Object Detection Based on the DETR Method

Yiru Wang

School of Statistics and Data Science, Capital University of Business and Economics, Beijing, China
Wangyr0221@outlook.com

Abstract. One of the computer vision research hotspots is object detection. Its aim is to accurately and quickly identify objects in images and locate their positions, converting visual information into understandable and actionable intelligence. With the success of the Transformer architecture in the field of natural language processing, the Transformer has also been gradually applied to object detection algorithms. DETR was proposed by Facebook as an end-to-end object detection framework. Although DETR shows great potential in the object detection task, it still has limitations such as slow training convergence, relatively weak performance in detecting small objects, and high computational complexity. This has prompted researchers to make improvements and refinements in subsequent works. This article aims to analyze and summarize the evolution stages of DETR, and divides the DETR method into four stages: the pioneering of the DETR method, the efficiency optimization of the DETR method, the improvement of the flexibility of the DETR method, and the breakthrough in the performance of the DETR method. At the same time, representative methods are introduced in each stage. Finally, the prospects of DETR in the object detection task are envisioned.

Keywords: Object Detection Task, DETR Algorithm, Visual Transformer, Computer Vision.

1 Introduction

One of the fundamental problems in computer vision is object detection. The goal of object detection serves to detect objects in an image and identify them, as well as to identify the item's category and precise location within the picture, and then complete different subsequent tasks. Nowadays, object recognition has been applied in various fields, such as security monitoring, autonomous driving, medical recognition, etc.

The Transformer model was first proposed by Ashish Vaswani et al. in their 2017 paper [1]. Its core innovation is the self-attention mechanism (Self-Attention) and adopts the encoder-decoder (Encoder-Decoder) structure. It replaces RNN and CNN in sequence modeling and completely relies on the self-attention mechanism to construct a sequence transformation model. DETR (Detection Transformer) is a landmark work in the field of object detection, proposed in the paper published by the Facebook team in 2020 [2]. This method regards object detection as a direct set prediction problem by

© The Author(s) 2026
K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_28

using bipartite graph matching (Hungarian algorithm), and uses the Transformer encoder-decoder architecture and a loss function based on bipartite graph matching. Without post-processing like NMS, anchor box design, etc., it achieves end-to-end training and inference, making the pipeline process more concise. In 2021, Google researchers published a paper proposing the ViT model [3]. The ViT model divides the input image into fixed-sized patches (such as 16x16 pixels), and then maps these patches through linear projection to a low-dimensional vector space to generate sequence input. These sequences are pre-added with a learnable token, and the final output state will go through an MLP classification head for image classification. ViT breaks the long-standing dominance of CNN in the territory of computer vision and lays the Basics for many subsequent Transformer-based visual models.

With the emergence of the Transformer architecture, more and more object detection algorithms based on DETR have been proposed and applied in different fields. This paper will summarize and analyze the evolution stages of the DETR algorithm, dividing it into four different stages, and presenting representative algorithms in each stage. Finally, it will propose the future outlook of the DETR object detection algorithm.

2 Initiating and Revealing Problems (2020)

In 2020, the paper [2] published by the Facebook AI team proposed a brand-new object detection framework called DETR. DETR views object detection as a direct set prediction problem. It draws on the end-to-end philosophy in complex structured prediction tasks, adopts a Transformer encoder-decoder architecture and a global loss function based on bipartite graph matching, simplifying the detection process and eliminating problems such as anchor box design and NMS post-processing in traditional object detection methods.

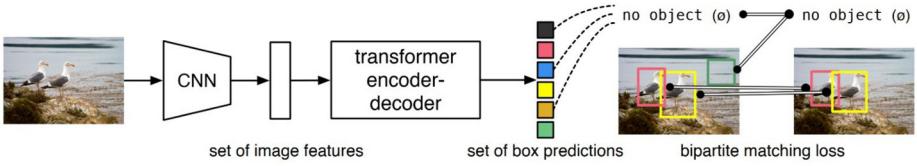

Fig. 1. The model framework of DETR [2]

Figure 1 depicts the model framework of DETR. The image is transmitted initially via a backbone network, which is a pre-trained CNN (usually ResNet-50 or ResNet-101), and the final pooling layer and classification layer are removed to obtain a grid-like feature map (The semantic information of different regions of the image can be captured by the feature map), which is the input image. Then, the grid-like feature map is converted into a sequence of feature map units and passed into the Transformer encoder (shown in the figure 1), where the encoder uses the self-attention mechanism to model the global relationships. After the information passes through all layers of the Transformer encoder, the representation of each grid unit is obtained. These

representations contain context information due to the use of the self-attention mechanism by the encoder. Then, it reaches the DETR decoder (shown in the figure 1), where the input is a series of object queries and the decoder is limited to generating N bounding boxes, resulting in N predicted boxes. Finally, through bipartite graph matching, it is determined which predicted boxes contain objects and which do not. After the matching is completed, the loss for the matching successful predictions is calculated, including classification loss and bounding box loss.

In terms of advantages, when comparing in terms of performance, DETR outperforms Faster R-CNN by achieving comparable performance to the finely-tuned Faster R-CNN baseline (measured by the AP metric - Average Precision). Thanks to the global attention mechanism of the Transformer, DETR performs exceptionally well in the detection of large objects.

Nevertheless, DETR still exhibited shortcomings in several aspects, such as poor performance in small object detection, long training periods, and high computational complexity of the encoder. Based on these issues, it provided guidance for researchers to gradually solve and improve the problems in their subsequent studies.

3 Convergence and Efficiency Optimization (End of 2020 - 2022)

3.1 Deformable DETR

In 2021, the paper published by Zhu Xizhou et al. Addressed the issues of sluggish convergence speed and subpar detection of small targets in the initial DETR model, and proposed the Deformable DETR model [4]. By introducing the deformable attention mechanism, the efficiency and performance of target detection were improved.

Deformable Convolution Networks (DCN) incorporate the ability to learn spatial geometric deformation in convolution, enabling it to handle more complex spatial geometric deformation tasks. Deformable DETR was inspired by deformable convolution and combined DCN with DETR to propose the deformable attention mechanism. The deformable attention mechanism concentrates solely on a small segment of key sampling points around the reference point (usually 4), significantly reducing the computational complexity through sparse sampling, allowing the model to dynamically focus its attention on sparse spatial positions that are more likely to contain information, thereby greatly improving the convergence speed of the model and solving the problem that the original DETR's convergence speed is much slower than traditional detectors such as Faster R-CNN.

Deformable DETR also proposed multi-scale deformable attention. Deformable DETR extended the deformable attention to multi-scale feature maps, forming a multi-scale deformable attention module. The module has the ability to extract information from multiple different scale feature layers extracted by the backbone network simultaneously, making it easier for the model to detect objects of different sizes, thereby addressing the shortcomings of the original DETR in small target detection.

Compared with the original DETR, Deformable DETR not only achieved remarkable improvements in convergence speed and performance, but also made progress in the average precision (AP) of small target detection.

3.2 Efficient DETR

In 2021, Yao et al. made improvements to Deformable DETR and proposed the Efficient DETR model to further enhance the training convergence speed and address the issues of the original DETR and Deformable DETR requiring complex multi-layer decoders [5].

Although Deformable DETR solved the convergence speed and small target detection problems of the original DETR, it still relied on a 6-layer encoder and a 6-layer decoder. According to experimental data, the 6-layer decoder still has its convergence speed. The authors of this paper proposed the Efficient DETR method, which made the single-layer decoder more high-quality, reducing the number of decoders and thereby reducing the computational load and improving the convergence speed.

Efficient DETR ingeniously combines "dense detection" and "sparse set detection", with both parts sharing the same detection head. In the dense part, the image is first input, and through the Backbone and Encoder (similar to the multi-scale feature backbone of Deformable DETR), multi-scale feature maps are obtained, which are passed to 3 deformable encoders. The encoded feature maps are then passed to the dense part, where dense predictions are generated by the dense prediction head, and the Top-K proposals are selected to initialize the object container. In the sparse part, the initialized object container is passed into the single-layer decoder, where the decoder interacts with the image features through the deformable attention mechanism, further refining the object container, and finally sent to the shared detection head to complete the detection.

This method requires only 3 encoder layers and 1 decoder layer, and the training period is significantly lower than that of the original DETR and Deformable DETR. On the MS COCO dataset, it was found that by training for only 36 epochs, Efficient DETR (3 encoders + 1 decoder) achieved an excellent performance of 44.2 AP, comparable to other state-of-the-art detection methods at that time, and with fewer parameters. In crowded scenes, its performance was even better than that of the original DETR and Deformable DETR.

3.3 Conditional DETR

In 2021, Meng et al. [6] discovered that DETR was highly dependent on high-quality content embeddings to locate the key points (extremities) of object boundaries. This part precisely corresponds to the key areas for object localization and recognition, resulting in a slow convergence speed of DETR. Therefore, Conditional DETR was proposed to solve the problem of slow convergence of DETR.

The core idea of this method is to decouple content and spatial queries and to decouple the function of cross-attention, proposing a conditional cross-attention mechanism. The content embedding is only responsible for searching for regions related to the object based on appearance features, while the spatial embedding explicitly locates the extremity area of the object through conditional spatial query,

narrowing the search range. This reduces the dependence of DETR on high-quality content embeddings and simplifies the training difficulty.

Based on the experimental data, Conditional DETR has increased the convergence speed of the backbone network Res-Net50 [7] and ResNet101 by 6 times and 7 times respectively. Under stronger backbone networks such as DC5-R50 and DC5-R101, Conditional DETR can reach the performance level of the original DETR after only 50 training cycles.

4 Enhancement of Flexibility (2022 - 2024)

4.1 DN DETR

In 2022, Li et al. discovered that the training method of DETR, which matches the predicted objects with the real annotations one by one through bipartite graph matching (Hungarian algorithm), was very unstable due to its random optimization characteristics [8]. That is, the queries in the same image might match different targets in different training cycles, which led to frequent changes in the optimization objective and a slow convergence speed. Based on this problem, this article proposed the DN DETR algorithm.

Due to the instability of bipartite graph matching, where the targets matched in different cycles are different, it makes model optimization difficult and significantly reduces the training efficiency. This paper introduces the denoising training technique details to solve this problem. This method adds noisy real bounding boxes to the input of the Transformer decoder and obtains the model's prediction of the original real boxes, thereby bypassing the unstable bipartite graph matching and avoiding the problem of different matching targets. This not only improves the convergence speed but also significantly enhances the model performance.

DN-DETR not only improves the training speed but also enhances the final detection accuracy. According to the experimental data, with ResNet-50 as the backbone network, its average accuracy reached 43.4 after 12 training times and 48.6 after 50 training times.

4.2 Group DETR

In 2023, Qiang Chen et al. discovered that the slow convergence of the original DETR was partly due to its one-to-one label assignment strategy (such as the Hungarian algorithm), where each real object was assigned to only one predicted query [9]. During training, the number of positive samples was small, resulting in sparse supervision signals, which led to slow convergence. Based on the hint of the traditional detector's one-to-many label assignment, this method, Group DETR, proposes a new strategy: "multi-group one-to-one assignment".

Group DETR decomposes "one-to-many label assignment" into "multi-group one-to-one assignment". Each group of queries is independently and one-to-one matched, and then input into the decoder for the Self-attention operation (parameter sharing), followed by Cross-attention and FFN. Compared with DETR, the number of query groups input into the decoder is increased, and then processed together in the decoder.

The advantage of this is that it retains the end-to-end characteristics of DETR while increasing the training speed, making it more efficient, more stable, avoiding the problem that traditional detectors heavily rely on NMS with a large amount of computation, and having more supervision to make bipartite graph matching more stable.

This paper conducted a large number of experiments on the COCO dataset, proving the effectiveness and generalization of Group DETR, performance improvement and convergence speed increase, and that Group DETR also has significant performance improvements on different DETR variants.

4.3 RT-DETR

In 2023, the method proposed by the Baidu team, RT-DETR [10], was the first to implement an end-to-end object detector in the Transformer architecture, surpassing the YOLO series in terms of accuracy and speed, and solving the problems of convergence speed, training cost, inference speed, and inability to perform real-time detection faced by Deformable DETR and DINO.

RT-DETR designed an efficient hybrid encoder, fully leveraging the global modeling capabilities of Transformer and the local feature extraction efficiency of CNN, achieving a balance between computational efficiency and feature representation capability.

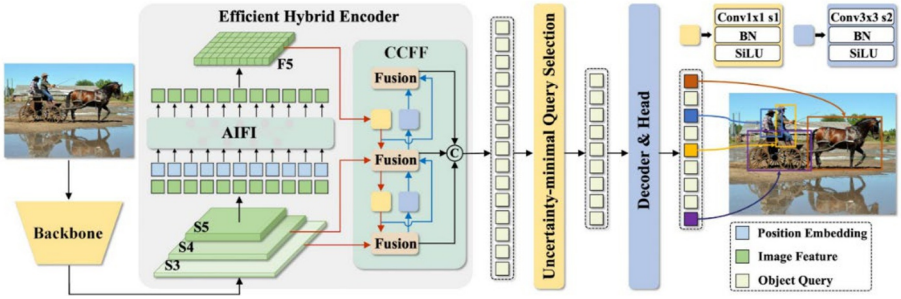

Fig. 2. The model framework of RT-DETR [10]

The images pass through the backbone and RT-DETR to reach the last three stages (as shown in S3, S4, S5 in Figure 2) and enter the efficient hybrid encoder. Through AIFI (the Attention-based Intra-scale Feature Interaction), the output is obtained and passed into CCFF (the CNN-based Cross-scale Feature Fusion) To fuse multi-scale features, thus enhancing the model's ability to detect objects of different scales.

Based on the dataset provided in this paper, RT-DETR-R50 achieved an AP of 53.1% and 108 FPS, while RT-DETR-R101 achieved an AP of 54.3% and 74 FPS. At that time, RT-DETR outperformed existing detectors in both speed and accuracy.

5 Performance Breakthrough

5.1 D-FINE

In 2024, the method D-FINE proposed by the team from the University of Science and Technology of China is a powerful real-time object detector, demonstrating advantages such as high accuracy, real-time performance, and high generalization [11]. It innovatively revolutionizes the regression approach in object detection from the task definition perspective.

D-FINE redefines the regression task of "bbox" and achieves more outstanding positioning accuracy. FDR: Fine-grained Distribution Refinement. The advantage of FDR lies in converting the previously fixed coordinates of the prediction in the regression process into iterative optimization of the probability distribution, which can better obtain fine-grained intermediate representations and significantly improve positioning accuracy. GO-LSD is a bidirectional optimization strategy that can transfer positioning knowledge from the fine distribution to the shallower layers through self-distillation and simplifies the residual prediction task at the deeper layers. At the same time, lightweight optimizations have been made in computationally intensive modules and operations, achieving a balance and improvement in speed and accuracy.

Based on the dataset provided in this article, it can be concluded that the average precision (AP) of D-FINE-L and D-FINE-X on the Pretrained on Objects365 dataset has achieved 57.1% and 59.3% respectively, which are higher than most detectors.

5.2 DEIM

In 2025, the method DEIM proposed by Shihua Huang et al. was presented at the beginning of this article with the following data: DEIM can achieve faster convergence speed, higher average precision (AP), and lower latency. It solves the issues of slow convergence speed, sparse supervision, and low-quality matching of previous methods [12].

Two reasons are behind the slow convergence speed of DETR: sparse supervision and low-quality matching. Reason 1: Sparse supervision: The O2O mechanism's limitation of positive samples causes the scarcity of positive samples to hinder effective model learning. To address sparse supervision, this method proposes Dense O2O, which, compared to methods like Group-DETR, does not require an additional decoder and has a better supervision level. Its method uses classic image enhancement techniques: Mosaic and Mixup, increasing the number of real boxes, thereby increasing the overall number of positive samples. Since Dense O2O causes the problem of low-quality matching, to solve this issue, the authors of this paper use the MAL (Match Ability-Aware Loss) function. MAL increases the focus on low-quality matching and improves the utilization efficiency of limited positive samples.

According to the experimental results of this article, it can be clearly seen that in a small number of epochs, its average precision (AP) is already on par with previous methods in high epochs, demonstrating its characteristic of fast convergence, and compared to D-FINE, DEIM improves the effect of small target detection.

6 Conclusions

Since its inception, the DETR model has gradually overcome issues such as slow convergence, insufficient detection of small targets, and high computational requirements through the efforts of researchers, The development of numerous important models, such as Deformable DEER and DN DETR, was a result, and DEIM. These models have continuously improved in terms of speed and accuracy. However, they still face challenges related to data dependency and annotation costs, as well as how to avoid low-quality matches and error accumulation. It is believed that the model will eventually progress towards more efficient architectures and attention mechanisms, stronger multimodal and open-world perception, self-supervised and weakly supervised learning. This paper hope is that these problems can be solved with the ongoing advancements in deep learning, AI, and computing power.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: arXiv preprint arXiv:1706.03762, pp. 5998–6008 (2017)
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Lecture Notes in Computer Science, pp. 213–229. Springer, Cham (2020)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
4. He, L., Zhou, Q., Li, X., Niu, L., Cheng, G., Li, X., Liu, W., Tong, Y., Ma, L., Zhang, L.: End-to-end video object detection with spatial-temporal transformers. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 1–15. ACM, New York (2021)
5. Tan, M., Pang, R., Le, Q.V.: EfficientDET: Scalable and efficient object detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1079–1087 (2020)
6. Meng, D., Chen, X., Fan, Z., Zeng, G., Li, H., Yuan, Y., Sun, L., Wang, J.: Conditional DETR for fast training convergence. In: IEEE/CVF International Conference on Computer Vision, pp. 363–372 (2021)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
8. Li, F., Zhang, H., Liu, S., Guo, J., Ni, L.M., Zhang, L.: DN-DETR: Accelerate DETR training by introducing query denoising. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1325–1334 (2022)
9. Chen, Q., Chen, X., Wang, J., Zhang, S., Yao, K., Feng, H., Han, J., Ding, E., Zeng, G., Wang, J.: Group DETR: Fast DETR training with group-wise one-to-many assignment. In: IEEE/CVF International Conference on Computer Vision, pp. 6610–6619 (2023)
10. Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Chen, J., et al.: DETRs beat YOLOs on real-time object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16965–16974 (2024)
11. Peng, Y., Li, H., Wu, P., Zhang, Y., Sun, X., Wu, F.: D-FINE: Redefine regression task in DETRs as fine-grained distribution refinement. In: arXiv preprint arXiv:2410.13842 (2024)

12. Huang, S., Lu, Z., Cun, X., Yu, Y., Zhou, X., Shen, X.: DEIM: DETR with improved matching for fast convergence. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 15162–15171 (2025)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.