail.com; drshishahmad@gmail.com

Abstract: Autonomous vehicles are emerging day by day and it is being used as a part of public transport like taxi in some developed countries. The technology Edge artificial Intelligence is used for implementation of the models that running directly on local hardware devices instead of using remote cloud server. In this process sensor fusion, object detections, path finding are some of the important tasks. Object detection takes major role because, all objects in the road must be identified. Based on the observation, decision are made like whether the vehicle needs to be moved right or left or front or back etc. In this proposed research work the novel object detection approach is proposed to improve the accuracy of this process. YOLOv8n model and CNN Transformer module are combined and implemented to achieve the task. Based on the experimentation, it is observed that proposed model has produced the better results in terms of various Evaluation Metrics like mAP (mean Average Precision), Precision, Recall, F1-Score, Inference Time (FPS) .

Keywords: Object detection; Driverless cars; YOLOv8n; CNN transformer

Introduction

Autonomous vehicles are one of the technological advancement in the transport now days. But the challenge behind the technology is very sensitive and collaborative. Starting from the sensor modules to the actuators, it consists of plenty of sensitive hardware's that are embedded. Though it is implemented and being used in some of the countries, still there are unaddressed challenges such as sensor fusion, object detection, path finding, decision making, prediction etc. In this research paper, the object detection is taken into consideration. During the uncertainty conditions, like heavy rain and fog, the objects in front of the vehicle will not be clear. Many objects can be there in front of the vehicle such as car, truck, bus, motorcycle, bicycle, pedestrian, traffic light, traffic sign, animals, road obstacles etc. Each object has its own properties. Pedestrian may move from one direction to another in the road needs immediate breaking. The cyclist may not move as per the prediction. The traffic lights may change periodically. The animals can cross the road unexpectedly. The vehicle needs to identify each object and based on that, the decision must be taken. It is very hard to make the decision with the available inputs with in a Nano seconds. The proposed approach, extracts important visual features such as human body, bicycle frame, road marking, and traffic signs. These features are mapped into numerical feature vectors. Detected object's learned characteristics are stored in this feature vector. Then the object localization determines the exact location of the object. The YOLO divides the image into various grids. This is to predict the height, width, center coordinates, confidence etc. To minimize the localization loss accurate bounding boxes are predicted. This process is important to estimate object position, lane occupancy, and distance with other vehicle.

Related works

In [1], the authors discussed about the object detection in low light scenarios. In this algorithms are failed to predict the low level features in the dark environments. LLD-YOLO based approach is proposed and it is designed for low light conditions. In [2], they introduced the scene information gathering and distributing network. They followed the slice and distribute method to prevent the information loss during fusion. Omni dimensional dynamic convolution is incorporated to redesign the network. In [3], a query based model with a new pipeline is proposed. Bidirectional feature pyramid network is implemented to detect the small object accurately and it different scales of features are integrated. It achieves the precision, speed of vehicles in high speed driving scenarios. In[4] a light weighted object detection is proposed with dynamic up scaling for remote sensing images. The convolution module based on improved multiscale attention is proposed instead of traditional convolution method. This improves the accuracy of object detection in remote sensing.

In[5] the authors discussed about the deep learning, CNN and vision transformers based object detection algorithm. A compete analysis of the anchor free, anchor based and transformers based detectors are taken for consideration and reviewed. Thy discussed the insights behind these approaches and quality metrics are analyzed through the experimental analysis. In [6] mutual assistance learning is proposed for object detection. During the head detection of the detector mutual assistance is manifested. And to provide a shared offsets and avoid inconsistency between prediction pairs, regression and decoupled classification are reintegrated. In [7] 3D object detections for autonomous vehicles are analyzed. This research work summarized the most commonly used 3D image detection techniques. In addition to that, they proposed two new object detection approaches. Feature lifting based methods, data lifting based methods, depth augmented feature learning method are some of the methods analyzed in this paper. In[8] they achieves accurate object detection by aggregating the effective features within the region dynamically. This method consists of two major modules such as dynamic region aggregation module and global multilevel perception module. This improves the object detection accuracy of the small objects.

Proposed Approach

In this work, the deep learning based object detection has been proposed which is based on the YOLOv8n model and CNN transformer method. The driverless vehicles always have multisensory inputs such as camera, ultrasonic sensors, LiDAR , Radar etc and each input source works independently in getting the input signals and all versatile input signals are combined together the process called as sensor fusion[1]. Camera detects lane marking, colors, traffic lights and pedestrians, LiDAR does the 3Dimention distance calculations, Radar measures the speed and distance perfectly, Ultrasonic sensors is detecting the short range obstacles, Inertial Measurement Unit (IMU) measures the orientation and acceleration, GPS provides the position of the vehicle. Each sensors operates in different frame rates like camera operates 30 frames per second, LiDAR operates 10 frames per second and Radar operates 20 frames per second and these information needs to be synchronized using time stamps with respect to the same instant. Sensor fusion involves various activities such as sensor calibration, preprocessing, feature extraction etc. Sensor calibration measures the spatial relationships between all sensors, preprocessing involves noise removal and data standardization and feature extraction extracts features from sensors [2]. During sensor fusion, all data's are combined in to unified representation. In this approach feature fusion approach is used, in which extracted features of each sensors are merged. The next phase is object detection and tracking phase. This provides the measurements of velocity, positions, acceleration and

predicted future location and it uses the kalman filter algorithm for object detections. YOLOv8n along with CNN transformer method is adopted to enhance the accuracy of the process. The architecture diagram of the proposed approach is given in Figure 1.

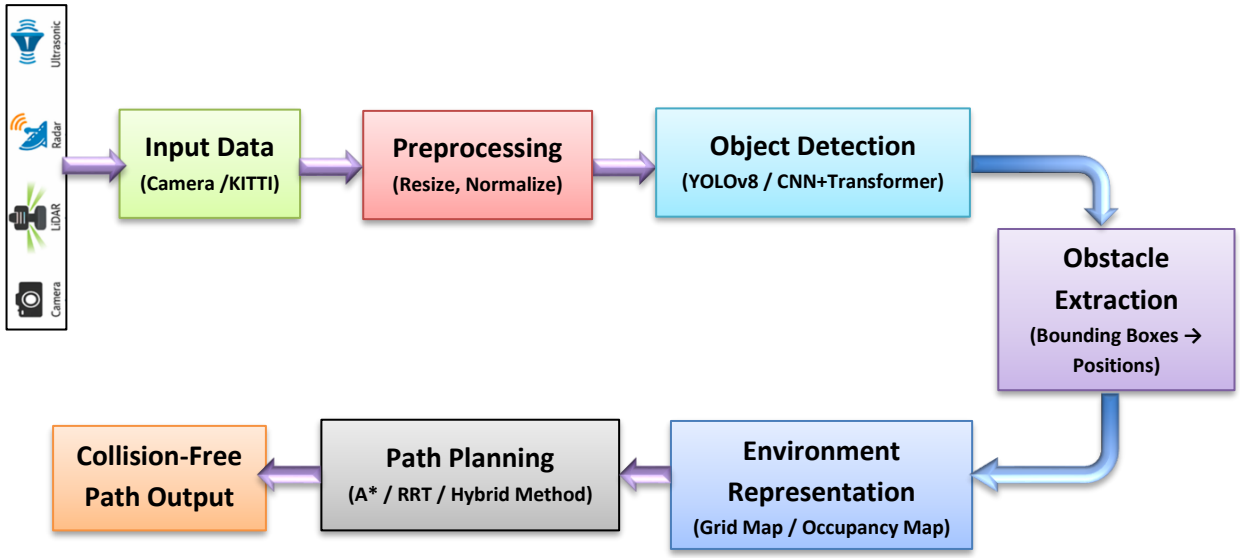

Figure 1. Proposed YOLOv8n-CNN based object detection model

YOLOv8n Model Effectuation

In a KITTI dataset, let X_d is the d_{th} input image, Z_d gives the ground truth annotations and V-is the total number of training samples, then the taken KITTI dataset can be represented (equation (1)) as $DS = \{(X_d, Z_d)\}_{d=1}^{V}$ ---- (1).

In this, each image could be defined $X_d \in \mathbb{Q}^{H\times W\times P}$, where H -denotes height, W-denotes width of an image, and P- gives 3(RGB channels). KITTI dataset is undergone for annotation. That will be converted to YOLO format. Here the original bounding box $(r_{min}, s_{min}, r_{max}, s_{max})$ is converted to YOLO format as

$$r_p = \frac{r_{min} + r_{max}}{2W} \quad \text{--(2)} \qquad w = \frac{r_{max} + r_{min}}{W} \quad \text{--(4)}$$

$$s_p = \frac{s_{min} + s_{max}}{2H} \quad \text{--(3)} \qquad h = \frac{s_{max} + s_{min}}{H} \quad \text{--(5)}$$

and the final annotation has become (p, r_p, s_p, w, h) ---(6)

Then the input data will pass through the YOLOv8 such as $G_{YOLO} = g_{YOLO}(X)$ where $G_{YOLO} \in \mathbb{Q}^{m\times n\times t}$

And the overall YOLO loss can be obtained as $M_{YOLO} = M_{box} + M_{cls} + M_{Oobj}$ --- (7) this denotes localization loss, classification loss and objectness loss.

CNN Feature Extraction Module

In this the local spatial features are extracted using CNN , the convolution operation of M_{YOLO} can be represented as $M_{CNN} = \sigma(\boldsymbol{R} * M_{YOLO} + \boldsymbol{b})$ --(8)

Where $\boldsymbol{b}$**- bias,** $\boldsymbol{R}$ **–** convolution kernel, σ – ReLU activation, and *-convolution. After multiple convolution layers as $M_{CNN} \in \mathbb{Q}^{q\times p\times d}$ --(9)

During experimentation, these CNN features are represented in to tokens and each tokens is projected. After that hybrid CNN-transformer integration is done [3]. Detection head predicts the hybrid loss

function. Finally the performance metrics precision, recall, Intersection over union and mean average precision and analyzed.

Experimentation:

Proposed hybrid YOLOv8n and CNN transformer approach is implemented in colab research platform and the necessary files such as numpy, pandas, pathlib, json etc are downloaded and the KITTI dataset is prepared for training and execution. The dataset preparation is done followed by the file structure id preparation. After the dataset is trained, YAML file is included. The YOLO model is adopted to the environment, the actual model training has started with respect to various aspects. Various objects such as car, pedestrian , van, truck, van, cyclist etc are taken for consideration and all are analyzed with respect to adopted model. Then the validation phase is executed for the taken images. E.g. Figure 2 represents the output of the proposed model with respect to training dataset , executing a model and validation of the results with some standard metrics like mAP(mean Average Precision), recall, precision, and Inference time. The results are observed and values are recorded. From the results it is observed that overall performance of the training process with the KITTI dataset, validation of a model and prediction results of the various objects given in values.

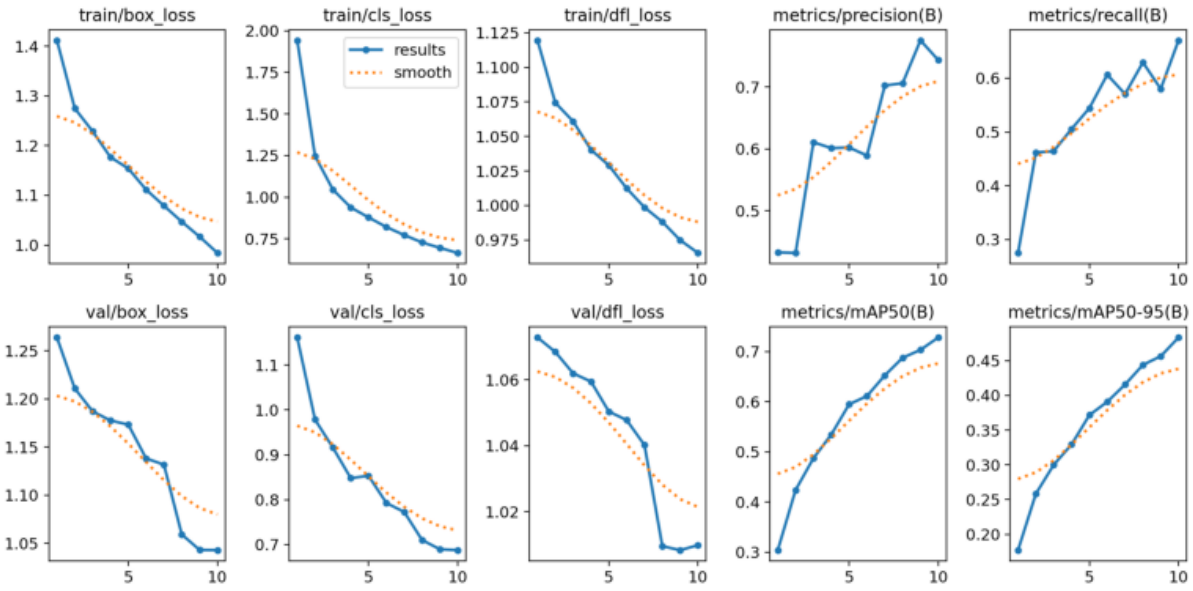

Figure 2. YOLOv8n and CNN transformer approach results

Confusion matrix of the proposed approach is given in figure 3. The matrix clearly says that the proposed approach got some improvements in predicting the objects like car, van, cyclist, truck, pedestrian, sitting person, bike, running person etc and the values are tabulated for future evaluation purpose which improves the accuracy of the process. At the end of YOLOv8n and CNN transformer approach in the environment, after running the model the prediction results in terms of images that contains the object identified. E.g. Figure 4 shows the output images and prediction results of the proposed approach. Figure 4(a), (b), (c), (d) identifies some cars in the road. Figure 4(e) identifies cyclist and Figure 4(f), 4(g), 4(h) and 4(i) identifies the van, pedestrian, truck etc.

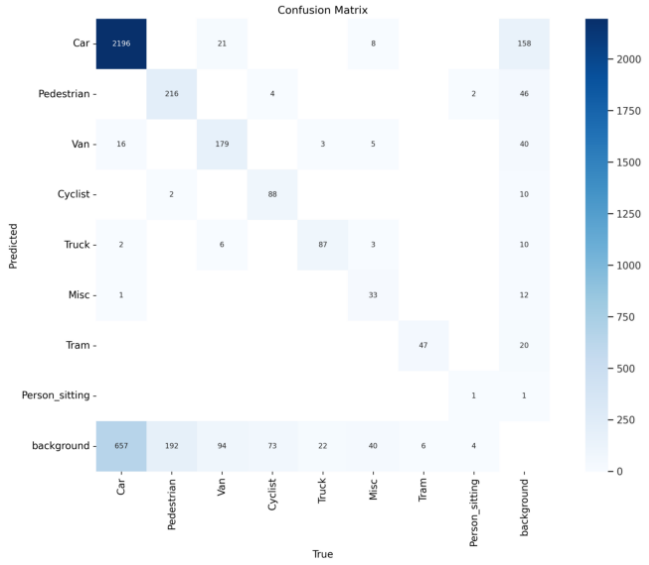

Figure 3. Confusion matrix of the proposed approach

Figure 4(a)

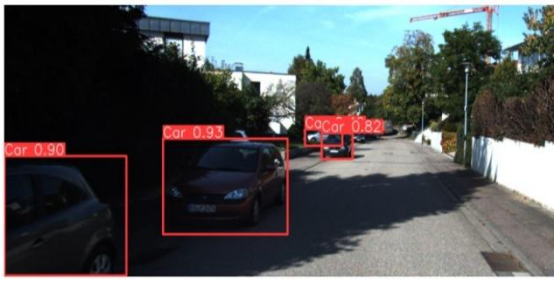

Figure 4(b)

The figures (4(a) to 4(i)) shows the output of the proposed model, Only 9 images given here out of 20 images generated as output. In each image some objects like car, cycle, pedestrian are analyzed and identified clearly. The proposed model improves the clarity in identifying the objects.

Figure 4(c)

Figure 4(d)

Figure 4(e)

Figure 4(f)

Figure 4(g)

Figure 4(h)

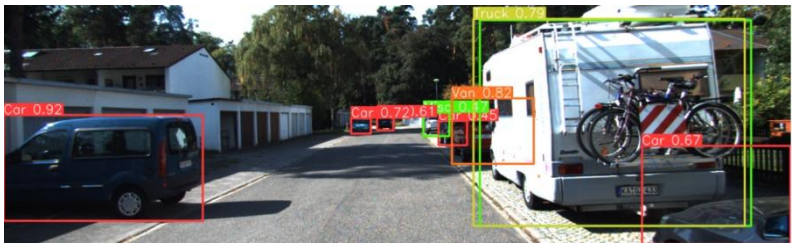

Figure 4(i)

Figure 4. The proposed YOLOv8n and CNN object detection

Conclusions

A novel object detection algorithm is proposed using hybrid YOLOv8n and CNN transformers in this work. In autonomous vehicles, the object detection process is a sensitive process as, within a micro or nano seconds, all objects in the road needs to be identified. The sensor fusion is taken place, after that object detection is done. Based on this, the vehicle creates 3D model of the environment in front of it. In order to improve the accuracy of the object detection process, a hybrid YOLOv8n and CNN transformer is

adopted to the model. The dataset KITTI is taken for consideration. After Converting annotations to YOLO format, the YOLOv8 baseline model is trained. CNN transformer module is developed and integrated to the hybrid architecture. Then the model is trained and fine-tuned, after the getting and analyzing the results, it is observed that the proposed approach improves the accuracy of the object detection process by 3.4%.

References

1. W. Cai, Y. Chen, X. Qiu, M. Niu and J. Li, "LLD-YOLO: A Low-Light Object Detection Algorithm Based on Dynamic Weighted Fusion of Shallow and Deep Features," in *IEEE Access*, vol. 13, pp. 69967-69979, 2025, https://doi: 10.1109/ACCESS.2025.3558574.
2. T. Zhao, R. Feng and L. Wang, "SCENE-YOLO: A One-Stage Remote Sensing Object Detection Network With Scene Supervision," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1-15, 2025, Art no. 5401515, https://doi: 10.1109/TGRS.2025.3539698
3. H. Wang, C. Liu, Y. Cai, L. Chen and Y. Li, "YOLOv8-QSD: An Improved Small Object Detection Algorithm for Autonomous Vehicles Based on YOLOv8," in *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1-16, 2024, Art no. 2513916, https://doi: 10.1109/TIM.2024.3379090.
4. X. Wang, Y. Wang, M. Bai and Q. Ma, "AUD-YOLO: A Lightweight Object Detection Algorithm Model Incorporating Dynamic Upsampling for Remote Sensing Images," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1-9, 2025, Art no. 2004809, https://doi: 10.1109/TGRS.2025.3635890 .
5. A. B. Amjoud and M. Amrouch, "Object Detection Using Deep Learning, CNNs and Vision Transformers: A Review," in *IEEE Access*, vol. 11, pp. 35479-35516, 2023, doi: 10.1109/ACCESS.2023.3266093.
6. X. Xie, C. Lang, S. Miao, G. Cheng, K. Li and J. Han, "Mutual-Assistance Learning for Object Detection," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 12, pp. 15171-15184, Dec. 2023, doi: 10.1109/TPAMI.2023.3319634
7. X. Ma, W. Ouyang, A. Simonelli and E. Ricci, "3D Object Detection From Images for Autonomous Driving: A Survey," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 5, pp. 3537-3556, May 2024, doi: 10.1109/TPAMI.2023.3346386.
8. Z. Zhu, R. Zheng, G. Qi, S. Li, Y. Li and X. Gao, "Small Object Detection Method Based on Global Multi-Level Perception and Dynamic Region Aggregation," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 10, pp. 10011-10022, Oct. 2024, doi: 10.1109/TCSVT.2024.3402097