Multimedia Tools and Applications (2023) 82:21353–21383
https://doi.org/10.1007/s11042-022-13801-3

A survey: object detection methods from CNN to transformer

Ershat Arkin[1] · Nurbiya Yadikar[1] · Xuebin Xu[1] · Alimjan Aysa[2] · Kurban Ubul[1,2]

Received: 11 March 2022 / Revised: 19 July 2022 / Accepted: 5 September 202 /
Published online: 21 October 2022

Abstract

Object detection is the most important problem in computer vision tasks. After AlexNet proposed, based on Convolutional Neural Network (CNN) methods have become mainstream in the computer vision field, many researches on neural networks and different transformations of algorithm structures have appeared. In order to achieve fast and accurate detection effects, it is necessary to jump out of the existing CNN framework and has great challenges. Transformer's relatively mature theoretical support and technological development in the field of Natural Language Processing have brought it into the researcher's sight, and it has been proved that Transformer's method can be used for computer vision tasks, and proved that it exceeds the existing CNN method in some tasks. In order to enable more researchers to better understand the development process of object detection methods, existing methods, different frameworks, challenging problems and development trends, paper introduced historical classic methods of object detection used CNN, discusses the highlights, advantages and disadvantages of these algorithms. By consulting a large amount of paper, the paper compared different CNN detection methods and Transformer detection methods. Vertically under fair conditions, 13 different detection methods that have a broad impact on the field and are the most mainstream and promising are selected for comparison. The comparative data gives us confidence in the development of Transformer and the convergence between different methods. It also presents the recent innovative approaches to using Transformer in computer vision tasks. In the end, the challenges, opportunities and future prospects of this field are summarized.

Keywords Computer vision · Object detection · Real-time system · CNN · Transformer

✉ Kurban Ubul
kurbanu@xju.edu.cn

[1] College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China

[2] The Key Laboratory of Multilingual Information Technology, Xinjiang University, Urumqi 830046, China

1 Introduction

Object detection is the most important research direction in many computer vision (CV) tasks. The task and goal correctly classify the objects category and location in the given picture with rectangular bounding box.

According to the development process, object detection can be divided into two stages, first is the traditional object detection method, and second is the current CNN algorithms. The traditional object detection models such as VJ detector [83], HOG [15] etc. are not good enough, calculation amount is huge, the calculation speed is slow, and it may produce multiple correct recognition results instead of the results we want, so they no longer meet the current needs.

Researchers have proposed many innovative new detection methods and these methods also brought new problems. For example, the detection speed of the two-stage object detection method is slow, the detection accuracy of the one-stage object detection method is low, and the transformer-based detection method requires a lot of training data. It is important for researchers to sort out how these limitations come about and how ideas for improvement come about. This is also the core idea of our survey paper. By reviewing, discussing and comparing these detection methods in the paper, hope to researchers can have better understanding on these detection methods and provide new ideas for proposing new methods.

At present, the CNN algorithm has an absolute position in both the general target detection method and the special object target detection method. Almost all object detection algorithms use CNN as their backbone until this article [18] was published. It is the first to use the transformer, which is widely used in NLP, directly in image processing without any modification. Since then, a new path has been opened up for computer vision, and it can be seen from recent research work that this method of transformer used in image processing has a tendency to completely replace convolutional neural networks.

The paper investigated total of 237 papers from the upstream task model of feature extraction to the downstream task model of object detection, observed the problems actually solved by these papers and classified them differently, after discussion, finally according to our purpose, that is, help readers to developed more efficient methods in future and consider the length of the paper, 98 papers are referenced. It also investigates different literatures, data and experiments, compare different detection algorithms, and conduct comparative experiments on different mainstream algorithms. Article structure shown in Fig. 1.

The historical development and technical challenges of object detection introduced in the second part of the article. Different state-of-the-art (SOTA) algorithms used CNN and Transformer listed and discussed in Section 3. The fourth section mainly introduces the comparative analysis between different types of object detectors. In the fifth section looks forward to the future research direction of target detection, and paper is concluded in last section.

2 Background

Object detection is an extension of object classification in CV. It not only needs to classify the target in an image but also locate the object. Therefore, computer vision tasks such as instance segmentation [30], object tracking based on frames [53] or temporal information [37] are all based on object detection.

Nowadays, with the continuous deepening of related technologies and increasingly powerful performance of hardware devices, the application of object detection can be seen in

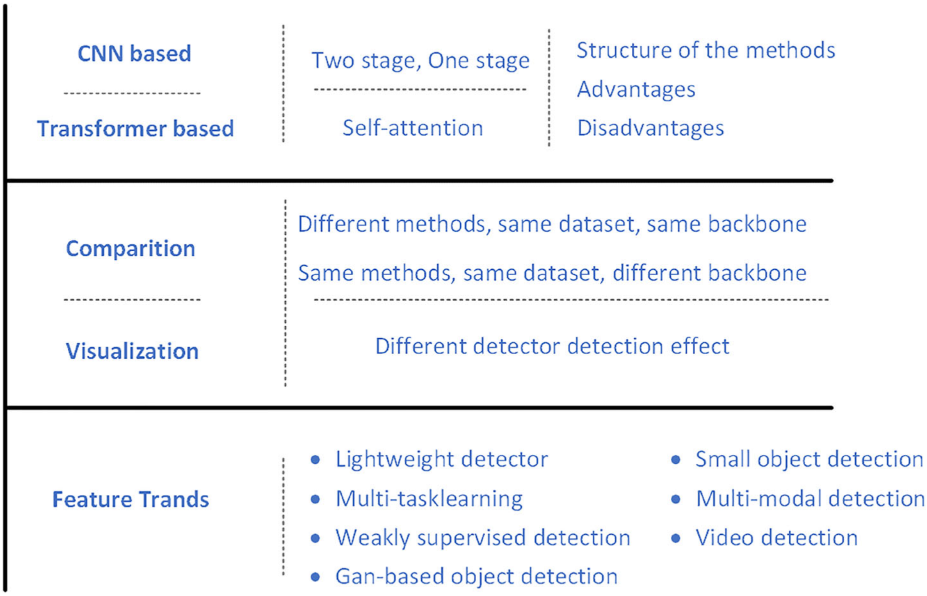


Fig. 1 Structure of the paper

industry [75] and daily life. The object detection can achieve such achievements, it is inseparable from the previous research work. Figure 2 shows key milestone of the object detection methods. This figure was made using the improvement of the figure in [1].

Although the existing object detection methods have made great advance, there are still many challenges [49] as follows:

- Multi-scale object detection model:

The scales of different object instances in real scene images often vary greatly. Among them, large-scale objects have the characteristics of large area and rich features and are easy to be detected, while dense object objects and small-scale objects are difficult to be detected due to

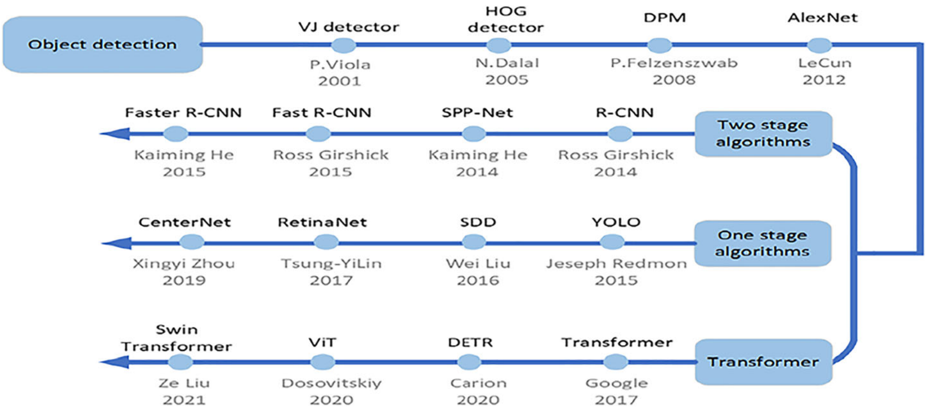


Fig. 2 Classical methods in the development history of target detection and their proposed time

fewer features available. The multi-scale object detection methods' [4, 55] goal is to detect all objects of different scales in an image.

- Real-time object detection model:

In industrial production such as self-driving cars, real-time requirements for object detection are put forward. The real-time object detection model needs to consider the memory requirements, computation cost, balance the detection performance and real-time requirements. In this regard, in except the YOLO series [2, 65–67] that have achieved good results, there are other algorithms that have made efforts in lightweight networks [8, 32, 33, 35, 71, 91].

- Weakly-supervised detection model:

This type of object detection tasks usually refers to the training samples that only give image-level labels and lack the object boundary annotation box. The performance of these detection models is largely determined by the number of labeled training samples. However, collecting bounding annotation boxes of objects to be detected in images is a time- and labor-intensive task, and in some special cases, it is hard to get supervise training samples. Moreover, in real life, due to the small number of rare object classes, there are few or no such samples in our training set, which makes the model unable to detect these classes. Therefore, some researchers have conducted Weakly Supervised Object Detection research [85], Few-Shot Object Detection [38] and Zero-Shot Object Detection research [64].

- Solve imbalance of training samples in object detection model:

Unbalanced training samples is a challenge in object detection research. Random sampling from an image will produce a lot of negative samples. Those negative and positive samples are extremely imbalanced. In addition, the difficulty of detection and recognition of training samples is not the same, so there will be an imbalance of difficult and easy samples. In the detection of multiple types of targets, some rarer samples are more difficult to obtain, causing the imbalance of sample classes. In recent years, many researches [5, 59, 73] try to solve the problem of unbalanced training samples in the field of target detection by improving the sampling method of training samples or adjusting the weight of samples in the loss function, and by studying the relation between samples.

3 Object detection

In this section, the paper will introduce different object detection methods. Object detection algorithms divided into two types according to the feature extraction methods used, one is a CNN-based object detection method, and the other is a transformer-based object detection method. Each type contains a different method of object detection, and in order to allow the reader to have a better understanding of these methods and more thinking about object detection, The paper list the advantages, disadvantages and areas for improvement of each method.

3.1 Convolutional neural networks used on object detection

After the great success of deep learning, many researchers have focused on object detection models using CNN. There have been many object detection models with simple structures and good effects, which make these models can be widely used.

Detection models need to solve two core problems, one is object classification, and the other is object positioning. The detection system first determines whether there are objects of interest and identifies the object category, and output the most likely category scores of these objects. Object positioning problem is based on the image classification. We also want to know where the object is in the image. Use rectangular boxes to frame the identified objects, for example, the rectangular box is given by determining the coordinates of the upper left and lower right corners, or by determining the upper left coordinate and rectangular boxes' length and width.

In CNN, definition of a feature map is a new image obtained by applying a convolution kernel on an image. There are as many feature maps as convolution kernels in a network layer. Size of the receptive field on the feature map is equivalent to how large the area of original image is affected by the pixels in the high-level feature map.

Object detection algorithms roughly divided into two types: one stage object detection and two stage object detection.

3.1.1 Two stage algorithms (candidate-based algorithms)

Two stage algorithms require two stages to perform object detection tasks. The two stage algorithms first extracts candidate regions from the image, these called region proposal, and then use CNN to classify and locate the region proposal to get the results. Two stage algorithms results' accuracy relatively high, but the detection speed is slow. R-CNN, SPP-Net, Fast R-CNN, Faster R-CNN are typical two stage object detection algorithms. It will also analyze these classic algorithms from different perspectives.

The detection steps of the two-stage algorithms are mainly as follows: first extract features, second extract region proposal, and third is classification, positioning.

1) R-CNN (Region-CNN)

After R-CNN [26] was launched, an improved R-CNN series appeared on this basis. We can say R-CNN is the earliest successful algorithm, which used deep learning in object detection. Although R-CNN has successfully used deep learning for object detection tasks, it is actually more accurate to say that it combined the deep learning methods and traditional methods. In the development of R-CNN, the traditional method of selective search [81] is used to extract region proposal, and SVM is used for this region proposal classification.

Figure 3 shows that R-CNN detection divided into four:

1. Receive input images.
2. Extract about 2000 region proposals.
3. Input region proposals and compute CNN features.
4. Use support vector machines (SVM) to determine classification.

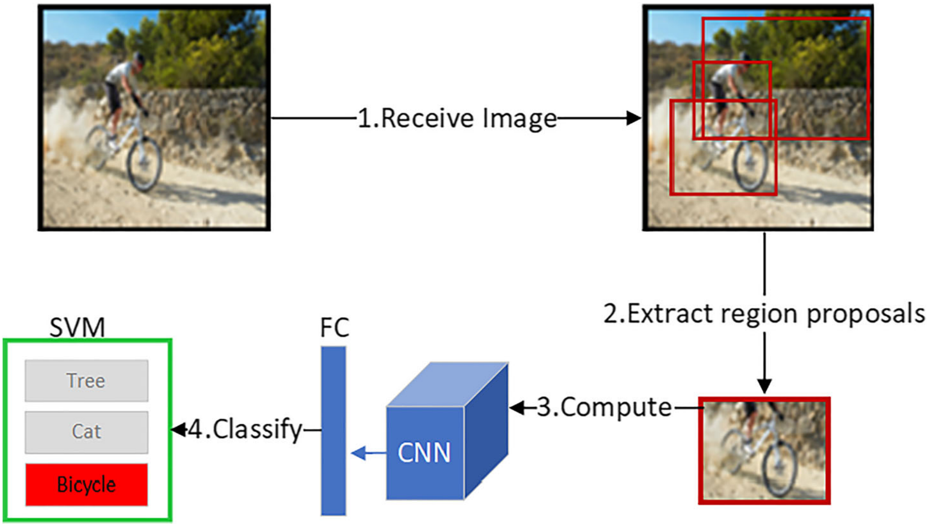


Fig. 3 The four steps of R-CNN [26]

About the extraction of feature maps, the training and testing of models and other issues will not be introduced in our paper.

Although R-CNN achieved good results at the time, it also had many disadvantages:

- Training time is long: Training is carried out in multiple stages, and the corresponding feature map of each different candidate region is computed separately.
- Space occupation is large: Feature map of each candidate area is saved for subsequent operations, which will lead to a large amount of space occupation.
- Long test time: The calculation of the feature map of each region proposal is calculated separately and not shared. Therefore, the amount of calculation for each picture during the test is also huge, and it takes a long time.

2) SPP-NET (Spatial Pyramid Pooling Network)

Aiming at a series shortcoming of R-CNN, for instance: input images are fixed size images, and the feature extraction is inefficient. He et al. proposed SPP-Net [29]. Figure 4 shows that SPP-Net does not require a fixed size for the input images, but feeds entire image to the convolutional layer.

The introduction of the SPP layer is as follows:

Figure 5 shows the structure of a SPP-Net. As shown in the figure above, a black feature map will be obtained after a picture passed a convolutional layer. Specifically, the input feature map is divided into blocks of different scales, such as 4*4, 2*2, and 1*1 in the three different scales (scale of the block is not solid), and then max pooling is used in each block, and then it becomes a vector with a fixed length of 16 + 4 + 1. Therefore, after images of different scales are input, they will get vectors of the same length after passing through the SPP layer.

Although SPP-net has made many improvements to R-CNN, it still inherits the multi-stage process of R-CNN. For example, use the selective search method to extract 2000 candidate regions, use CNN network to extract the feature map, and input extracted features into the

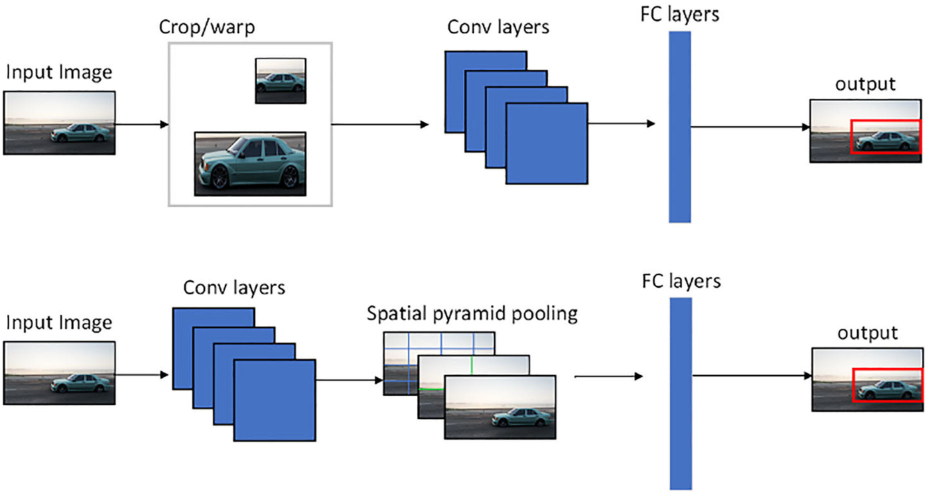


Fig. 4 SPP-Net directly inputs the image to the convolutional layer without cropping or wrapping [29]

SVM for classification, etc. However, it also pointed out some shortcomings of R-CNN, as summarized at the beginning, and proposed new methods for these shortcomings. The difference in their operating procedures can be seen more intuitively in Fig. 6.

The innovation of SPP-Net:

- Input a whole picture directly into the convolutional neural network, so as to avoid the repetition of feature extraction separately for every region proposal.
- After using the SPP-Net layer, model can accept different scale of images.

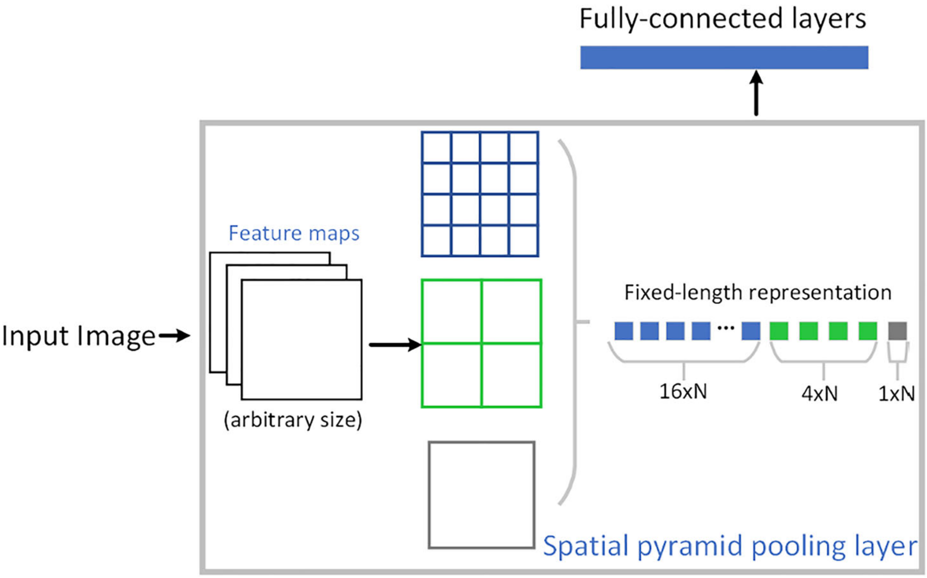


Fig. 5 Structure of a spatial pyramid pooling layer. Here N represent for the filter number of the convolutional layer. [29]

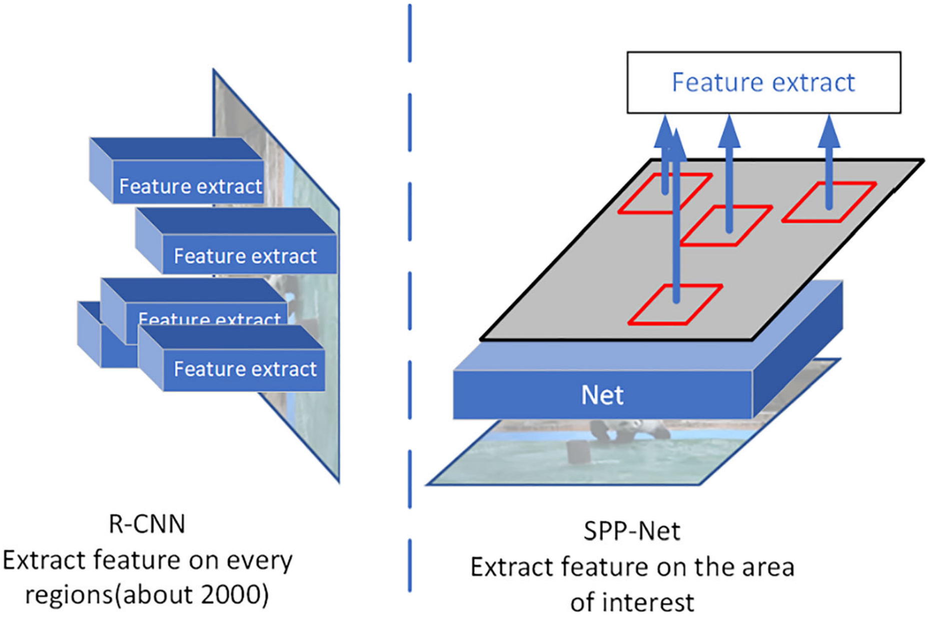


Fig. 6 Difference between R-CNN and SPP-Net is that the former applies feature extraction on 2000 candidate regions, while the latter directly apply feature extraction on images [29]

Limitations of SPP-Net:

SPP-Net is difficult to fine-tune the parameters of the network before the SPP-layer, so that the efficiency becomes very low, when doing fine-tuning each training sample (RoI: region of interest) comes from a different image, the backpropagation efficiency of the SPP layer become very low, so that it re-produce a new feature map for each image, which is inefficient, and Fast-RCNN has improved it.

3) Fast-RCNN

Although SPP-Net solves some of the shortcomings of R-CNN, SPP-Net inherits many methods of R-CNN, so it still has some problems to be solved, such as training steps are too many (including train the SVM classifier) and occupation of hardware space is still big. So as to solve the problems of slow R-CNN training speed and large space requirement, the author of R-CNN, improved R-CNN and proposed Fast R-CNN [25], which absorbed the characteristics of SPP-Net. These modifications greatly improve the detection speed. Its architecture shown in Fig. 7.

Fast R-CNN algorithm achieved amazing results. Compared with Faster R-CNN and SPP-Net, the training speed is 9 times faster than the former and 3 times faster than the latter, test speed also much faster than both.

Fast RCNN has the following improvements over RCNN:

- Change the last convolutional layer to region of interest pooling Layer.
- Multi-task Loss is proposed, which directly adds the bounding box regression to the network for training, and includes the region proposal classification loss and position regression loss.

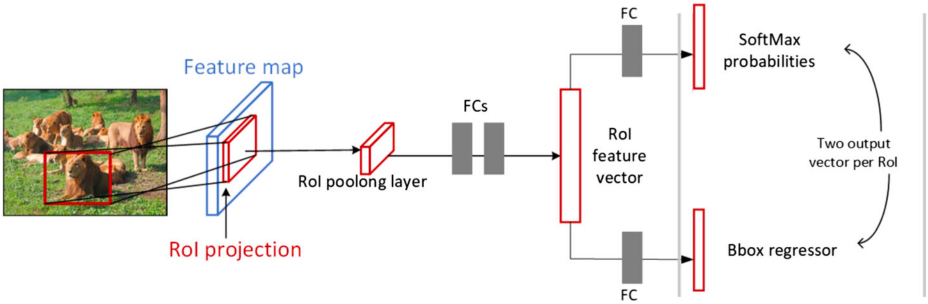


Fig. 7 Architecture of Fast R-CNN [25]

- Fast RCNN no longer uses support vector machine (SVM) but uses SoftMax for classification, and at the same time uses Multi-task Loss to add bounding box regression to the network, so that entire training only includes the two stages of extracting candidate regions and convolutional neural network training.

Limitations of Fast R-CNN:

The main limitation is that the extraction of the region proposal uses selective search, which is still obtained from outside the network. Most of the object detection time is spent on region proposal, which is also one of the improvement directions of the Faster RCNN.

4) Faster R-CNN

From the release of Faster R-CNN [68] in 2015 to the present, its performance is still very good. Instead of using a selective search algorithm, Region Proposal Network (RPN) used in this algorithm. This RPN helps to reduce irrelevant region proposals to a large extent, and this makes the image processing speed of the model greatly improved.

RPN will let this model directly use convolutional neural networks to generate region proposal. Object boundary and score of each boundary box will be predict at the same time. RPN can be trained with Fast R-CNN at the same time to achieve end-to-end optimization. Process of Faster R-CNN shown in Fig. 8 [1].

Faster R-CNN even compared with recent algorithms, it still a good algorithm, but there are also the following improvements:

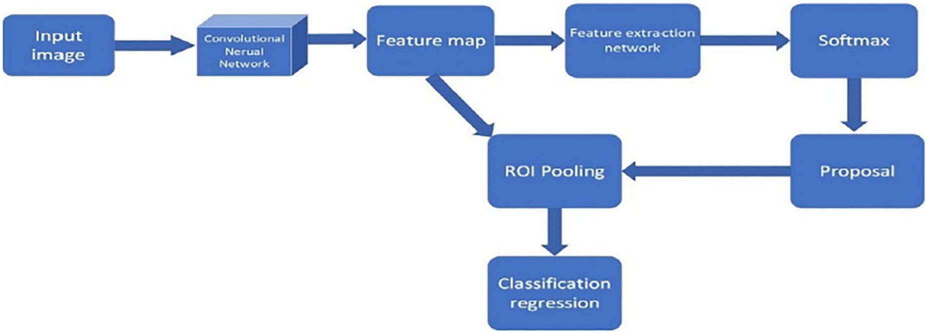

Input image
Convolutional Nerual Network
Feature map
Feature extraction network
Softmax
Proposal
ROI Pooling
Classification regression

Fig. 8 Detection process of Faster R-CNN [1]

1. To classify each region proposal, large calculation is still required.
2. Compared with the previous object detection algorithms, although its speed has been improved a lot, it still cannot achieve real-time detection results.

Limitations of Faster R-CNN:

- Relative to the one-stage algorithms, the speed is slower.
- The background false detection rate is high; only some negative samples are used for training, and the background learning is relatively not very sufficient.
- The Non-Maximum Suppression (NMS) is not friendly to occluded objects.

3.1.2 One stage algorithms (regression based algorithms)

The one-stage algorithms directly generate the positioning coordinates and classification probability of objects in the image, without the need to generate region proposal in advance. Because there is one less computationally intensive step, their detection speed is very fast, and the detection result can be obtained directly. Their representative algorithms will be introduced below.

1) YOLO series

This series is just like its name, You Only Look Once (YOLO) [67], this algorithm can detect the target in the image at an extremely fast speed by just looking at the image once.

Its central idea is to combine the two steps of generate region proposal and detection in the object detection task, and then treat it as a regression problem to solve. Therefore, the YOLO series of algorithms can know the target classification and positioning in the image at one time.

- YOLOv1

If look at the structure of Yolo as a whole, we find that the structure is not very complicated, that is, convolution, pooling, and full connections are added at the end. So, it is similar to the ordinary CNN object classification model. The most obvious difference is that linear function used as activation function at final output layer, because it not only to predict the class of the object, but also to locate the position of the bounding box. The structure shown in Fig. 9.

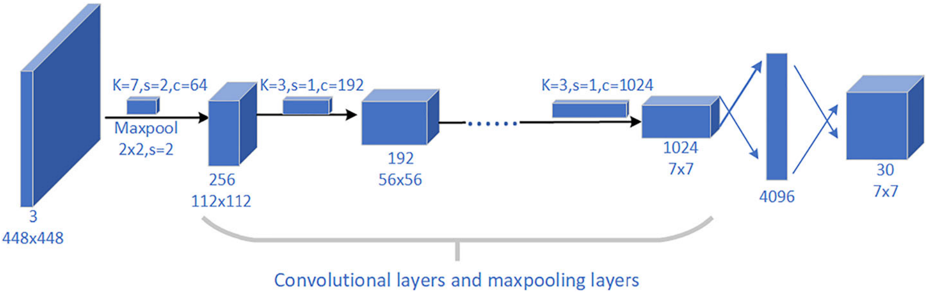


Fig. 9 Structure of YOLOv1 almost has no difference from other ordinary classification network [67]

YOLOv1 has the following advantages:

1. Pioneeringly transform the detection task into a regression problem. Classification and positioning proceed at the same time.
2. The model has strong generalization and can be extended to other fields.

YOLOv1 also has the following shortcomings:

1. If there are multiple objects that are close together, or the objects are relatively small, the detection effect will be poor.
2. Poor generalization ability for different shapes of the same object.
3. Because of the loss function, the main reason that affects the detection effect is positioning error.

- YOLOv2

Redmon et al. [65] released a new version of YOLOv2. They designed a brand new backbone network called Darknet-19 which included 19 convolutional layers and 5 max pooling layers. YOLOv2 no longer uses dropout and added batch normalization layer after each convolutional layer. A union training method for detection and classification is proposed. A different model called YOLO9000, which can detect around 9000 types of objects is trained on COCO and ImageNet dataset under this union training method. Therefore, this article actually contains two models: YOLOv2 and YOLO9000, but the latter is based on the former, and the main structure of the two models is the same.

- YOLOv3

The biggest changes in YOLOv3 [66] include two points: the use of residual module and FPN architecture. The backbone of YOLOv3 is a residual model which called Darknet-53, because it contains 53 convolutional layers. From perspective of the network structure, compared with Darknet-19 network, it can be built deeper. Another point is to use FPN architecture to achieve multi-scale detection.

- YOLOv4

The YOLOv4 [2] algorithm follows the original YOLO framework and adopts the best optimization strategy in recent years. It improves in many ways, like data processing, network training, activation function, loss function, etc. Although there is no theoretical innovation, it is welcomed by many engineers.

The following three points are its contribution:

1. Contributed a high-efficiency and high-accuracy object detection model that can also be trained on ordinary consumer-grade GPU.
2. Verify the impact of a series of SOTA target detector training methods.
3. The effect has reached a new benchmark for target detection that achieves a balance between FPS and Precision.

Limitations of different versions of the YOLO series:

- YOLOv1: Detection of small targets is not very good; Each cell can only generate 2 boxes and can only have one class; Low recall rate.
- YOLOv2: Although many tricks are used to improve performance, detection effect on small objects is still not well improved.
- YOLOv3: Compared with the R-CNN series, the accuracy is still lacking; At the same output layer, the same anchor, the data that is filled first, will be overwritten by the latter. For example, if there is a cat and dog of similar size in the same location, yolo may only detect one.
- YOLOv4: The performance improvement of this method mainly uses a lot of tricks.

2) SSD (Single Shot Multibox Detector)

After this algorithm [47] was released in 2016, its performance surpassed YOLOv1. When this algorithm applies the innovative multi-scale feature map to the detection task, the new target detection algorithms [23, 45] proposed since then have more or less used the concept of multi-scale.

The positioning of the object is related to the local information, and the local information is expressed by shallow feature maps. Object classification is related to the segmentation information, and segmentation information is expressed by deep feature maps. Therefore, convolution on multi scale feature maps can raise the detection accuracy.

SSD maintains the accuracy of Fast-RCNN while maintaining the detection speed of YOLO. This is because SSD put regression idea of YOLO and the anchor mechanism of Fast-RCNN in one model, and uses multi-scale regions in different positions of the image for regression.

Based on VGG-16, using the first five convolutions of VGG, 5 convolution structures starting from Conv6 are added, and the input image requires 300*300. Figure 10 shows the structure of SSD.

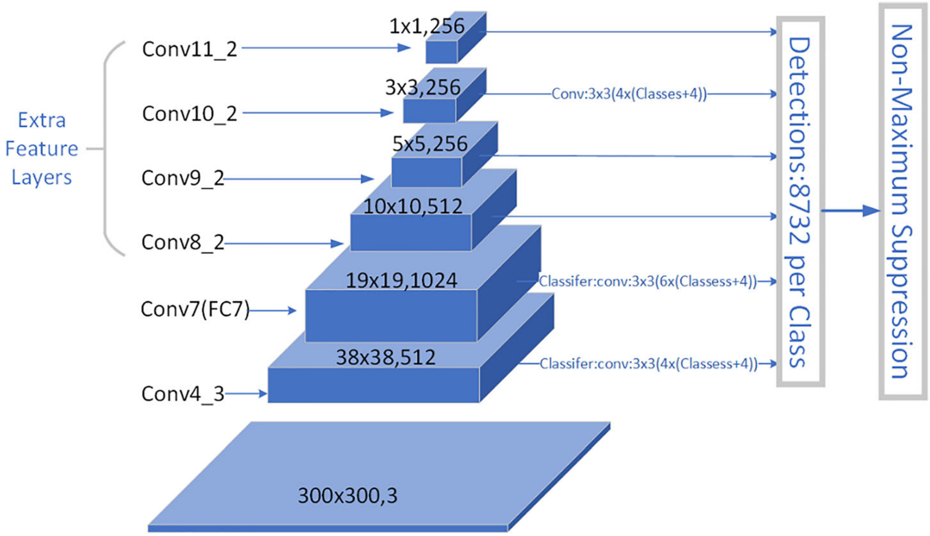


Fig. 10 Network structure of SSD [47]

The results achieved by SSD are very good. When mean average precision (mAP) reaches 74.3%, the speed is also very fast, reaching 59FPS. Although mAP of Faster R-CNN is also very high, the speed is only 7FPS. The mAP of YOLOv1 is 63.4%, and the speed is 45FPS. Even if the resolution of the input picture is low, it can still achieve good results. Experimental results show that the method used by the author does not reduce the accuracy due to the increase in speed. Of course, there are some shortcomings. For example, the default box needs to be manually set, and the detection performance of small objects are not better than Faster R-CNN.

Limitations of SSD:

- The min size, max size and aspect ratio values of the prior box need to be set manually;
- It is still not effective in detecting small objects.

3) RetinaNet [46]

The two stage object detection methods have high accuracy and slow speed. When we speed up the detection speed of one-stage object detection methods, the accuracy will decrease. Therefore, the author brings a new method that can achieve the accuracy of two-stage detection in one stage algorithm.

Low accuracy problem in one-stage object detection methods is caused by the imbalance of positive and negative samples. The Focal loss proposed by the author is a modification of the cross-entropy loss function, which can balance the imbalance of negative and positive samples. Through the experimental data, we can see that Focal loss effectively solves the imbalance of positive and negative samples problem in the one stage object detection, and improves the accuracy of the model.

RetinaNet reached the high accuracy as Faster R-CNN. The formula of Focal Loss is not fixed, it can also have other forms, the performance has no big difference, so the expression of Focal Loss is not critical.

For verify the focal loss, a simple one-stage detector is designed. The network structure is called RetinaNet. It is more like the combination of Resnet + FPN. For feature extraction ResNet selected as the backbone. The role of FPN is to enhance the utilization of multi-scale features formed in Resnet to obtain more expressive feature maps containing multi-scale target region information. Figure 11 shows the basic structure of RetinaNet:

Focal loss has largely alleviated the imbalance between positive and negative samples, but there are also some areas to be improved:

It is susceptible to noise interference, so the correct labeling of samples is very demanding.

4) CenterNet

CenterNet [19], released in April 2019, a new anchor-free detection method proposed on the basis of CornerNet [40], which constructs triples for object detection, and greatly exceeds all existing one-stage methods on the MSCOCO dataset.

In fact, it is difficult to match the upper left corner and lower right corner of the object frame, because for most objects, these two corner points are outside of the object, and the embedding vector of the two corner points can't perceive the internal information of objects very good. Therefore, Cascade Corner Pooling is used instead of the original Corner Pooling. First it extracts the maximum value of the object boundary (Corner Pooling), then continously

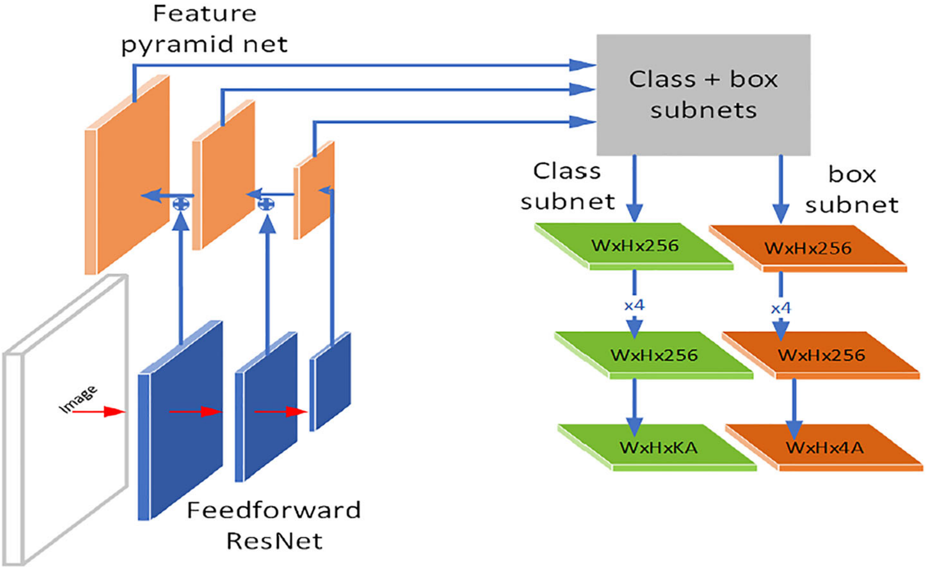


Fig. 11 RetinaNet basic structure [46]

extract the maximum value from the boundary, and adds it to the maximum value of the boundary to combine more internal information.

Compared the AP with other one-stage and two-stage methods in the COCO dataset, we can find that the optimal performance surpasses the current one-stage method and most two-stage methods. There are also some improvements [50, 97] on the basis of CenterNet, and achieved good results. Figure 12 is the Structure of CenterNet.

CenterNet also has disadvantages:

In the training and prediction process, if the center point of two objects overlaps after downsampling, then CenterNet can only detect one center point and identify the two objects as one object.

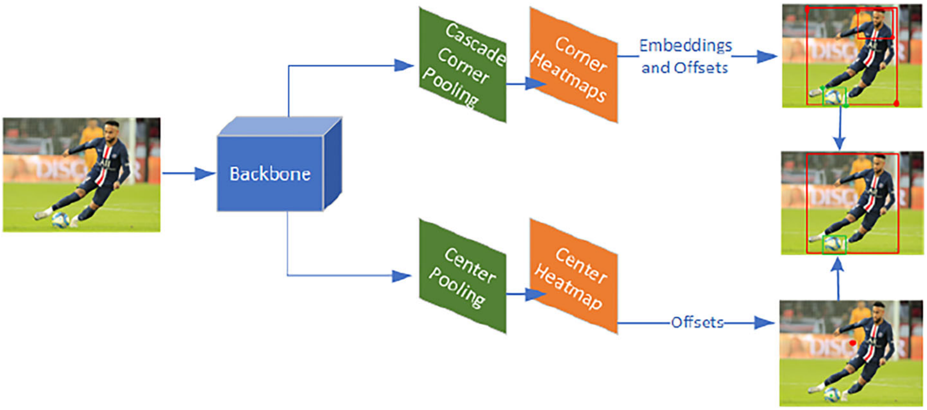


Fig. 12 Structure of CenterNet [19]

3.2 Object detection with transformer

Since AlexNet in 2012, CNN has dominated all directions in the field of CV tasks. With the deepening of research, various methods have emerged. The relationship between the attention mechanism in NLP field and the CNN in CV field is becoming more and more obvious. The use of Transformer [82] to handle various computer vision tasks has become more and more mainstream, and some of the characteristics of Transformer have made up for the shortcomings of CNN.

Several works related to Transformer in the field of object detection will introduce in below:

- The earlier work that used combination of Transformer and CNN for classification and instance segmentation includes DETR [6] and D-DETR [98].
- ViT [18], which used pure Transformer directly for classification, pioneered a Transformer-based computer vision model.
- With the idea of ViT, a new universal backbone, Swin Transformer [51], that uses Transformer to handle computer vision tasks is proposed.

1) Detection Transformer (DETR)

DETR [6] can be regarded as the pioneering work of Transformer in the field of target detection. The result of DETR is very good, and the DETR based on ResNet50 matches the performance of Faster-RCNN after various finetunes.

In fact, the entire DETR architecture is easy to understand. It contains three main components:

a) CNN backbone network
b) Encoder-decoder transformer
c) A simple feedforward networks

The CNN backbone network generates feature maps from the input images. Then, the output of the CNN backbone network is converted into a one-dimensional feature map and passed to Transformer encoder as input. The output of this encoder is N fixed-length embeddings (vectors), where N is the number of objects in the image assumed by the model. The Transformer decoder use the encoder-decoder attention mechanism to decodes these embeddings into bounding box coordinates.

Finally, normalized center coordinates, height and width of the bounding box predicted by feedforward neural network, and linear layer uses the SoftMax function to predict the category label. The above flow and DETR structure shown in Fig. 13.

Advantages of DETR:

1. Propose different path for object detection and need less prior information.
2. Although the accuracy and efficiency are not the highest, they are comparable to the highly optimized Faster R-CNN on COCO dataset. The detection effect will be better on large object.
3. DETR does not require any custom layers, so the model is easy to rebuild, and the modules involved can be found in any deep learning framework.

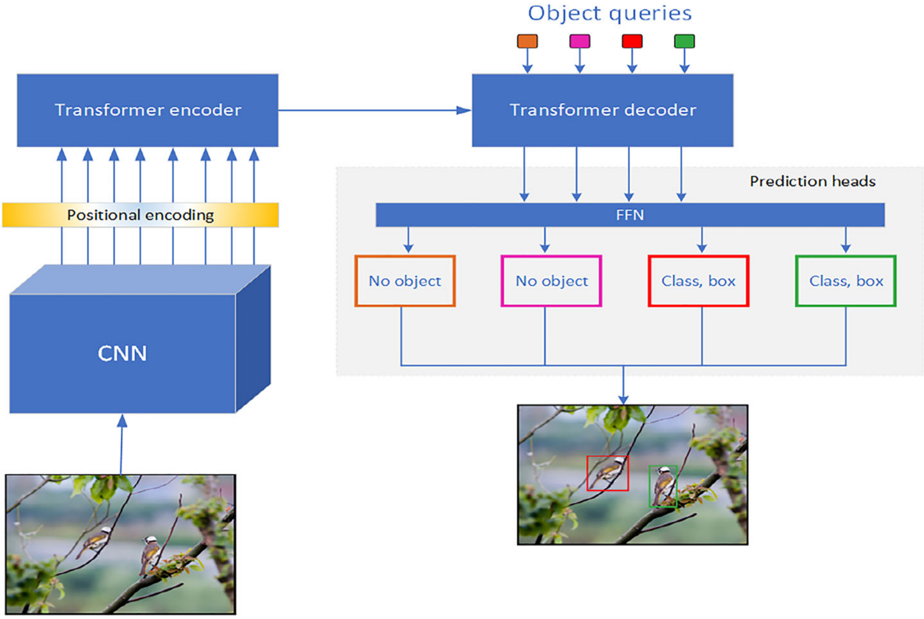


Fig. 13 The process of DETR and its structure [6]

The disadvantages are also obvious:

DETR has problems such as slow convergence speed, poor detection accuracy, and low operating efficiency

2) Vision Transformer (ViT)

The author tries to apply the standard Transformer directly to the image with minimal modification. The framework of ViT shown in Fig. 14.

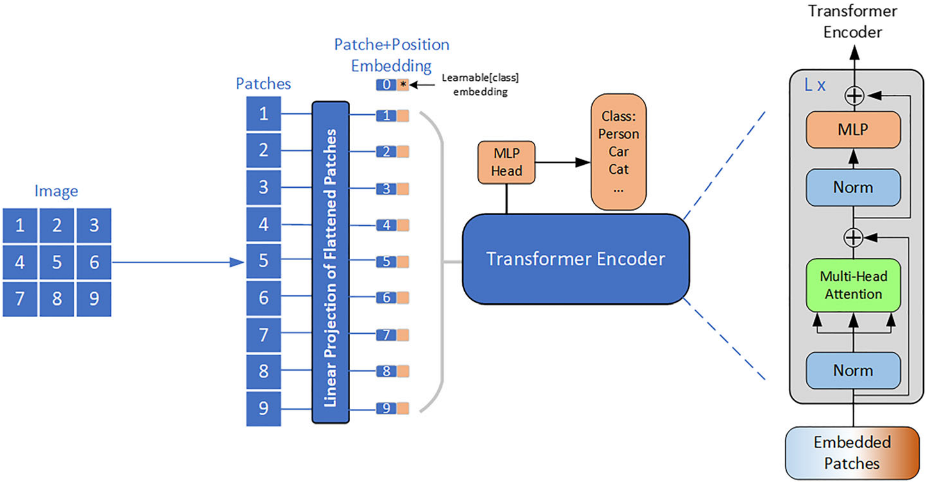


Fig. 14 Vision Transformer structure [18]

The main task of the ViT is to do classification tasks. The main idea is to use the Transformer Encoder part to do classification. Like NLP, the classification token will be added to the picture sequence. The picture sequence is obtained by cutting a picture into multiple patches. Number of patches $N = HW/P^2$,where the H, W is stand for height and width of input image, and P^2 is the resolution of each image patch.

In the model design, the author follows the original transformer as much as possible. One advantage of this deliberately simple setup is that the scalable NLP transformer architecture and its effective implementation can be used almost immediately.

when image sliced into patches, location information of each patch must be added before it is sent to the Transformer encoder, because in the linear projection process this information will be lost. Insert another vector, which is independent of the analyzed image and represents global information about the entire image. In fact, the output corresponding to the patch is passed to the MLP, and the MLP will return the prediction class. However, the loss of information in this process is very serious. In fact, during the conversion from patch to vector, any information about the position of the pixel in the patch will be lost. Therefore, some researchers have proposed new solutions [27] to this problem. In addition, ViT also has many problems such as large demand for data, large amount of computation, and inability to encode location embedding.

3) Swin ViT

In addition to the ViT mentioned above, there are also work such as iGPT [9] that use Transformer in image classification. But these methods have the following two serious problems:

1) A picture needs at least a few hundred pixels to express the content. To process so many pixels in a sequence is not what Transformer good at.
2) The previous methods are all to find the solution to object classification, but theoretical, Transformer is better at finding answer to detection problem, but the ability to solve the dense prediction scene of instance segmentation needs to be improved.

So as to solve the above problems, the author present Swin Transformer, and successfully made the model achieve SOTA results in classification, detection and instance segmentation. At present, Swin Transformer has become the universal BackBone when Transformer is used in visual tasks.

Swin Transformer does not use small-resized images as input images like the previous ViT and iGPT, but directly inputs the original images, so that there will be no information loss caused by resize. One advantage of Swin Transformer is the use of the most commonly used hierarchical network structure in CNN. With the deepening of the hierarchical network structure, the receptive field of the node expands. The receptive field of the node will expand with the deepening of the hierarichical network structure. This characteristic is also given to the Swin Transformer, and this give Swin Transformer the ability to detect and segment objects like FPN [45] and U-Net [69] structures. The hierarchical network structure of ViT and Swin Transformer is shown in Fig. 15.

Its advantages are shown in the following three points:

1. Achieved faster landing of Transformer in CV field.

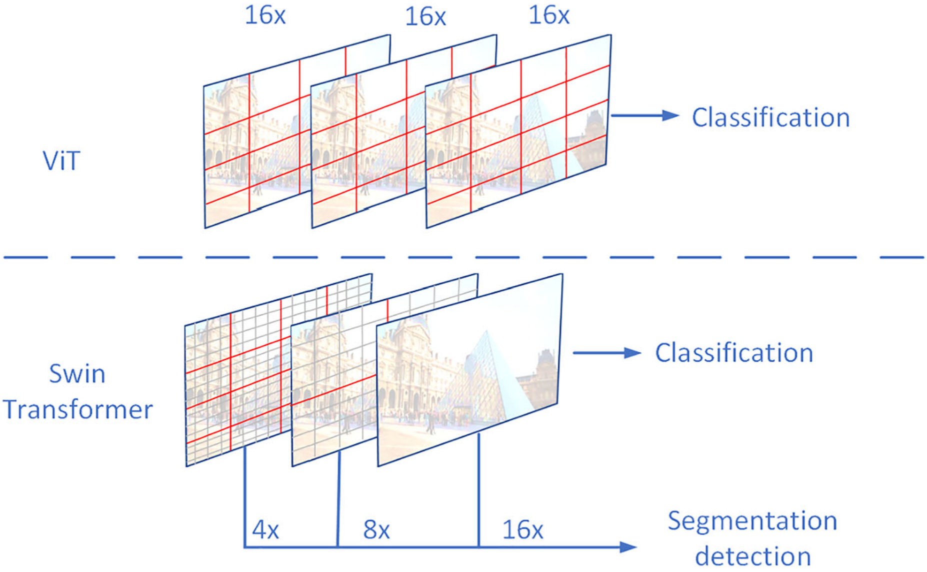


Fig. 15 Main difference between Swin Transformer and ViT [51]

2. Swin Transformer combines the advantages of Transformer and CNN, and a hierarchical structure used to reduce the resolution and increase the number of channels.
3. Achieved SOTA results in different visual tasks.

There are the following disadvantages:

- The computational coast is significant.
- High demands on GPU memory.
- Precise fine tuning is required in some project applications.

4) Twins

Twins [11] proposed two new architectures, named Twins-PCPVT and Twins-SVT.

The first architecture, Twins-PCPVT, structure shown in Fig. 16, replaces the positional coding in PVT [87] (the same fixed-length learnable positional coding as DeiT [80]) with the Conditional Positional Encodings proposed by the team in CPVT [12]. Large performance improvements can be directly obtained on classification and downstream tasks, especially on dense tasks. Since the conditional position coding CPE [12] supports variable length input, the visual Transformer can flexibly process features from different spatial scales. This architecture shows that PVT can match or surpass the performance of Swin only through CPVT's conditional position coding enhancement. This finding confirms that the reason why PVT performance is inferior to Swin is the use of inappropriate position coding. It can be seen that position codes such as CPE, which can flexibly handle varying resolutions, have a great impact on downstream tasks.

The second architecture, Twins-SVT, optimizes and improves the attention strategy based on a detailed analysis of the current global attention. The new strategy integrates the local-global attention mechanism. The author compares it to the depth-wise separable convolution in the convolution neural network is named Spatially Separable Self-Attention (SSSA). Different

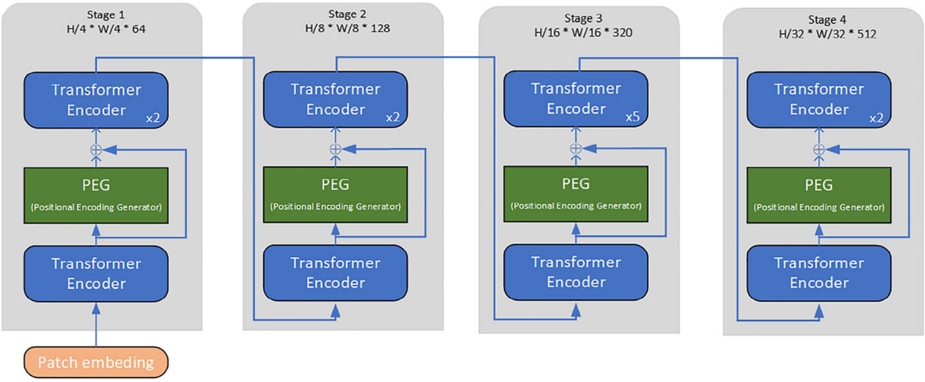


Fig. 16 Structure of Twins-PCPVT [11]

from the depth separable convolution, the spatially separable self-attention (Fig. 17) proposed by Twins-SVT is to group the spatial dimensions of the features to calculate the self-attention of each group, and then perform grouping attention results from the global fusion.

This approach needs to improve as follows:

- The training phase requires processing images of a fixed input size

Some modules designed for CNNs are not available in models.

3.3 The latest object detection research

In recent years, based on the ideas of the above-mentioned classic object detection algorithms, new object detection algorithms have continuously emerged.

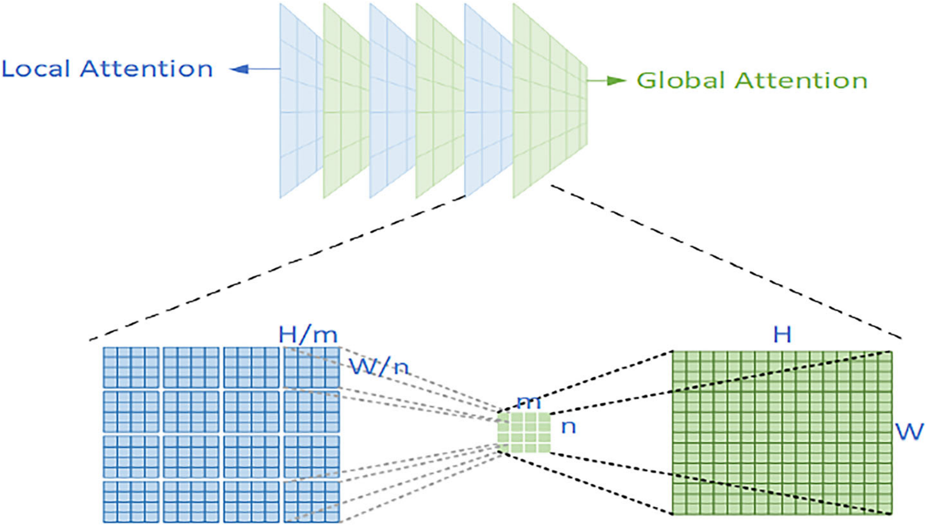


Fig. 17 Application of LSA and GSA in Twins-SVT. LSA: locally-grouped attention. GSA: global sub-sampled attention [11]

Whether it is based on CNN or transformer, all thinking that detection system can be faster and more accurate.

The accuracy of the convolution-based object detection methods largely depends on its feature extraction backbone network. A good backbone network can design an excellent object detection method. The detailed introduction of the classic backbone network is in this article [1].Also used the new backbone Densenet [34] to design some novel target detection models like STDN [96], DSOD [72] and TinyDSOD [41]. There are also lightweight networks designed to adapt to devices with limited performance, such as SquezzeNet [35], MobileNets [32], MobileNetv2 [71], MobileNetv3 [33], ThunderNet [62], ShuffleNet [95], ShuffleNetv2 [54], PeleeNet [86] and MnasNet [78].

In addition, there are some excellent CNN-based classical backbone networks and detectors that researchers are constantly exploring with different techniques, methods and combinations, such as ConvNet [52], YOLOX [24]. And some academics use their own unique methods to explore new paths in object detection [63].

Although the application of pure Transformer [18] in cv has hardly attracted the attention of others, recent researches on target detection has been almost entirely occupied by transformers.

On the basis of the above-mentioned transformer-based object detection method, different methods with greatly improved effects have appeared, such as, Focal Transformer [92], CSWin Transformer [17], CBNetV2 [43] and Mobile ViT [57], the first to apply Transformer to mobile Moreover the research in the direction of object detection, many methods of using transformers to solve problems have also appeared in other directions in CV field, such as, SegFormer [90], MaskFormer [10], TransGAN [36], TNT [27], DVT [88], YOLOS [22].

Object detection based on CNN and Transformer methods is also gaining more and more attention for specific detection tasks outside the mainstream field. For example, [76] used for classification of skin disease, CenterPoint [93] for 3D object detection, O2DETR[56]for remote sensing monitoring, SSPNet [31] for small object detection, and so on.

According to recent papers published, transformers have great potential in processing CV tasks. The combination of CNN and transformers has also improved efficiency. This trend can also be seen in some of the newly proposed methods, such as the use of Transformer with the concept of convolution [28]. Therefore, transformers will go further in computer vision field.

4 Comparison

In this section the paper will compare different angles of the different object detection algorithms discussed above. We will not discuss and compare the algorithm optimization and learning rate during training in detail, and the discussion ideas for this can refer to [84]. The evaluation indicators used for the comparison will also be introduced. In Section 4.2, the operating results of different mainstream object detection algorithms are visualized and the computational complexity of these algorithms is compared. Our purpose is not to tell the reader which one is the best algorithm, but to be able to see the advantages and disadvantages of different algorithms.

4.1 Data comparison

In this part, the paper compared several object detection methods in detail and comprehensively, and show the comparative data.

First, it briefly introduces the data sets and metrics that are often used in object detection tasks.

- PASCAL VOC [20, 21]: This dataset has two versions of commonly used as standard benchmarks. The datasets are divided into 4 categories: vehicle, household, animal, and person, for a total of 20 subclasses. VOC 2007 has a total of 9963 images, train+val has 5011 images, test has 4952, and the total objects are 24,640. VOC 2012 has a total of 23,080 images, Train+Val has 11,540 images, and the total objects are 54,900.
- ILSVRC [70]: From 2010 to 2017, one of the most sought after and authoritative academic competitions in the field of CV has represented the highest level in the field of imaging. It contains the ImageNet [16] images.
- MS-COCO [44]: The largest dataset with semantic segmentation so far. And it also the most difficult and complicated object detection dataset available today. This dataset aims at scene understanding, mainly intercepted from complex daily scenes, and the objects in the image are calibrated by precise segmentation. It is the largest dataset for semantic segmentation so far, there are 80 categories provided, there are more than 330,000 images, 200,000 of which are annotated, and the number of individuals in the entire dataset exceeds 1.5 million.
- Open Images [39]: About 9 million images span approximately 6000 categories, and these tags contain more real-life entities than ImageNet.

To estimate the performance of an object detector, there must be a unified standard. Next, several norms will introduce, those are used to evaluate the performance of a model.

- Precision: precision is calculated from the perspective of the prediction result, which is the ratio of true positive samples among the positive samples predicted by the model
- This is the ratio of true positive samples among the positive samples predicted by the model.
- Recall: recall rate is calculated from the real sample set, which means the number of positive samples recovered by the model from the total positive samples.
- IoU: Intersection over Union is calculated as the percentage of the overlap between the ground truth and the predicted bounding box and the area of their union. As shown in Fig. 18.
- FPS: FPS is how many pictures the target network can detect per second. FPS is simply to understand the refresh rate of the image.
- AP: AP is the average detection precision rates at different recall points.

  The confidence level interval greatly from model to model, and may be 50% confidence interval in our model equivalent to 80% confidence interval in other models, which makes

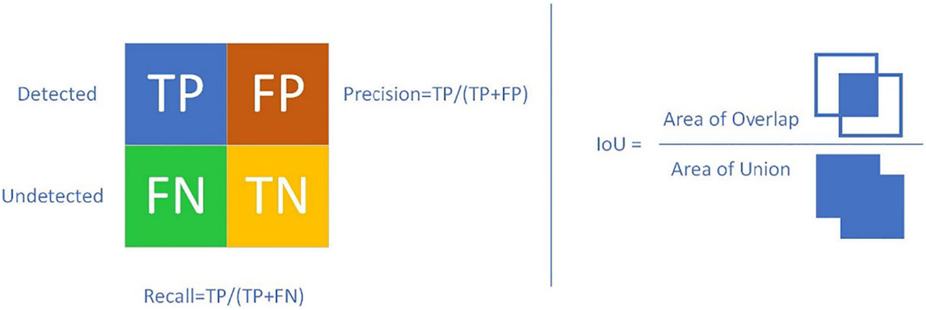


Fig. 18 The calculation of precision and recall and the expression form of IoU

it difficult to evaluate between different models. To eliminate this assessment discrepancy, experts proposed AP metrics.

- mAP: mean Average precision measures the quality of the learned model in all categories, and it is one of the most important indicators in object detection.

  The AP value calculation is only for one class, and the calculation of mAP after getting the AP becomes very simple, that is, take the average after the AP_i of all classes (K) is calculated. This process can be formulated as:

$$\text{mAP} = \frac{\sum_{i=1}^{K} AP_i}{K} \tag{1}$$

The above is most of the indicators used by most existing target detection models to test and compare performance. Dissimilar models may be different in the use of datasets and the goals they want to achieve. For example, some models are optimized for small object detection, some models want to increase the ability of feature extraction, etc., but final goal is to quickly and accurately identify and locate objects in view.

The object detection technique has a long development history [22], and it may be some improprieties in comparing the detection technique of different years together. Therefore, the paper tries to compare the different methods proposed in recent years using the same dataset and backbone shown as in Table 1, in order to comparatively see advantages and disadvantages of these methods. Although there are some unfair factors in this method, for example, the backbone is not VGG when the method proposed, such as, the backbone used by SPP-Net is ZF5 [94]. Some data is removed from VOC07 for comparison. It also did not compare the best mAP results of each target detection method, although mAP is the most important performance indicator, the performance of the model needs to be considered comprehensively. For example, R-CNN when uses [74] the mAP result is obviously improved, but the calculation time consumption will also be longer.

Although the object detection methods shown in Table 2 are not explained in detail in our paper, these methods contribute to the existing detection methods, such as Mask R-CNN [30] integrated the object detection and instance segmentation functions, and the detection and segmentation are calculated in parallel. The paper also conducts mAP comparisons under relatively fair standards. AP [0.5,0.95] is added to the performance indicators of the MS COCO dataset, which means the average mAP at different IoU thresholds (from 0.5 to 0.95, step size 0.05).

Table 2 shows that as the object detection technology continues to update and innovate, some of our understanding is also refreshed. Mask R-CNN is a completely innovative two-stage target detection method, and its mAP is also very high. Of course, its FPS is low. This is because of the characteristics of the two-stage object detection method. But we found that EfficientDet maintained a high FPS when mAP surpassed Mask R-CNN. This is what we are pursuing, high FPS and high precision.

Table 1 Capability comparison of different methods discussed in our paper in our paper.bb stands for bonding box regression and /diff is without "difficult" examples in voc07

Framework	Backbone	Dataset	mAP
R-CNN BB	VGG16	VOC07	66.0
SPP-Net BB	VGG16	VOC07\diff	63.1
Fast R-CNN	VGG16	VOC07	66.9
Faster R-CNN	VGG16	VOC07	69.9

Table 2 Performance comparison between different methods. * refers to this method using Faster R-CNN on FPN

Framework	Backbone	Dataset	$AP_{0.5}$	AP $_{[0.5,0.95]}$
Faster R-CNN	ResNet-101	MS COCO	48.5	27.2
R-FCN [14]	ResNet-101	MS COCO	48.9	27.6
FPN [45]*	ResNet-101	MS COCO	59.1	36.2
Mask R-CNN	ResNet-101-FPN	MS COCO	62.3	39.8
DetectoRS [61]	ResNeXt-101[89]	MS COCO	71.6	53.3
TridentNet [42]	ResNet-101	MS COCO	69.7	48.4
EfficientDet [79]	EfficentNet[77]-B6	MS COCO	71.4	52.2
HTC [7]	ResNet-50	MS COCO	62.6	43.6

After Transformer is applied in field of CV, there have been many related studies. The methods of object detection in these studies are compared in Table 3. It should be noted that most of these transformer methods appear as backbones (Table 4). DETR is a combination of Transformer and CNN methods, so the existing backbone is used. In order to make the comparison as fair as possible, the pure Transformer method uses the same framework, and some adjustments have been made to the framework used [11].

4.2 Visualization results and model complexity

Five mainstream object detectors from the detection methods selected to discussed above to show their detection effect as shown in Fig. 19, and compared their calculation cost, parameters and detection speed, as shown in Table 5. We can more intuitively see the characteristics of these mainstream detection methods from these test results and data.

5 Future trends

In this part, the paper looks forward to the future development trend of object detection:

- Lightweight detector

Aim to accelerate the compute speed of detection and enable it to run on mobile devices, some researchers have made great efforts in recent years. But the current lightweight detection

Table 3 Performance comparison of one stage algorithms

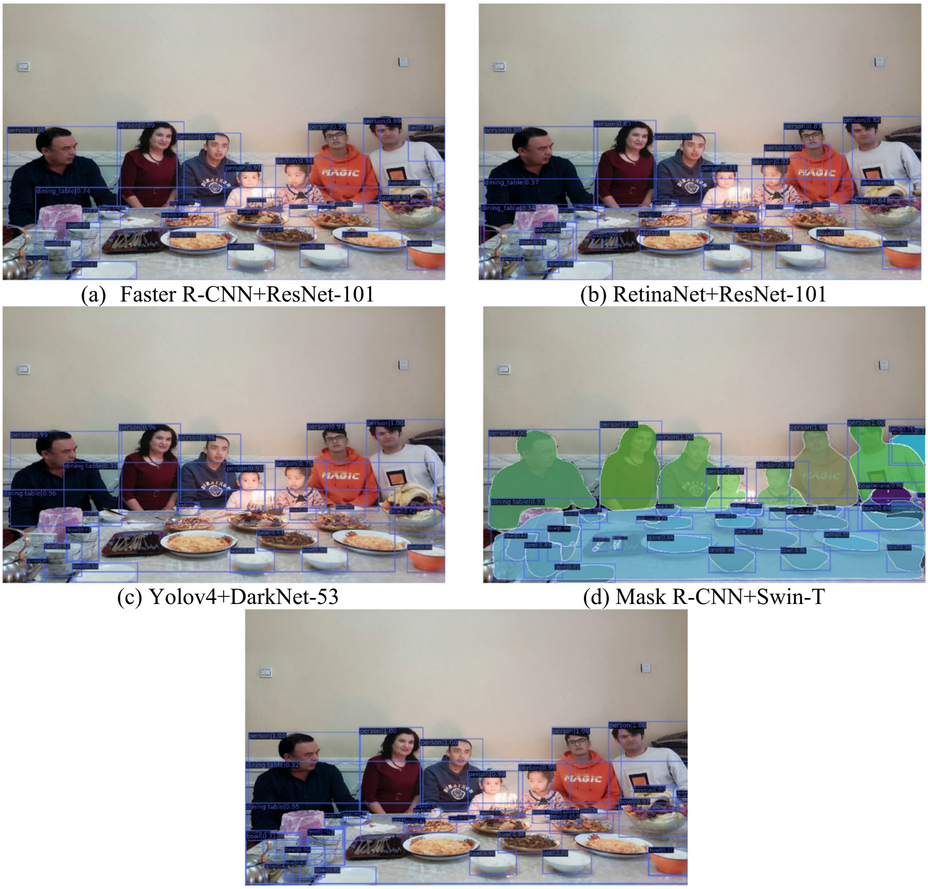
Framework	Backbone	Dataset	$AP_{0.5}$	AP $_{[0.5,0.95]}$
SSD512	VGG16	VOC07	71.6	
SSD512	VGG16	MS COCO	46.5	26.8
YOLO	Modified GoogLeNet	VOC07+VOC12	63.4	
YOLOv2	DarkNet-19	MS COCO	44.0	21.6
YOLOv3	DarkNet-53	MS COCO	51.5	28.2
YOLOv4	CSPDarknet53	MS COCO	64.9	43.0
RetinaNet	ResNet-101	MS COCO	53.1	34.4
Center Net	Hourglass-104[58]	MS COCO	62.4	44.9

Table 4 Transformer based detection methods performance comparison on the COCO val2017

Framework	Backbone	$AP_{0.5}$	AP $_{[0.5,0.95]}$
RetinaNet	ResNet-101	53.1	34.4
RetinaNet	PVT-Medium	63.1	41.9
RetinaNet	Twins-PCPVT-B	65.6	44.3
RetinaNet	Twins-SVT-B	66.7	45.3
RetinaNet	Swin-S	65.7	44.5
DETR	ResNet-101	64.9	44.9

algorithm is still not satisfactory. Therefore, in a future development trend of detection algorithms, lightweight, fast and high-precision is the eternal theme of target detection.

- Multi-task learning


(a) Faster R-CNN+ResNet-101

(b) RetinaNet+ResNet-101

(c) Yolov4+DarkNet-53

(d) Mask R-CNN+Swin-T

(e) DETR+ResNet-50

Fig. 19 The detection effect of five different mainstream detection methods. **a** Faster R-CNN + ResNet-101, **b** RetinaNet+ResNet-101, **c** Yolov4 + DarkNet-53, **d** Mask R-CNN + Swin-T, **e** DETR+ResNet-50

Table 5 Model complexity of mainstream detection methods trained on COCO datasets

Methods	Flops (G)	Param (M)	Inference time (fps)
Faster R-CNN + ResNet-101	283.14	60.52	15.6
RetinaNet+ResNet-101	315.39	56.74	15.0
Yolov4 + DarkNet-53	195.55	61.95	66.0
Mask R-CNN + Swin-T	263.78	47.79	15.3
DETR+ResNet-50	91.64	41.3	–

Aiming at the current problem of low single-task learning and detection performance, a multi-task learning method that combines multiple tasks in the network and multi-level features of the network is proposed to improve detection performance and perform multiple computer vision tasks at the same time [48].

- Weakly supervised object detection

  The training of detection algorithms based on deep learning depends on great number of high-quality images with noted datasets, and model training process is often time-consuming and inefficient. The use of weakly-supervised object detection can make the detection algorithm use part of the bounding box labeled dataset for training. Therefore, Weakly supervised techniques are important to reduce labor costs and improve detection flexibility.
- GAN-based object detection

Whether it is based on convolutional neural network or Transformer, it requires great amount of image data to train. Using generative adversarial networks to generate fake images to produce a large number of data samples to achieve data expansion. The real scene data is mixed with the simulation data generated by the GAN training target detector, so that the detector has stronger robustness and generalization ability [3].

- Small object detection

Detecting small objects in scene images has been a long-standing challenge in the field of object detection, and some potential applications of small object detection research directions include: the use of remote sensing images to count the number of wild animals, and detect the status of some important military targets, so how to solve the problem of small targets has always been a hot topic for researchers.

- Multi-modal detection

With the popularization and development of different sensors, the use of different sensors in the field of target detection, such as depth cameras, lidars and other equipment to obtain target information, has made some progress in the past few years. In the field of autonomous driving, cars are often equipped with a complex array of sensors for accurate and robust environmental perception. How these large numbers of different types of sensors complement each other and fuse them to facilitate perception is still an open question.

- Video detection

Real-time object detection/tracking in high-definition video is of great significance for video surveillance and autonomous driving. Existing object detection algorithms are usually designed for object detection in a single image, while ignoring the correlation between video frames. Improving detection performance by exploring the spatial and temporal correlations between sequences of video frames is an important research direction.

6 Conclusion

Classic detection methods, key technique, datasets and indicators are intruded in our paper. In the past ten years, the key research content of object detection tasks has been CNN. However, the new method that has appeared recently, that is, the application of Transformer in CV, has opened up a new path for object detection and has become a recent research hotspot. Some people even think that Transformer will completely replace CNN in the future.

The following are our thoughts on the application of CNN and Transformer on object detection tasks:

1. CNN uses the convolution kernel to continuously extract high-level abstract features. In theory, its receptive field should cover the entire image, but many studies have shown that its actual theoretical receptive field is larger than the receptive field, which is not good for to feature extraction with making full use of contextual information. On the other hand, the advantage of the transformer is that it uses attention to capture global contextual information and build long-range dependencies on objects to extract more useful features.

From the experimental results of ViT, Transformer has also learned the same receptive field from small to large as CNN, which further illustrates the relationship between Transformer and CNN [13]. Through the comparative test of CNN-based and Transformer-based object detection methods as shown in Tables 3 and 4, some Transformer-based detection methods such as DETR and PVT, twins as the backbone network of RetinaNet have surpassed RetinaNet with ResNet-101 as the backbone network.

2. Because CNN has the characteristics of inductive bias and transitional invariance, it is easier to deal with image problems. Transformer does not have this feature, so the training model requires much more data or stronger data enhancement [85] to learn image features.
3. A large part of the current work is the mixed use of CNN and Transformer. The hybrid structure of ViT also uses CNN to input feature maps into ViT, achieving better results than pure Transformer.

The paper summarizes the characteristics and shortcomings of mainstream detection methods, and through comparative experiments, readers can have a deeper thinking about these test methods. How to improve the shortcomings of these detection algorithms, how to combine the advantages of CNN-based and Transformer-based object detection methods, we hope that our paper can give readers some help.

For the CNN that has been developed for more than ten years, it is very mature in tasks such as object detection, segmentation, and classification. It is difficult for the Transformer method to completely replace CNN in a short time. The focus of future work will be to add more excellent features of CNN to Transformer [60].

Acknowledgments This work was supported by National Key Research and Development Program of China (No.2021YFB2802100) and National Science Foundation of China under Grant (61862061, 62061045).

Data availability All data generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interests The authors declare that they have no conflict of interest.

References

1. Arkin E, Yadikar N, Muhtar Y, Ubul K (2021) "A Survey of Object Detection Based on CNN and Transformer," 2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML), pp. 99–108, https://doi.org/10.1109/PRML52754.2021.9520732.
2. Bochkovskiy, A, Wang, CY, Liao, HYM (2020) Yolov4: Optimal speed and accuracy of object detection. https://doi.org/10.48550/arXiv.2004.10934.
3. Brock, A, Donahue, J, Simonyan, K (2018) Large scale GAN training for high fidelity natural image synthesis. https://doi.org/10.48550/arXiv.1809.11096.
4. Cai, Z, Fan, Q, Feris, RS, Vasconcelos, N (2016) A Unified Multi-scale Deep Convolutional Neural Network for Fast Object Detection. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture notes in computer science(), vol 9908. Springer, Cham. https://doi.org/10.1007/978-3-319-46493-0_22.
5. Cao Y, Chen K, Loy CC, Lin D (2020) "Prime Sample Attention in Object Detection," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11580–11588, https://doi.org/10.1109/CVPR42600.2020.01160.
6. Carion, N, Massa, F, Synnaeve, G, Usunier, N, Kirillov, A, Zagoruyko, S (2020) End-to-End Object Detection with Transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, JM. (eds) Computer Vision – ECCV 2020. ECCV 2020. Lecture notes in computer science(), vol 12346. Springer, Cham. https://doi.org/10.1007/978-3-030-58452-8_13.
7. Chen K et al. (2019) "Hybrid Task Cascade for Instance Segmentation," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4969–4978, https://doi.org/10.1109/CVPR.2019.00511.
8. Chen C, Liu M, Meng X, Xiao W, Ju Q (2020) "RefineDetLite: A Lightweight One-stage Object Detection Framework for CPU-only Devices," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2997–3007, https://doi.org/10.1109/CVPRW50498.2020.00358.
9. Chen, M, et al. (2020) “Generative Pretraining From Pixels.” ICML 2020: 37th International Conference on Machine Learning, vol. 1, 2020, pp. 1691–1703
10. Cheng, B, Schwing, A, Kirillov, A (2021) Per-pixel classification is not all you need for semantic segmentation Advances in Neural Information Processing Systems, 34
11. Chu, X, et al. (2021) "Twins: Revisiting the design of spatial attention in vision transformers." Advances in Neural Information Processing Systems 34 (NeurIPS 2021)
12. Chu, X, Tian, Z, Zhang, B, Wang, X, Wei, X, Xia, H, Shen, C (2021) Conditional positional encodings for vision transformers. https://doi.org/10.48550/arXiv.2102.10882.

13. Cordonnier, J-B, et al. (2020) “On the Relationship between Self-Attention and Convolutional Layers.” ICLR 2020 : Eighth International Conference on Learning Representations. https://doi.org/10.48550/arXiv.1911.03584
14. Dai J, Li Y, He K, Sun J. (2016) R-FCN: object detection via region-based fully convolutional networks. In proceedings of the 30th international conference on neural information processing systems (NIPS'16). Curran associates Inc., red hook, NY, USA, 379–387
15. Dalal N, Triggs B (2005) "Histograms of oriented gradients for human detection," 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), pp. 886–893 vol. 1, https://doi.org/10.1109/CVPR.2005.177.
16. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255, https://doi.org/10.1109/CVPR.2009.5206848.
17. Dong, X, Bao, J, Chen, D, Zhang, W, Yu, N, Yuan, L, ..., Guo, B. (2021) Cswin transformer: A general vision transformer backbone with cross-shaped windows. https://doi.org/10.48550/arXiv.2107.0065.
18. Dosovitskiy, A, et al. (2020) “An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale.” https://doi.org/10.48550/arXiv.2010.11929.
19. Duan K, Bai S, Xie L, Qi H, Huang Q, Tian Q (2019) "CenterNet: Keypoint Triplets for Object Detection," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6568–6577, https://doi.org/10.1109/ICCV.2019.00667.
20. Everingham M et al (2010) The Pascal Visual Object Classes (VOC) Challenge. Int J Comput Vis 88(2): 303–338
21. Everingham M et al (2015) The Pascal Visual Object Classes Challenge: A Retrospective. Int J Comput Vis 111(1):98–136
22. Fang, Y, Liao, B, Wang, X, Fang, J, Qi, J, Wu, R, ..., Liu, W (2021) You only look at one sequence: rethinking transformer in vision through object detection. Adv Neural Inf Proces Syst, 34. https://doi.org/10.48550/arXiv.2106.00666
23. Fu, CY, Liu, W, Ranga, A, Tyagi, A, Berg, AC (2017) Dssd: Deconvolutional single shot detector. https://doi.org/10.48550/arXiv.1701.06659.
24. Ge, Z, Liu, S, Wang, F, Li, Z, Sun, J (2021) Yolox: Exceeding yolo series in 2021. https://doi.org/10.48550/arXiv.2107.08430.
25. Girshick R (2015) "Fast R-CNN," 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1440–1448, https://doi.org/10.1109/ICCV.2015.169.
26. Girshick R, Donahue J, Darrell T, Malik J (2014) "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587, https://doi.org/10.1109/CVPR.2014.81.
27. Han, K, et al. (2021) "Transformer in transformer." Advances in Neural Information Processing Systems 34 (NeurIPS 2021)
28. Hassani, A, Walton, S, Li, J, Li, S, Shi, H (2022) Neighborhood Attention Transformer. https://doi.org/10.48550/arXiv.2106.03146.
29. He K et al (2015) Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. IEEE Trans Pattern Anal Mach Intell 37(9):1904–1916
30. He K et al (2020) Mask R-CNN. IEEE Trans Pattern Anal Mach Intell 42(2):386–397
31. Hong M, Li S, Yang Y, Zhu F, Zhao Q, Lu L (2022, Art no 8018505) SSPNet: Scale Selection Pyramid Network for Tiny Person Detection From UAV Images. IEEE Geosci Remote Sens Lett 19:1–5. https://doi.org/10.1109/LGRS.2021.3103069
32. Howard, AG, Zhu, M, Chen, B, Kalenichenko, D, Wang, W, Weyand, T, ..., Adam, H. (2017) Mobilenets: Efficient convolutional neural networks for mobile vision applications. https://doi.org/10.48550/arXiv.1704.04861.
33. Howard A et al. (2019) "Searching for MobileNetV3," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1314–1324, https://doi.org/10.1109/ICCV.2019.00140.
34. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) "Densely Connected Convolutional Networks, " 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261–2269, https://doi.org/10.1109/CVPR.2017.243.
35. Iandola, FN, Han, S, Moskewicz, MW, Ashraf, K, Dally, WJ, Keutzer, K (2016) SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. https://doi.org/10.48550/arXiv.1602.07360.
36. Jiang, Y, Chang, S, Wang, Z (2021) Transgan: two pure transformers can make one strong Gan, and that can scale up. Adv Neural Inf Proces Syst, 34
37. Kang K, Li H, Yan J, Zeng X, Yang B, Xiao T, Zhang C, Wang Z, Wang R, Wang X, Ouyang W (Oct. 2018) T-CNN: Tubelets with convolutional neural networks for object detection from videos. IEEE Trans Circuits Syst Vid Technol 28(10):2896–2907. https://doi.org/10.1109/TCSVT.2017.2736553

38. Karlinsky L et al. (2019) "RepMet: Representative-Based Metric Learning for Classification and Few-Shot Object Detection," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5192–5201, https://doi.org/10.1109/CVPR.2019.00534.

39. Kuznetsova A, Rom H, Alldrin N, Uijlings J, Krasin I, Pont-Tuset J, Kamali S, Popov S, Malloci M, Kolesnikov A, Duerig T, Ferrari V (2020) The open images dataset V4. Int J Comput Vis 128:1956–1981. https://doi.org/10.1007/s11263-020-01316-z

40. Law H, Deng J (2020) CornerNet: detecting objects as paired Keypoints. Int J Comput Vis 128:642–656. https://doi.org/10.1007/s11263-019-01204-1

41. Li Y, Li J, Lin W, Li J (2018) Tiny-DSOD: lightweight object detection for resource-restricted usages. https://doi.org/10.48550/arXiv.1807.11013

42. Li Y, Chen Y, Wang N, Zhang Z-X (2019) "Scale-Aware Trident Networks for Object Detection," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6053–6062, https://doi.org/10.1109/ICCV.2019.00615.

43. Liang T, Chu X, Liu Y, Wang Y, Tang Z, Chu W, ... Ling H (2021) Cbnetv2: a composite backbone network architecture for object detection. https://doi.org/10.48550/arXiv.2107.00420

44. Lin, TY. et al. (2014) Microsoft COCO: Common Objects in Context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture notes in computer science, vol 8693. Springer, Cham. https://doi.org/10.1007/978-3-319-10602-1_48.

45. Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) "Feature pyramid networks for object detection," 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 936–944. https://doi.org/10.1109/CVPR.2017.106

46. Lin T-Y, Goyal P, Girshick R, He K, Dollár P (2020) Focal Loss for Dense Object Detection. IEEE Trans Pattern Anal Mach Intell 42(2):318–327. https://doi.org/10.1109/TPAMI.2018.2858826

47. Liu, W et al. (2016) SSD: Single Shot MultiBox Detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture notes in computer science(), vol 9905. Springer, Cham. https://doi.org/10.1007/978-3-319-46448-0_2.

48. Liu S, Johns E, Davison AJ (2019) "End-to-end multi-task learning with attention," 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 1871–1880. https://doi.org/10.1109/CVPR.2019.00197

49. Liu L, Ouyang W, Wang X, Fieguth P, Chen J, Liu X, Pietikäinen M (2020) Deep learning for generic object detection: a survey. Int J Comput Vis 128:261–318. https://doi.org/10.1007/s11263-019-01247-4

50. Liu Z, Zheng T, Xu G, Yang Z, Liu H, Cai D (2020) Training-time-friendly network for real-time object detection. Proceedings of the AAAI Conference on Artificial Intelligence 34(07):11685–11692. https://doi.org/10.1609/aaai.v34i07.6838

51. Liu Z et al. (2021) "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9992–10002, https://doi.org/10.1109/ICCV48922.2021.00986.

52. Liu, Z, Mao, H, Wu, CY, Feichtenhofer, C, Darrell, T, Xie, S (2022) A ConvNet for the 2020s. https://doi.org/10.48550/arXiv.2201.03545.

53. Ma C, Huang J-B, Yang X, Yang M-H (2015) "Hierarchical Convolutional Features for Visual Tracking," 2015 IEEE International Conference on Computer Vision (ICCV), pp. 3074–3082, https://doi.org/10.1109/ICCV.2015.352.

54. Ma, N, Zhang, X, Zheng, HT, Sun, J (2018) ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds) Computer Vision – ECCV 2018. ECCV 2018. Lecture notes in computer science(), vol 11218. Springer, Cham. https://doi.org/10.1007/978-3-030-01264-9_8.

55. Ma W et al (2020) MDFN: Multi-Scale Deep Feature Learning Network for Object Detection. Pattern Recog 100:107149

56. Ma, T, Mao, M, Zheng, H, Gao, P, Wang, X, Han, S, ..., Doermann, D. (2021) Oriented object detection with transformer. https://doi.org/10.48550/arXiv.2106.03146.

57. Mehta, S, Rastegari M (n.d.) "Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer." https://doi.org/10.48550/arXiv.2110.02178.

58. Newell, A, Yang, K, Deng, J (2016) Stacked Hourglass Networks for Human Pose Estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture notes in computer science(), vol 9912. Springer, Cham https://doi.org/10.1007/978-3-319-46484-8_29.

59. Pang J, Chen K, Shi J, Feng H, Ouyang W, Lin D (2019) "Libra R-CNN: towards balanced learning for object detection," 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 821–830. https://doi.org/10.1109/CVPR.2019.00091

60. Peng Z et al. (2021) "Conformer: Local Features Coupling Global Representations for Visual Recognition," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 357–366, https://doi.org/10.1109/ICCV48922.2021.00042.
61. Qiao S, Chen L-C, Yuille A (2021) "DetectoRS: Detecting Objects with Recursive Feature Pyramid and Switchable Atrous Convolution," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10208–10219, https://doi.org/10.1109/CVPR46437.2021.01008.
62. Qin Z et al. (2019) "ThunderNet: Towards Real-Time Generic Object Detection on Mobile Devices," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6717–6726, https://doi.org/10.1109/ICCV.2019.00682.
63. Qiu H et al. (2021) "CrossDet: Crossline Representation for Object Detection," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3175–3184, https://doi.org/10.1109/ICCV48922.2021.00318.
64. Rahman S, Khan SH, Porikli F (2020) Zero-shot object detection: joint recognition and localization of novel concepts. Int J Comput Vis 128:2979–2999. https://doi.org/10.1007/s11263-020-01355-6
65. Redmon J, Farhadi A (2017) "YOLO9000: Better, Faster, Stronger," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6517–6525, https://doi.org/10.1109/CVPR.2017.690.
66. Redmon, J, Farhadi A (n.d.) "YOLOv3: An Incremental Improvement." https://doi.org/10.48550/arXiv.1804.02767.
67. Redmon J, Divvala S, Girshick R, Farhadi A (2016) "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788, https://doi.org/10.1109/CVPR.2016.91.
68. Ren S et al (2017) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans Pattern Anal Mach Intell 39(6):1137–1149
69. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A (eds) Medical image computing and computer-assisted intervention – MICCAI 2015. MICCAI 2015. Lecture notes in computer science, vol 9351. Springer, Cham. https://doi.org/10.1007/978-3-319-24574-4_28
70. Russakovsky O et al (2015) ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vis 115(3): 211–252
71. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C (2018) "MobileNetV2: Inverted Residuals and Linear Bottlenecks," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4510–4520, https://doi.org/10.1109/CVPR.2018.00474.
72. Shen Z, Liu Z, Li J, Jiang Y, Chen Y, Xue X (2017) "DSOD: learning deeply supervised object detectors from scratch," 2017 IEEE international conference on computer vision (ICCV), pp. 1937-1945, https://doi.org/10.1109/ICCV.2017.212.
73. Shrivastava A, Gupta A, Girshick R (2016) "Training Region-Based Object Detectors with Online Hard Example Mining," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 761–769, https://doi.org/10.1109/CVPR.2016.89.
74. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015). OpenReview.net, : 1–14
75. Singh S, Ahuja U, Kumar M, Kumar K, Sachdeva M (2021) Face mask detection using YOLOv3 and faster R-CNN models: COVID-19 environment. Multimed Tools Appl 80:19753–19768. https://doi.org/10.1007/s11042-021-10711-8
76. Srinivasu PN, SivaSai JG, Ijaz MF, Bhoi AK, Kim W, Kang JJ (2021) Classification of skin disease using deep learning neural networks with MobileNet V2 and LSTM. Sensors 21:2852. https://doi.org/10.3390/s21082852
77. Tan, M, Le Q (2019) "Efficientnet: rethinking model scaling for convolutional neural networks." International conference on machine learning. PMLR, https://doi.org/10.48550/arXiv.1905.11946
78. Tan M et al. (2019) "MnasNet: Platform-Aware Neural Architecture Search for Mobile," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2815–2823, https://doi.org/10.1109/CVPR.2019.00293.
79. Tan M, Pang R, Le QV (2020) "EfficientDet: Scalable and Efficient Object Detection," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10778–10787, https://doi.org/10.1109/CVPR42600.2020.01079.
80. Touvron, H, et al. (2021) "Training Data-Efficient Image Transformers & Distillation through Attention." ICML 2021: 38th International Conference on Machine Learning, pp. 10347–10357.
81. Uijlings JR et al (2013) Selective search for object recognition. Int J Comput Vis 104(2):154–171

82. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. 2017 Attention is all you need. In proceedings of the 31st international conference on neural information processing systems (NIPS'17). Curran associates Inc., red hook, NY, USA, 6000–6010
83. Viola P, Jones M (2001) "Rapid object detection using a boosted cascade of simple features," proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001, pp. 511–518, https://doi.org/10.1109/CVPR.2001.990517.
84. Vulli A, Srinivasu PN, Sashank MSK, Shafi J, Choi J, Ijaz MF (2022) Fine-tuned DenseNet-169 for breast Cancer metastasis prediction using FastAI and 1-cycle policy. Sensors 22:2988. https://doi.org/10.3390/s22082988
85. Wan F, Liu C, Ke W, Ji X, Jiao J, Ye Q (2019) "C-MIL: Continuation Multiple Instance Learning for Weakly Supervised Object Detection," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2194–2203, https://doi.org/10.1109/CVPR.2019.00230.
86. Wang RJ et al (2018) "Pelee: a real-time object detection system on mobile devices." NIPS'18 Proceedings of the 32nd international conference on neural information processing systems, vol 31, pp 1967–1976
87. Wang W et al (2021) "Pyramid vision transformer: a versatile backbone for dense prediction without convolutions," 2021 IEEE/CVF international conference on computer vision (ICCV), 2021, pp 548–558. https://doi.org/10.1109/ICCV48922.2021.00061
88. Wang Y, Huang R, Song S, Huang Z, Gao H (n.d.) Not All Images Are Worth 16x16 Words: Dynamic Vision Transformers with Adaptive Sequence Length. Adv Neural Inf Process Syst 34. https://doi.org/10.48550/arXiv.2105.15075
89. Xie S, Girshick R, Dollár P, Tu Z, He K (2017) "Aggregated Residual Transformations for Deep Neural Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5987–5995, https://doi.org/10.1109/CVPR.2017.634.
90. Xie, E, Wang, W, Yu, Z, Anandkumar, A, Alvarez, JM, Luo, P (2021) SegFormer: simple and efficient design for semantic segmentation with transformers. Adv Neural Inf Proces Syst, 34
91. Xiong Y et al. (2021) "MobileDets: Searching for Object Detection Architectures for Mobile Accelerators," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3824–3833, https://doi.org/10.1109/CVPR46437.2021.00382.
92. Yang, J, Li, C, Zhang, P, Dai, X, Xiao, B, Yuan, L, Gao, J (2021) Focal self-attention for local-global interactions in vision transformers. https://doi.org/10.48550/arXiv.2107.00641.
93. Yin T, Zhou X, Krähenbühl P (2021) "Center-based 3D Object Detection and Tracking," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11779–11788, https://doi.org/10.1109/CVPR46437.2021.01161.
94. Zeiler, MD, Fergus, R (2014) Visualizing and Understanding Convolutional Networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture notes in computer science, vol 8689. Springer, Cham https://doi.org/10.1007/978-3-319-10590-1_53.
95. Zhang X, Zhou X, Lin M Sun J (2018) "ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6848–6856, https://doi.org/10.1109/CVPR.2018.00716.
96. Zhou P, Ni B, Geng C, Hu J, Xu Y (2018) "Scale-Transferrable Object Detection," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 528–537, https://doi.org/10.1109/CVPR.2018.00062.
97. Zhou, X, Koltun, V, Krähenbühl, P (2021) Probabilistic two-stage detection. https://doi.org/10.48550/arXiv.2103.07461.
98. Zhu, X, Su, W, Lu, L, Li, B, Wang, X, Dai, J (2020) Deformable detr: Deformable transformers for end-to-end object detection. In Proc. ICLR, 2021 Oral, PP. 1–16