Research of Improved DETR Models and Transformer Applications in Computer Vision

Ruoyu Li

Petroleum Institute, China University of Petroleum, Beijing, 834000, China
2022015511@st.cupk.edu.cn

Abstract. Researchers in the domain of computer vision have increasingly turned their attention towards harnessing the power of Transformer models for visual tasks. This paradigm shift has led to the emergence of pioneering models such as Detection Transformer (DETR) and Vision Transformer (ViT), opening up new frontiers for advancement in computer vision research. In this article, the significance of this transition and its implications for target detection are explored. Specifically, light is shed on the inherent limitations of the DETR model in effectively identifying targets in visual data, paving the way for a comprehensive discussion on strategies for enhancing its performance. Through an exploration of various DETR models and their innovative approaches, readers are provided with a nuanced understanding of the challenges and opportunities in target detection within the context of Transformer-based methodologies. By elucidating the guiding principles driving the evolution of DETR models, valuable insights into the future trajectory of computer vision research and the transformative potential of Transformer technology in visual perception tasks are offered.

Keywords: DETR Models, Transformer, Computer Vision.

1 Introduction

The field of computer vision first gained traction in the 1970s and 1980s. In the field of artificial intelligence, it has gained popularity after 40 to 50 years of technological development. One of the most crucial applications of computer vision technology is target detection. Its goal is to find and identify helpful targets in images or movies. To complete this challenge, one must locate the target image, determine its boundaries, and categorize it.

At present, there are two types of classic algorithm structures for target detection: the two-stage method represented by Faster Region-based Convolutional Neural Networks (Faster R-CNN) [1]. The first stage is mainly to extract the frame selection suggestion area, and the second stage is to extract the first stage. Classification and regression, this method has high accuracy but has the disadvantage of slow speed; the one-stage method represented by You Only Look Once (YOLO) and Single Shot MultiBox Detector (SSD) only performs regression classification of the frame selection

Y. Wang (ed.), *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, Advances in Computer Science Research 115,
https://doi.org/10.2991/978-94-6463-540-9_98

suggestions [2, 3]. This method is fast but has the disadvantage of accuracy and Bad shortcomings.

The Transformers model, which the Google team proposed in 2017 and which was based on the self-attention mechanism design, demonstrated clear benefits and made significant advancements in the field of natural language processing [4]. The capacity to employ a vast amount of training data such that the model can learn target properties from the original data is known as the self-attention mechanism. It enables the model to focus more on the location and characteristics of the target item, which effectively raises the model's efficiency and accuracy.

A large number of researchers in computer vision have also started to look into ways to use the Transformer model in computer vision technology. Transformer's attention mechanism was first used in 2020 to solve the target identification task using the Detection Transformer (DETR) model. Since then, several enhanced models based on DETR have emerged one after the other. These include Deformable DETR, which enhances the attention mechanism; Efficient DETR, which enhances the target query initialization; Sparse Multi-Head Cross-Attention Transformer-based Detection Transformer (SMCA DETR), which enhances the encoder's feature interaction; and Set Abstraction Model Detection Transformer (SAM DETR), which enhances the decoder's features. improvements in label matching techniques by Grouped Detection Transformer (Group DETR), improvements in interaction, and improvements in unsupervised training by Upsample Detection Transformer (UP DETR).

This article aims to elaborate on the improvement measures of attention mechanism in target detection tasks and analyze and summarize related models.

2 Principles of Transformer and Visual Transformer

2.1 Transformer

The self-attention mechanism, which uses the multi-head attention mechanism to extract information from the input sequence, is the fundamental building block of the Transformer model [4]. The model structure of the Transformer is shown in Fig. 1. The model's input is processed through multiple self-attention layers and feedforward neural network layers, and residual connections and layer normalization are included in each layer to speed up training and improve model accuracy. In each self-attention layer, each position in the input sequence undergoes attention computation with other positions to capture dependencies between sequences. At the same time, Transformer also introduces positional encoding to process the order information of the input sequence so that the model can better understand the semantic relationships in the sequence. By stacking multiple self-attention layers, the Transformer model can effectively learn long-term dependencies, thereby achieving better performance in sequence-to-sequence tasks.

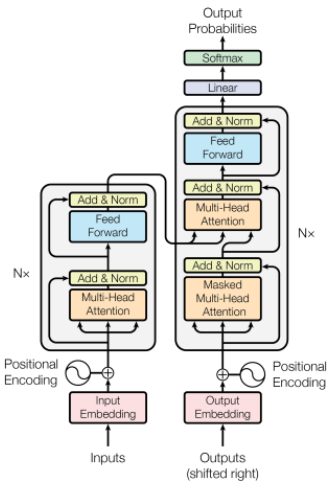

Fig. 1. Transformer model framework

2.2 ViT

A Vision Transformer (ViT) is a visual model based on a Transformer [5]. The model structure of ViT and the encoder of its Transformer part are shown in Fig. 2. The principle of the ViT model is as follows: ViT divides the input image into a series of image blocks, and then maps each image block into a vector through a linear transformation. These vectors are called "image block embeddings". To preserve the position information of the image, ViT introduces position encoding. After inputting the position encoding and vector into the Transformer-encoder, the Transformer-encoder can learn the relationship between image blocks through the self-attention mechanism, and extract and represent features of each image block through the feedforward neural network. ViT adds a fully connected layer after the Transformer encoder to map the output of the last Transformer encoder into a class probability distribution. In this way, the image classification task can be achieved through training.

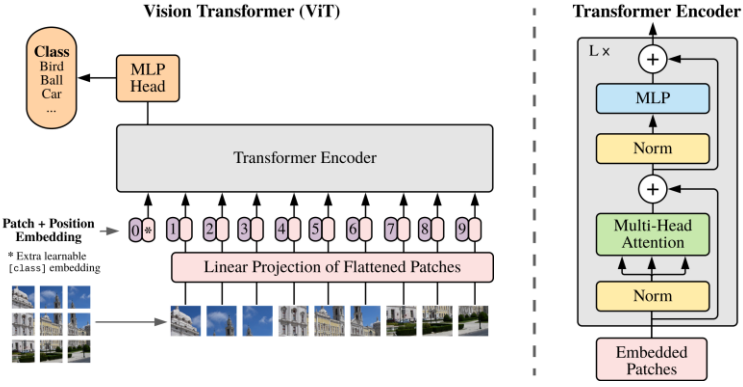

Fig. 2. The model structure of ViT and the encoder of its Transformer part

2.3 DETR

The structure of the DETR model is composed of three parts: backbone, encoder and decoder of special Transformers, and Feedforward Neural Network(FNN) [6]. The model structure of DETR is shown in Fig. 3. The function of the backbone part is to use CNN to extract local features of the data, and then convert the local features into a one-dimensional sequence as the input of Transformer-encode. The function and process of the second part of the encoder are the same as those of the Transformer's encoder, but what is different from the Transformer's decoder is that it does not use masked multi-head attention, because the image processing uses a non-autoregressive model, and requires parallel output. The role of FNN is to classify and predict the output of the Transformer structure.

DETR uses the Hungarian algorithm to select an effective prediction frame algorithm as follows:

$$\hat{\sigma} = \arg\min \sum_{k}^{N} \mathrm{L}_m \left(y_k, \hat{y}_{\sigma_k} \right) \tag{1}$$

Matching takes into account both category prediction and the similarity between the predicted box and the real box. Subsequently, the Hungarian loss matched in the previous step is calculated and the gradient is passed back. Because DETR has the disadvantages of slow training convergence speed, redundant target query, and poor small target detection results, subsequent algorithms are also developed around these aspects.

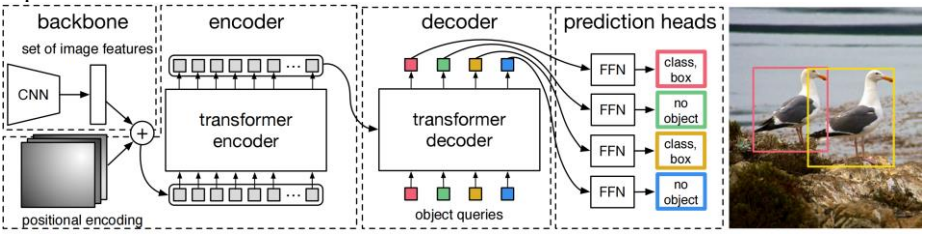

Fig. 3. The model structure of DETR

3 Improved Algorithm of DETR

3.1 Deformable DETR

Deformable DETR draws on the idea of Deformable Convolutional Networks (DCN) and applies it to the attention mechanism [7]. The multi-scale deformable attention module of DCN (as shown below) replaces the ordinary multi-head attention module in DETR. Only attention is sampled by the attention module. A tiny group of crucial sampling locations within a range. The Deformable Attention Module of the Deformable DETR Model is shown in Fig. 4.

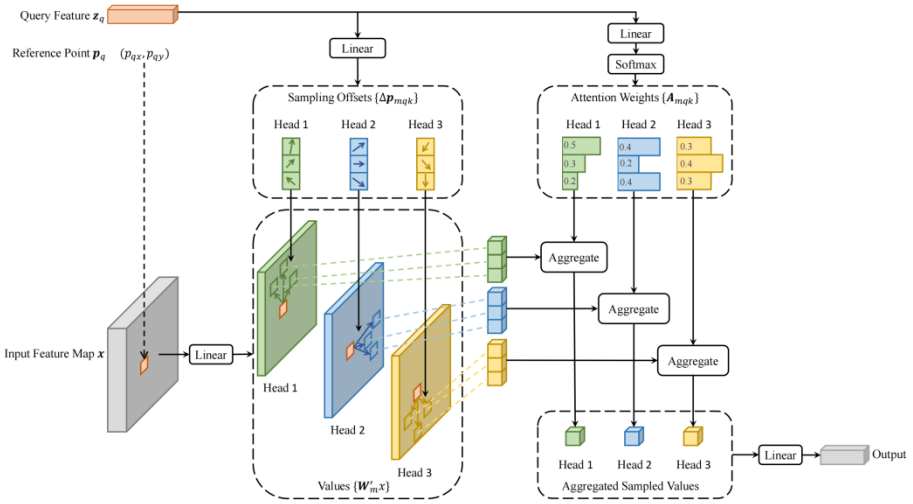

Fig. 4. Deformable attention module of the Deformable DETR model

3.2 Dynamic DETR

To speed up training convergence, dynamic DETR approximates the attention mechanism of the DETR encoder using convolution-based dynamic encoders with different attention kinds in the encoder link [8]. In the decoder link, the dynamic attention mechanism based on RoI pooling operation is used to replace the cross-attention module, so that DETR focuses on Region of Interest(ROI) from coarse to fine, effectively reducing the learning difficulty and accelerating the convergence speed of training. The architecture of the Dynamic DETR Model is shown in Fig. 5.

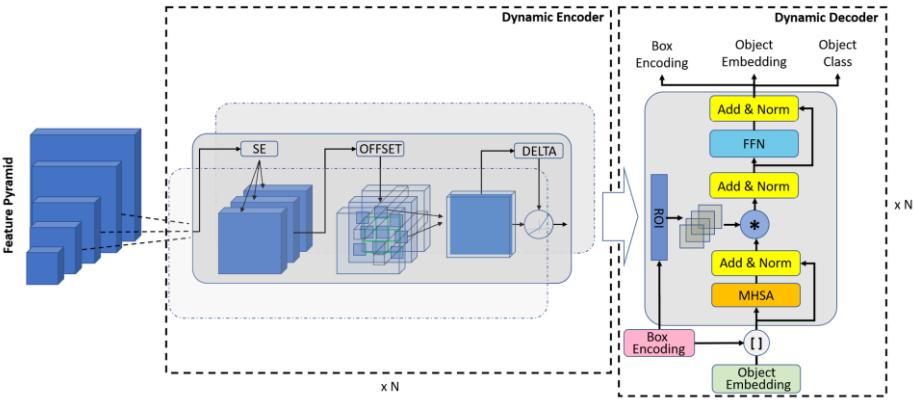

Fig. 5. Architecture of the Dynamic DETR Model

3.3 Efficient DETR

After improving the original DETR's six encoder levels and six decoder layers—which included three encoder layers and just one decoder layer—and removing the cascade inside the decoder, an efficient DETR is created. Organization [9]. It was discovered that the decoder is more crucial than the encoder after doing several tests and examining DETR. This is mainly because the cascade of decoders shows hierarchical auxiliary loss characteristics, and when updating query characteristics, the auxiliary decoding loss adds a strict monitoring mechanism, thereby improving the decoder's working efficiency. As the number of cascades gradually decreases, its performance has declined significantly.

The two primary components of an efficient DETR are the sparse and dense components, which share data on the same detection head. Within a dense region, the system predicts dense features from the encoder, and creates a candidate box, and then uses a top-k selection strategy to filter out a series of solutions from the dense prediction set. The 4-D suggestions provided by the decoder along with its 256-D properties are adopted as reference points and initial settings for object querying. In areas where data is sparse, object containers, reference points, and object queries are initialized with dense priors and fed into the first-layer decoder, interacting with the features of the encoder to achieve deeper detail processing. The final prediction results are based on more granular object containers. Whether it is a dense part or a sparse part, target matching is achieved through one-to-one allocation rules. In all encoder and decoder layers, the deformable attention module described in Deformerable-DETR is used. The workflow of the Efficient DETR Model is shown in Fig. 6.

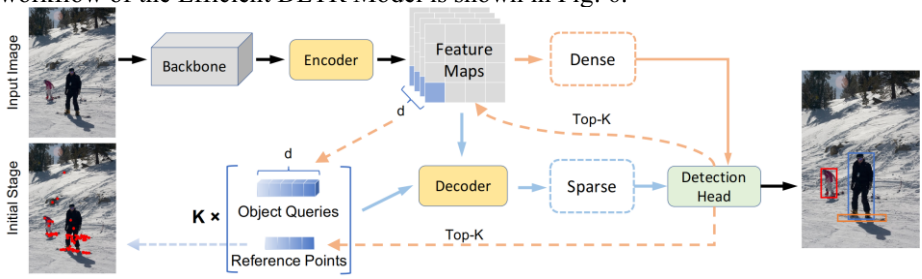

Fig. 6. Workflow of the efficient DETR model

By designing two parts, the dense part, and the sparse part, Efficient DETR can gradually optimize the object query initialization through multiple iterations of training to handle the target prediction task more efficiently.

3.4 SMCA DETR

Several query vectors in DETR are in charge of locating targets at various locations [10]. To adaptively choose the key characteristics from this geographical information and estimate the position and category, each object query interacts with the spatial location features that CNN has acquired. Co-attention is employed throughout this

process. Nevertheless, it is possible that the co-attended zone in the decoder for each object query does not match the box that the query anticipated. Therefore, larger training rounds are required for DETR's decoder to correctly identify the matched target by co-attending areas.

To achieve quick convergence and enhance the effect, the co-attention module in DETR is replaced by the Spatially Modulated Co-attention (SMCA) module in SMCA.

The architecture of the SMCA DETR Model is shown in Fig. 7.

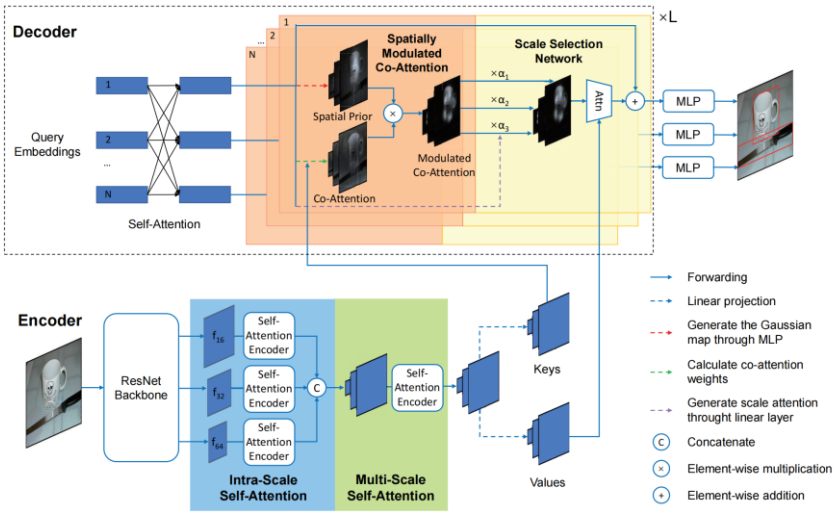

Fig. 7. Architecture of the SMCA DETR model

The Gaussian distribution model is incorporated into Spatially Modulated Co-attention to enhance the speed of the model's convergence by modifying the attention mechanism's search range to a specific region close to the object's center. For each given object query vector, SMCA first predicts the center position and scale of the object;

$$\begin{aligned} c_h^{\text{norm}}, c_w^{\text{norm}} &= \text{sigmoid}\left(\text{MLP}\left(O_q\right)\right) \\ s_h, s_w &= \text{FC}\left(O_q\right) \end{aligned} \tag{2}$$

And create the object's two-dimensional Gaussian distribution using the anticipated value, which is then utilized to modify the Gaussian distribution's bandwidth;

$$G(i,j) = \exp\left(-\frac{\left(i-c_w\right)^2}{\beta s_w^2} - \frac{\left(j-c_h\right)^2}{\beta s_h^2}\right) \tag{3}$$

Finally, the two-dimensional Gaussian distribution of the object is combined with the attention matrix in the joint attention model to obtain the spatially modulated multi-head attention matrix as follows:

$$C_i = \text{softmax}\left(K_i^T Q_i / \sqrt{d} + \log G_i\right) V \tag{4}$$

3.5 SAM DETR

SAM DETR's fundamental concept revolves around leveraging the twin network's superior efficiency in diverse matching activities, simplifying the cross-attention object query process to concentrate on particular domains [11]. Parties involved in the matching process will assess similarity using identical semantic frameworks, thus simplifying the matching process and enhancing its precision. Before the Transformer's decoder layer's cross-attention, SAM DETR adds a Semantics Aligner. Directly perform convolution + MLP operation on the regional features obtained by RoIAlign to predict salient points, and use resampling to achieve semantic alignment matching. To preserve semantic alignment between the output object query and image features, feature reweighting is carried out when a new object query is constructed using the reweighting parameters produced by the preceding object query. The Semantic Matching Process of the SAM DETR Model is shown in Fig. 8.

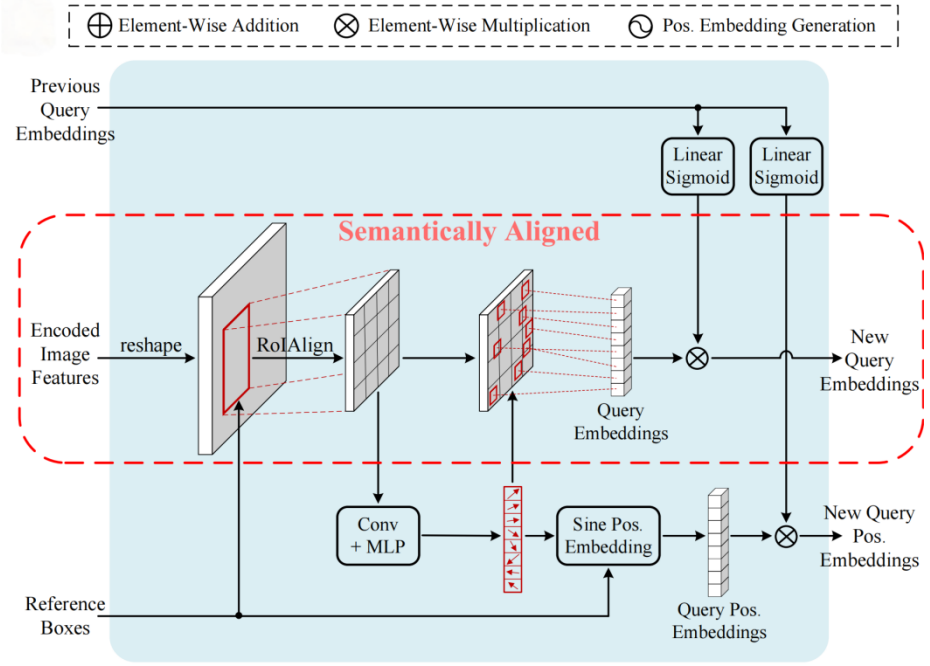

Fig. 8. Semantic matching process of the SAM DETR model

By introducing a semantic alignment module and salient point sampling, the matching problem is solved and the convergence speed of DETR is accelerated. At the same time, the "plug and play" design allows SAM DETR to be easily combined with existing DETR convergence solutions and achieve better results.

3.6 UP DETR

Because the backbone link of DETR has been trained, and the transformer module used as a detection module has not been pre-trained, this part mainly focuses on the positioning of spatial positions and the suppression between output boxes, so the current unsupervised training learning. The method is not suitable for pre-training of transformers in DETR.

A novel pre-training technique dubbed UP DETR suggests a novel agent task for target detection known as random query patch detection [12]. The Random Queries of the UP DETR Model are shown in Fig. 9. This job randomly selects many query patches from the image to pre-train the Transformer on detection to predict the bounding boxes of these query patches in the given picture. The graphic depicts the UP-DETR training procedure. The allocation issue between query patch and object query is resolved by creating the Shuffled object queries and attention masking, which also resolves the multi-query positioning issue.

Under unsupervised pre-training, UP-DETR has a faster convergence speed and higher accuracy than DETR.

During the training process, two problems are solved: multi-task learning and multi-query positioning. By freezing the reconstruction of the pre-trained backbone network and patch features to maintain the feature recognition of the Transformer, multi-task learning is completed.

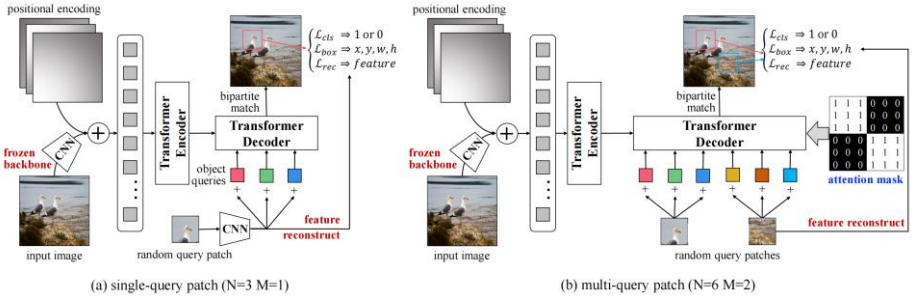

Fig. 9. Random queries of the UP DETR model

3.7 Group DETR

In the DETR series of algorithms, label allocation consists of one-to-one label allocation and one-to-many label allocation. Through multiple experiments, it was found that with one-to-one tags, regardless of whether NMS is used or not, increasing the number of object queries alone has little effect on improving model performance, while one-to-many tags significantly improve model performance when using NMS. Group DETR proposes a new label-matching strategy for Group-wise One-to-Many assignments [13]. The Group Matching Process of the Group DETR Model is shown in Fig. 10.

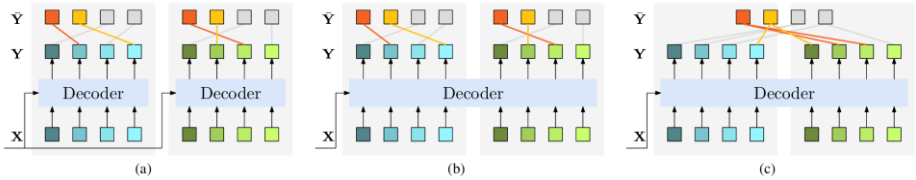

Fig. 10. Group matching process of the group DETR model

Group DETR divides all object queries into K groups, uses the self-attention of the same encoder to calculate the object queries of each group, performs one-to-one label matching in each group, and has a benchmark object in each group Matches an object query such that in K groups, a benchmark object is predicted by K query objects.

3.8 Summary of DETR Improved Model

Deformable DETR uses a deformable attention network to make training converge faster, but the number of encoder tokens is 20 times greater than DETR; Dynamic DETR's dynamic attention provides lower resolution and speeds up training convergence. , but relies on the CNN network as a convolution-based encoder and a Rol-based decoder; Efficient DETR reduces the number of decoding layers and encoding layers, but increases the number of GFLOPs; SMCA DETR uses a regression-aware mechanism to improve the convergence speed, SAM DETR Semantic alignment is used to speed up the matching process, but the performance of both in detecting small objects is reduced; UP DETR uses pre-training of multi-task learning and multi-query localization to improve the convergence speed and accuracy, but pre-training for patch positioning requires integration Pre-training of convolutional neural network and Transformer; Group DETR adopts "multi-group one-to-one Allocation" label allocation improves training speed, but requires higher space.

4 Conclusion

Transformer has demonstrated remarkable success in natural language processing, revolutionizing the field and introducing novel concepts to target detection. This article elucidates the DETR model, along with other contemporary models, which are actively employed to mitigate challenges such as subpar identification of small targets, redundancy in target queries, and sluggish training convergence rates. Through the comprehensive review, readers are provided with a thorough understanding of the research conducted in this domain, encompassing both the methodologies employed and the outcomes achieved. It is believed that this examination will afford readers a clear and insightful perspective on the advancements made in target detection using Transformer-based models.

References

1. Ren, S., He, K., Girshick, R. B., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: IEEE Transactions on Pattern Analysis and Machine Intelligence 39, 1137-1149 (2015)
2. Joseph, R., Santosh, D., Once, F. A. Y. O. L.: Unified, real-time object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
3. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., Berg, A. C.: Ssd: Single shot multibox detector. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14 pp. 21-37. Springer International Publishing (2016)
4. Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I. Attention is all you need. Neural Information Processing Systems (2017)
5. 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. ArXiv, abs/2010.11929 (2020)
6. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-End object detection with transformers. ArXiv, abs/2005.12872 (2020)
7. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: Deformable transformers for end-to-end object detection. ArXiv, abs/2010.04159 (2020)
8. Dai, X., Chen, Y., Yang, J., Zhang, P., Yuan, L., Zhang, L.: Dynamic detr: End-to-end object detection with dynamic attention. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 2988-2997 (2021)
9. Yao, Z., Ai, J., Li, B., Zhang, C.: Efficient DETR: Improving end-to-end object detector with dense prior. ArXiv, abs/2104.01318 (2021)
10. Gao, P., Zheng, M., Wang, X., Dai, J., Li, H.: Fast convergence of detr with spatially modulated co-attention. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 3601-3610 (2021)
11. Zhang, G., Luo, Z., Yu, Y., Cui, K., Lu, S.: Accelerating DETR Convergence via Semantic-Aligned Matching. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 939-948 (2022)
12. Dai, Z., Cai, B., Lin, Y., Chen, J.: Unsupervised Pre-Training for Detection Transformers. In: IEEE Transactions on Pattern Analysis and Machine Intelligence 45, 12772-12782 (2020)
13. Chen, Q., Chen, X., Wang, J., Feng, H., Han, J., Ding, E., Zeng, G., Wang, J.: Group DETR: Fast DETR Training with Group-Wise One-to-Many Assignment. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 6610-6619 (2022)