Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 7
TPCEE 2024

Comparative Study of Vision-Based and LiDAR-Based Object Detection Algorithms for Autonomous Driving

Wenxiao Chen [1], Zihan Gao [2, *]

[1] School of Computer Science and Technology, Jiangnan University, Wuxi, China

[2] School of Physics and Optoelectronics, South China University of Technology, Guangzhou, China

* Corresponding author: 202320130616@mail.scut.edu.cn

Abstract. Autonomous vehicles achieve motion perception and path planning through object recognition. With the integration of deep learning into object detection, numerous practical and innovative studies have emerged in this field. Considering both hardware costs and recognition accuracy, object detection algorithms can be categorized into vision-based and Light Detection and Ranging (LiDAR)-based approaches. This paper reviews classic literature and cutting-edge technologies of both methods, detailing how vision-based approaches utilize computer vision to implement the You Only Look Once (YOLO) series and Faster Region-Convolutional Neural Network (Faster-RCNN). It also explains LiDAR-based approaches, focusing on classic models of point cloud and image fusion, including early fusion, deep fusion, and late fusion architectures. The advantages and limitations of both methods are compared, with experiments on the Karlsruhe Institute of Technology and Toyota Technological Institute (KITTI) dataset showing that vision-based single-stage detection algorithms perform excellently, while the LiDAR-based Cascade Bi-directional Fusion algorithm stands out. The paper concludes with promising research directions, highlighting that continuous hardware advancements and constant updates in object detection algorithms are bringing us closer to fully realizing the goal of autonomous driving.

Keywords: Autonomous Driving; Object Detection; Convolutional Neural Network (CNN); Light Detection and Ranging (LiDAR).

1. Introduction

Autonomous driving technology, as a significant breakthrough in today's technological realm, is leading us into a new era of transportation [1]. However, with the development of autonomous driving technology, more and more challenges and difficulties have emerged in the process of research. The core of autonomous driving is to judge the road conditions and make timely decisions and actions in various complex road environments and different weather conditions. The complexity of these factors sometimes makes it difficult for autonomous driving systems to make correct decisions, thus sparking concerns about their safety and reliability. Accurate and timely detection is necessary to guarantee the stability and safety of autonomous driving. Therefore, research and improvement in object detection technology have become one of the crucial directions in the advancement of technologies for autonomous vehicles [2,3].

In the realm of autonomous driving, strategies for object detection are broadly categorized into two approaches: vision-based approach and Light Detection and Ranging based (LiDAR-based) approach. The vision-based approach utilizes cameras as the primary sensor, augmented by affordable sensors like millimeter-wave radar for assistance. Currently, mainstream algorithms employed in vision-based approach include Faster Region-based Convolutional Neural Networks (Faster-RCNN) and the You Only Look Once (YOLO) series [4-6], they have gained widespread popularity for their efficient real-time detection capabilities and high accuracy. Additionally, in conjunction with cameras and other sensors, the LiDAR-based approach generally centers on leveraging LiDAR as the foundational sensor. In terms of data, fusion of point cloud data from LiDAR and image data from cameras is required for robust object detection. Therefore, point cloud fusion algorithms are divided into three

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.

categories as the continuous improvement of technology: early fusion, deep fusion and late fusion. Using image detector results to narrow down the range of point clouds, late fusion such as Frustum PointNets (F-PointNet) have the ability to combines 2D image information with 3D point cloud data to enhance the detection process [7]; Deep fusion algorithms as multi-View 3D (MV3D) combine point cloud and picture characteristics [8]; And early fusion algorithms like Enhancing Point Net++ (EPNet++) that process the two types of data separately [9].

This paper aims to explore and analyze the core concepts and applications of various vision algorithms in autonomous driving. This paper is organized as follows: initially, the background, significance, and evolution of research in autonomous driving and object detection algorithms are introduced. Following this, Section 2 delves into the fundamental principles of CNN, with an in-depth examination of Faster-RCNN, YOLOv5, and YOLOv8, as well as the core concepts of point cloud image fusion algorithms. Section 3 presents a range of experimental results, offering a thorough discussion and analysis of the future prospects and potential applications of these methods. The paper concludes with a summary in Section 4, reinforcing the main points and insights presented. Through this comprehensive structure, the paper systematically addresses the topic, highlighting the significance of the algorithms and their practical implications in autonomous driving.

2. Methodology

2.1. Dataset Description and Preprocessing

For general object detection, Common Objects in Context (COCO) [4-6] datasets are commonly utilized because of its large scale and diversity. However, the assessment of autonomous driving mostly rely on Karlsruhe Institute of Technology and Toyota Technological Institute (KITTI)[7-9] dataset. This benchmark dataset consists of real-world data collected from a vehicle equipped with various sensors including cameras, LiDAR, and Global Positioning System (GPS). KITTI's dataset is famous for its realism, capturing diverse weather conditions, lighting variations, and urban scenarios. Varying object scales, occlusions, and complex traffic interactions in the dataset let it become a useful evaluation criterion in this field. SUN Red-Green-Blue Depth (RGB-D) dataset [7] is a large-scale dataset that has both RGB images and depth information receiving from indoor environment. It enables the development of algorithms that demand understanding poor lighting environment.

2.2. Proposed Approach

The research goal of this paper is to study the core ideas of two vision algorithms, namely the deep learning-based convolutional neural network algorithm and the image and point cloud fusion algorithm, and their applications in the area of self-driving cars. In this section, the core ideas of the two types of algorithms through the model structure will be introduced. Faster-RCNN, YOLOv5 and YOLOv8 in the convolutional neural network algorithms as well as F-PointNet, MV3D and EPNet++ in the point cloud and image fusion algorithms will be explained in detail.

2.2.1. CNN-based on deep learning.

A neural network is mainly composed of individual neurons, each of which receives input, weights, and then processes them through an activation function, Neural network is shown in Fig. 1. Data from external inputs must be received by the input layer. The hidden layer is responsible for feature extraction and classification through complex calculations and processing of data. The output layer converts the processed result into the desired output form.

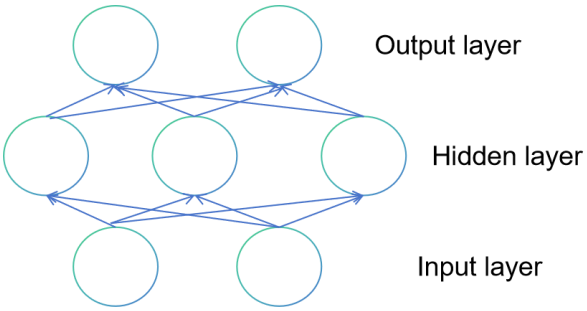

Figure 1. Neural network

This paper will first introduce the two-stage object detection algorithm Faster-RCNN. The structure of Faster-RCNN is shown in Fig. 2. From Fig. 2, Faster-RCNN can be broken down into the following four parts: Conv layers are used to extract features in the image; Region Proposal Networks is used to generate candidate regions and generate precise candidate boxes; Region of Interest Pooling is used to process the target region of interest and pool it for feeding into the subsequent fully connected layer. Classification calculates the category of the proposal and obtains the exact position of the final detection box.

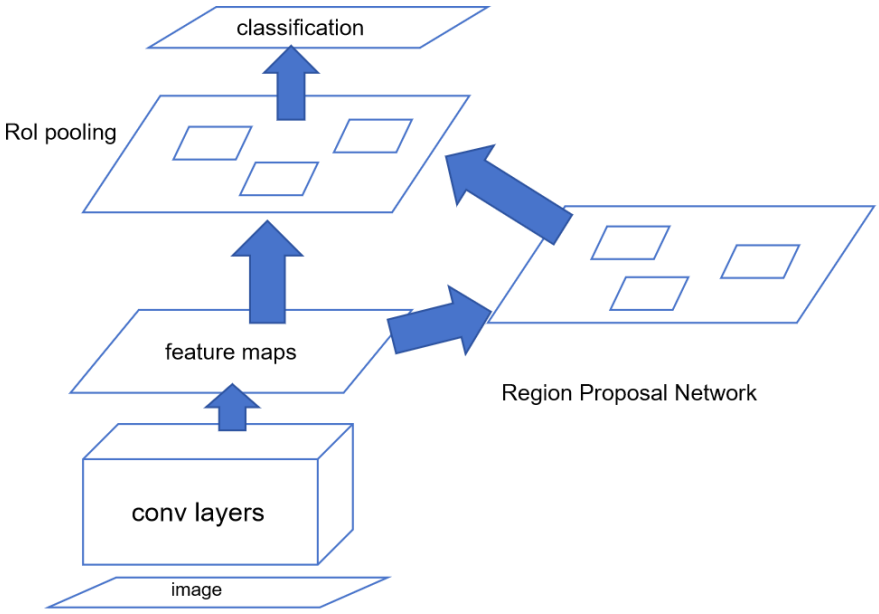

Figure 2. Structure diagram of Faster-RCNN

Then, compared with the two-stage detection algorithm, the single-stage object detection algorithm that has less process to generate candidate regions will be introduced. This paper will start with YOLOv5, which is widely used in the YOLO series. The structure of YOLOv5 is shown in Fig. 3. The network structure of YOLOv5 consists of three main parts: Backbone, Neck, and Head. First of all, the Backbone part is mainly responsible for extracting multi-level features from the input image. Its input is a raw image, which is processed through a module and a spatial pyramid pooling layer composed of multiple convolutional layers, batch normalization layers, and activation functions to generate multi-scale feature maps. The main function of the Neck part is to further process and fuse the features extracted from the Backbone, and capture the target information at different scales through upsampling and downsampling. In the Head section, the coupling head is used to further process the feature map output by Neck to generate the final detection result, including the location and category of the target.

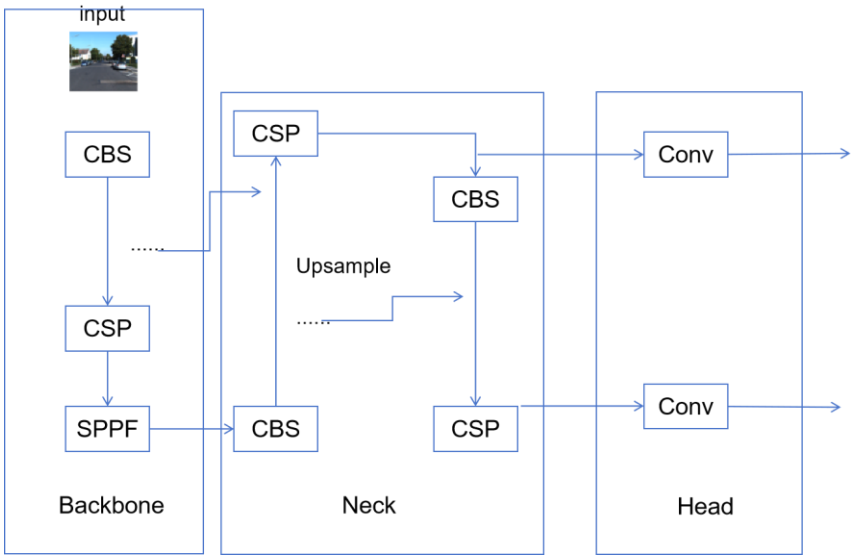

Figure 3. Structure diagram of YOLOv5

Finally, this paper will introduce the newer and more mature YOLOv8 in the YOLO series. Its network structure is not much different from that of YOLOv5. YOLOv8 is also composed of Backbone, Neck, and Head, where feature map extraction is handled by Backbone, but compared with YOLOv5, only the nonlinear activation function and convolution structure have changed. The Neck part still uses upsampling and downsampling to further process and fuse the extracted features to capture target information at different scales. The Head part uses a decoupled head to generate the location and category of the target Instead of the coupling head used in the YOLOv5.

2.2.2. Point cloud and image fusion algorithm for object detection.

The data structure called point cloud is frequently used to show a group of points in three dimensions. Each point's data includes three-dimensional coordinates (X, Y, Z) and, depending on the sensor, additional feature data such as color and laser reflection intensity. Point cloud data can be acquired from various sensors including laser scanners, photogrammetry devices, and radar. In the field of autonomous driving, point cloud data is primarily obtained from LiDAR sensors and is frequently used for environment perception and object detection. Images are common data obtained from cameras. While Image data is ordered and discrete, point cloud data is unordered and continuous. Therefore, different types of effective methods should be introduced in solving the problem. Currently, early fusion, deep fusion and lately fusion are used for better fusion of these two types of data.

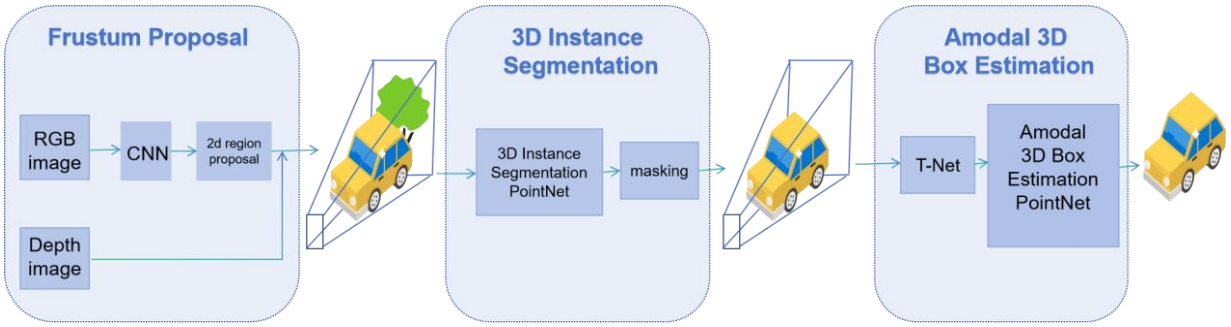

Figure 4. The structure of F-pointnet

Early fusion means use an image detector to refine the point cloud area. The data of the F-pointnet, a typical early fusion. As Fig. 4 shows, inputting RGB data into the CNN, a proposal 2D region will be selected. For example, if the region of interest (ROI) is a car, it will be easy to choose from a 2D image through CNN. Apart from that, the depth image will be imported to form the 3D frustum. This procedure is named as frustum proposal. To separate the car from the tree, it is essential to do segmentation. By using 3D instance segmentation pointnet and masking, a frustum which only

contain the car will be obtained. The final process is amodal box estimation. T-Net, a small scale of pointnet, align points by translation and get the amodal 3D bounding box for the object at last. The output is box parameters.

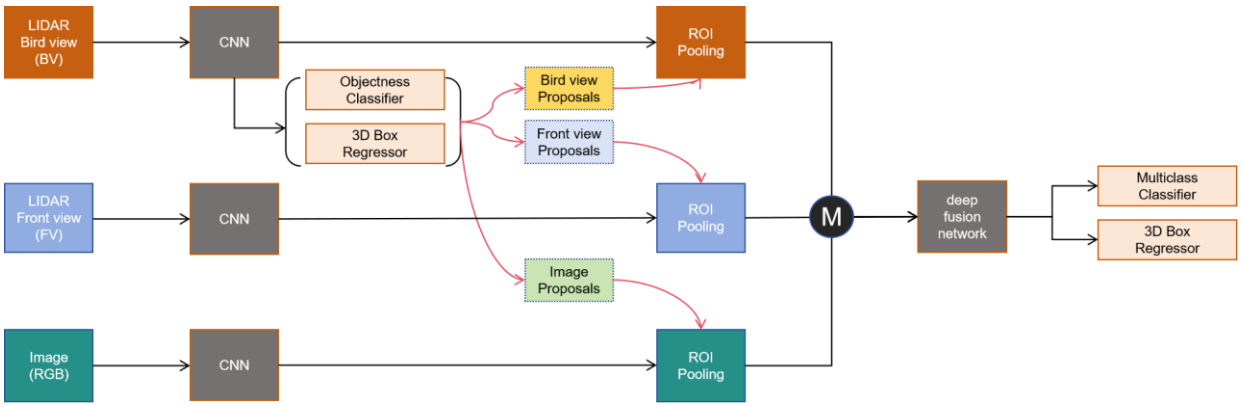

Figure 5. Simplify network of MV3D

The core of deep fusion is deep integration of image features and point cloud features which is illustrate in Fig. 5. For instance, MV3D perfectly satisfy this characteristic. The original data of MV3D is RGB image and LiDAR. The input includes three parts with the most important part of MV3D-the bird's eye (BV) view. It is simple to attain a 3D proposal from deep learning of CNN in bird view. What is more, a deep fusion network fuses region-wise features from ROI pooling across multiple views. These combined features are utilized to collectively predict object classes and estimate oriented 3D box regression.

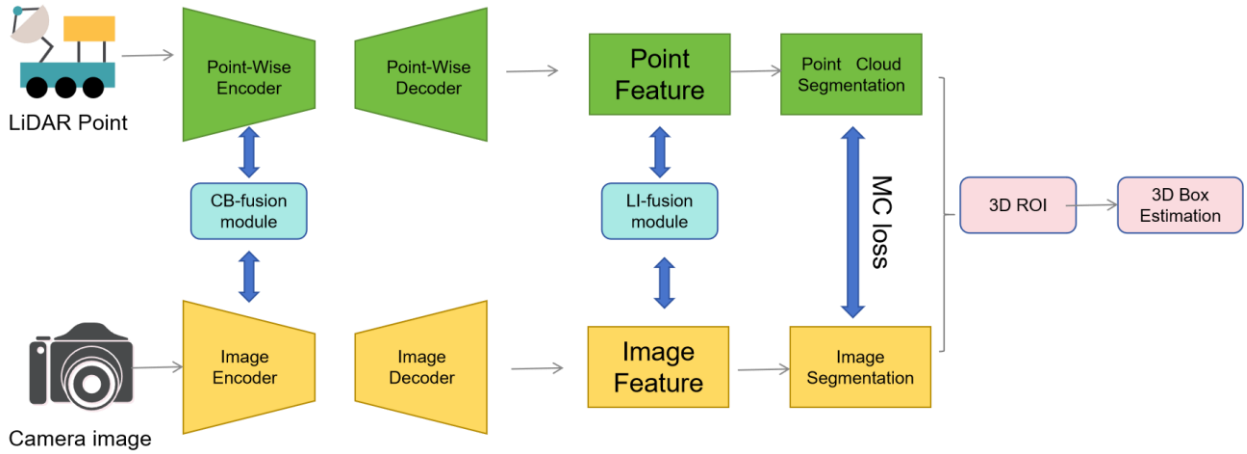

Figure 6. Mind map for EPNet++ for 3D object detection

Lately fusion emphasis on separately deals with point cloud and image which ultimately decide the 3D box estimation refer to both types of data as depicted in Fig. 6. Lately fusion algorithm such as EPNet++ includes a two-steam Region Proposal Network (RPN) stage and a refinement stage. In two-steam RPN stage, both Cascade Bi-directional Fusion (CB-Fusion) and LiDAR-guided Image Fusion (LI-Fusion) are utilized in order to enhance the semantic information. In refinement stage, Multi-Modal Consistency loss (MC loss) is invented to optimize the 3D box estimation.

3. Result and Discussion

This section will illustrate the experimental results of the specific applications of the above algorithms, analyze the advantages and disadvantages of their effects, and then discuss the advantages and disadvantages of these algorithms and the future development direction of computer vision algorithms in the field of autonomous driving.

3.1. Result Analysis

Table 1 shows the comparison of the accuracy evaluation of YOLOv5 and Faster-RCNN on the KITTI dataset, and the dataset is divided into easy, moderate and hard datasets, and its detection effect on cars, pedestrians and cyclists is measured separately. It can be seen that the detection effect of the two models is better, and Faster-RCNN has higher detection accuracy than YOLOv5 on datasets of different complexity, and the accuracy does not decrease much, which proves that Faster-RCNN is a more stable detection model.

Table 1. Accuracy of Faster-RCNN and YOLOv5 [10]

	Car	Pedestrian	Cyclist	Level
Faster-RCNN	84.81%	76.52%	74.72%	easy
YOLOv5	88.17%	60%	55.00%	
Faster-RCNN	86.18%	59.98%	56.83%	moderate
YOLOv5	78.70%	44%	39.29%	
Faster-RCNN	78.03%	51.84%	49.60%	hard
YOLOv5	69.45%	43%	32.58%	

Table 2 shows the performance indicators of Faster-RCNN and YOLOv8 training and evaluation on the customized STAMFORD automotive dataset, and the mean average of the model in the training set, validation set and test set is listed and analyzed The accuracy values of 0.5 and 0.95 are taken for the Mean Average Precision (mAP) threshold, respectively, and it can be seen that YOLOv8 has higher accuracy than Faster-RCNN, which proves that the YOLOv8 model is better and more suitable for object detection.

Table 2. Accuracy of Faster-RCNN and YOLOv8 [11]

	YOLOv8		Faster-RCNN	
	AP0.5	AP0.5-0.95	AP0.5	AP0.5-0.95
Train	93%	71%	71%	50%
Validation	93%	68%	71%	50%
Tset	90.6%	68.6%	74%	51%

Table 3 gives information about three algorithm's Average Precision (in percentage) of detecting cars on KITTI benchmark. MV3D, which proposed the bird's eye view opinion for the first time, performs the worst. The performance of F-Pointnet is not so prominent as well. According to the statistics, EPNet++ performs better than the others in easy, moderate and hard tasks. The main reason for EPNet++ outstanding performance is CB-Fusion structure. The cascade bidirectional feature enhances not only point-wise semantics but also image features, letting the weights of the two different data structure be the same. In addition, the strength of EPNet++ includes the capability of adaptable to various architectures like point or voxel.

Table 3. Comparisons on KITTI benchmark for cars

	Easy	Moderate	hard
F-Pointnet[7]	81.20	70.39	62.19
MV3D[8]	71.29	62.68	56.56
EPNet++[9]	91.37	81.96	76.71

3.2. Discussion

Crucial role of object detection. There are three fundamental tasks in the realization of autonomous driving: motion perception, motion formulation and motion control. Object detection play an

important role in motion perception which means the comprehension of surroundings to ensure the safety of the driving experience and the efficiency of path selection. Autonomous vehicle will constantly meet with various obstacles when driving. To detect these obstacles like pedestrians, animals, along with other stationary objects, different strategies are utilized. Recognizing various traffic signals, the vehicle learns to obey traffic rules and make appropriate driving decisions. In addition, lane lines recognition enables vehicles accurately identify the position and drive steadily.

Advantage and disadvantage of image detection. The traits of image detection include huge database, deep learning, and in-time react, leading to its merits and shortcomings. Image data is the only origin of collected data which means that the demand for the sensors like camera is quite affordable. By deep learning, the network has the ability to handle traffic signals, lane markings, and road signs, guaranteeing the safety of daily commuting. At the same time, driving at night or in severe weather conditions, the image data frequently fall to recognize obstacles. Therefore, more sensors should be involved in complex situations. Accurate Localization and Navigation is requested to take into consideration as well.

Strength and drawback of point cloud and image fusion. High-quality LiDAR and cameras are expensive, increasing the overall expense of the autonomous system. Therefore, the drawback could influence the large-scale factory manufacturing which impede the progress of LiDAR-based approach autonomous vehicles. However, with the development of technology, this shortage may be overcome in the future. The strong point of point cloud and image fusion is obvious. Point cloud and image fusion surpass image detection in localization, and navigation capabilities. By merging point cloud and image data, autonomous systems are good at deal with variation in dynamic environments, such as the movement of pedestrians and other vehicles.

4. Conclusion

The objective of this paper was to explore the application of several mainstream and mature computer vision algorithms in the field of autonomous driving. Initially, the paper reviewed the evolution of autonomous driving and the development of computer vision algorithms, highlighting their significance in this domain. Detailed examinations of the Faster-RCNN, YOLOv5, and YOLOv8 algorithms within convolutional neural networks were presented, along with F-PointNet, MV3D, and EPNet++ in point cloud image fusion models. Specific experimental results were then listed and analyzed. The advantages and disadvantages of these methods were discussed, along with prospects for future advancements. The study found that single-stage detection algorithms in convolutional neural networks generally outperform two-stage detection algorithms. In the realm of point cloud and image fusion, late fusion techniques demonstrated the best performance. Moreover, all the aforementioned algorithms, when applied to autonomous driving, effectively meet real-time requirements. Looking ahead, future research will continue to monitor developments in the YOLO series, shifting focus from single-modality image models to multimodal models, with the aim of applying these advancements across various fields.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Yan Z. and Wei Q. Traffic sign recognition based on deep learning. Multimedia Tools and Applications, 2022, 81 (13): 17779 – 17791.

[2] Sarthak B. and Jatin B. Real-time traffic, accident, and potholes detection by deep learning techniques: a modern approach for traffic management. Neural Computing and Applications, 2023, 35 (26): 19465 – 19479.

[3] Ya T. Z. Tian H. Song G. and Martin R. Incorporating multimodal context information into traffic speed forecasting through graph deep learning. International Journal of Geographical Information Science, 2023, 37 (9):1909 – 1935.

[4] Ren S. He K. Girshick R. and Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39 (6): 1137 - 149.

[5] Jia X. Tong Y. Qiao H. et al. Fast and accurate object detector for autonomous driving based on improved YOLOv5. 2023, 13: 9711.

[6] Glenn J. “Ultralytics yolov8”, 2023, https://github.com/ultralytics/ultralytics.

[7] Qi C. Liu W. Wu C. Su H. and Guibas L. Frustum PointNets for 3D Object Detection from RGB-D Data. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, 918 - 927.

[8] Chen X. Ma H. Wan J. Li B. and Xia T. Multi-view 3D Object Detection Network for Autonomous Driving. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, 6526 - 6534.

[9] Liu Z. Huang T. Li B. Chen X. Wang X. and Bai X. EPNet++: Cascade Bi-Directional Fusion for Multi-Modal 3D Object Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45 (7): 8324 - 8341.

[10] Yang C. Object Detection in the KITTI Dataset using YOLO and Faster R-CNN, 2023.

[11] Mulia D.A. Safitri S. & Negara, I.G.P.K. YOLOv8 and Faster R-CNN Performance Evaluation with Super-resolution in License Plate Recognition. International Journal of Computing and Digital Systems, 2024, 15 (1); 365 - 375.