Exploring The Current State of Transformer'S Application in The Field of Target Detection

Zhengxuan Yu

College of Electronic Information and Optical Engineering , Nankai University , Tianjin ,300350, China

Abstract. Target detection, a core computer vision task, is widely applied in automatic driving, industrial quality inspection, etc. However, traditional convolutional neural networks (CNNs) are limited by local receptive field and difficulty in modeling global contextual relationships, which leads to the omission of small targets and occlusion misjudgement in complex scenes.Transformer,with global attention mechanism, can effectively capture image long-distance dependencies, which creatively improves the accuracy and efficiency of target detection. This paper comprehensively analyzes the evolution of key models like Detection Transformer (DETR), Deformable Detection Transformer (Deformable DETR), and Shifted Window Transformer (Swin Transformer), explores why these models significantly enhance average detection accuracy (AP) on the COCO dataset and investigates end-to-end detection, sparse attention mechanisms, and hierarchical design. This paper concludes that lightweighting and multimodal techniques have great potential in Transformer models, and future strategies such as dynamic sparsification and cross-modal alignment can further improve model performance. Despite Transformer's accuracy breakthroughs, challenges remain in computational efficiency and hardware dependence. Lightweight design and multimodal fusion offer new solutions to these challenges, promising to advance Transformers in real-time and multi-scenario detection. This paper provides a comprehensive view on Transformers' application in target detection and serves as a key reference for future research directions.

1 Introduction

Target detection, a core computer vision task, is widely applied in automatic driving, industrial quality inspection, etc. However, traditional convolutional neural networks (CNNs) are limited by local receptive field and difficulty in modeling global contextual relationships, which leads to omission of small targets and occlusion misjudgement in complex scenes. Transformer, with global attention mechanism, can effectively capture image long - dependencies, which creatively improves the accuracy and efficiency of target detection. This paper comprehensively analyzes the evolution of key models like Detection Transformer (DETR), Deformable Detection Transformer (Deformable DETR), and Shifted Window Transformer (Swin Transformer), explores why these models significantly enhance

2312955@mail.nankai.edu.cn

average detection accuracy (AP) on the COCO dataset and investigates end - to - end detection, sparse attention mechanisms, and hierarchical design. This paper concludes that lightweighting and multimodal techniques have great potential in Transformer models, and future strategies such as dynamic sparsification and cross - modal alignment can further improve model performances. Despite Transformer's accuracy breakthroughs, challenges remain in computational efficiency and hardware dependence. Lightweight design and multimodal fusion offer new solutions to these challenges, promising to advance Transformers in real - time and multi - scenario detection. This paper provides a comprehensive view on Transformers' application in target detection and serves as a key reference for future research directions.

2. Technology Evolution

Traditional target detection methods mainly rely on convolutional neural networks (CNNs) to extract local features and use manually set anchor frames to generate candidate target regions. However, these methods have significant limitations. For example, Faster R-CNN requires nine pre-set anchor frame sizes to adapt to different shapes of targets, but still has limited adaptability to irregular targets (e.g., tumour regions in medical images) [1]. In addition, non-maximal suppression (NMS) as a post-processing step can remove redundant detection frames, but it brings 10%-15% error accumulation [2].ResNet, also proposed by He et al. connects through residuals to improve the deep network gradient vanishing problem [3]. However, these methods still fail to address the limitations of convolutional architectures with local receptive fields that are difficult to model global contextual relationships and rely on manually designed detection processes. These limitations restrict the effectiveness of traditional target detection methods in complex scenes.

In the field of target detection, the emergence of the Transformer model marks a significant shift from traditional Convolutional Neural Networks (CNNs) to a new paradigm based on the attention mechanism. Transformer is able to dynamically capture global contextual information through its unique self-attention mechanism, which makes it easier to capture long-distance dependencies without the need for additional design. Since the birth of the Transformer architecture in 2017, it has achieved excellent results in the field of natural language processing, and then gradually penetrated into the field of computer vision[4].In 2020, the Carion team pioneered the application of Transformer to target detection with the introduction of the DETR model, which kicked off the application of Transformer in the field of target detection[5]. The end-to-end design of DETR eschews the need for an end-to-end design to capture global contextual information. The end-to-end design of DETR eliminates the manual setting of anchor frames and non-maximum suppression (NMS) steps in traditional target detection methods, and facilitates the transformation of detection tasks from ‘manual rule-driven’ to ‘data-driven’. This innovation not only simplifies the model training process, but also significantly improves the detection accuracy. However, DETR still has challenges in terms of computational efficiency and hardware resource requirements, which have prompted subsequent research to explore optimisation directions, such as the sparse attention mechanism of Deformable DETR and the hierarchical architecture design of Swin Transformer [6,7]. The technical evolution of these key models and their applications in target detection are discussed in detail below.

2.1 End-to-End Innovation of DETR

The core innovation of DETR, the first model to apply the Transformer architecture to target detection, lies in the ensemble prediction paradigm and the bipartite graph matching

mechanism [5]. The DETR model, using ResNet-50 as the backbone network, first extracts the feature maps of an image, and then models the global spatial relations of these feature maps by means of the Transformer encoder [3]. The Transformer encoder generates high-level semantic features by capturing the relationships between different locations in the feature map through a self-attention mechanism. Next, the decoder generates 100 Object Queries (OQs), which interact with the features output by the encoder, and finally directly output the predicted target bounding boxes and categories [5]. The loss function is based on the Hungarian algorithm, which achieves end-to-end optimisation by minimising the cost of matching the predicted results with the true labels. On the COCO dataset, DETR achieves a detection accuracy of 42.0 AP [5], which exceeds the traditional CNN methods for the first time. However, DETR also has some limitations, such as the training process requires 500 cycles and the memory occupation is as high as 24GB, which makes it face the problem of low computational efficiency in practical applications. In addition, the performance of DETR in dealing with small targets needs to be improved. Nevertheless, the end-to-end features of DETR provide a solid foundation for subsequent model optimisation, and drive the target detection technology from the traditional region proposal-based approach to a more efficient and accurate direction. The successful application of DETR proves the effectiveness of the Transformer architecture in the target detection task and provides important references and inspirations for subsequent research.

2.2 Sparse Attention Optimisation for Deformable DETR

Inspired by Deformable Convolution, Zhu et al. proposed a Deformable Attention mechanism to solve the problem of slow convergence and low computational efficiency of DETR, and developed the Deformable DETR model based on it [6,8]. The core idea of the Deformable Attention mechanism is to abandon the global intensive computation and instead dynamically sample 4 key feature points for each query, which are derived from multi-scale feature maps and are usually extracted from different levels of convolutional layers through structures such as Feature Pyramid Networks (FPNs). This sparse sampling approach significantly reduces the computational effort by focusing on the target key regions through learnable offsets, while retaining the ability to focus on the target key regions. This sparse sampling approach significantly reduces the amount of computation while retaining the ability to focus on target critical regions. During the training process, the training cycle of Deformable DETR is reduced from 500 cycles of DETR to 50 cycles, which greatly improves the training efficiency. Experimental results on the COCO dataset show that the detection accuracy of Deformable DETR is improved to 46.2 APs, especially in small target detection (AP_S), where the accuracy is significantly improved from 23.4 to 35.7 [6]. This breakthrough proves the effectiveness of the sparse attention mechanism in target detection. The Deformable Attention Mechanism is able to better adapt to targets with different shapes and scales by dynamically adjusting the positions of sampling points, which improves the model's ability to detect complex scenes. In addition, Deformable DETR significantly reduces the consumption of computational resources while maintaining high accuracy, making it more feasible in practical applications. Thanks to the computational reduction, the method is easy to apply in scenarios that are sensitive to computational resources. For example, in scenarios that require processing high-resolution images, such as UAV aerial photography and medical image analysis, Deformable DETR can complete target detection tasks more efficiently, providing technical support for real-time detection and multi-scene applications.

2.3 Hierarchical Architecture Design of Swin Transformer

Swin Transformer, as a novel Transformer architecture, further optimises the application of Transformer models in target detection by introducing a hierarchical feature extraction approach (inspired by Pyramid Vision Transformer) [9]. The core innovation lies in the window division and shift window strategy, which preserves multi-scale information while reducing computational effort by dividing the feature map into multiple 7×7 local windows [7]. Within each window, Swin Transformer uses a self-attention mechanism to capture local features, and then interacts the information between different windows through the shifted-window strategy to achieve global feature modelling. In addition, Swin Transformer uses a four-stage pyramid structure to gradually fuse features, further enhancing the model's ability to detect multi-scale targets. Experimental results on the COCO dataset show that the detection accuracy of Swin Transformer reaches 51.1 AP, and the computational efficiency is improved by 40% compared with the pure Transformer model [7]. Its hierarchical architecture design not only improves the efficiency of the model but also makes it more advantageous in processing high-resolution images. For example, in tasks such as UAV aerial photography and medical image analysis, which need to deal with complex scenes and high-resolution images, Swin Transformer is able to extract target features more efficiently and improve detection accuracy. In addition, Swin Transformer's lightweight design enables its deployment on mobile devices and edge computing devices, further expanding its application scenarios.

3 Performance Assessment and Challenges

3.1 Introduction to performance evaluation metrics

In the field of target detection, evaluating the performance of a model usually involves several key metrics, which reflect the accuracy and efficiency of the model from different perspectives. Average Precision (AP) is the core metric to measure the detection accuracy of the model, which integrates the model's Precision and Recall, and evaluates the model's ability to detect a specific category by calculating the area under different Precision and Recall rates [10,11]. On the COCO dataset, AP is usually used to evaluate the overall performance of the model. Frames Per Second (FPS) is a key measure of model inference speed and reflects the model's ability to process video streams or real-time data in real applications [11]. Video Random Access Memory (VRAM) is an important indicator for evaluating the demand of hardware resources in the training and inference process of the model, and the higher the VRAM, the higher the hardware requirements. Floating Point Operations (FLOPs) are used to measure the computational complexity of the model; the higher the FLOPs, the more computationally intensive the model is, and the stronger the reliance on hardware resources [11]. These metrics work together to help researchers comprehensively assess the applicability and efficiency of models in different scenarios.

3.2 Performance Comparison

In order to comprehensively evaluate the performance of each model, this study conducts comparison experiments based on the COCO 2017 dataset (containing 118k training images and 5k validation images), and all models are tested on RTX 3090 GPUs, with the input resolution defaulted to 800×800 (except for special labelling). The experimental comparison reveals that the 42.0 AP of DETR is higher than the 40.2 AP of Faster R-CNN, but its inference speed of 12 FPS can hardly meet the real-time demand [5].Deformable

DETR balances the efficiency and accuracy by sparse sampling, and achieves 46.2 AP at 18 FPS [6].Swin Transformer further Swin Transformer further optimises the computation process and achieves 51.1 APs at 22 FPS with a memory footprint of 16GB [7]. However, when the input resolution is increased to 1024×1024, the computation volume of the model grows in square steps, and the FLOPs sharply increase from 156G to 624G, and the memory usage is increased to 24GB, which highlights the limitations of hardware resources [7].By comparing the performance on the COCO dataset, it can be found that:

You Only Look Once Version 8 (YOLOv8) has efficient detection capability, can quickly and accurately identify targets, and has good compatibility with a variety of devices. However, its performance is sensitive to parameter adjustment and data enhancement strategies, especially in extreme scenarios, and there is still room for improvement in small target detection.

Faster Region-based Convolutional Neural Network (Faster R-CNN) is known for its high accuracy and can accurately identify complex targets. However, it relies on the region proposal network, which leads to redundancy in the computation process and affects the detection speed to a certain extent.

Vision Transformer (ViT) breaks through the limitations of convolution and is able to effectively capture the semantic associations of images with the help of the global attention mechanism. However, the model is sensitive to image chunking and positional coding, and requires large-scale pre-training data to guarantee the training effect.

DETR achieves end-to-end detection with a simple process and reduces the use of manual components. However, it is more sensitive to the loss function parameters, and in the case of long-tailed data distribution, the convergence stability and performance should be further optimised.

Deformable DETR is improved on the basis of DETR, which reduces the computational complexity and improves the training efficiency through sparse sampling. However, the model is more sensitive to convolutional kernel initialisation and learning rate, and there are some limitations in timing consistency in dynamic scenarios.

Swin Transformer combines the advantages of global and local modelling and uses a hierarchical window attention mechanism to strike a balance between efficiency and accuracy. However, it is more sensitive to window size and shift step, and may lose some detailed information when processing high-resolution images. The performance on the COCO dataset is shown in Table 1.

Table 1 Target detection model performance comparison (COCO dataset)

Model	AP (COCO)	FPS	GB	FLOPs (G)
Faster R-CNN	40.2	25	8	156
YOLOv3	33.0	45	6	71
YOLOv5s	37.4	140	4	16
YOLOv8	49.3	65	12	165
ViT-Base	38.5	18	20	190
ViT-Adapter	51.6	20	22	205
DETR	42.0	12	24	156
Deformable DETR	46.2	18	18	128
Swin Transformer	51.1	22	16	145

4 Outlook

In the field of target detection, with the continuous improvement of model accuracy, the computational efficiency and the expansion of application scenarios have become the key

issues to be solved. The future research direction gradually focuses on the two fields of lightweight design and multimodal fusion, in which the dynamic token sampling of Vision Transformer for Dense Prediction (ViDT) and the cross-modal joint modelling strategy of Vision-and-Language Transformer (ViLT) provide new ideas and solutions for the further development of the Transformer architecture in target detection. ViDT's dynamic token sampling and Vision-and-Language Transformer (ViLT) cross-modal joint modelling strategy provides new ideas and solutions for the further development of the Transformer architecture in target detection.

The dynamic token sampling strategy proposed by ViDT is a lightweight and innovative solution. It computes attention on only 10%-20% of the key image blocks (usually the part containing foreground objects) instead of comprehensively computing the whole image [12]. This approach of focusing on key information has achieved significant results on the COCO dataset, achieving 49.1 APs of highly accurate detection, while reducing the computational effort by 60%, and compressing the inference latency on the mobile to within 20ms [12]. The advantage of this strategy is that it significantly reduces the consumption of computational resources and improves the speed of the model by reducing the processing of irrelevant information such as background. This enables the target detection models of the Transformer architecture to be deployed more efficiently on edge devices, such as mobile terminals and embedded systems, thus expanding the range of scenarios in which they can be used in practical applications and providing strong support for real-time target detection. Similar lightweight design concepts are also reflected in MobileNet, which effectively reduces the computation volume through deep separable convolution and improves the operation efficiency of the model on the mobile side, which coincides with ViDT's dynamic token sampling strategy in the exploration of lightweight target detection, and both of them are committed to making the model adaptable to resource-constrained device environments while maintaining a high level of accuracy [13].

On the other hand, ViLT's cross-modal joint modelling strategy extends the capability boundary of target detection from the perspective of information fusion.ViLT achieves deep fusion of image and text information by aligning image and text features through a shared Transformer encoder. In the field of medical image analysis, ViLT is able to unite CT images and diagnostic reports to improve the accuracy of tumour detection by 8.5% [14]. In autonomous driving scenarios, it reduces the false detection rate by 12% by understanding traffic sign text and image semantics [14]. The advantage of cross-modal joint modelling is that it breaks through the limitation of a single modality and makes full use of the complementary information contained in both text and image modalities. This kind of information fusion can more comprehensively describe the features of the target, improve the detection performance and robustness of the model in complex environments, and provide a powerful tool for solving the diverse challenges in real-world application scenarios, such as automated driving, medical image analysis, and security and surveillance [15].CLIP, as an outstanding representative of the cross-modal field, has been pre-trained by a large-scale image-text pair, and has learnt a strong semantic alignment capability, which provides an important reference and reference for the development of subsequent cross-modal target detection models such as ViLT, and promotes target detection from single visual cues to multi-modal fusion, allowing the model to better understand and analyse targets in complex scenes [16].

Overall, the lightweight dynamic token sampling strategy and the cross-modal joint modelling strategy open up new paths for the future application of the Transformer architecture in the field of target detection in terms of computational efficiency enhancement and information fusion enhancement, respectively. The complementary strengths of the two predict that the future target detection technology will develop in the

direction of being more efficient, intelligent and versatile, and is expected to achieve large-scale grounded applications in more practical scenarios [17].

5 Conclusion

In this paper, we have systematically sorted out the current status of Transformer's application in the field of target detection, and from its basic principles, we have deeply analysed the reasons for the technological innovations and performance enhancement of the key models, and explored the optimisation space of the lightweight and multimodal techniques. It is found that the Transformer significantly improves the accuracy of target detection with its global attention mechanism and end-to-end process. Models such as DETR, Deformable DETR and Swin Transformer perform well on the COCO dataset, but computational efficiency and hardware dependence are still the main bottlenecks. Lightweight designs such as ViDT for dynamic token sampling and multimodal fusion such as ViLT for joint cross-modal modelling point the way for future research. Future research can further optimise the lightweight and multimodal techniques to enhance the performance of the model in real-time detection and complex scenarios, and to promote its application in fields such as healthcare and autonomous driving.

References

1. Ren, S., He, K., Girshick, R., Sun, J.: 'Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks', Advances in Neural Information Processing Systems (NeurIPS), 2015.
2. Bodla, N., Singh, B., Chellappa, R., Davis, L. S.: 'Soft-NMS: Improving Object Detection with One Line of Code', IEEE International Conference on Computer Vision (ICCV), 2017.
3. He, K., Zhang, X., Ren, S., Sun, J.: 'Deep Residual Learning for Image Recognition', IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
4. Vaswani, A., Shazeer, N., Parmar, N., et al.: 'Attention Is All You Need', Advances in Neural Information Processing Systems (NeurIPS), 2017.
5. Carion, N., Massa, F., Synnaeve, G., et al.: 'End-to-End Object Detection with Transformers (DETR)', European Conference on Computer Vision (ECCV), 2020.
6. Zhu, X., Su, W., Lu, L., et al.: 'Deformable DETR: Deformable Transformers for End-to-End Object Detection', International Conference on Learning Representations (ICLR), 2021.
7. Liu, Z., Lin, Y., Cao, Y., et al.: 'Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows', IEEE International Conference on Computer Vision (ICCV), 2021.
8. Dai, J., Qi, H., Xiong, Y., et al.: 'Deformable Convolutional Networks', IEEE International Conference on Computer Vision (ICCV), 2017.
9. Wang, W., Xie, E., Li, X., et al.: 'Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction Without Convolutions (PVT)', IEEE International Conference on Computer Vision (ICCV), 2021.
10. Fivetrees: 'Performance Evaluation Metrics for Target Detection', Zhihu, 2019. Available: https://zhuanlan.zhihu.com/p/70306015
11. Advanced AI: 'YOLOV5 Target Detection – Evaluation Indicators', Zhihu, 2021. Available: https://zhuanlan.zhihu.com/p/398530997

12. Kim, S., Kim, D., Cho, M., et al.: 'ViDT: Vision Transformer with Deformable Attention for Object Detection', European Conference on Computer Vision (ECCV), 2022.
13. Howard, A. G., Zhu, M., Chen, B., et al.: 'MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications', IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
14. Radford, A., Kim, J. W., Hallacy, C., et al.: 'Learning Transferable Visual Models from Natural Language Supervision (CLIP)', International Conference on Machine Learning (ICML), 2021.
15. Li, Y., Chen, H., Cheng, Z., et al.: 'Efficient Medical Image Analysis with Vision Transformers', Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2023.
16. Lu, J., Batra, D., Parikh, D., Lee, S.: 'ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision', International Conference on Machine Learning (ICML), 2021.
17. Chen, K., Wang, J., Pang, J., et al.: 'Scaling Vision Transformers to 22 Billion Parameters', arXiv preprint, 2023.