Application and Development of DETR and Its Variants in Traffic Vehicle Detection

Shaohua Luo[1*]

[1]School of Computer Science, Beijing Information Science and Technology University, 102200, Beijing, China

Abstract. With the development of intelligent transportation and autonomous driving, higher requirements are placed on the accuracy, robustness, and real-time performance of vehicle detection in traffic scenarios. Traditional CNN based methods have shortcomings in global relationship modeling and prior dependencies. DETR, as the first end-to-end Transformer detection framework, provides a new path for vehicle detection. However, the original DETR still has limitations in training efficiency, small object detection, and high-resolution computational overhead, which are more prominent in complex traffic scenarios. In response to the above issues, researchers have proposed methods such as multi-scale feature enhancement, lightweight and real-time improvement to enhance detection performance and inference efficiency. The article further explores the research direction of integrating geometric priors, multimodal information, and adaptive reasoning to enhance model robustness and cross scene adaptability. I hope this review can provide reference for the research of intelligent transportation perception systems and promote further application and breakthroughs of Transformer based detection technology in the field of vehicles.

1 Introduction

In recent years, with the widespread application and rapid development of intelligent transportation systems and autonomous driving technology, the accuracy, real-time performance, robustness, and other requirements of vehicle target detection algorithms in road scenes have significantly improved. However, traditional convolutional neural network–based detection frameworks, such as YOLO, generally rely on a large number of manually designed anchor boxes and non-maximum suppression (NMS). This reliance leads to performance bottlenecks in scenarios with severe occlusion and large variations in object scale.

With the tremendous success of Transformer in the field of natural language processing, researchers have begun to introduce it into the field of object detection. Among these efforts, Detection Transformer (DETR), proposed by Carion et al., reformulates object detection as a sequence prediction problem and pioneeringly realizes an end-to-end detection paradigm without NMS or anchor box design [1]. The self attention mechanism introduced by DETR

* Corresponding author's email: tukuai1109@gmail.com

can capture global contextual information, which has advantages for road scenes with high vehicle density and complex traffic element interactions, thus bringing new development directions for traffic vehicle detection.

However, the original DETR has problems such as slow convergence speed, poor small object detection ability, and low efficiency in processing high-resolution images. In order to further enhance the applicability of DETR in vehicle detection scenarios, researchers have developed a large number of improved models, including Deformable DETR, DN-DETR, DINO, and models for traffic scenarios. These models have made progress in feature alignment ability, training stability, inference speed, and small object detection performance, making DETR an important candidate solution in fields such as traffic monitoring and lane scene analysis.

Therefore, this review aims to systematically review the research progress of DETR and its improved models in traffic vehicle detection, with a focus on analyzing their technical mechanisms, performance improvement strategies, and application scenarios. It discusses the challenges they face in the field of road traffic and looks forward to their future development trends.

2 Fundamental principles and Issues of DETR

2.1 Theoretical framework of DETR

DETR proposes an end-to-end object detection paradigm that achieves a structurally unified detection process by treating the detection task as a set prediction problem [1]. The model first uses Convolutional Neural Network (CNN) as the backbone network to extract image features and obtain a two-dimensional feature map with a certain semantic hierarchy. Since the Transformer structure was originally designed to process sequential data [2], DETR flattens feature maps into sequences and adds positional encoding to supplement spatial information, enabling it to understand the relative relationships between different regions in the image.

Then, the feature sequence is fed into a multi-layer Transformer encoder. The self attention mechanism of the encoder can model the dependency relationship between any position on a global scale. Compared to traditional methods, it can more effectively handle complex backgrounds and long-distance correlations. On this basis, DETR uses a fixed number of learnable object queries, interacts with encoded features through decoders, and actively retrieves features related to a potential target from the image [1]. Each vector output by the decoder corresponds to a candidate target, and parallel prediction is achieved by simultaneously generating category and bounding box parameters through the prediction head.

During the training process, DETR uses Hungarian matching to one-to-one match the predicted set with the true annotated set [1], thereby avoiding candidate boxes and non maximum suppression (NMS) in traditional methods. The matching results are further used to supervise classification errors and bounding box regression errors, allowing the model to directly learn object localization and semantic information within a globally consistent optimization framework. This end-to-end training strategy makes the DETR architecture more concise, yields more consistent predictions, and reduces the influence of manually designed priors.

Overall, DETR integrates object detection into a unified Transformer architecture through global attention mechanisms, object query design, and set-based matching strategies, promoting a shift from modular pipelines toward holistic modeling [1]. This framework lays the foundation for subsequent improvements in multi-scale feature fusion, training

acceleration, and small object performance enhancement. However, to be practically applied to traffic scenarios, its inherent early-stage limitations must be addressed, which will be discussed in the next section.

2.2 Issues of DETR in traffic vehicle detection

DETR has revolutionary significance in methodology, but in its application in the field of transportation, the original DETR has several obvious flaws. These defects have also become the fundamental motivation for the subsequent improvement of DETR models.

The original DETR training is very slow, mainly due to the difficulty in matching object queries and feature maps in the Transformer decoder. The initial query learns random vectors without prior knowledge of the real target, making it difficult for the network to effectively match the query with the target in the early stages of training. Currently, there are papers that can verify the above issues. For example, conditional DETR introduces conditional space queries to provide stronger localization priors, significantly accelerating convergence while maintaining end-to-end properties [3].

In traffic scenes, small targets such as distant vehicles, pedestrians, and motorcycles are very common. The original DETR used the last layer feature map of the backbone, resulting in lower spatial resolution, which limits the model's ability to recognize small objects in the image. Due to the fact that Transformer's global attention focuses on a large number of spatial positions, the attention is very scattered and difficult to focus on small objects. At the same time, DETR essentially relies only on single scale features, which reduces compatibility with targets of different sizes. In traffic scenarios, vehicle sizes vary greatly, such as vehicles in the same lane that are far apart. These issues have a significant impact on traffic detection [4].

In addition, the self attention mechanism of Transformer has a huge computational load in high-resolution scenes. The encoder and decoder of the original DETR require dense attention interaction for all tokens, resulting in high inference latency. In practical applications, it is usually not possible to provide sufficient computing power to achieve real-time obstacle avoidance or traffic monitoring functions. All the above issues provide motivation for continuous improvement in subsequent DETR variants.

3 Development of DETR in traffic scenarios

3.1 Multi-Scale Feature Enhancement Improvements

In traffic scenarios, the target scale span is extremely large, and the pixels of long-distance vehicles on highways are very small. However, in urban streets and intersections, there are both large vehicles at close range and small targets with severe occlusion, which brings researchers an improvement idea of multi-scale feature enhancement. The core idea of multi-scale feature enhancement is to improve the attention mechanism, introduce multi-level feature fusion and positional priors, so that the Transformer detector can maintain the ability to model global relationships while also considering high-resolution local details, thereby enhancing the detection performance of near and far vehicles, dense obstructions and other scenes. Below, the paper will explain the technical points and their roles in transportation scenarios one by one according to the representative work.

Firstly, there is Deformable DETR. Its primary contribution lies in using deformable attention instead of traditional dense attention. The core principle is to sample a small number of key points K around the reference point of each query and perform weighted summation, thereby reducing the attention computational complexity from O(N^2) to a linear relationship

with the number of sampled points, O(NK) [4]. Then, Deformable DETR natively supports multi-scale feature mapping, allowing the decoder to obtain complementary information between high-resolution details and low resolution semantics at different resolutions. For traffic images, this directly brings two advantages. One is the ability to accurately locate small vehicles at a distance using shallow high-resolution features; Secondly, for occlusion and dense traffic flow, local adaptive sampling can be used to focus attention on key pixels, avoiding dilution by global attention. The experiment and ablation fully demonstrate the advantages of deformable attention in convergence speed, small-object average precision (AP), and computational efficiency.

In order to accelerate training and improve stability, DN-DETR proposes injecting noisy ground truth boxes into the decoder during training and training the model to recover them, in order to reduce the severe oscillation of bipartite matching in the early stages of training and accelerate convergence [5]. On this basis, DETR with Improved deNoise Anchor Boxes (DINO) combines the strategies of mixed query initialization, comparative denoising, and "Look Forward Twice", and explicitly uses multi-scale features as decoder inputs to stably generate candidate boxes at different scales, taking into account both long-range and short-range object detection [6]. These improvements enable the model to learn to locate small objects faster, while balancing fine-grained and coarse-grained targets, which is an important improvement for detecting vehicles of multiple scales in traffic scenes.

For scenes with extremely mixed scales, severe occlusion and perspective distortion such as intersections and crossroads, FlowDet recently proposed two collaborative modules, Scale Aware Attention (SAA) and Geometric Deformable Unit (GDU), to specifically address traffic geometry and scale issues. SAA introduces scale aware weight allocation at the attention mechanism level, allowing cross scale features to be dynamically weighted in attention, enabling it to maintain high recognition ability when facing both long-distance small vehicles and close range large vehicles. GDU explicitly models perspective and geometric relationships such as road vanishing lines, traffic direction, and perspective scaling. By using deformable sampling guided by geometry, the attention sampling points are further corrected, improving the positioning accuracy and occlusion recovery ability from the perspective of intersections. On its newly established Intersection-Flow-5K dataset, FlowDet demonstrated a significant improvement in its ability to recognize intersection monitoring scenarios. Improving AP while maintaining or reducing computational costs has proven to be a feasible and effective improvement direction in engineering [7].

3.2 Lightweight and Real-Time improvements

Traffic detection in real scenarios has extremely high requirements for real-time performance and resource consumption. In order to make the DETR detector truly practical, researchers have proposed a series of lightweight and real-time improvements. The core goal of these improvements is to significantly reduce latency, decrease model parameters, and adapt to edge devices while maintaining Transformer detection performance. Below, several representative models will be introduced.

Real Time Detection Transformer (RT-DETR) was proposed by teams such as Baidu to improve the inference efficiency of Transformer object detectors in real-time scenarios. This model adopts a lightweight backbone and designs a hybrid encoder, which decouples the same scale interaction and cross scale fusion, while reducing computational redundancy and preserving multi-scale feature information. In addition, the decoder of RT-DETR supports adjustable layers, allowing it to add or remove decoder layers without retraining, in order to strike a balance between speed and accuracy. The paper shows that it achieves low latency inference on a general dataset, which provides feasibility for potential in vehicle or edge deployments [8].

Mamba RT-DETR (MRD) was proposed by Li&Huang et al. It has been optimized for lightweight and real-time vehicle detection tasks. MRD introduces a linearly complex State Space Model into the encoder to replace traditional self attention and reduce computational overhead. At the same time, its encoder combines multiple collaborative learning frameworks to optimize channel dimension processing and enhance spatial dependency modeling capabilities. According to publicly available experimental results, MRD achieved a high mAP on the BDD100K dataset, while reducing the parameter count by over 55% compared to some RT-DETR versions [9].

Overall, the design of RT-DETR and MRD demonstrates that these improvements have the potential to be deployed in real-time scenarios and resource constrained environments both theoretically and experimentally through structural optimization and lightweight modifications.

4 Discussion

Although significant progress has been made in traffic vehicle detection with many improvements of DETR, there are still many challenges in practical applications. In the traffic environment, as mentioned earlier, distant targets occupy small pixel areas, and scenes such as intersections can lead to problems such as scattered attention and missed targets. These issues have achieved significant improvements in multi-scale improvement models such as Deformable DETR and DINO. But there is still room for further optimization in feature fusion and attention design to achieve a better balance between accuracy and computational efficiency.

Secondly, real-time performance and computational limitations remain the main bottlenecks for edge deployment. In the above improvements, lightweight models such as RT-DETR and MRD have significantly improved inference speed compared to the original scheme, but in vehicle systems or large-scale traffic monitoring still need to further reduce latency and memory usage. This requires researchers to explore more efficient attention mechanisms, dynamic adjustments, and hardware friendly parallel strategies.

Finally, the lack of adaptability to diverse scenarios is also a problem that needs to be addressed. The existing models are mainly validated on general datasets or specific city road datasets. In the face of extreme weather conditions such as rain, fog, snow, sunrise, sunset, nighttime, etc., the detection accuracy may decrease. Therefore, improving the robustness and generalization ability of the model in different times and environments is also an important direction for future research.

In the future, the improvement of DETR in traffic vehicle detection may develop into combining geometric information and scene priors, improving the accuracy of near and far vehicle positioning by introducing mechanisms such as scale perception and perspective geometry, while achieving lightweight and adaptive reasoning.At present, research has begun to put it into practice, such as PETR, which encodes camera geometry into a Transformer based detection framework, indicating that geometric priors can enhance robustness in complex traffic environments [10]. On this basis, techniques such as sparse attention or state space modeling can be added to ensure that it balances speed and accuracy in edge environments. In addition, semi supervised and self supervised training methods are used to reduce reliance on large-scale annotated data and improve its generalization in various scenarios.

In summary, DETR and its variants have shown potential for engineering applications in traffic vehicle detection, but to achieve widespread deployment in vehicle and urban transportation, continuous exploration and improvement are needed in terms of efficiency, cross distance recognition, and cross scenario adaptability.

5 Conclusion

This article systematically reviews the development and application of DETR and its variants in traffic vehicle detection. Firstly, the basic framework and principles of DETR were introduced, and the limitations of the original DETR in terms of training speed, small target recognition ability, scale adaptability, and high-resolution computational overhead were analyzed. Subsequently, the focus was on two directions: multi-scale feature enhancement and lightweight, real-time improvement, including Deformable DETR、DN-DETR、DINO、RT-DETR、MRD Waiting for models and explaining their improvement principles, as well as the effectiveness enhancement in experimental verification.

By analyzing these improvement methods, multi-scale feature fusion and attention mechanism optimization have effectively improved the detection accuracy of small targets and dense scenes, while accelerating the convergence speed. Lightweight design and adjustable decoder layers have achieved feasibility in edge environments and real-time scenes. It can be seen that these improvements have the potential for engineering applications in traffic vehicle detection.

Looking ahead to the future, the improvement direction combining geometric priors and adaptive reasoning is expected to further enhance the robustness and scene adaptability of the model, providing more efficient and reliable object detection solutions for intelligent transportation systems and in vehicle perception.

References

1. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, End-to-end object detection with transformers, In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, pp. 213–229 (2020)
2. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, Attention is all you need, Advances in Neural Information Processing Systems, 30, 5998–6008 (2017)
3. D. Meng, X. Chen, Z. Fan, G. Zeng, H. Li, Y. Yuan, and L. Sun, Conditional DETR for fast training convergence, In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, Canada, pp. 3651–3660. (2021)
4. X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, Deformable DETR: Deformable transformers for end-to-end object detection, arXiv preprint arXiv:2010.04159 (2020)
5. F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, DN-DETR: Accelerate DETR training by introducing query denoising, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, USA, pp. 13619–13627 (2022)
6. H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, and H. Y. Shum, DINO: DETR with improved denoising anchor boxes for end-to-end object detection, arXiv preprint arXiv:2203.03605 (2022)
7. Z. Wang and Y. Zhao, FlowDet: Overcoming perspective and scale challenges in real-time end-to-end traffic detection, arXiv preprint arXiv:2508.19565 (2025)
8. Y. Lv, W. Chen, B. Chen, Z. Zhang, and J. Wang, RT-DETR: DETRs beat YOLOs on real-time object detection, In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Paris, France, pp. 2420–2430. (2023)
9. K. Li and X. Huang, MRD: A linear-complexity encoder for real-time vehicle detection,World Electric Vehicle Journal, 16(6), 307 (2025)

10. Y. Liu, T. Wang, X. Zhang, and J. Sun,PETR: Position embedding transformation for multi-view 3D object detection,In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, pp. 531–548. (2022)