Depiction Of Traffic Situation From Imagery For Vehicles

Atakan Sucu, Mert Yürekli
Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi, 34220 Istanbul, Türkiye
{atakan.sucu, mert.yurekli}@std.yildiz.edu.tr

***Özetçe* —Arabalar ilk ortaya çıktığından beri olası kazaları en aza indirmek için çeşitli kurallar ortaya koyulmuştur. Gelişen teknoloji ile bu kurallara uymada yardımcı olan hatta kurallara kendisi uyan taşıtlar geliştirmek mümkün olmuştur. Bu durum başlıca trafik durumunu anlayabilen makineler ile mümkün hâle gelmiştir. Trafik durumu tasviri, trafikte bulunan levhaların ve trafik ışıklarının tespit edilerek ne anlama geldiğinin belirlenmesi için görüntü işleme metodlarıyla yapılır. Bu makalede ViT mimarisini kullanan DETR ve CNN mimarisini kullanan YOLOv8 metodu ile iki farklı model geliştirilmiştir. Ayrıca iki modelin kıyaslaması yapılmıştır. YOLOv8 Modeli 93.6 doğruluk oranında sonuç verirken DETR modeli 89.4 doğruluk oranında sonuç vermektedir.**

Anahtar Kelimeler—Görüntü işleme, Trafik, Otonom araçlar, Trafik Tasviri, YOLOv8, DETR, ViT.

***Abstract*—Since cars first appeared, various rules have been put forward to minimize possible accidents. With the developing technology, it has been possible to develop vehicles that help to comply with these rules and even follow the rules themselves. This has been made possible mainly by machines that can understand the traffic situation. Traffic situation representation is done by using image processing methods to detect traffic signs and traffic lights and determine their meaning. In this paper, two different models are developed with DETR using ViT architecture and YOLOv8 using CNN architecture. A comparison of the two models is also made. YOLOv8 model gives 93.6 accuracy rate while DETR model gives 89.4 accuracy rate.**

Keywords—Image processing, Traffic, Autonomous cars, Traffic description, YOLOv8, DETR, Vision Transformer.

I. Introduction

With advancing technology, autonomous vehicles are on the verge of a revolution that will radically change the paradigm of urban transportation. As driverless and automated vehicles, autonomous vehicles offer potential solutions for safety, traffic efficiency and transportation accessibility. Analyzing imagery for vehicles can play an important role in traffic situation depiction, helping autonomous vehicles to better understand environmental conditions. In this context, the development of autonomous vehicle technology offers an exciting perspective towards the goal of making urban transportation systems smarter, safer and more efficient. However, rainy, foggy and dusty weather makes traffic situation description very difficult. In addition, fast reaction is required to ensure the safety of people in traffic. This brings the problem of fast object detection. In this paper, we propose a real-time traffic description model using transformers and CNN architectures and focus on comparing these two different architectures.

II. PRELIMINARY EXAMINATION

This paper investigates and compares two different real-time traffic depiction models implemented with CNN-based YOLOv8 and transformer-based DETR. It attempts to find a model that can effectively depict traffic conditions in the face of adverse weather conditions. To do this, it pulls images from the Road Signs dataset which includes traffic sign and traffic light images.

A. Literature Review

Traffic situation depiction research, which likely began in the 2000s, has seen considerable advancements. Many different models have been developed and some concrete results have started to be obtained in recent years. [1] In a 2021 study, a R-CNN method used to achieve more accurate and faster detection of small and hidden objects in complex traffic conditions. Achieving precision and recall of 86.2 and 98.6 respectively. [2] In A 2023 study, YOLOv8 architecture was used and achieved a mAP50-95% improvement of 14% and 13% compared to Yolov5 and YOLOv7 architectures, respectively. [3] Another 2023 study, A model was developed with a transformer architecture using the GTSRB dataset. As a result of this method, AP increased by 2.1% compared to Cascade CNN and 2.4% compared to Faster R-CNN. [4] A study in 2023, focused on the performance of YOLOv8 versions in day and night conditions. The accuracy rates of YOLOv8 models were calculated as YOLOv8n 0.76%, YOLOv8s 0.78%, YOLOv8m 0.82%, YOLOv8l 0.85%, YOLOv8x 0.89%. [5] In 2021, CNN and ViT models are compared. It was observed that training the ViT model takes 30 times longer than training the CNN model and the accuracy of the CNN model is 75.71% while for ViT it is 75.46%. [6] In 2022, a Pyramid transformer model was implemented with pyramid shrinkage layers that shrink and embed data into rich tokens which uses GTSDB as the dataset. This method achieved a mAP value of 77.8%. [7] In a 2020 study. Facebook presented a model that uses CNN and transformer architectures together. Although the accuracy rate produced by the model was not specified, the approach that transformers can be used with CNN was put forward for the first time.

III. Data Design

While selecting the dataset, many sources were searched. Road Signs dataset was selected considering the training time of the model. There are 2093 images in this dataset. Train set contains 1376 images, valid set contains 488 images and test set contains 229 images. The images in the classes in the dataset are distributed evenly and each class has an average of 100 images. The images are given in 640x640 resolution. The dataset contains examples such as traffic signs and traffic lights in various conditions.

Figure 1 Example For Training Images 1

Figure 2 Example For Training Images 2

IV. Traffic Detection Algorithm

A. YOLOv8

YOLOv8 (You Only Look Once) is an object recognition algorithm that provides powerful performance in image-based object detection. Unlike other object detection algorithms, the main purpose of YOLOv8 is to perform object detection from a single image instead of performing object detection multiple times in different parts of the image. This makes YOLOv8 fast and efficient in object detection.

YOLOv8 uses a deep neural network architecture called EfficientDet that provides fast and high accuracy. This network is trained by utilizing a large amount of image data to learn the patterns and properties of real-world objects. Once training is complete, the network can be used to perform object detection on new images. This can be used to quickly and accurately determine the presence and positions of objects in the image.

The algorithm consists of two main stages: the feature extractor (backbone) and the object detection head. In the first stage, features of the image are extracted, usually using pre-trained models such as convolutional neural networks (CNN). Then, more specific features are extracted using a specialized feature extractor (neck). In the head part of the

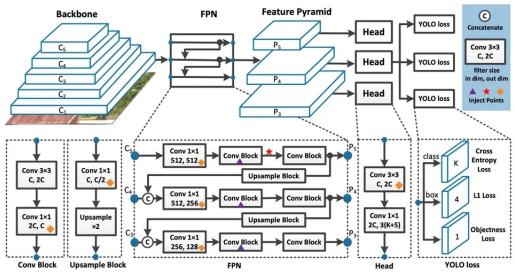

Figure 3 YOLOv8 Model Structure

algorithm, object detection is performed over feature maps of different sizes. YOLOv8 is characterized by its ability to detect all objects in an image simultaneously and in a single pass. This enables high-speed object recognition in real-time applications. The outputs are in the form of a list of detected objects with their positions, classes and confidence scores. YOLOv8 offers improved performance and ease of use compared to previous versions and is preferred by a wide range of users in object recognition applications.

B. DETR

Facebook's DETR (DEtection TRansformers) algorithm is a prominent image processing model for object detection. Unlike traditional object detection algorithms, DETR takes an approach that focuses on end-to-end learning. DETR transforms object detection into a ranking problem using a network of transformers. This ensures that all objects in the image are evaluated simultaneously. The algorithm includes a pre-trained image feature extractor and a transformer-based object detection head. The image feature extractor works on the basis of a convolutional neural network (CNN) and extracts feature maps from the image. Then, the transformer-based head performs object detection using the feature maps. DETR can dynamically handle the number of objects and performs the classification and localization of objects in a single pass. This allows DETR to provide more flexible and efficient object detection. DETR is an open source project developed by Facebook and has been widely adopted by the industry.

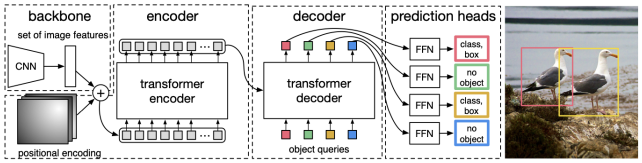

Figure 4 DETR Structure

V. Experimental Results

The criteria used to measure the performance of the models are "IoU", "AP", "mAP", "Precision", "Recall", "F1 Score". After training the model, it was tested with 229 different images. Descriptions and results of the tests are described below.

1) YOLOv8: Confusion matrix is a table used to evaluate the performance of a classification model. It can also be used to evaluate object recognition algorithms such as YOLOv8. The confusion matrix contains the number of true classes and predicted classes, allowing to analyze the accuracy, precision, specificity and other performance metrics of the model. In figure 5 we can see the accuracy rate distribution of our model by class. The figure 6 shows the change in the loss functions as the number of epochs increases. As the number of epochs increases, mAP and Recall values increase and loss functions decrease.

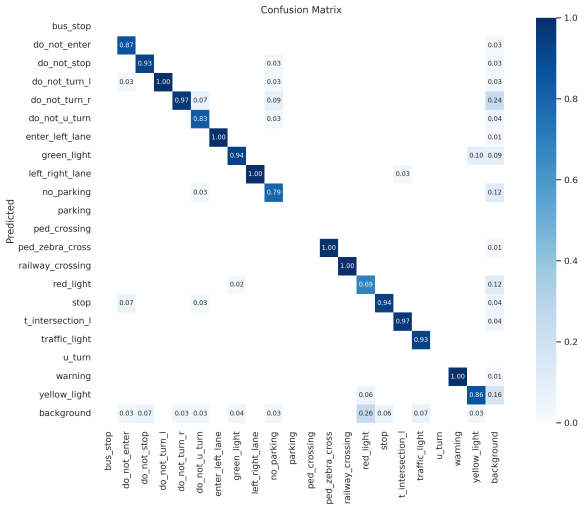

Figure 5 100 Epoch

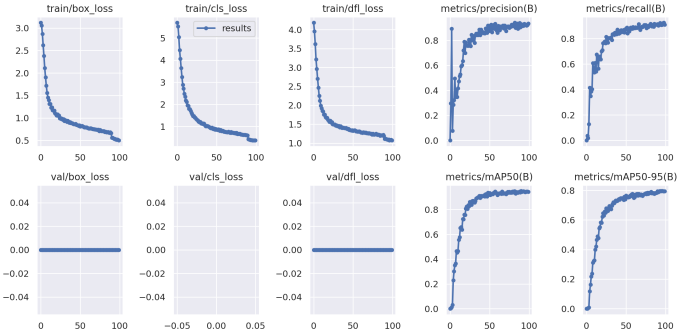

Figure 6 100 Epoch

2) DETR: Graphs of the output of the DETR algorithm play an important role in evaluating and optimizing the performance of the algorithm. These graphs include metrics such as mAP, lossbox and class error. mAP indicates the detection accuracy of different classes, while low lossbox values indicate that the model predicts object locations more accurately. Likewise, lower class error values indicate that the model classifies classes more accurately. These graphs visualize changes in the model's performance during the training process and provide guidance for training and tuning the algorithm more effectively.

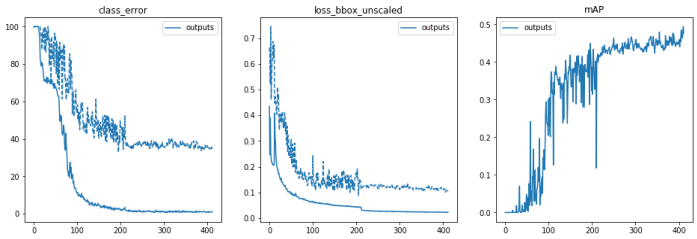

Figure 7 DETR Results

VI. PERFORMANCE ANALYSIS

Trained with 1376 different images, the model was trained with 50 and 100 epoch values. The model trained with 50 epochs took 4 hours while the model trained with 100 epochs took 8 hours. The DETR model was trained in 10 hours with 100 epochs.

Figure 8 YOLOv8 Object Detection

Mean Average Precision values of YOLOv8 models trained with 50 and 100 epochs, respectively. While the mAP50 value of the model trained with 50 epochs is 0.936, the mAP50 value of the model trained with 100 epochs is calculated as 0.942.

Figure 9 DETR Object Detection

The ViT-based DETR model was trained with 100 epochs and achieved the highest average accuracy of 0.894. Based

on this data, it is determined that the DETR model achieves a lower accuracy rate than the YOLOv8 model.

VII. CONCLUSION

The project focuses on developing a model that can depict the traffic situation for vehicles from images. Within the scope of the project, multiple models were produced with two different methods. These models were developed with the CNN-based YOLOv8 method, which was released in 2023, and the DETR method, which uses transformers with CNN, which is used in natural language processing and has recently started to be used in image processing. The models were trained with different epoch values and the time difference between 50 and 100 epochs was doubled, but the accuracy did not change enough to affect performance. The YOLOv8 model has a very good detection rate with an average accuracy of 93.6%. The DETR model achieved an average accuracy of 89.4%. It was observed that YOLOv8 achieved a better detection rate than DETR. It was concluded that YOLOv8 is better than DETR in real-time detection, where DETR was insufficient in the experiments conducted on videos. It was also observed that the training time of DETR model was considerably longer than YOLOv8 model. For this reason, the YOLOv8 model can be considered as a more effective solution than the transformer-based DETR model in terms of both speed and accuracy.

REFERENCES

[1] C.-j. Li, Z. Qu, S.-y. Wang, and L. Liu, "A method of cross-layer fusion multi-object detection and recognition based on improved faster r-cnn model in complex traffic environment," *Pattern Recognition Letters*, vol. 145, pp. 127–134, 2021.

[2] F. N. Ortataş and M. Kaya, "Performance evaluation of yolov5, yolov7, and yolov8 models in traffic sign detection," in *2023 8th International Conference on Computer Science and Engineering (UBMK)*. IEEE, 2023, pp. 151–156.

[3] J. M. Kaleybar, H. Khaloo, and A. Naghipour, "Efficient vision transformer for accurate traffic sign detection," *arXiv preprint arXiv:2311.01429*, 2023.

[4] A. Afdhal, K. Saddami, S. Sugiarto, Z. Fuadi, and N. Nasaruddin, "Real-time object detection performance of yolov8 models for self-driving cars in a mixed traffic environment," in *2023 2nd International Conference on Computer System, Information Technology, and Electrical Engineering (COSITE)*. IEEE, 2023, pp. 260–265.

[5] K. Lu, Y. Xu, and Y. Yang, "Comparison of the potential between transformer and cnn in image classification," in *ICMLCA 2021; 2nd International Conference on Machine Learning and Computer Application*. VDE, 2021, pp. 1–6.

[6] O. N. Manzari, A. Boudesh, and S. B. Shokouhi, "Pyramid transformer for traffic sign detection," in *2022 12th International Conference on Computer and Knowledge Engineering (ICCKE)*. IEEE, 2022, pp. 112–116.

[7] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.