Review Article

Review of Typical Vehicle Detection Algorithms Based on Deep Learning

ABSTRACT

Object detection is the crucial task in the field of computer vision. In recent years, intelligent driving technology and intelligent transportation system have set off a boom. Therefore, vehicle object detection has also become a hot research task in the field of computer vision and deep learning. With the rapid development of deep learning, the current mainstream vehicle detection algorithms are Convolutional Neural Networks (CNN)-based two-stage and one-stage object detection algorithms. Because of the local nature of the image presented by CNN, the global receptive field of the network is limited. At the same time, Transformer shows a strong long-distance dependence characteristic, and opens up a new idea of combining images with Transformer. Therefore, the research of object detection algorithm based on Transformer gradually causes a boom. This paper mainly introduces the advantages and disadvantages of several representative algorithm models, and makes a summary and prospect.

Keywords: vehicle detection;deep learning;convolutional neural networks;Transformer.

1. INTRODUCTION

In the development of intelligent driving tech-nology and intelligent transportation system, vehicle object detection, as an important part of intelligent driving environment perception, provides strong support for subsequentvehi-cle decision-making planning, behaviorcont-rol and other tasks. It is of great significance and value in intelligent driving, building intel-ligent transportation system and smart city.T- he main problems are still the accuracy,spe- ed and accuracy of the detection algorithm. Therefore, how to achieve efficient vehicle detection has become a hot researchcont-ent. The task of object detection includes two parts: object classification and position-ing. Its essence is to locate and classify the object, locate the object of interest in the image, and then correctly identify thecateg-ory of the object according to the feature information, and then use the detected frame to locate the object, so as to complete the detection task. The traditional vehicle detection algorithm[1]. [2]. [3]. is based on artificial featureextrac-tion. Although it has achieved certain results, it also shows its inherent drawbacks. In the artificial feature extraction, a large number of redundant Windows will be generated by sliding Windows, which has a direct impact on the accuracy of the algorithm. Moreover, there are a large number of redundantcalcu-lations, which will make it difficult to improve the running speed. Therefore, the traditional vehicle detection algorithm has been difficult to meet the needs of high performance detection. The vehicle detection algorithm based on deep learning aims at the defects of traditional vehicle detection, a neural net- work model capable of self-learning image features is

proposed. Since the emergence of AlexNet[4]. , the object detection algorithm has opened an era of deep learningdomin-ated by convolutional neural networks[5]. [36]. **Error! Reference source not found.**(CNN), making a newbreakthrou-gh in the accuracy and real-time of vehicle detection, which has aroused widespread concern. At present, a large number of network models with simplified architecture and good training effect have been proposed. The dominant vehicle object detection algor-ithms are mainly divided into one-stage and two-stage detection algorithms.

Convolutional neural network has always been considered as the basic model of computer vision. Due to the outstanding performance of Transformer [6] in the field of Natural Language Processing[55]. [38]. [41]. , it has been highly concerned by researchers, and has gradually been introduced into visual tasks and gained competitiveness. Convolutional neural network has translation invariance, local sensitivity and other induct-ive biases, which can well capture the local feature information of the image. However, CNN has limited receptive field and does not have the ability to obtain global information. It needs to stack convolution layerscontinu-ously to extract the image from local to global information. In contrast, Transformer's advantage lies in its global receptive field[56]. [57]. , which can focus on global information. Therefore, Transformerprovid- es a new possibility[42]. [43]. [44]. [45]. [46]. for visual feature learning. The visual model based on Transformer has achieved a comparable or even leading effect of convo-lutional neural network in imageclassificat- ion[47]. [25]. , object detection[26]. [27]. , image segmentation[48]. [49]. , video understanding[50]. [51]. , image generation[52]. and other fields, making the accuracy and real-time of vehicle detection reach a new height.

2. TWO-STAGE VEHICLE DETECTION ALGORITHM

The two-stage algorithm divides thedetecti-on problem into two stages. First, the algorithm, including selective search or regional proposalion network, is used to extract the region proposal, and then the candidate region is put into the classifier SVM for secondary correction to get the detection results. At present, typicalalgori-thms include R-CNN[7]. , SPP-Net[8]. , Fast R-CNN[9]. , Faster R-CNN[10]. , and Feature Pyramid Networks (FPN)[11]. , Mask R-CNN[34]. .

2.1 R-CNN

In 2014, Girshick[7]. et al. proposed R-CNN model and made a great breakthrough by using convolutional neural network in object detection tasks. As shown in Figure 1 ,The main process is roughly divided into four steps: 1) input image; 2) selective search (SS) algorithm was used to extract about 2000 candidate domains that may contain objects; 3) Send the regions proposal to CNN for feature extraction; 4) SVM was used for classification and borderadjustm-ent of extracted features.

Compared with the traditional vehicle detection algorithm, R-CNN uses selective search to solve the problem of too much computation when sliding window generates candidate boxes. In addition, CNN is used to extract the features of the region of interest, which solves the defect of the limited feature characterization ability of the traditional detection algorithm. SVM classifier was use-d for classification, and regression algorithm was introduced to adjust the object boundary box to improve the inconsistency between the region of interest and the actual object.

Through these methods, the performance of R-CNN algorithm is significantly improved, and the mAP on Pascal VOC 2007[12]. dataset reaches 58.5%. However, there are still some problems. It takes too long to generate candidate boxes by using the selective search algorithm, which affects the detection speed. Candidate areas need to be trimmed to a fixed size, which will cause information loss or partition of too much background; The training of the network should be carried out in multiple steps, which leads to a long time and other problems.

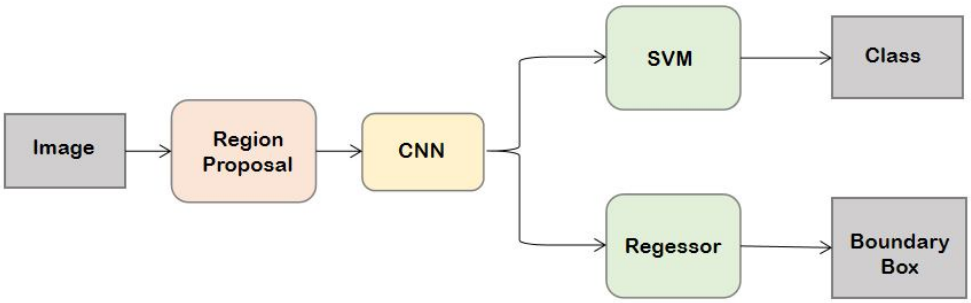

Figure 1: Architecture of R-CNN

2.2 SPP-NET

In 2015, Kaiming He et al. proposed the Spatial Pyramid Pooling Network [8]. (SPP-Net) object detection algorithm , which is an improvement of R-CNN[7]. . In this algo-rithm, a spatial pyramid pool structure is ad-ded between the last convolution layer and the full connection layer, so that the features of any proportion region can be extracted without scaling the region proposal, avoiding the information loss caused by scaling and cutting the image in the region of interest. This algorithm sends the whole image into the convolutional neural network to extract features without repeating the convolution operation. While ensuring the detection acc-uracy, the detection speed is greatly improv- ed, which is 24--102 times higher than that of R-CNN algorithm. Meanwhile, the mAP on VOC 2007 dataset[12]. is increased to 59.2%. Although SPP-Net optimizes the time cons-uming problem of R-CNN algorithm, there are still some problems. Just like R-CNN, multi-step training is required, and the multi-scale of the spatial pyramid pool model cannot fine-tuned all the previousconvoluti- on layers.

2.3 FAST R-CNN

In 2015, Girshick et al. proposed Fast R-CNN[9]. algorithm and improved R-CNN by adopting the mthod of SPP. Architecture of Fast R-CNN shown in Figure 2. VGG-16[13] backbone network is used to replace AlexNet to adjust the multi-scale pyramid Pooling model in the spatial pyramid algorithm into a single-scale ROI Pooling layer, so that parameters of all convolution layers can be fine-tuned. Multi-task Loss function is also proposed. SoftMax classifier is used to replace SVM classifier, and classification and regression tasks are carried out at the same time, so that classification and positioning tasks can not only share convolution features, but also promote each other to improve detection effect. Compared with R-CNN[7]. and SPP-Net algorithms, Fast R-CNN integrates multiple steps into a model, and the training process is no longer divided into multiple steps, which improves the network performance and speeds up the training

speed. However, Fast R-CNN still has some problems, such as the slow generation of regions proposal.

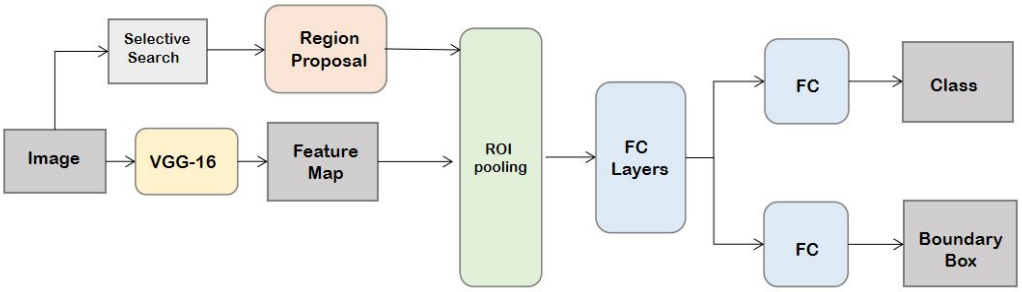

Figure 2: Architecture of Fast R-CNN

2.4 FASTER R-CNN

In 2015, Ren et al. proposed the algorithm framework Faster R-CNN[10]. , which introduced Region Proposal Networks (RPN) to replace the original Selective search method. As shown in Figure 3. The main contribution of this algorithm is RPN, which generates a large number of object regions proposal based on Anchor mechanism and greatly improves the speed of regional proposal. RPN takes the feature map of arbitrary size as input, and generates some candidate regions which may contain the object through convolution operation. Faster R-CNN can carry out multiple steps of region proposal, feature extraction, classification and positioning in the same network, so as to achieve end-to-end training and greatly improve the training efficiency. The mAP of Faste R-CNN on PASCAL VOC2007[12]. dataset was improved to 78%. Although it has a high detection accuracy, it has poor detection effect on small objects.

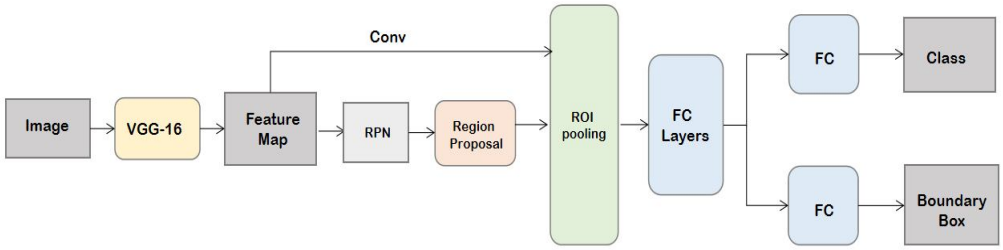

Figure 3: Architecture of Faster R-CNN

2.5 FPN

In 2017, Lin et al. improved Faster R-CNN[10]. and proposed the Feature Pyramid Networks [11]. (FPN) detection algorithm. FPN constructs a top-down architecture characterized by the addition of multi-scale features and feature fusion. The detection algorithm introduced above only detects the top-level feature, but the object location information of the top-level feature is less. However, the underlying feature has little semantic information, but it just has exact location information. Therefore, the top-down structure of FPN is adopted. By fusing the spatial information of the semantic-rich feature maps of different resolutions, the prediction is carried out on the feature maps of multiple scales at the

same time. The FPN has not only strong semantic information, but also richgeome-tric information, and the detection effect of small objects is improved.

2.6 MASK R-CNN

In 2017, He et al. proposed Mask R-CNN[34]. , which is an extension of Faster R-CNN. The RoI Align layer is used to replace the RoI Pooling layer of Faster R-CNN, and the bilinear interpolation algorithm is used to adjust the deviation caused by integerquan-tization, so that the features obtained by each receptive field can be aligned with the receptive field of the original image, thus improving the accuracy of object detection branches. Mask R-CNN adds a mask branch on the basis of classification and regression, and combines multi task loss withclassifica-tion error, regression error, andsegmentati- on error to achieve a network for image segmentation and object detection. The mAP of the algorithm on the MS COCO dataset reaches 39.8% However, it is difficult to meet the needs of real-time detection, because of the addition of segmentation branches, resulting in a large amount of computation.

3. ONE-STAGE VEHICLE DETECTION ALGORITHM

One-stage algorithm, that is, end-to-end, one-stage detection of vehicle objects. The stage of candidate region generation is omitted and the object classification and position coordinates are obtained directly. This detection method greatly improves the running speed of the algorithm and meets the requirement of real-time object detection. Typical algorithms include YOLO(You Only Look Once) series[14]. [15]. [16]. [17]. [18]. [19]. [20]. and SSD(Single Shot Multibox Detector) series [21]. [22]. .

3.1 YOLO

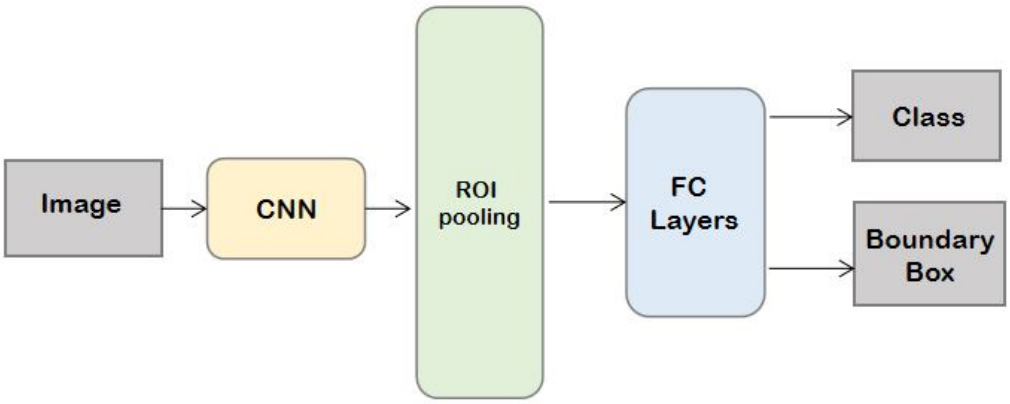

Figure 4: Architecture of YOLO

In 2016, Redmon et al. proposed a one-stage object detector, YOLO[14]. (You Only Look Once). The difference between YOLO and two-stage detection algorithm is that YOLO eliminates the step of candidate box extraction and directly uses regression to classify objects and predict candidate boxes. The network structure is simple, and the detection speed is improved to about 10 times that of Faster R-CNN[10]. . Structure of YOLO shown in Figure 4. The algorithm first divides the input image into S×S grids, and each grid unit is only responsible forpredic- ting the object in the center of the grid. The result of each prediction includes theproba-bility of the boundary box and the objects in the boundary box belonging to various

cate-gories. Finally, the non-maximum suppress-ion algorithm (NMS) [23]. is used to remove the excess boundary boxes and obtain the detection result. Although YOLO algorithm is fast, it also has several disadvantages :1) for multiple adjacent objects, it is easy to miss detection; 2) YOLO does not solve the problem of multi-scale Windows, so the detection effect of small-scale objects is not good.

In 2017, Redmon et al. improved the YOLO network structure and proposed the YOLOv2[15]. algorithm. This model uses DarkNet-19 as the backbone network, and Batch Normalization (BN) is added after each convolutional layer, thus solving the problem of low detection accuracy of YOLO v1 algorithm. Higher resolution classifiers are used to adapt to high resolution inputs. Two different scale features were used to enhance the prediction robustness of the model for multi-scale images. The number and shape of boundary boxes were generated by K-Means clustering to improve the confidence score. The Binary Cross Entropy loss function was used to replace the Softmax function, which improved the recall rate and accuracy, and the mAP was increased to 78.6% on VOC 2007[12]. dataset . However, the detection effect of YOLOv2 on small objects is still poor, and the detection accuracy is not high enough.

In 2018, Redmon et al. improved YOLOv2 and proposed YOLOv3[16]. . The algorithm uses a more complex backbone network, the residual network model Darknet-53, to extract features. Moreover, FPN structure is used for multi-scale prediction to obtain more effective information of small objects, so as to improve the detection accuracy of small objects. The 1 ×1 convolution and Logistic activation function were used to replace the Softmax classification layer for more effective data fitting. The YOLOv3 algorithm obtained 33.0% AP and 57.9% mAP in MS COCO[24]. data set. This algorithm can improve the detection performance of small objects obviously, but the detection accuracy and real-time performance are still poor.

In 2020, Bochkovskiy et al. improved on YOLOv3 and proposed YOLOv4[17]. . This algorithm uses a CSP Darknet-53 backbone Network combined with Spatial Pyramid Pooling (SPP) and Path Aggregation Network (PAN) for feature fusion. Mish activation function is also used to achieve higher performance. mAP on the MS COCO dataset[24]. achieved 43.5% and a speed of 65 FPS.

In 2020, Jocher et al. proposed YOLOv5[18]. . Focus structure and CSP structure are added to the backbone network, and FPN+PAN structure is used in Neck to enhance the network feature fusion. The YOLOv5 is slightly weaker than the YOLO v4[17]. in performance, but far more flexible and faster.

The YOLOv6[19]. algorithm uses the more efficient main network EfficientRep, and neck also builds ReP-PAN based on Rep and PAN. In terms of training strategy, Achor-free is adopted. At the same time, SimOTA label assignment strategy and SIOU bounding box regression loss were used to further improve the detection accuracy. In MS COCO[24]. dataset, the accuracy reached 43.1% AP.

YOLOv7[20]. algorithm, focusing on the optimization of training process, designed several trainable bag-of-freebies, so that real-time object detection can greatly improve the detection accuracy without increasing the reasoning cost. A new label assignment method, coarse-to-fine lead guided lable assignment, is proposed. A compound scaling method of extend and compound scaling for real-time object

detector is also proposed to make effective use of parameters and calculations.

3.2 SSD

In 2016, Liu et al. proposed the SSD [21]. (Single Shot Multibox Detector) model based on the advantages of fast detection speed of YOLO and accurate positioning of Faster R-CNN[10]. . As shown in Figure 5. In this algorithm, VGG-16[12]. was used as the backbone network to extract image features, and multiple convolution layers were added after VGG-16 to obtain multi-scale feature maps for predicting results. Based on the Anchor mechanism in the Faster R-CNN[10]. algorithm, candidate regions are obtained from the feature map through prior boxes with different sizes, so as to detect objects with different sizes better. It has betterdete-ction effect for objects with overlapping areas or close distances and improves the recall rate. In addition to ensuring thedetec-tion accuracy, the algorithm speed was also accelerated. The mAP on VOC 2007[12]. dataset reached 79.8%, which was 3 times faster than that of Faster R-CNN[10]. . However, SSD has many duplicate boxes, and the detection effect of small objects is not as good as that of the two-stage algorithm.

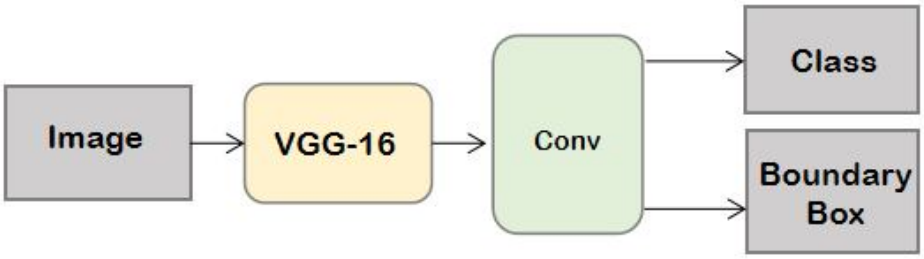

Figure 5: Architecture of SSD

In 2017, FU C Y et al. proposed DSSD[22]. algorithm, because the weak detection effect of SSD algorithm on small objects is mainly due to the weak representation ability of feature map, so DSSD is mainly aimed at improving the representation ability of shallow stage. ResNet-101 was used as the backbone network, replacing VGG. The deep features and shallow features obtained by multiple deconvolution are fused to enrich the context information of small objects, thus effectively improving the detection performance of small objects. Compared with SSD[21]. , DSSD extracts more robust features, which improves the accuracy.

In 2017, Li Z X et al. proposed the FSSD[35]. algorithm. By using the idea of FPN algorithm for reference, multi-scale features and information are fused. Although the detection accuracy of small object is reduced, the detection speed is significantly improved.

4. VEHICLE DETECTION ALGORITHM OF TRANSFORMER-BASED

The typical feature of CNN is locality, which is an inductive bias feature based on the strong correlation of adjacent pixels. Unlike CNN, Transformer's learning process is based on the interaction of globalinformati-on. Therefore, the combination of CNN and Transformer will help improve the network's ability to learn and represent features. The structure of Transformer shown in Figure 6. This section mainly introduces several com-mon Transformer based vehicle detection models, such as DETR series [28].

[27]. [26]. [30]. [54]. , Vision and FPT[33]. . Transformer(ViT) series[25]. [32]. [53].

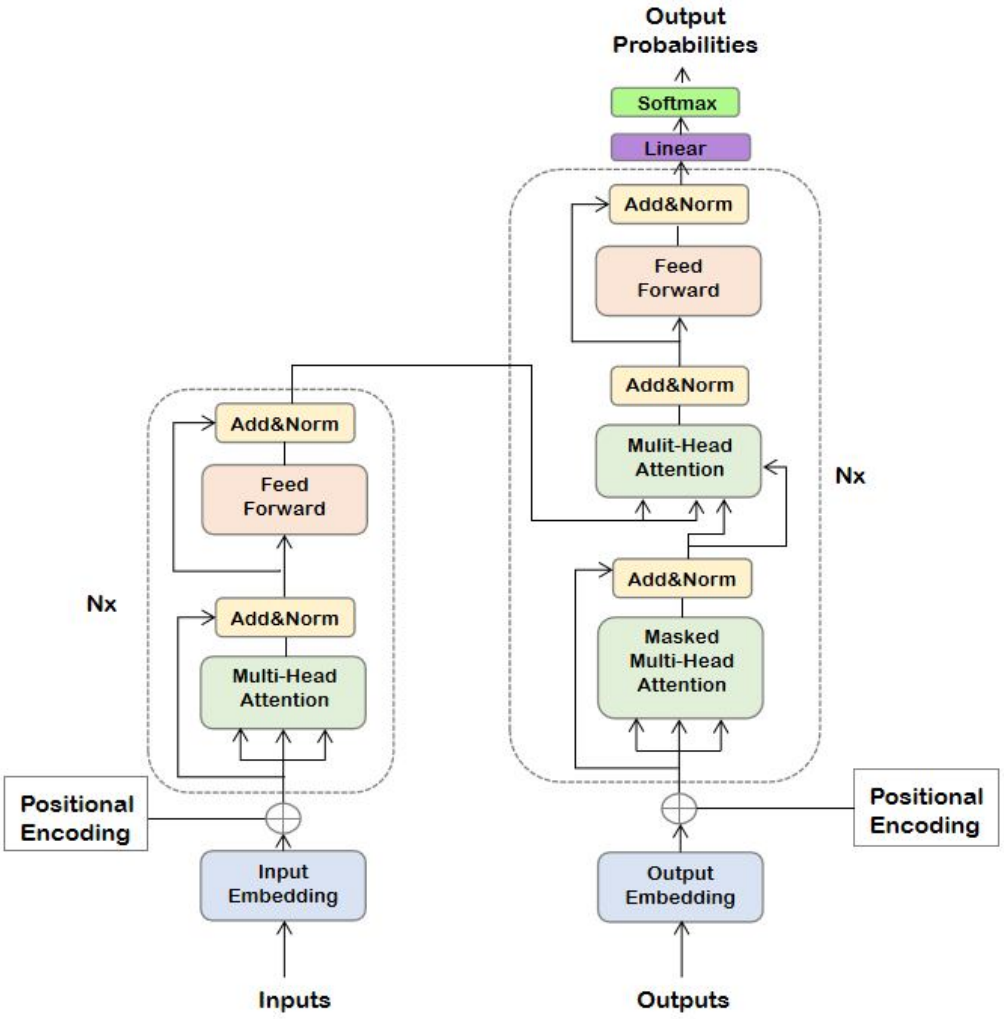

Figure 6: Architecture of Transformer

4.1 DETR SERIES

In 2020, Carion et al. proposed DETR[26]. model, which is the first end-to-end object detection algorithm based on Transformer. Architecture of DETR shown in Figure 7. The model adopts a structure combining ResNet-50 feature extraction andTransfor-mer to realize object detection, and the detection task is divided into two parts: feature extraction and object detection. During the training, the bipartite matching loss function that forces unique matching between ground-truth boxes and prediction, eliminating the post-processing performance of a hand-designed NMS[23]. . The DETR extracts the image features through the encoder, and then interacts with the image features using the randomly initialized object query mechanism.The object query vector contains the location information and feature information of potential objects. It extracts the object information using the self attention mechanism. After multi-layer interaction, it uses the full connection layer to predict the

target information from each object query to form the final detection result. The AP of this model on MS COCO dataset is similar to that of Faster R-CNN[10]. , but due to the complexity of attention quadratic computing and other factors, the convergence speed is slow, It takes up to 500 epochs of training to get a more stable effect. The main reason for its slow convergence is the design of the object query mechanism,10-20 times slower than that of Faster R-CNN, and the detection effect of small objects is poor.

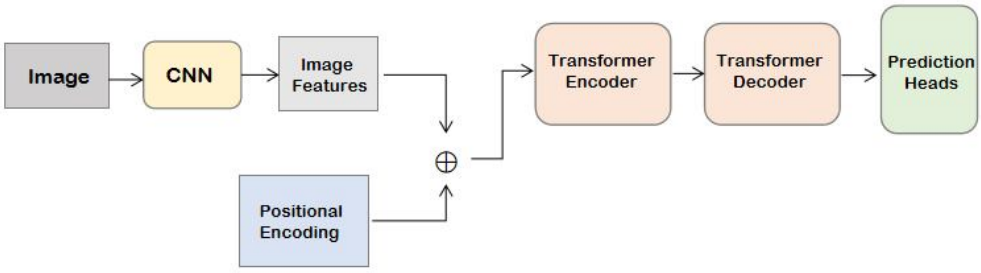

Figure 7: Architecture of DETR

In 2020, Zhu et al proposed Deformable DETR[27]. mode to solve the DETR problem. This model draws on the idea of deformable convolution[31]. and carries out sparse sampling on feature maps of different levels to accelerate the convergence rate of key positions that the model focuses on learning. Meanwhile, the multi-scale deformable attention mechanism is used to aggregate the information between multi-scale feature maps to improve the detection accuracy of small objects. Compared with DETR[26]. model, the convergence rate is 10 times faster, so the training speed and small object detection are improved, but the effect is not good for occluded objects.

Zheng et al proposed a new Adaptive Clustering Transformer[28]. (ACT) based on DETR[26]. , Locality Sensitive Hashing (LSH) [29]. is used to reduce the complexity of the model,to compress the number of object queries, and ACT replaces the self-attention[38]. [39]. [40]. [41]. module in the DETR model without any retraining. Indeed, Multi-Task Knowledge Distillation (MTKD) was used to reduce the performance decline. This meth- od reduces the computation cost andachie-ves a good balance between performance and computation.

Zhigang Dai et al proposed UP-DETR[30]. (Unsupervised Pre-trained DETR), a object detection algorithm with unsupervised pre-training. Pre-training Transformer in an unsupervised way to give it good visual representation; In order to solve the problem of multi-query positioning, multiple single query blocks were allocated to different object queries, in which each query block was independently predicted by the attention mask and object query shuffle mechanism, thus simulating the multiple objectsdetecti- on task and accelerating the convergence rate of the DETR model. 42.8% AP was implemented on the MS COCO[24]. dataset. UP-DETR has higher precision and faster convergence rate than DETR[26]. , which proves the feasibility and effectiveness of the unsupervised pre-training strategy.

Zhuyu Yao et al. proposed the Efficient DETR[54]. model, which is a simple and effective pipeline for end-to-end objectdete-ction. This model predicts the proposals of top score on the deny feature

map from CNN. The 2D or 4D coordinates of these proposals are used to initialize the reference point In addition, select the top-k feature to initialize the object query In order to improve the problem of arrogant convergence speed caused by initialization, the model uses the features learned by the encoder network based on Transformer to make intensive prediction, obtain the position, size and cate-gory information corresponding to theposs-ible object information, and select the results with high confidence as the initial state of the target query, and then use the decoder to make sparse prediction to match the final detection results. It mainly uses densedete-ction and sparse set detection at the same time to reduce the gap between one decoder structure and six decoders structure before initializing the target container. With only three encoders and one decoder, training 36 epochs can achieve 44.2% mAP on MS COCO. It greatly surpasses the modern detectors on CrowdHuman datasets.

4.2 VISION TRANSFORMER (VIT) SERIES

In 2020, Dosovitskiy et al. proposed Vision Transformer[25]. (ViT) model, whichcomple-tely replaces convolution structure tocompl-ete image classification task. Architecture of ViT shown in Figure 8. Firstly, the input image is cut into small blocks of fixed size, which is linearly mapped and then thepositi- on code is added and input into a standard Transformer encoder.The advantage of ViT is that it constructs a global information interaction mechanism, which helps toesta-blish more adequate feature representation. The best results are achieved on large datasets.

Therefore, Beal et al. used the ViT[25]. model as the feature extraction network andpropo- sed the ViT-FRCNN[32]. model for object detection. By adding a ViT with a detection task-specific header to detect and locate objects in the image, it is shown that the ViT can transfer the learned classificationrepre-sentation to the object detection task.

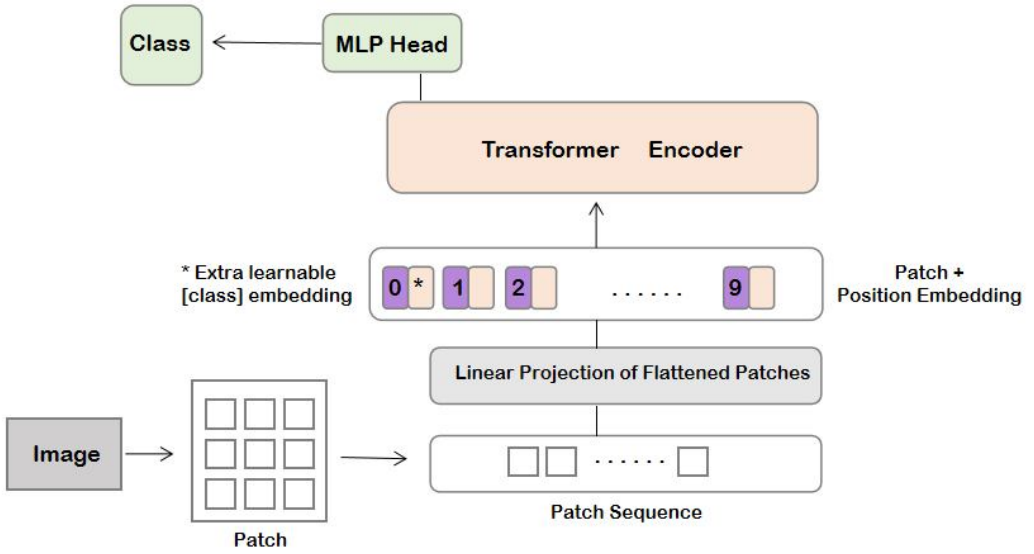

Figure 8: Architecture of ViT

In 2021, Ze Liu et al. Proposed SwinTransf-ormer[53]. model, which is characterized by the introduction of hierarchy, locality and translation invariance

into Transformer network structure design. Compared with the previous application of transformer in images, Swin transformer has more Conv's shadow. The ViT model divides images into patches, and the dimensions of feature ma-ps in the whole network will not change. A key design of the SwinTransformer model is to partition continuous self attention layers with moving windows. The shifted windows connect the windows on the upper layer, providing the connection between them and effectively enhancing the modeling capability. At the same time, the self - attention is calc-ulated in local non-overlapping windows; This design makes the complexity change from the previous square relationship with image size to a linear relationship, andmak- es it possible to design a hierarchical overall structure and introduce local priori; Because non overlapping windows are used, different queries will share the same key set during self attention computing, which is more hardware friendly and practical.

4.3 FPT

In 2020, Dong Zhang et al. proposed a Feat-ure Pyramid Transformer[33]. (FPT) model for intensive prediction tasks, and applied Transformer to feature fusion, drawing on the idea of FPN[11]. and combining non-local features and multi-scale features. Three Transformer modes are designed, ST (Self Transformer): feature enhancement for the current layer is the same as non local opera-tion; GT (Grounding Transformer): This is a non local operation in the form of top-down. High level features (small size) are used to enhance low level features respectively; RT (Rendering Transformer): This is a non local operation in the form of bottom up, which uses low level features (large size) toenhan-ce high level features. With top-down and bottom-up interaction, any feature pyramid can be transformed into another feature pyramid with the same size and richersem- antic information.On the MS-COCO test-dev dataset, the percentage gain for object boxes detection is 8.5%, and the mask AP value gain for mask instances is 6.0%.

Table 1 shows the detection results for different transformer-based object detectors mentioned earlier on COCO 2017 val set. Compared with the feature extraction network based on CNN, Transformer can be applied to the object detection model as a new feature extraction network to achieve betterobject detection results.

Table 1. Comparison of different transformer-based object detectors on COCO 2017 val set.

Method	Epochs	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
DETR-R50	500	42.0	62.4	44.2	20.5	45.8	61.1
Deformable DETR	50	43.8	62.6	47.7	26.4	47.1	58.0
ACT-MTKD(L=32)	-	43.1	-	-	22.2	47.1	61.4
UP-DETR	300	42.8	63.0	45.3	20.8	47.1	61.7
Efficient DETR	36	45.1	63.1	49.1	28.3	48.4	59.0
ViT-B/16-FRCNN	21	36.6	56.3	39.3	17.4	40.0	55.5
Swin-T+RetinaNet	12	41.5	62.1	44.2	25.1	44.9	55.5
Swin-T+ATSS	36	47.2	66.5	51.3	-	-	-
FPT+BFP	-	42.6	62.4	46.9	24.9	43.0	54.5

5.CONCLUSION

As an important research task in the field of computer vision, object detection has a wide range of application scenarios. This paper lists the classical vehicle detectionalgorith-ms, and expounds their advantages and problems. As an emerging architecture,visio-n Transformer has its unique advantages compared with CNN, but it still has many limitations, such as large number of model parameters, high computational complexity, high hardware requirements during training, and long training time, so it has a hugeimpr-ovement space and development potential.

Under the support of deep learning, vehicle detection has made great progress, but the current detector performance has not reached the ideal state, how to strike a better balance between detection accuracy and speed, etc., is the future research direction of detection technology.

REFERENCES

[1]. Viola P，Jones M．Rapid object detection using a boosted cascade of simplefeatu-res［C］//Proc.of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition．Kauai:IEEE Press，2001:511－518．

[2]. Dalal N，Triggs B．Histograms of oriented gradients for human detection［C］//Proc．of the2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition．San Diego:IEEE Press，2005:886－893．

[3]. Felzenszwalb P F，Mcallester D A，Ramanan D A．Discriminatively trained，multi scale，deformable part model［C］//Proc．of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Anchorage:IEEEPress，2008:1－8．

[4]. Krizhevsky A，Sutskever I，Hinton G E．ImageNet classification with deep convolutional neural networks［C］//Proc．Of the International Conference on Neural Information Processing Systems，2012:1097－1105．

[5]. ZHOUFei-yan，JINLin-peng，DONGJun．Review of convolutional neural network［J］．Chinese Journal of Com-puters，2017，40(6) :1229－1251．(in Chinese)

[6]. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[7]. GIRSHICK R，DONAHUE J，DARRELL T，et al.Richfeature hierarchies for accurate object detection and semanticsegmenta-tion[C]//Proceedings of the IEEEConfere-nce on Computer Vision and Pattern Recognition，2014：580-587.

[8]. HE K，ZHANG X，REN S，et al.Spatial pyramid poolingin deep convolutional networks for visual recognition[J].IEEE Transactions on Pattern Analysis and Machine Intelligence，2015，37（9）：1904-1916.

[9]. GIRSHICKR.FastR-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision，2015：1440-1448.

[10]. REN S，HE K，GIRSHICK R，et al.Faster R-CNN：Towardsrealtime object detection with region proposal networks[J]. Advanc-es in Neural Information Processing Syst-ems，2015，28：91-99.

[11]. Lin T Y，Dollar P，Girshick R，et al. Feature pyramid networks for object detection［C］//Proc．of the 2017 IEEE Computer Society Conference on Compu-ter Vision and Pattern Recognition.Honolulu:IEEE Press，2017:936 －944．

[12]. EVERINGHAMM，VANGOOLL，

WILLIAMSCK，etal．The pascal visual object classes(VOC) challenge[J]. Intern-ational Journal of Computer Vision，2010，88(2) :303－38 .

[13]. SIMONYANK，ZISSERMANA. Very deep convolutional networks for large scale image recognition［A］. International Conference on Learning Representations［C］. USA:IEEE，2015. 714－723.

[14]. Redmon J，Divvala S，Girshick r，et al．You only look once:unified，real-time object detection［C］//Proc. of the 2016IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Las Vegas:IEEE Press，2016:779－788.

[15]. Redmon J，Farhadi A. Yolo9000:Better，faster，stronger[C]//Pro. of the 2017 IEEE Computer Society ConferenceonComput-er Vision and Pattern Recognition. Honolulu: IEEE Press，2017:6517 －6525.

[16]. Redmon J, Farhadi A. Yolov3: An incremental improvement[J].arXiv preprint arXiv:1804.02767, 2018.

[17]. Bochkovskiy A，Wang C Y，Liao H. Yolov4:optimal speedand accuracy of object detection［J］. arXiv preprint arXiv:2004.10934，2020.

[18]. Heng Ge,SongtaoLiu,Feng Wang, Zeming Li,and Jian Sun. YOLOX:Exceeding YOLO series in 2021. arXiv preprint arXiv:2107.08430,2021.

[19]. Chuyi Li, Lulu Li, Hongliang Jiang, et al. YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications. arXiv preprint arXiv:2209.02976,2022.

[20]. Chien-Yao Wang, Alexey Bochkovski-y,Hong-Yuan Mark Liao. YOLOv7:Traina-ble bag-of-freebies sets new state-of-the-art for real-time objectdetec- tors. arXiv preprint arXiv:2207.02696, CVPR 2022.

[21]. Liu W，Anguelov D，Erhan D，et al．SSD: single shot multi-box detector［C］//Proc． of the European Conference onComputer Vision．Amsterdam: Springer Press，2016:21－37.

[22]. Jeong J，Park H，Kwak N．Enhancement of ssd by concatenating feature maps for object detection［C］//Proc． of the British Machine Vision Conference． London: BMVAPress，2017.

[23]. NEUBECKA，VANGOOLL．Efficient non-maximum suppression[A]. Proceedi-ngs of the International Conference on Pattern Recognition[C]，USA:IEEE，2006．850－855.

[24]. LIN T Y，MAIRE M，BELONGIE S，et al.Microsoft COCO：common objects in context［C］//Proceedings of European Conference on Computer Vision.Berlin，Germany：Springer，2014：740-755.

[25]. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[26]. Carion N，Massa F，Synnaeve G，et al．End-to-end object detection with transformers［C］//Proc． of the European Conference on Computer Vision. Glasgow:Springer Press，2020:213 －229.

[27]. Zhu X，Su W，Lu L，et al． Deformable detr:Deformable transformers for end-to-end object detection［J］. arXiv preprint arXiv:2010．04159，2020.

[28]. Zheng M，Gao P，Wang X，et al．End-to-end object detection with adaptive clustering transformer[J]. arXiv preprint arXiv:2011．09315，2020.

[29]. Datar, M., Immorlica, N., Indyk, P., and Mirrokni, V. S. Locality sensitive hashing scheme based on pstable distributions. In Proceedings of the twentieth annual symposium on Computational geometry (2004), ACM, pp. 253-262.

[30]. Dai Z, Cai B, Lin Y, et al. Up-detr:

Unsupervised pretraining for object detection with transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 1601-1610.

[31]. Dai J, Qi H, Xiong Y, et al. Deformable convolutional networks[C]//Proceedings of the IEEE international conference on computer vision. 2017: 764-773.

[32]. Josh Beal, Eric Kim, Eric Tzeng,Dong Huk Park,et al. Toward Transformer-Based Object Detection. arXiv preprint arXiv:2012.09958,2020.

[33]. Zhang D, Zhang H, Tang J, et al. Feature pyramid transformer[C]//European Conference on Computer Vision. Springer, Cham, 2020: 323-339.

[34]. HE K M，GKIOXARRI G，DOLL R P，et al．Mask R-CNN［C］//Proceedings of the IEEE International Conferenceon Compu-ter Vision，2017:2961－2969.

[35]. LI Z X，ZHOU F Q．FSSD:Feature fusion single shot multibox detector［J］. arXiv preprint，arXiv:1712．00960，2017.

[36]. LeCun Y, Boser B, Denker J S,Henders-on D, Howard R E,Hubbard W, et al. Backpropagation applied to handwritten zipcode recognition. Neural Computation, 1989, 1(4): 541−551.

[37]. He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learningfor image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).Las Vegas, USA: IEEE, 2016. 770−778.

[38]. Bahdanau D, Cho K, Bengio Y. Neural machine translation byjointly learning to align and translate. In: Proceedings of the3rd International Conference on Learning Representations. SanDiego, USA, 2015.

[39]. Gehring J, Auli M, Grangier D, Yarats D, Dauphin Y N. Convolutional sequence to sequence learning. In: Proceedings of the34th International Conference on Machine Learning. Sydney,Australia: JMLR.org, 2017. 1243−1252.

[40]. Jozefowicz R, Vinyals O, Schuster M, Shazeer N, Wu Y H. Exploring the limits of language modeling. arXiv preprint arXiv:1602.02410, 2016.

[41]. Luong T, Pham H, Manning C D.Effec-tive approaches to attention-based neural machine translation. In: Proceedings ofthe 2015 Conference on Empirical Meth-ods in Natural Language Processing. Lisbon, Portugal: The Association for Computational Linguistics, 2015. 1412−1421.

[42]. Han K, Wang Y H, Chen H T, Chen X H, Guo J Y, Liu Z H,et al. A survey on vision transformer. IEEE Transactions onPattern Analysis and Machine Intelligence, DOI: 10.1109/TPAMI.2022.3152247.

[43]. Liu Y, Zhang Y, Wang Y X, Hou F, Yuan J, Tian J, et al. A survey of visual transform-ers. arXiv preprint arXiv:2111.06091, 2021.

[44]. Khan S, Naseer M, Hayat M, Zamir S W, Khan, F S, Shah M.Transformers in vision: A survey. arXiv preprint arXiv:2101.01169, 2021.

[45]. Selva J, Johansen A S, Escalera S, Nasrollahi K, Moeslund TB, Clapés A. Video transformers: A survey. arXiv preprint arXiv: 2201.05991, 2022.

[46]. Shamshad F, Khan S, Zamir S W, Khan M H, Hayat M, KhanF S, et al. Transformers in medical imaging: A survey. arXiv preprint arXiv: 2201.09873, 2022.

[47]. Wang W H, Xie E Z, Li X, Fan D P, Song K T, Liang D, et al.Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the 2021IEEE/CVF

International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021. 548−558.

[48]. Xie E Z, Wang W H, Yu Z D, Anandkumar A, Alvarez J M,Luo P. SegFormer: Simple and efficient design for semantic segmentation with transformers. arXiv preprint arXiv:2105.15203, 2021.

[49]. Cheng B W, Misra I, Schwing A G, Kirillov A, Girdhar R.Masked-attention masktran-sformer for universal image segmentation. arXiv preprint arXiv: 2112.01527, 2021.

[50]. Zhou L W, Zhou Y B, Corso J J, Socher R, Xiong C M. End-to-end dense video captioning with masked transformer. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE,2018. 8739−8748.

[51]. Zeng Y H, Fu J L, Chao H Y. Learning joint spatial-temporaltransformations for video inpainting. In: Proceedings of the16th European Conference on Comp-uter Vision. Glasgow, UK:Springer, 2020. 528−543.

[52]. Jiang Y F, Chang S Y, Wang Z Y.Trans-GAN: Two transformers can make one strong gan. arXiv preprint arXiv:2102. 07074, 2021.

[53]. Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada:IEEE, 2021. 9992−10002.

[54]. Yao Z Y, Ai J B, Li B X, Zhang C. Efficient DETR: Improving end-to-end object detector with dense prior. arXiv preprint arXiv: 2104.01318, 2021.

[55]. Han Q, Fan Z J, Dai Q, Sun L Cheng M M, Liu J Y, et al. Demystifying local vision transformer: Sparse connectivity, weightsharing, and dynamic weight. arXiv preprint arXiv: 2106.04263,2021.

[56]. Buades A, Coll B, Morel J M. A non-local algorithm for image denoising. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR"05). San Diego, USA: IEEE, 2005. 60−65.

[57]. Wang X L, Girshick R, Gupta A, He K M. Non-local neuralnetworks. In:Proceedin-gs of the 2018 IEEE/CVF Conferenceon Computer Vision and Pattern Recognition. Salt Lake City,USA: IEEE, 2018.7794- 7803.