A Review of Deep Learning Based Object Detection Algorithms

Abstract: With the rapid development of artificial intelligence technology, the traditional manual feature-based object detection method has been gradually replaced by deep learning-based object detection technology.Object detection is the basis and prerequisite for many computer vision tasks aimed at recognizing targets in an image and determining their class and location.The purpose of object detection algorithm research based on deep learning is to improve the detection accuracy and detection speed of the detection algorithm, so that the object detection algorithm can be more safe and convenient applied to People's Daily production and life.This article reviews the evolution of object detection algorithms, focusing on two-stage, one-stage and object detection algorithms based on the transformer architecture.In addition, the performance, advantages and disadvantages of different algorithms are compared, and the development trend of object detection algorithms based on deep learning is summarized in the end.
Key words:Object detection; Computer vision; CNN; Transformer

1. INTRODUCTION

The evolution of object detection algorithms has undergone a remarkable transformation from traditional methods to modern deep learning techniques.Initially, object detection algorithms relied on manual extraction of features, which were represented by VJ[1], HOG[2], and DPM[3].The traditional detection methods are characterized by a complex detection process, slow speed, and low accuracy.With the rise of deep learning technology, the field of object detection has been revolutionized. Efficient object detection algorithms such as R-CNN, Fast R-CNN, Faster R-CNN, YOLO and SSD have been proposed one after another.These algorithms not only optimize the feature extraction process, but also simplify the network architecture of the model through an end-to-end training approach.Currently, with the transformer model shining in the field of NLP, the transformer architecture is also used in the framework of object detection algorithms, such as Vision Transformer (VIT), DETR, and Swin Transformer.These object detection algorithms draw on the overall architecture of the transformer model and capture global context information through the self-attention mechanism, providing a new perspective for object detection tasks in different scales and complex scenes[4].In this review, the development history of object detection is comprehensively reviewed, the main methods and key technologies in the current object detection field are summarized, the advantages and limitations of various methods are analyzed, and the development trend of future object detection algorithms is summarized and prospected.The development of the field of object detection is comprehensively reviewed in this review.The main methods and key technologies in the field of current object detection, as well as the advantages and limitations of various methods are summarized successively. At the end of this paper, the future development trend of object detection algorithm is summarized and prospected.

2. THE DEVELOPMENT OF OBJECT DETECTION TECHNOLOGY

2.1 Development of Object Detection Algorithms

The evolution of object detection techniques can be summarized in two stages: the first stage is the period dominated by traditional algorithms, from 1998 to 2014; and the second stage is the period of modern algorithms led by deep learning techniques, from 2014 to present. The object detection technology at this stage

is further subdivided into three parts: two-stage detection algorithm requiring regional suggestion network, one-stage detection algorithm directly predicting target class and location, and object detection algorithm with the transformer architecture.The specific development is shown in figure 1.

Fig. 1. Timeline of the evolution of object detection algorithms

2.2 Architecture of the Object Detection Algorithm

In the early stages of object detection, the algorithm relies on manually designed features.This process usually consists of three main steps: first, identifying the region of interest in the image that may contain the target object;Secondly, features are extracted from within these regions, which usually involves the application of various hand-defined feature descriptors;Finally, these features are used for object detection and classification.With the proposed and continuous development of convolutional neural networks, object detection has entered the era of deep learning algorithms, which can automatically extract image features.The two-stage algorithm first needs to generate candidate regions and then perform classification and bounding-box accurate regression on these regions; The one-stage algorithm predicts the class and location of the target directly on the image without the need of generating candidate regions; The object detection algorithm based on transformer firstly extracts image features using CNN backbone network, and then adds location coding to retain spatial information.Next, the feature set is input into the Transformer encoder, which is processed through a multi-head self-attention mechanism to capture the global context;Subsequently, the decoder combines the encoder output with the learned target query for decoding, with each query focusing on a different region of the image; The decoder output generates the category and bounding box predictions through the prediction header; Finally, the dichotomous matching loss and the Hungarian algorithm are used to loss is computed to achieve end-to-end training, avoiding post-processing steps such as NMS in traditional detection.

3. TWO-STAGE OBJECT DETECTION ALGORITHMS

R-CNN:R-CNN was proposed by Ross Girshick in 2014, firstly R-CNN extracts about 2000 candidate regions from the input image by the selective search algorithm[5]. These regions may contain the target object are scaled to a fixed size required by the CNN input layer, then these candidate regions are fed into the CNN model for feature extraction, and finally the extracted features for each region are fed into a classifier for classification to determine whether the region contains objects of the category that needs to be detected. For each category, the best detection results need to be retained by removing overlapping detection frames by NMS. In terms of performance, R-CNN achieved significant results on the VOC2012 dataset with an average accuracy of 53.3%, which is a qualitative leap in detection

accuracy for deep learning-based object detection algorithms compared to traditional detection algorithms.

SPP-Net: SPP-Net Proposed in 2014 by Kaiming He et al. SPP Net proposes a spatial pyramid pooling that breaks the limitation of traditional convolutional neural networks to fixed-size inputs by the network generating fixed-scale feature representations prior to the fully connected layers[6]. It enables the network to accept images of arbitrary size and scale. This design not only enhances the robustness of the model to object deformation, but also improves the adaptability to scale changes through the multi-size training method and effectively reduces overfitting. In the image classification task, SPP-Net achieves significant performance improvement on multiple public datasets. In the object detection task, it achieves fast pooling of features by computing the feature map of the whole image at once, dramatically improving the detection speed while maintaining high accuracy. In terms of performance, SPP-Net achieves 58.0 mAP on the VOC2007 dataset when using single-scale features, and the mAP improves to 59.2 when using multi-scale features. On the ILSVRC2014 dataset, the SPP-Net single model achieves 31.84 mAP on the test set, and the mAP improves to 35.11 after using the model combination strategy.

Fast R-CNN: In 2015 Ross Girshick proposed Fast R-CNN[7]. Fast R-CNN is an improvement on R-CNN, which simplifies the training process by adopting a single-stage training method and introduces a multi-task loss function to optimize object classification and bounding box localization. At the same time, it improves the accuracy of detection. Fast R-CNN innovatively uses an ROI pooling layer to flexibly handle objects of different sizes. Fast R-CNN significantly improves the training speed. In terms of training speed, Fast R-CNN is 9 times faster than R-CNN. In the testing phase, Fast R-CNN is 213 times faster than R-CNN in terms of processing speed blocks.

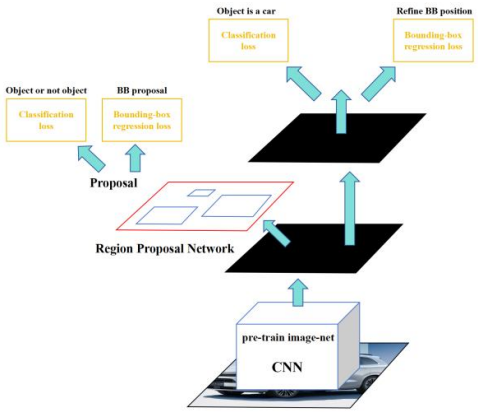

Fig. 2. Structure of Faster R-CNN model

Faster R-CNN: After Ross Girshick proposed Fast R-CNN, in 2015 Ross Girshick proposed Faster R-CNN again[8]. Fast R-CNN is the first end-to-end deep learning detection algorithm that comes closest to real-time performance. The specific model structure is shown in figure 2.The innovation of Faster R-CNN is the introduction of Region Proposal Network (RPN), which is a fully convolutional network that shares convolutional features with the object detection network to generate high-quality region proposals at almost zero cost. Through an end-to-end training approach, the RPN is able to specifically learn how to generate proposals for object detection while sharing features with the Fast R-CNN through an alternating optimization strategy, which significantly improves detection speed. It also reached 73.2 mAP on the VOC2007 dataset.

Mask R-CNN: In 2017, Kaiming He et al. proposed the Mask R-CNN model, which is innovative in predicting the exact segmentation mask of an object by adding a parallel branch, while maintaining fast object detection[9]. The core innovations include the introduction of the RoI Align layer, which enables pixel-level

accurate alignment, and the independent processing of mask prediction and category prediction. It also improves the efficiency and accuracy of model learning. In addition, Mask R-CNN employs a fully convolutional network to maintain the spatial structure of objects and optimizes classification, localization, and segmentation tasks simultaneously with a multi-task loss function. This framework not only excels in instance segmentation, but also can be easily extended to other visual tasks such as human pose estimation.

FPN: In 2017, Tsung-Yi Lin et al. proposed FPN (Feature Pyramid Networks) , which achieves effective detection of objects at different scales by constructing a feature pyramid in a deep convolutional network[10]. Compared to traditional methods, FPN generates high-level feature maps with rich semantic information at all scales at a very small additional computational cost. This top-down architecture incorporates side-by-side connectivity, which not only improves the accuracy of the detection, but also maintains the efficiency of the system. It is capable of running at 5 frames per second on the GPU. The FPN model with ResNet-101 as the backbone network achieves good detection results on the COCO dataset. It reaches $mAP_{0.5}$ = 59.1and $mAP_{0.5,0.95}$ = 36.2. The proposed FPN greatly contributes to the detection accuracy of the object detection network, especially for dealing with datasets with significant differences in object sizes that show superior detection performance.

Cascade R-CNN: Cai et al. proposed Cascade R-CNN model in 2018[11]. This model is an improvement on Faster R-CNN. When Faster R-CNN performs object detection, the Region Proposal Network (RPN) is first used to generate candidate regions, and then the detector is used to make a refined prediction. Cascade R-CNN builds on this foundation by repeatedly stacking the detector part into multiple cascaded modules, and each module trains a series of detectors with progressively increasing IOU thresholds. It effectively solves the overfitting problem in object detection and achieves a stricter screening of false positives during inference. The architecture is unique in that it utilizes the output of the previous stage as training data, which ensures that each detection stage obtains a set of positive samples matching its quality, thus improving the overall detection accuracy.

4. SINGLE-STAGE OBJECT DETECTION ALGORITHMS

YOLOv1: YOLO (You Only Look Once) is an innovative single-stage object detection method proposed by Joseph Redmon et al. in 2016[12]. It unifies and simplifies the detection process by treating the detection task as a regression problem, directly predicting from image pixels to bounding box coordinates and category probabilities. YOLO's end-to-end architecture significantly speeds up detection, with real-time processing speeds of 45 frames per second for the base model and 155 frames per second for the Fast YOLO version. In terms of detection performance, it achieved 63.4 mAP on the VOC2007 dataset. Although YOLOv1 has a great improvement in detection speed, it still has the problem of low detection accuracy relative to the detection of some small objects.

SSD: Liu et al. proposed the SSD algorithm in late 2016, which achieves fast and accurate object detection in images through a single deep neural network[13].The difference between the SSD algorithm and some detection algorithms that detect only at the deepest branches of the network is that SSD uses a set of default boxes with different aspect ratios and scales, which are distributed over the feature maps at multiple resolutions, allowing the network to simultaneously predict objects of multiple sizes and shapes. That said, SSD has several different detection branches, and different detection branches can detect Objects at multiple scales. So SSD has greatly improved the accuracy of multi-scale object detection. SSD achieves 74.3 mAP on the VOC2007 dataset and outperforms both Fast R-CNN and Faster R-CNN. In addition, SSD employs effective data augmentation techniques to enhance the model's ability to detect small objects.

YOLOv2: In the same year , Joseph Redmon et al. proposed the YOLOv2 model[14]. The model has a significant overall improvement in that it achieves real-time detection of over 9000 object classes, introduces

batch normalized BN and high resolution classifier pre-training, as well as the use of anchor frames to improve the accuracy of bounding box predictions. In addition, YOLOv2 employs a multi-scale training approach to enable the model to adapt to different sizes of input images and achieve a balance between speed and accuracy. It reached 76.8mAP at 67FPS on the VOC2007 dataset. Through the joint training method, YOLOv2 is able to train on both the COCO detection dataset and the ImageNet classification dataset at the same time. It uses the hierarchical model of Word Tree to integrate the categories of different data sets effectively, and enhances the generalization ability of the model to new categories. In addition, YOLOv2 also proposes a model Darknet-19 for classification, which further optimizes the computational efficiency while maintaining the accuracy.

RetinaNet: Tsung-Yi Lin et al. proposed RetinaNet in 2018, which employs a new algorithm of focal loss to solve the category imbalance problem in object detectionFocal Loss reduces the loss weight on correctly categorized samples by adding a moderator to the standard cross-entropy loss, allowing the training of the model to be more focused on difficult samples[15]. RetinaNet utilizes a feature pyramid network as the backbone and combines two sub-networks for object categorization and bounding-box regression, which is able to achieve high detection accuracy while maintaining fast detection speeds .

YOLOv3: 2018 Joseph Redmon et al. proposed the YOLOv3 model[16]. YOLOv3 is an innovative improvement on v2 by replacing the feature extraction network with DarkNet53 and borrowing the FPN idea of using three branches to detect objects with different sizes, which can more effectively recognize objects with different scale sizes. In addition, YOLOv3 improves the prediction method of bounding box by using anchor frames and logistic regression to improve the accuracy of prediction, and adopts multi-scale training and data enhancement techniques in the training process to avoid difficult negative sample mining.

YOLOv4: 2020 Alexey and others co-developed the YOLOv4 model, which is a combination of a series of optimization strategies based on v3[17]. It employs new residual connectivity (WRC) and feature fusion techniques (CSP), as well as an improved batch normalization method (CmBN), which optimize model performance and reduce computational effort. YOLOv4 also introduces SAT data augmentation techniques and the Mish activation function that enhance the generalization ability and nonlinear representation of the model. Besides, in the detection head, the SPP module is introduced, which draws on the FPN+PAN structure, Mosaic data enhancement and Drop Block regularization further improve the model robustness. The accuracy of the bounding box regression is significantly improved by using CIOU loss as a new loss function, which integrates the overlap area, centroid distance and aspect ratio.

YOLOv5: In June of the same year, Jocher et al. proposed the YOLOv5 model. Similar to YOLOv4, v5 integrates a large number of excellent technologies in the field of Object detection. It significantly improves the detection performance of small objects by introducing Mosaic data enhancement technology, and at the same time, it optimizes the model's processing ability for images of different sizes by adopting adaptive anchor frame computation and adaptive image scaling strategies. In terms of network structure, YOLOv5 innovatively integrates the Focus structure and CSP structure to strengthen the efficiency of feature extraction and information flow. In addition, it adopts the FPN+PAN structure to enhance the fusion of multi-scale features, which further improves the accuracy of detection. Compared with v4, v5 has strong advantages in model flexibility and rapid deployment.

5. OBJECT DETECTION ALGORITHMS BASED ON TRANSFORMERSTRUCTURE

Since Transformer's huge success in natural language processing, people have wondered if it would be equally effective in the image domain. Since 2020, Transformer has been making its mark in image recognition and object detection tasks.

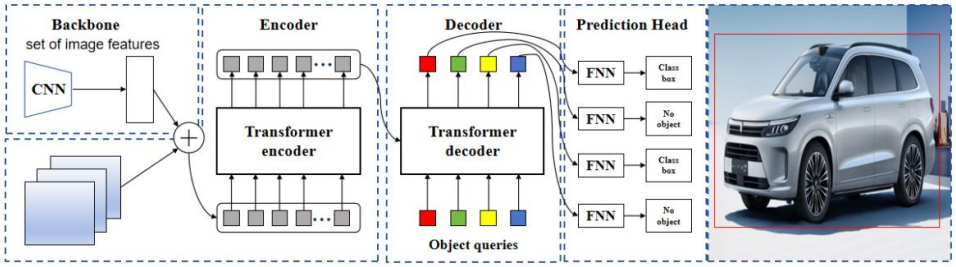

Fig. 3. Structure of the DETR model

DETR: Nicolas Carion et al. proposed an innovative object detection framework DETR in 2020[18]. It breaks the limitations of traditional methods and simplifies the whole detection process by transforming the object detection problem into a direct ensemble prediction problem. DETR employs a Transformer encoder-decoder architecture that utilizes a self-attention mechanism to efficiently process the interrelationships among all pixels in an image, eliminating the dependence on hand-designed components such as non-extremum suppression and anchor point generation. It uses a set prediction method based on global loss to force the unique prediction through bipartite graph matching, which realizes the uniqueness and permutation of prediction results, and allows the model to output the final prediction set in parallel. The specific model structure is shown in figure 3. Furthermore, DETR exhibits performance comparable to Faster R-CNN during end-to-end training, while performing better on large-size objects and can be easily scaled to panoramic segmentation tasks.

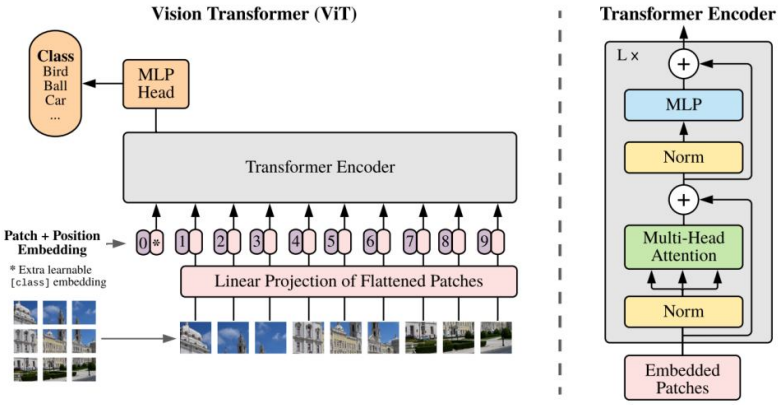

Fig. 4. Structure of the VIT model

VIT: Alexey Dosovitskiy et al. proposed the VIT model around 2020 with the aim of successfully applying the Transformer architecture to large-scale image recognition tasks[19]. The VIT model works by dividing the image into small pieces. After that, these small pieces are mapped to a higher-dimensional space using linear embedding and positional embedding is introduced to preserve spatial information. Finally, it is input to Transformer for processing. The specific model structure is shown in figure 4. The authors discarded the previous practice of using Transformer in combination with convolutional networks and process image block sequences directly with a pure Transformer model. Their model, pre-trained on a large amount of data, is able to achieve better results on multiple image recognition benchmarks while significantly reducing computational resource requirements. In addition, the article also explores the self-supervised learning in image recognition and demonstrates the scalability and effectiveness of Vision Transformer on datasets of different sizes. These findings not only advance the development of image recognition technology, but also open up new possibilities for the application of the Transformer model in computer vision.

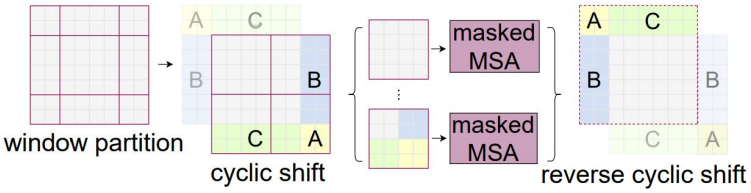

Fig. 5. Structure of the sliding window

Swin Transformer: in 2021 Liu et al. proposed an innovative hierarchical visual Transformer. It efficiently solves the challenges of high computational complexity and large scale variations faced by traditional Transformers in visual tasks by means of a shift window[20]. Shift window improves efficiency by restricting the self-attention computation to non-overlapping local windows while also allowing cross-window connections. The sliding window structure is shown in figure 5. This structure not only achieves linear complexity processing of the image, but also maintains the ability to connect across windows, thus enhancing the representational power of the model. In computing self-attention, Swin Transformer employs a relative position bias, which helps the model better understand pixel relationships within and across windows. Swin Transformer demonstrates superior performance on several visual recognition tasks, including image classification, object detection, and semantic segmentation, with results that surpass previous state-of-the-art methods.

YOLOs: In the same year, Fang et al. proposed an innovative object detection model YOLOs[21]. It is based on a pure Transformer architecture and achieves 2D object and region level recognition with minimal modifications.YOLOs avoids the a priori dependence on 2D spatial structure by introducing DET tokens instead of CLS tokens in the VIT model, and uses bipartite matching loss for ensemble prediction instead of classification loss in the VIT. And it also proves that the object detection task can be performed using a sequence-to-sequence approach to realize it[22]. After pre-training on the medium-sized ImageNet-1k dataset, YOLOs are able to efficiently deliver the pre-trained features to the object detection task and show excellent performance on the COCO dataset. In addition, the review provides insights into different model scaling strategies and evaluates current pre-training schemes, providing unique insights into the generalization and transferability of Transformer to visual tasks.

PVT: Inspired by ViT, Wang et al. proposed a PVT network model in 2021[23]. It breaks through the limitations of traditional convolutional neural networks (CNN) by introducing a convolution-free pyramid structure to generate multi-scale feature maps suitable for dense prediction tasks. Unlike Vision Transformer, which usually produces low-resolution outputs, PVT is able to achieve the capability of high-resolution outputs by training on dense partitions of an image, which will help in dense prediction. In addition, PVT employs an incremental shrinkage pyramid to reduce computational cost, as well as the application of Spatially Reduced Attention Layers (SRA) to further minimize resource consumption. These designs not only allow PVT to excel in vision tasks such as object detection, instance segmentation and semantic segmentation, but also offer the possibility of an end-to-end convolution-free object detection system.

Mobile-Former: In order to enable the effective combination of convolution and Vision Transformer, Chen et al proposed Mobile-Former in 2022[24]. The model adopts a parallel structure that skillfully combines the strengths of Mobile Net and Transformer by introducing an efficient bi-directional bridging mechanism[25]. They exchange information through a bidirectional bridge structure, and use a lightweight cross-attention mechanism as the bridge to realize the deep fusion of local features and global features. This design not only retains the efficiency of Mobile Net in processing local information, but also utilizes the powerful global information encoding capability of Transformer to significantly enhance the representational power of the model. In addition, Mobile-Former uses only a small number of learnable tokens in the Transformer part, which

significantly reduces the computational cost, and at the same time demonstrates performance beyond existing efficient CNN and Transformer models in ImageNet classification and COCO object detection tasks.

BiFormer: In order to solve the problem of computationally intensive and memory-intensive self-attention mechanism in Transformer, in 2023 Zhu et al proposed the BiFormer model[26]. BiFormer proposes a dynamic sparse attention mechanism, i.e., Bi-Level Routing Attention (BRA) mechanism. Unlike traditional fully-connected attention, BRA first filters out the most relevant key-value pairs to the query at the coarse-grained region level, and then applies fine-grained attention within the union of these regions to achieve a dynamic and query-aware efficient allocation of computational resources, which saves computation and reduces the memory footprint. BiFormer's model design allows it to improve computational efficiency while ensuring superior model performance for intensive prediction tasks. This gives BiFormer the flexibility to tackle all kinds of computer vision challenges, such as image classification, object detection and semantic segmentation.

BEVFormer v2: In order to solve the problem of co-limitation between image backbone networks over bird's-eye view detectors, Yang et al proposed the BEVFormer v2 model in 2023[27]. BEVFormer v2 is an innovative two-stage bird's-eye view (BEV) detector that significantly improves the performance of the image backbone network in the 3D object detection task by introducing perspective view supervisory signals. The BEVFormer v2 combines perspective 3D detection head and BEV detection head to provide direct 3D learning guidance for image backbone networks, thus solving the complexity and gradient flow problems of traditional BEV detectors during optimization[28]. The improvements to the original components and the introduction of new components in the BEVFormer v2 model allow the model to converge more quickly and to better adapt to a variety of different image backbone networks. In addition, the model employs a total objective loss function that combines perspective loss and BEV loss, which further facilitates the optimization of the model. Through extensive experiments on the nuScenes dataset, BEVFormer v2 proves its effectiveness and superiority, bringing new advances in 3D object detection techniques in the field of autonomous driving.

RT-DETR: The YOLO model suffers from difficult model optimization, poor robustness, and delays in NMS post-processing and detector detection speed. However, due to the lack of anchor pre-processing and NMS post-processing, DETR model has slow training, convergence and inference, which still cannot achieve the requirements of real-time detection. To solve the above problems, Yian Zhao et al. proposed an innovative real-time end-to-end object detector RT-DETR[29]. It significantly improves the accuracy and speed of detection by introducing an efficient hybrid encoder for fast processing of multi-scale features and employing a minimum uncertainty query selection mechanism to provide high-quality initial queries. In addition, the RT-DETR model can be flexibly adapted to different real-time detection scenarios by using different decoder layers to adjust the model size and inference speed without retraining. The review also provides an in-depth analysis of the impact of non-maximal suppression (NMS) on the speed and accuracy of existing object detectors and establishes an end-to-end speed benchmark to further advance the development of real-time object detection technology.

Tables 1 and 2 summarize the backbone network, detection rate, and detection accuracy on the VOC2007 dataset, VOC2012 dataset, VOC07+12 dataset, and MS COCO dataset used for the publication of the two-stage and single-stage object detection algorithms, respectively. "-" indicates no relevant data, and parentheses in the mAP values indicate that they were used as the training set. Table 3 summarizes the detection accuracy of the object detection algorithms with the Transformer architecture on the backbone network and MS COCO dataset used for the experiments. As the object detection algorithms continue to evolve, their detection accuracy and speed continue to improve.

Table 1 Performance comparison of two-stage object detection algorithms

Method	Backbone	FPS	mAP (VOC2007)	mAP (VOC2012)	mAP (VOC07+12)	mAP@0.5% (MS COCO)	mAP@[0.5,0.95]% (MS COCO)
R-CNN	AlexNet	0.03	-	53.3	-	-	-
SPP-Net	ZF-5	-	59.2	-	-	-	-
Fast R-CNN	VGG-16	-	66.9	65.7	68.4	-	-
Faster R-CNN	VGG-16	5	73.2	70.4	70.4	-	-
Mask R-CNN	ResNeXt-101-FPN	5	-	-	-	62.3	39.8
FPN	ResNet-101	5	-	-	-	59.1	36.2
Cascade R-CNN	ResNet-101	-	-	-	-	62.1	42.8

Table 2 Performance comparison of single-stage object detection algorithms

Method	Backbone	FPS	mAP (VOC2007)	mAP (VOC2012)	mAP (VOC07+12)	mAP@0.5% (MS COCO)	mAP@[0.5,0.95]% (MS COCO)
SSD300	VGG-16	46.0	68.0	-	72.4	41.2	23.2
YOLOv1	GoogleNet	45.0	63.4	57.9	63.4	-	-
YOLOv2	Darknet-19	40.0	-	-	78.6	44.0	21.6
RetinaNet	ResNet-1001-FPN	-	-	-	-	59.1	39.1
YOLOv3	Darknet-53	78	-	-	-	57.9	33.0
YOLOv4	CSPDarknet53	65	-	-	-	65.7	43.5
YOLOv5n	CSP-v5	45	-	-	-	45.7	28.0
YOLOv5s	CSP-v5	98	-	-	-	56.8	37.4

Table 3 Performance comparison of object detection algorithms using Transformer as architecture

Method	Backbone	mAP@0.5% (MS COCO)	mAP@[0.5,0.95]% (MS COCO)
DETR	DETR-DC5-R101	64.7	44.9
Swin Transformer	Swin-S	70.4	51.8
YOLOs	YOLOs-Base	-	42.0
PVT	PVT+RetinaNet	-	40.4
Mobile-Former	MF-508M	61.8	43.3
Biformer	BiFormer-Base	68.5	47.1
RT-DETR	ResNet-50	71.3	53.1
RT-DETR	ResNet-101	72.7	54.3

6. CONCLUSION AND OUTLOOK

Nowadays, object detection technology is experiencing continuous innovation and progress, and its accuracy and response speed have made a significant leap compared with traditional algorithms. However, in the context of the continuous expansion and diversification of application scenarios, the technology still faces many challenges and problems that need to be solved urgently. These challenges include, but are not limited to, further optimization of model algorithms, efficient preprocessing of image data, design of more efficient detection networks, and in-depth research on model lightweighting. Synthesizing the current research problems and challenges in object detection, the following outlook for future research is made:

（1）Lightweight model: With the rapid development of object detection algorithms, their network architectures are becoming more and more complex, which leads to the inability to meet the real-time detection needs of existing detection devices. In order to ensure high accuracy while increasing the detection rate and making the model more streamlined, researchers are actively studying model compression and quantization, network pruning, knowledge distillation and other technologies, aiming to reduce the computational requirements of the model and storage space. So that the object detection technology is more adaptable to the actual application scenarios.

（2）Small object detection: With the wide application of deep convolutional neural networks, object detection techniques based on such networks have gradually become the leading direction of research. Nevertheless, existing techniques still face the problem of underperformance in recognizing smaller-sized objects, especially when deep networks perform feature extraction on these objects, which often leads to the loss of key semantic information. To address this problem, researchers have begun to experiment with super-resolution reconstruction techniques to enhance the detail information of small objects. For example, EDSR[30] , VDSR, SRCNN,FSRCNN，ESPCN,DRCN and other techniques learn the conversion from low-resolution images to high-resolution images by applying convolutional layers, which not only improves the visual effect of the reconstructed image, but also enhances the ability to capture the detailed features of small objects. In addition, in order to overcome the technical challenges of small object detection, researchers have proposed a variety of solutions, including but not limited to: multi-scale feature fusion, introduction of attentional mechanism, data enhancement, and innovative network structure.

（3）Occlusion and Complex Environment Detection: In the field of object detection, occlusion and complex environment are two important challenges. For the occlusion problem, some methods try to reduce the impact of occlusion on the detection performance by learning the local features of the object, such as the YOLO family of algorithms, which improves the robustness of the model in complex environments through data augmentation and multiscale training. In addition, some studies have combined Transformer with conventional CNN to utilize the advantages of CNN in feature extraction and the ability of Transformer to handle long range dependencies to enhance the understanding of the interactions between different regions in an image. For complex environmental problems, multimodal sensor fusion techniques can enable models to extract more feature information in complex environments and improve the perception of the environment.

References:

1. Viola, P.; Jones, M.J. Robust Real-Time Face Detection. International Journal of Computer Vision 2004, 57, 137-154.
2. Dalal, N. Histograms of oriented gradients for human detection. Proc of Cvpr 2005.
3. Felzenszwalb, P.; McAllester, D.; Ramanan, D. A discriminatively trained, multiscale, deformable part model. In Proceedings of the 2008 IEEE conference on computer vision and pattern recognition, 2008; pp. 1-8.
4. Vaswani, A. Attention is all you need. arXiv preprint arXiv:1706.03762 2017.
5. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2014; pp. 580-587.
6. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence 2015, 37, 1904-1916.
7. Girshick, R. Fast r-cnn. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2015; pp. 1440-1448.

8. 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 2015, 28.
9. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2017; pp. 2961-2969.
10. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017; pp. 2117-2125.
11. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018; pp. 6154-6162.
12. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016; pp. 779-788.
13. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, 2016; pp. 21-37.
14. Redmon, J.; Farhadi, A. YOLO9000: better, faster, stronger. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017; pp. 7263-7271.
15. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2017; pp. 2980-2988.
16. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 2018.
17. Bochkovskiy, A. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv preprint arXiv:2004.10934 2020.
18. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European conference on computer vision, 2020; pp. 213-229.
19. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations, 2021.
20. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021; pp. 10012-10022.
21. Fang, Y.; Liao, B.; Wang, X.; Fang, J.; Qi, J.; Wu, R.; Niu, J.; Liu, W. You only look at one sequence: Rethinking transformer in vision through object detection. Advances in Neural Information Processing Systems 2021, 34, 26183-26197.
22. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. Advances in neural information processing systems 2014, 27.
23. Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021; pp. 568-578.

24. Chen, Y.; Dai, X.; Chen, D.; Liu, M.; Dong, X.; Yuan, L.; Liu, Z. Mobile-former: Bridging mobilenet and transformer. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022; pp. 5270-5279.
25. Howard, A. Mobilenets: Efficient convolu-tional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 2017.
26. Zhu, L.; Wang, X.; Ke, Z.; Zhang, W.; Lau, R.W. Biformer: Vision transformer with bi-level routing attention. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023; pp. 10323-10333.
27. Yang, C.; Chen, Y.; Tian, H.; Tao, C.; Zhu, X.; Zhang, Z.; Huang, G.; Li, H.; Qiao, Y.; Lu, L. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023; pp. 17830-17839.
28. Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Qiao, Y.; Dai, J. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In Proceedings of the European conference on computer vision, 2022; pp. 1-18.
29. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. Detrs beat yolos on real-time object detection. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024; pp. 16965-16974.
30. Lim, B.; Son, S.; Kim, H.; Nah, S.; Mu Lee, K. Enhanced deep residual networks for single image super-resolution. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2017; pp. 136-144.