Vol. 1, No. 1, September 2024

IJAIS International Journal of Artificial Intelligence and Science

Publsihed By Asosiasi Doktor Sistem Informasi Indonesia

Vehicle Detection on The Traffic Using Detection Transformer (DETR) Algorithm

Rofiatul Khoiriyah[1], Aria Hendrawan[2]

[1,2]Faculty of Information Technology and Communication, Semarang University, Semarang, Indonesia
Email: 211190094@student.usm.ac.id[1], ariahendrawan@usm.ac.id[2]

Received:
August 3, 2024
Revised:
August 31, 2024
Accepted:
September 20, 2024
Published:
September 28, 2024

Corresponding Author:
Author Name*:
Rofiatul Khoiriyah
Email*:
rofiatulk292@gmail.com

DOI: 10.63158/IJAIS.v1.i1.4

Abstract. Object detection is a computer vision technique aimed at detecting and identifying objects in images or videos. In recent years, with advancements in Machine Learning and Deep Learning, object detection has made significant progress in various fields such as healthcare, security, and transportation. The DETR algorithm is a novel approach in object detection that combines transformer architecture with attention techniques to address object detection challenges. This research applies the DETR algorithm with ResNet backbone for vehicle detection on the roads, involving 6 object classes: Car, Truck, Bus, Motorcycle, Pickup Car, and Truck Box. Four training experiments were conducted: DETR-ResNet50, DETR-ResNet101, DETR-DC5-ResNet50, and DETR-DC5-ResNet101. The implementation results show that DETR-DC5 improves the accuracy of vehicle detection. DETR-DC5 with ResNet-101 achieved the highest score for AP50, which is 0.957. However, it should be noted that DETR-DC5 with ResNet-50 managed to maintain overall AP stability, with a lower parameter of 35.5. The model's outcomes in this study can be effectively applied for vehicle detection on the roads.

Keywords: DETR, object detection, vehicle detection

1. INTRODUCTION

Computer vision approach called "Object Detection" seeks to find and recognize objects in images or videos. With the advancements in Deep Learning technology, object detection has achieved remarkable levels of accuracy and speed, opening up new opportunities in various fields. This makes object detection highly interesting for development and research. According to the analysis of 2000 scientific journals using bibliometric techniques [1], "Object Detection" has been the most researched and developed topic in the last three years : 2021, 2022, and 2023 Figure 1. In Indonesia, the development of machine learning technology for detecting objects in traffic, such as in autonomous vehicle applications [2], traffic management [3], and predicting transportation service demands [4], is highly needed. To achieve accurate vehicle object detection, the use of appropriate algorithms is crucial. Some algorithms that can be used for object detection include Faster-RCNN [5], Single Shot MultiBox Detector (SSD) [6], and DETR [7].

DETR [7] (Detection Transformer) is a powerful object detection algorithm that combines transformer architecture with attention mechanisms to address object detection challenges. DETR offers several advantages over traditional object detection methods by utilizing a transformer architecture that consists of three main components: a backbone CNN, an Encoder-Decoder transformer, and feedforward networks (FFN) [8]. This makes DETR a pioneering application of transformers in the field of object detection. However, DETR also has its limitations. It may struggle with effectively recognizing small-sized objects and can be time-consuming in processing images [9]. These limitations are important considerations when applying DETR in practical scenarios. Researchers and developers are continuously working to improve the performance and efficiency of DETR and other object detection algorithms to overcome these challenges.

DETR offers a solution by providing a straightforward workflow and requiring a relatively small network architecture [10]. It has demonstrated state-of-the-art performance on the COCO benchmark for object detection. It represents a promising new approach for simple, efficient, and effective object detection [7], making it suitable for detecting vehicle objects in this research. However, DETR may face challenges in accurately recognizing small-sized vehicle objects and can exhibit slow convergence [8].

Residual Network (ResNet) [11] is one of the architectures of Convolutional Neural Networks (CNNs) that has shown significant improvements compared to previous CNN architectures. By incorporating ResNet into the DETR algorithm, it is possible to leverage its ability to capture more complex features and improve the overall performance of

IJAIS International Journal of Artificial Intelligence and Science

Publsihed By Asosiasi Doktor Sistem Informasi Indonesia

vehicle detection. ResNet's deep residual connections allow for better information flow and can enable more accurate representation learning, leading to enhanced object detection capabilities.

This research will apply ResNet as the backbone in the DETR algorithm to detect vehicles in traffic. The aim of this research is to maximize the accuracy of vehicle detection in traffic using the selected algorithm. Therefore, to generate a model that can be applied in the future to assist in the development of more efficient and reliable vehicle object detection systems.

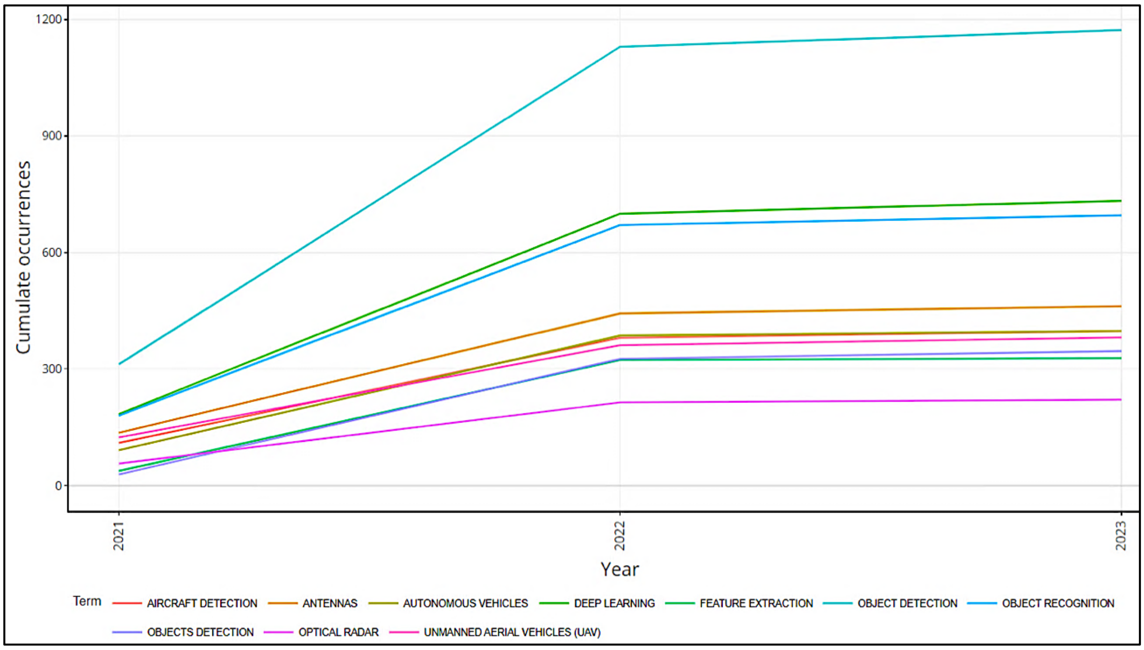


Figure 1. Analysis of Bibliometrix in the Last 3 Years [12]

2. METHODS

2.1 Dataset

This research utilized the COCO dataset to train our model [13], which consists of 1,104 (One Thousand One Hundred Four) images with a frame size of 480x480 pixels. The dataset has been processed by previous researchers using Mosaic data augmentation [13], which is a Machine Learning technique used to increase the size and diversity of the dataset by combining four images into one. The dataset includes six object classes related to vehicles: Car, Truck, Bus, Motorcycle, Pickup Car, and Truck Box, as shown in Figure 2.

Vol. 1, No. 1, September 2024

IJAIS International Journal of Artificial Intelligence and Science

Publsihed By Asosiasi Doktor Sistem Informasi Indonesia

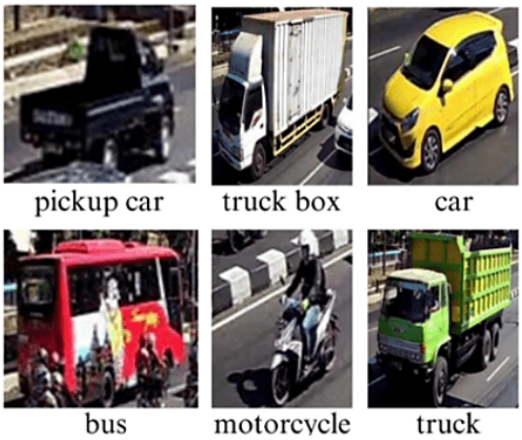


Figure 2. Dataset Classes [13]

2.2 Pre-Processing

This research utilizes the ResNet backbone and the DETR algorithm to train the vehicle detection model. Following [7], the concept of Inference and NMS (Non-Maximum Suppression) is applied in this study. The hyperparameter settings and training strategy adhere to the DETR approach [7]. The results from two different backbones, ResNet-50 and ResNet-101, are reported. Additionally, the performance of DETR-DC5 [7], [14], which improves the detection of small objects, is also evaluated. The base model is trained for 50 epochs using an A100 GPU on 1,104 images as the training data. The training process typically takes around 2-5 hours, depending on the specific model used. The architecture as shown in Figure 3.

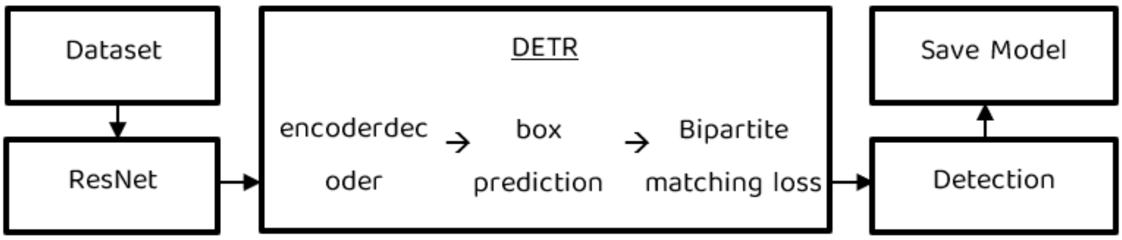


Figure 3. Object Detection Pre-Processing

The overlap between predicted bounding boxes and the ground truth bounding boxes is measured by DETR using IoU. IoU is calculated by subtracting the intersection area from the union area of two bounding boxes, as shown in Equation 1. A higher IoU provides a stricter and more conservative evaluation of object prediction accuracy.

$$IoU = \frac{intersection\ area}{union\ area} \tag{1}$$

IJAIS International Journal of Artificial Intelligence and Science
Published By Asosiasi Doktor Sistem Informasi Indonesia

AP (Average Precision) is a metric commonly used in object detection and information retrieval tasks to evaluate the performance of a model in terms of precision and recall. The formula as shown in Equation 2.

$$AP = \int_0^1 p(r)dr \tag{2}$$

Here, $p(r)$ represents the precision at a given recall value (r). The integral calculates the area under the precision-recall curve, which provides a single scalar value that summarizes the model's performance across different recall levels. A higher AP indicates better precision-recall trade-off and thus a more accurate and reliable model.

3. RESULTS AND DISCUSSION

3.1 Performance Evaluation

The evaluation results of DETR and DETR-DC5 models with ResNet-50 and ResNet-101 backbones on the COCO dataset are presented in Table 1.

Table 1. Results of Vehicle Detection

Method	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	$AP_{\bar{x}}$	Params
DETR	R50	0.626	0.915	0.779	0.455	0.604	0.791	0.695	41.5
DETR	R101	0.620	0.946	0.704	0.461	0.561	0.801	0.628	60.5
DETR-DC5	R50	**0.665**	0.951	**0.845**	0.464	**0.654**	0.768	**0.725**	**35.5**
DETR-DC5	R101	**0.665**	**0.957**	0.744	**0.494**	0.626	**0.820**	0.718	60.5

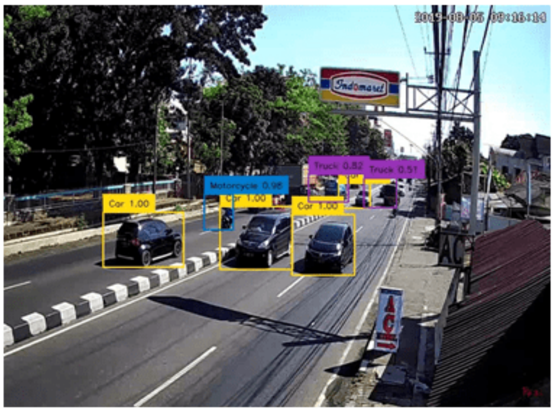
This research conducted four training sessions with the DETR-R50, DETR-R101, DETR-DC5-R50, and DETR-DC5-R101 models Table 1. The analysis results indicate that $(AP_{\bar{x}})$ represents the average of six types of Average Precision (AP), namely $AP_{50\text{-}95}$, AP_{50}, AP_{75}, AP_S, AP_M, and AP_L. DETR-DC5 with ResNet-50 achieved the highest $(AP_{\bar{x}})$ value of 0.725. to calculate AP using Equation 3.

$$AP_{\bar{x}} = \frac{x_1 + x_2 + \ldots + x_n}{n} \tag{3}$$

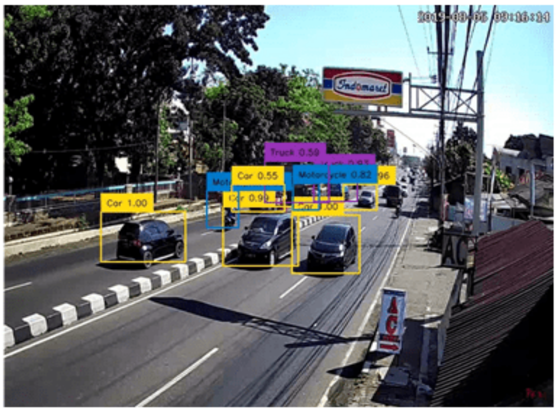
IJAIS International Journal of Artificial Intelligence and Science
Published By Asosiasi Doktor Sistem Informasi Indonesia

DETR-DC5 with ResNet-50 excelled in AP_{50-95} (0.665), AP_{75} (0.845), and AP_M (0.654). On the other hand, DETR-DC5 with ResNet-101 excelled in AP_{50-95} (0.665), AP_{50} (0.957), AP_S (0.494), and AP_L (0.820). These results indicate that DETR-DC5 with ResNet-101 is more effective in detecting small-sized vehicle objects. However, DETR-DC5 with ResNet-50 still provides good performance.

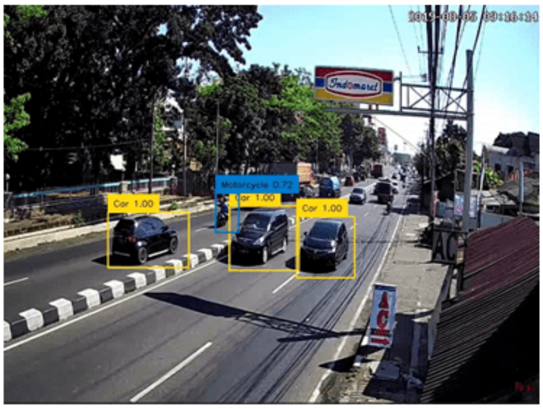
This research attempted to implement the generated model to detect vehicle objects on the highway in real- Two videos were used as a comparison material to assess the extent to which the model recognizes vehicle objects. Figure 4 represents the outcome from Video-1, which served as the training data in this study, and Figure 5 is the result of Video-2 acquired from pixabay.com.



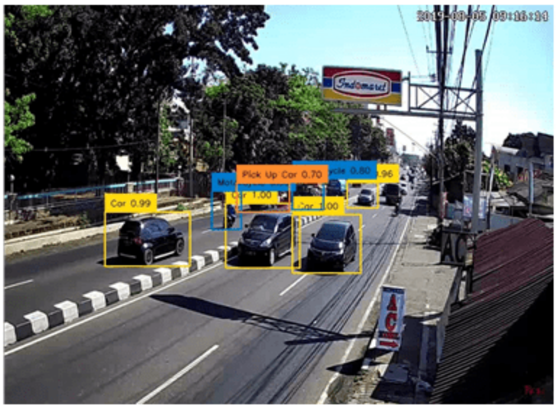
DETR-ResNet50



DETR-ResNet101



DETR-DC5-ResNet50



DETR-DC5-ResNet101

Figure 4. Vehicle detection in video-1

From Figure 4 and Figure 5, it can be observed that DETR-DC5 with ResNet-50 as the backbone exhibits a more cautious approach in object detection, aiming to reduce the

IJAIS International Journal of Artificial Intelligence and Science
Publsihed By Asosiasi Doktor Sistem Informasi Indonesia

potential for detection errors. On the other hand, using higher parameters, DETR-DC5 with ResNet-101 as the backbone is capable of detecting more vehicles and displays higher sensitivity to smaller-sized vehicles. Based on the results of experiments using all four models (DETR-ResNet50, DETR-ResNet101, DETR-DC5-ResNet50, and DETR-DC5-ResNet101), it can be concluded that the performance of DETR in terms of accuracy can be considered good.

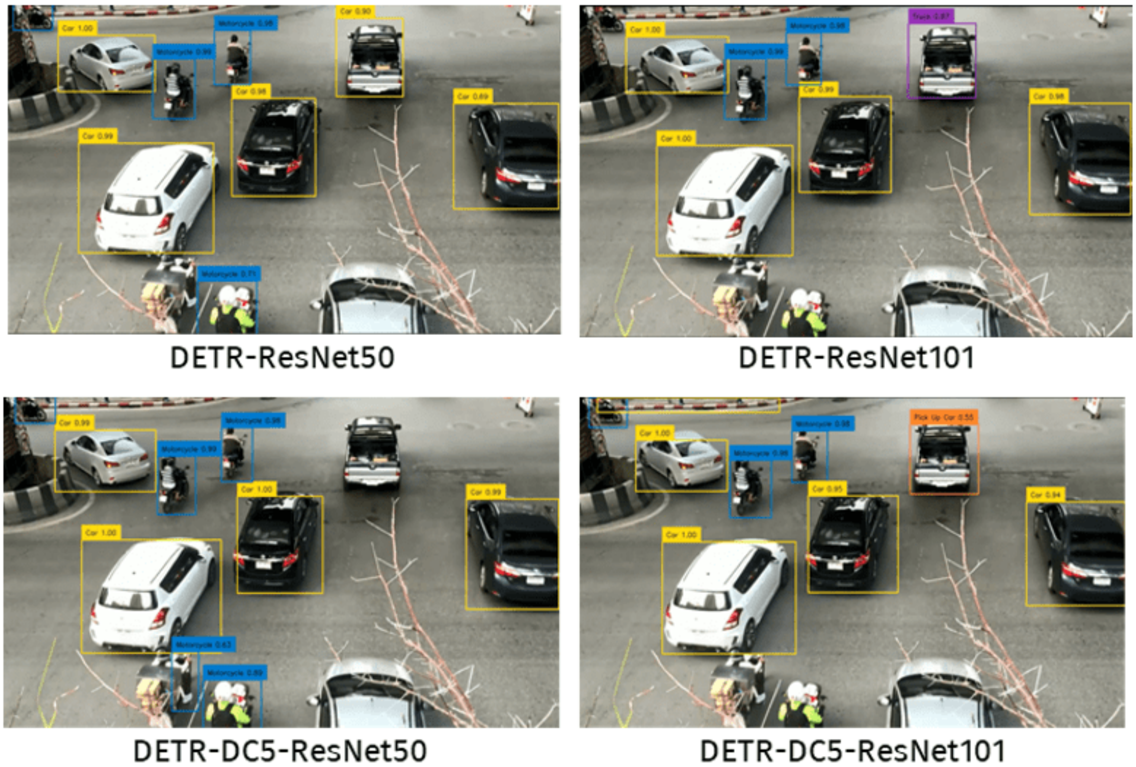


Figure 5. Vehicle detection in video-2

In Figure 5 the model is unable to recognize or misidentifies the pickup car/transporter. This is influenced by the different video/image capture positions between video-1 and video-2, and in this study, video-2 was not included in the training data used to create the model. Training the model with a diverse dataset is a highly effective strategy to enhance detection accuracy. Therefore, the model can be adapted and applied in various situations and conditions, such as vehicle object detection in both daytime and nighttime scenarios. Several factors can impact object detection scores, including data quality and quantity, model architecture, label quality, parameters and hyperparameters, data preprocessing, and computational capacity. All these factors interact with each other and contribute to the object detection outcomes.

Vol. 1, No. 1, September 2024

IJAIS International Journal of Artificial Intelligence and Science

Publsihed By Asosiasi Doktor Sistem Informasi Indonesia

In the context of the training dataset used in this study, there are several aspects that have the potential to significantly affect detection scores. First, the size of objects (large, medium, small) within the training data can have a notable impact on prediction scores. Additionally, overlapping object positions also have the potential to affect prediction scores. This underscores the importance of considering variation and complexity in the training data to achieve accurate prediction scores.

3.2 Discussion

The results of this study provide insightful analysis into the performance of the DETR and DETR-DC5 models with ResNet-50 and ResNet-101 backbones on vehicle detection tasks using the COCO dataset. The evaluation results, as detailed in Table 1, offer a comprehensive view of the models' capabilities across various metrics, including AP, AP50, AP75, APS, APM, and APL.

The DETR-DC5 model with the ResNet-50 backbone achieved the highest average precision ($\overline{AP\_x}$) of 0.725, demonstrating its superior performance across most metrics compared to other models. Specifically, DETR-DC5 with ResNet-50 excelled in overall AP (0.665), AP75 (0.845), and APM (0.654). This indicates that DETR-DC5 with ResNet-50 is particularly adept at detecting medium-sized vehicle objects, which could be crucial in scenarios where vehicles of moderate size are predominant.

On the other hand, the DETR-DC5 model with ResNet-101 backbone showed exceptional performance in AP50 (0.957), APS (0.494), and APL (0.820), suggesting it is more sensitive to small-sized and large-sized vehicle objects. This sensitivity makes DETR-DC5 with ResNet-101 more effective in scenarios where vehicle objects vary significantly in size or when the detection of smaller objects is critical.

When applying these models to real-world scenarios, such as detecting vehicles in highway videos, the DETR-DC5 with ResNet-50 backbone exhibited a more cautious approach, reducing the likelihood of detection errors. In contrast, DETR-DC5 with ResNet-101, due to its higher parameter count, demonstrated a higher sensitivity to detecting a larger number of vehicles, including smaller-sized ones. This difference in performance highlights the trade-off between detection accuracy and sensitivity depending on the backbone and model parameters used.

IJAIS International Journal of Artificial Intelligence and Science
Publsihed By Asosiasi Doktor Sistem Informasi Indonesia

The study also identified limitations in the model's performance when applied to different datasets. For instance, in Video-2, the model struggled to correctly identify certain vehicles, such as a pickup car/transporter. This issue can be attributed to differences in video capture conditions between the training data (Video-1) and the test data (Video-2). The model's inability to generalize effectively to Video-2 underscores the importance of training with a diverse dataset. Incorporating a wider range of scenarios, including various angles, lighting conditions, and vehicle types, is crucial to improving the model's robustness and generalizability.

Several factors have been identified that could potentially impact the object detection scores, including the quality and quantity of data, model architecture, label quality, parameters and hyperparameters, data preprocessing, and computational resources. Among these, the variation in object sizes (large, medium, small) within the training data is particularly influential, as it directly affects the model's prediction scores. Additionally, overlapping object positions in the dataset could also skew prediction results, further emphasizing the need for a well-rounded and complex training dataset.

The DETR-DC5 models, particularly when paired with the ResNet-50 and ResNet-101 backbones, show promising results in vehicle detection tasks. While DETR-DC5 with ResNet-50 provides a balanced performance across various metrics, DETR-DC5 with ResNet-101 excels in scenarios requiring high sensitivity to object size variation. The study highlights the importance of a diverse training dataset to improve the model's generalization capability and suggests that model performance can be significantly influenced by the interplay of various factors such as data quality, model architecture, and computational resources.

4. CONCLUSION

This study implements the DETR algorithm for vehicle detection on the road. The main objective is to demonstrate the performance of the DETR algorithm in achieving state-of-the-art results in object detection tasks based on COCO metrics. Training was conducted to compare the performance of the DETR method with DETR-DC5, using ResNet-50 and ResNet-101 as backbones. The overall results show that DETR-DC5 improves vehicle detection accuracy. The overall AP50-95 remains consistent between

IJAIS International Journal of Artificial Intelligence and Science
Publsihed By Asosiasi Doktor Sistem Informasi Indonesia

DETR-DC5 with ResNet-50 and DETR-DC5 with ResNet-101, at 0.665. DETR-DC5 with ResNet-101 achieves the highest value of 0.957 in AP50 and proves to be more effective in detecting small objects, with an APS value of 0.494. However, it should be noted that DETR-DC5 with ResNet-50 as the backbone achieves a good performance combination with a relatively low number of parameters, namely 35.5, making it an efficient choice. Finally, DETR-DC5 is proven to be more effective in improving object detection accuracy. However, it's important to note that the training computation time for DETR-DC5 is approximately twice as long as regular DETR. The choice of method depends on specific needs and resource availability. It's also important to consider other factors such as training time and implementation complexity when selecting the appropriate object detection method.

ACKNOWLEDGMENT

The Authors would like to extend our gratitude to the Department of Informatics Engineering, Semarang University, for their full support in this research.

REFERENCES

[1] M. Aria and C. Cuccurullo, "bibliometrix: An R-tool for comprehensive science mapping analysis," *J. Informetr.*, vol. 11, no. 4, pp. 959–975, Nov. 2017, doi: 10.1016/j.joi.2017.08.007.

[2] Z. Chen et al., "Fast vehicle detection algorithm in traffic scene based on improved SSD," *Measurement*, vol. 201, p. 111655, Sep. 2022, doi: 10.1016/j.measurement.2022.111655.

[3] D. Lorencik and I. Zolotova, "Object recognition in traffic monitoring systems," in *DISA 2018 - IEEE World Symposium on Digital Intelligence for Systems and Machines, Proceedings*, Oct. 2018, pp. 277–282, doi: 10.1109/DISA.2018.8490634.

[4] M. A. Bin Zuraimi and F. H. Kamaru Zaman, "Vehicle detection and tracking using YOLO and DeepSORT," in *ISCAIE 2021 - IEEE 11th Symposium on Computer Applications and Industrial Electronics*, Apr. 2021, pp. 23–29, doi: 10.1109/ISCAIE51753.2021.9431784.

IJAIS International Journal of Artificial Intelligence and Science
Publsihed By Asosiasi Doktor Sistem Informasi Indonesia

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[6] W. Liu et al., "SSD: Single Shot MultiBox Detector," in *Lecture Notes in Computer Science* (including subseries *Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*), vol. 9905 LNCS, Dec. 2015, pp. 21–37, doi: 10.1007/978-3-319-46448-0_2.

[7] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," in *Lecture Notes in Computer Science* (including subseries *Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*), vol. 12346 LNCS, May 2020, pp. 213–229, doi: 10.1007/978-3-030-58452-8_13.

[8] X. Zhu et al., "Deformable DETR: Deformable Transformers for End-to-End Object Detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 13177–13186, doi: 10.1109/CVPR46437.2021.01298.

[9] X. Dai, Y. Chen, J. Yang, P. Zhang, L. Yuan, and L. Zhang, "Dynamic DETR: End-to-End Object Detection with Dynamic Attention," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 2968–2977, doi: 10.1109/ICCV48922.2021.00298.

[10] E. Arkin, N. Yadikar, X. Xu, A. Aysa, and K. Ubul, "A survey: object detection methods from CNN to transformer," *Multimed. Tools Appl.*, Oct. 2022, doi: 10.1007/s11042-022-13801-3.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Dec. 2015, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[12] M. Aria and C. Cuccurullo, "bibliometrix: An R-tool for comprehensive science mapping analysis," *J. Informetr.*, vol. 11, no. 4, pp. 959–975, Nov. 2017, doi: 10.1016/j.joi.2017.08.007.

[13] A. Hendrawan, R. Gernowo, O. D. Nurhayati, B. Warsito, and A. Wibowo, "Improvement Object Detection Algorithm Based on YoloV5 with BottleneckCSP," in *Proc. IEEE Int. Conf. Commun. Netw. Satell. (COMNETSAT 2022)*, 2022, pp. 79–83, doi: 10.1109/COMNETSAT56033.2022.9994461.

[14] F. Yu, V. Koltun, and T. Funkhouser, "Dilated Residual Networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR 2017)*, May 2017, vol. 2017-Jan., pp. 636–644, doi: 10.1109/CVPR.2017.75.