# Research on location target detection algorithm based on neural network

**Long He[1,a,*], Zhou Yang[2,b], Xiaoyu Feng[2,c]**

*[1]School of Intelligent Technology and Engineering, Chongqing University of Science & Technology, Chongqing, China*
*[2]School of Intelligent Technology and Engineering, Chongqing University of Science & Technology, Chongqing, China*
*[a]913074415@qq.com, [b]972389201@qq.com, [c]1247656916@qq.com*
*[*]Corresponding author*

***Abstract:*** *Object detection is a key problem in the field of computer vision, which has a wide range of applications, such as automatic driving, security monitoring, medical image analysis and so on. With the rapid development of deep learning technology, object detection algorithm based on neural network has become one of the research hotspots. In this paper, the object detection algorithm based on neural network is deeply studied, and its principle, method and application field are discussed. Firstly, this paper introduces the difference between traditional target detection algorithm and neural network-based target detection algorithm. Traditional algorithms usually rely on hand-designed feature extraction methods, while neural network-based algorithms automatically learn image features through convolutional neural networks (CNN), which has stronger generalization ability. In addition, we discuss in detail the basic concepts of deep learning techniques, including convolutional layers, pooling layers, fully connected layers, etc., and commonly used neural network structures. Secondly, this paper focuses on the key problems in object detection algorithms. We analyze the differences in the performance of various neural network architectures in solving these problems, and these methods have made significant progress in improving detection accuracy and reducing false detection rates. Further, this paper studies the data sets and evaluation indicators in the field of target detection. We cover some commonly used evaluation metrics such as accuracy, recall, F1 score, and mAP. These data sets and metrics are essential for the training and evaluation of algorithms, helping researchers compare the performance of different algorithms. Finally, this paper discusses the potential of target detection algorithm based on neural network in practical application. We highlight its wide application in several fields and discuss future research directions. With the improvement of hardware performance and the continuous development of deep learning technology, target detection algorithms based on neural networks will continue to make breakthroughs, bringing more innovation to various fields.*

***Keywords:*** *neural networks, CNN, YOLO, feature extraction*

## 1. Introduction

Urban location target detection has important application value in modern urban management, traffic monitoring, military reconnaissance and automatic driving. Object detection is a core problem in the field of computer vision, which aims to automatically identify and locate various objects, such as cars, pedestrians, buildings, etc., from complex urban landscapes. With the rapid development of deep learning technology, the city location target detection algorithm based on neural network has become one of the focuses of research[1].

Urban location target detection plays an important role in modern urban management and planning. The development of this technology has profound implications for the safety, efficiency, sustainability and social development of cities. First of all, urban location target detection is very important for traffic management and safety. The density of vehicles and pedestrians on urban roads is high, and a good object detection system can help traffic management departments better monitor and manage traffic flow, reduce traffic accidents and congestion, and improve road safety. This is essential for the city's transportation system and the quality of life of its residents. Secondly, city location target detection is the key to realize intelligent transportation. By monitoring vehicles and pedestrians on the road, traffic signals can be intelligently adjusted according to traffic conditions, improving the efficiency of road traffic. This not only improves traffic mobility, but also helps to reduce energy waste and reduce environmental pollution[2]. Urban security also benefits from urban location target detection. This technology contributes to security monitoring in cities and can be used to prevent criminal activities. It can monitor suspicious behavior, respond quickly and take measures to improve the safety of the city. Emergency rescue is another area where urban location target detection plays a key role[3]. In emergency situations, such as fires, natural disasters, etc., target detection systems can be used to find trapped people, provide accurate location information, and assist rescue operations, which can save lives and reduce disaster losses. Urban planning also benefits from urban location target detection. This technology helps to gather information on urban population movements, building use and urban trends for better planning of infrastructure and resources. Cities can respond more intelligently to different needs, improving urban sustainability. In addition, urban location target detection can also be used for environmental monitoring[4]. It can detect air quality, waste disposal, green space coverage, and more, helping cities achieve environmental sustainability and improve residents' quality of life. Commercial and retail businesses can use this technology to monitor customer traffic, improve product placement and promotion strategies, and increase economic efficiency. Finally, urban location target detection is a key component of building smart cities. Through large-scale data collection and analysis, cities can respond more intelligently to different needs and improve the quality of life.

In short, the importance of urban location target detection is self-evident[5-6]. Continued development and innovation in this area will help build smarter and more livable cities that meet the growing needs and challenges of urbanization. In urban management, planning and development, urban location target detection technology will continue to play an irreplaceable role[7].

The research of urban location target detection involves many fields of knowledge, such as computer vision, machine learning, neural network and remote sensing image processing. Challenges in this area include, but are not limited to, complex urban backgrounds, variations in illumination under different weather conditions, uncertainties in target scale, and changes in target shape from different viewing angles. Therefore, the study of urban location target detection algorithm needs to constantly explore innovative methods to overcome these challenges.

Object detection has a wide range of applications in various fields. According to its development history, object detection algorithms can be divided into two categories: traditional object detection algorithms and deep learning-based object detection algorithms [8].

Traditional target detection algorithms usually include several steps: input image, candidate box selection, feature extraction, classification, correction and optimization results and output. The specific process is as follows: Firstly, multi-scale image scanning is carried out by sliding window technology to generate multiple candidate frames; Then, traditional feature extraction methods such as HOG or SIFT are used to extract features from each candidate box[9-10]. Next, the Adaboost or SVM classifier is trained to classify the features of the candidate boxes to determine the category of the target; Finally, the non-maximum suppression (NMS) algorithm is used to compare the position of the candidate box with that of the real box to filter out the final detection result. However, the traditional target detection algorithm has many shortcomings, including time-consuming selection of candidate frame, slow recognition speed, limited feature expression and low recognition accuracy. Therefore, there are some limitations in practical application, and it is difficult to meet the high requirements of performance and speed.

With the continuous development of neural networks, object detection algorithms based on deep learning have emerged and become the mainstream algorithms in the field of object detection, abandoning the weaknesses of poor robustness and lack of real-time feature extraction in traditional methods. This paper discusses in detail the basic concepts of deep learning technology, including convolutional layer, pooling layer, fully connected layer, etc., and common neural network structures. Secondly, this paper focuses on the key problems in object detection algorithms. We analyzed the performance differences of various neural network architectures in solving these problems. Aiming at the problems of missing detection and false detection in the detection of small target traffic signs in real scenes, we constructed a target detection layer with more shallow feature information and introduced an attention module to enhance the effect of target detection.

## 2. Related work

In recent years, with the continuous improvement of computer hardware conditions, the number of problems that can be handled by computers has also increased, which has also set off a wave of exploration by researchers. The existing detection methods can be classified into two categories:

algorithms based on traditional models and algorithms based on deep learning.

Traditional significance target detection algorithm. Early approaches in this field mainly used bottom-up strategies to achieve their goals. Zhou et al. analyzed the advantages and disadvantages of different visual cues, such as compactness, uniqueness and target, and found that the significance regions affected by compactness could be accurately detected by local comparison method, thus creating a bottom-up significance detection architecture [11]. Early salient object detection was mainly based on low-level information such as color, direction and brightness, and the mapping relationship between visual features of high-level information such as human visual awareness and prominent areas collected by human eyes. It was Itti et al., who first proposed the relevant concept of salient object detection algorithm, and proposed the "center-surrounding" difference method in 1998 [12]. Based on the encircling mechanism and biological structure, researchers selected a common network framework to design a model, formed a Gaussian pyramid with various features, calculated each feature using the difference mechanism, and then linearized them to form a significance graph, and then obtained the most significant position by using the biological structure [13].

The most outstanding effect of this study is that it is the first time to assign visual attention machine to a computational model and generate the most obvious salient prediction graph through the model, which also makes a great contribution to the research of the field and related fields, and its central idea has been repeatedly speculated and constantly innovated by later generations. Cheng et al. for the first time modeled and calculated the principle of significance prediction graph in the form of center-surrounding comparison and proposed the RC algorithm, which can divide an image into multiple superpixels, weigh the gap between an image block and the spatial distance, and use it to analyze and calculate the color-weighted contrast image of each image block [14]. Shi et al. found in the study of human psychology that human attention system needs to complete multi-level analysis when making judgments.

Compared with other schemes, this process is much more complicated, but at the same time, its accuracy is much better, so a new model was proposed based on this change. Ji et al. obtained the final saliency map in three steps. In the first step, the affinity matrix and the pull Laplace matrix of the image were calculated to extract simple features; in the second step, the initial saliency map was obtained using the popular sorting method; and in the third step, the multilayer cellular automata was used to predict the final prediction map. Wei et al. proposed boundary connectivity evaluation in 2014 to isolate the significance targets. Klein et al. chose K-L divergence in information theory to weigh the difference between the central orientation of image features and its peripheral features. These algorithms are all based on the local pixel comparison Angle to complete the final selection, it is not very good in detecting the whole object, but it is excellent in processing the edge of the object. In order to improve the accuracy and robustness of significance detection based on boundary priors, Yuan et al. proposed significance inversion correction to remove foreground superpixels near the boundary.

Significant object detection algorithm based on deep learning. With the emergence of a large amount of data and the continuous improvement of artificial intelligence, deep learning has far-reaching influence in the field of computer vision. Deep learning and machine learning also have excellent performance in processing image problems, and researchers have also introduced deep learning into the significance detection algorithm. Compared with traditional algorithms, it can be seen that the significance detection model based on deep learning no longer relies on manual extraction of features and prior knowledge, and its performance has been greatly improved. So far, its performance is closest to the level observed by the human eye. Among them, convolutional neural network is one of the most representative algorithms based on deep learning. Li et al proposed a significance detection model based on convolutional neural network, which relies on convolutional neural network to extract multi-scale features and obtain initial image features by using spatial information, and finally implement feature fusion of information to obtain final results. Wang et al. used convolutional neural networks combined with significant prior knowledge to supplement the output of the previous round through iterative loops to obtain significant features. With the advent of RGBD images with color information and depth information, more resources are provided for tasks such as saliency detection and image segmentation. Good depth images contain high-quality depth information, and their information can highlight significant objects under interference. Ge et al. proposed a weighted group integration network using RGBD images. The algorithm integrates the depth information and color information of the RGBD image for the two-stream network structure, and then detects the image according to the residual thought.

## 3. Model

### *3.1 Convolutional Neural Networks*

Convolution neural network (CNN) is a highly efficient identification technology has attracted much attention in recent years. This network can directly input image data without pre-processing, and is the most widely used neural network at present. Convolutional neural networks are mainly divided into three parts: input layer, hidden layer and output layer, among which the hidden layer generally consists of convolutional layer, pooling layer, activation function layer and fully connected layer, as shown in Figure 1

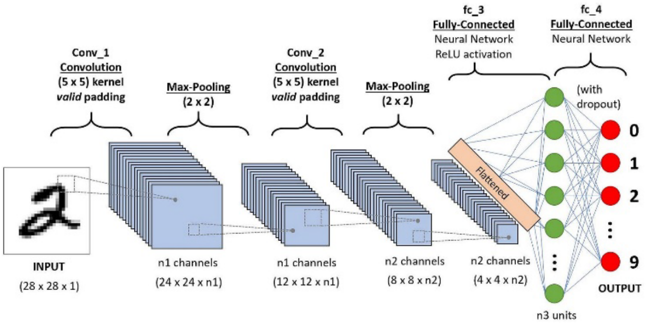

*Figure 1: The main component of convolutional neural network*

Color images usually contain RGB three color channel information, can use the pixel matrix to represent the image features, usually h, w, d to represent the length pixel, width pixel and depth pixel of the input image. The convolution layer contains multiple convolution nuclei, which are equivalent to multiple filters. Each convolution kernel can be regarded as a weight matrix with the size of $N \times N$, where each element corresponds to a coefficient and deviation of connecting weights. The convolution operation of the convolutional layer can extract various features from the input image, and the coverage area of the convolutional kernel is the "receptive field". Unlike low-level convolution, which can extract edge features, the receptive field corresponding to the upper layer of the convolutional layer is larger, and more complex semantic information can be extracted from the input. The calculation process of single-layer convolution is shown in Figure 2.

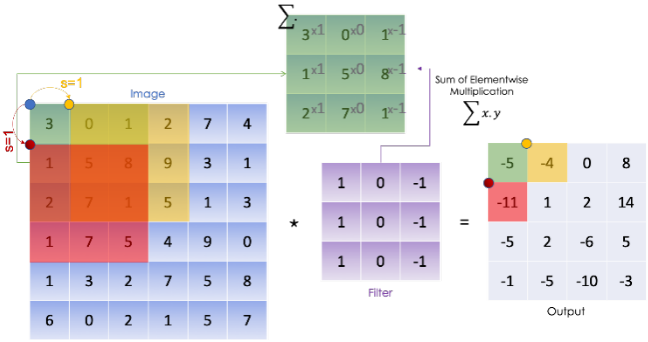

*Figure 2: Single-layer convolution calculation process*

The leftmost 5×5 grid in the figure is the input image, where the values in each grid represent pixel values in the range [0,255]. The middle 3×3 grid is the convolution kernel, and each value in the grid represents the weight of the convolution kernel. On the far right is the output matrix obtained after the convolution operation of the convolution layer on the input image, which is called the output feature map. First of all, it is decided whether to fill zeros around the input image according to the actual situation. Then, the convolution kernel starts from the top left of the input image and slides horizontally and vertically according to specific step sizes to carry out convolution operations and obtain the output feature graph with the size shown in formula (1).

$$N = \frac{W - F + 2P}{S} + 1 \tag{1}$$

Where W represents the size of the feature layer of the input image, F, S, and P respectively represent the size of the convolution kernel, the sliding step length, and the number of filled pixels.

In order to meet the requirements of the model under different conditions and improve the nonlinear expression ability of the model, different activation function layers are used to nonlinear activate the features extracted from the convolutional layer. Common activation functions include Sigmoid, Hard-Sigmoid, Tanh, ReLU, ReLU6, Swish, and Hard-Swish. Sigmoid activation function is a more commonly used activation function, belonging to saturation activation function with output interval of (0,1). When the input is large or small, the function value tends to be fixed, and the gradient is likely to disappear. In the process of backpropagation, it is difficult to update the weight, the loss function is difficult to decrease, and the output center is not zero, which will lead to low efficiency of weight update, and the network cannot be trained or the training effect is poor. Its mathematical expression is shown in equation (2).

$$Sigmoid(x) = \frac{1}{1+e^{-x}} \tag{2}$$

Tanh activation function has zero value at the origin and output interval is (-1,1). Compared with sigmoid function, TANH activation function has faster convergence speed and improved weight update efficiency. However, it still has the problem of gradient disappearance and exponential calculation, which is time-consuming. The function expression is shown in equation (3).

$$\mathrm{Tanh}(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{3}$$

Some activation function curves are shown in the figure 3.

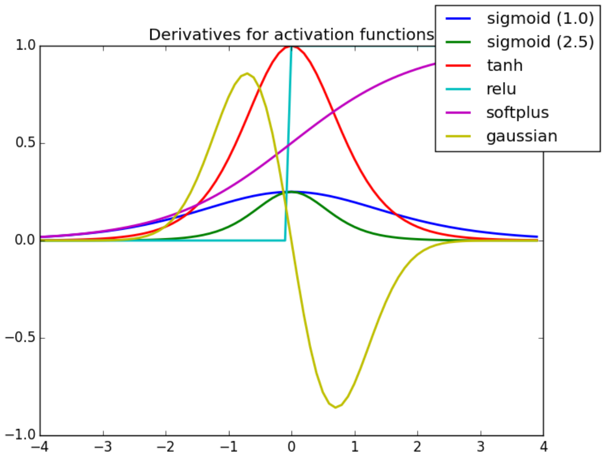

*Figure 3: Activation function graph*

### *3.2 YOLOv5 Network*

YOLOv5 algorithm is an algorithm in the current YOLO series, which belongs to a single-stage target detection algorithm. The detection performance of this algorithm is not the same as that of YOLOv4, but the model size is only one ninth of YOLOv4. It has the advantages of simplicity and easy deployment, and is widely used in target detection, tracking, segmentation and other scenarios. On the whole, the YOLOv5 network can be divided into four parts: Input (input), Backbone (backbone feature extraction network), Neck (multi-scale feature fusion module) and Head (detection head). Figure 4 shows the network structure of YOLOv5 network in Generation 6.0.

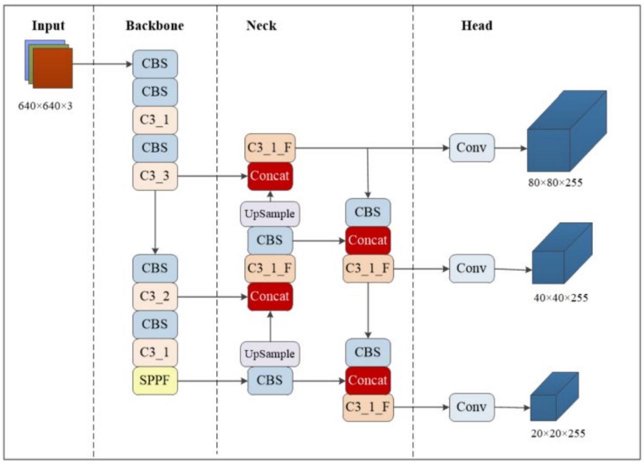

*Figure 4: YOLOv5 network diagram*

The YOLOv5 algorithm uses adaptive anchor frame calculation to obtain the optimal anchor frame value. The YOLO series algorithm presets an initial anchor frame with a certain size and aspect ratio. In the training, the network will output the prediction box based on the initial anchor box, and compare it with the real box, reverse update by reducing the error, and iteratively update the network parameters, and finally get the detection model with the most appropriate parameters. Unlike previous generations of algorithms that calculate the anchor frame size independently, YOLOv5 algorithm can automatically learn the anchor frame size during each training, and adaptively calculate the best anchor frame value in different training sets by using K-means and genetic learning.

### *3.3 Attention Mechanism*

The Attention Mechanism is inspired by the way the human brain processes image information. When observing the global information of the image, attention can quickly lock the focus area, automatically mask part of the background and redundant information, analyze the existing network feature map, and make more accurate correction. Extended to neural networks, the attention mechanism is a way for the network to automatically focus on areas that require more attention. Compared with convolutional neural networks, such as increasing the size of convolutional nuclei and increasing the depth and width of the network, the attention mechanism can bring better network performance through less computing cost. Attention mechanism can be divided into channel attention mechanism, spatial attention mechanism, and the combination of the two. Given a set of queries Q, a set of keys K, and a set of values V, attention is calculated as follows.

$$AS = soft\max(\frac{QK^T}{\sqrt{d_k}}) \tag{4}$$

$$AW = AS \bullet V \tag{5}$$

$$Out = AW \tag{6}$$

Where $d_k$ the key dimension and $QK^T$ represents the inner product of the query and the key. The attention weight is calculated by multiplying the attention score with the value, and the final output can be weighted values that indicate how much the model focuses on different values.

## 4. Experiments

### *4.1 Datasets*

In order to train and evaluate the capabilities and performance of the designed collaborative saliency object detection model, researchers have continuously adjusted and proposed datasets tailored for the collaborative saliency object detection task to meet specific requirements. These datasets account for variations in image data influenced by factors such as deformation and blurriness in the foreground region, but generally maintain uniformity in the background region. The most commonly used datasets include MSRC, iCoSeg, CoSal2015, CoSOD3K, and CoCA. As shown in Figure 5, partial examples from these datasets are presented.

In recent years, Microsoft Research has proposed MSRC datasets for collaborative salience target detection tasks. It contains 8 combinations of 240 manually labeled pixel-level real labels, with approximately 30 to 53 images per category. MSRC data sets are not consistent in color and appearance of consensus salient information, but consistent in semantic information. Therefore, due to this feature, MSRC data sets are widely used in collaborative salient object detection. Of course, it can be seen from the figure that this data set is relatively simple, and the appearance difference of the same significant target is small, and there are fewer interference factors. Therefore, the performance of this data set in co-significant target detection is not as good as that in significance detection.

The CoSal2015 dataset was proposed in 2015 by Zhang et al., which contains 2015 images with a total of 50 categories, and each type of image is also composed of 26 to 52 images. At the same time, this dataset is widely used in cooperative significance target detection tasks. It contains many deceptively similar non-cooperative significance targets, which is more challenging and larger than other datasets.

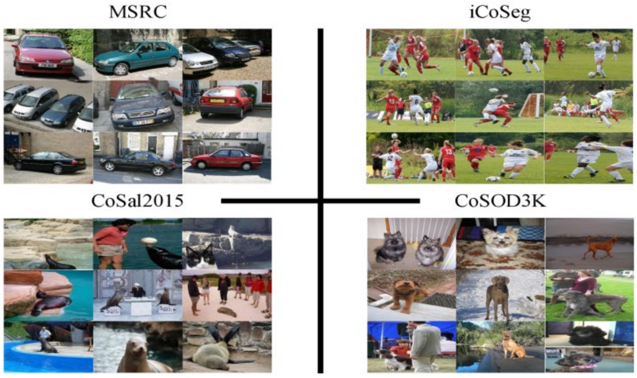

*Figure 5: YOLOv5 network diagram*

### *4.2 Evaluation index*

The data set of the simulation experiment in this article has a to In daily life, people usually use their eyes to assess the quality of things, but the premise of being able to make judgments is that people have accumulated rich experience or have a reference. In contrast, for the collaborative significance target detection algorithm, there should be a unified standard to predict the ability of the algorithm. Here, a fair and objective evaluation index is proposed. The unified evaluation index can quickly and objectively evaluate the quality of the cooperative significance object detection algorithm in the processing of problems. The principle is to measure the performance of the network model by comparing the coincidence degree between the collaborative significance result graph predicted by the proposed network model and the pixel-level truth value graph. Up to now, there are six evaluation indexes used to evaluate the co-salience object detection algorithm. They are precision recall curve (PR curve), Average Precision score (AP), Mean Absolute Error value (MAE). The following is a detailed explanation of the above three evaluation criteria.

The accuracy rate is an indicator used to calculate the proportion of the forecast plot in the sample result plot. The principle is to quantize the value of each pixel in the cooperative significance graph by calculating the threshold between 0 and 1 to obtain significant and non-significant values. The accuracy-recall curve is drawn.

$$PR = \frac{TP}{TP + FP} \tag{7}$$

AP is obtained by statistical calculation of the graph area composed of horizontal and vertical coordinates under the PR curve, so it is similar to the principle of PR.

Mean Absolute Error translates to mean absolute error value. Its purpose is to avoid the positive and negative balance in the sum. It presents the error between the predicted value and the real value objectively by calculating the deviation between the cosignificance result graph and the target observation value and taking the average.

$$MAE = \frac{1}{N}\sum_{t=1}^{N} | f_t - y_t | \tag{8}$$

### *4.3 Experimental result*

In this paper, the proposed model is compared with YOLO v3 and YOLOv4, and the results are shown in Table 1.

*Table 1. Comparison between YOLO v3 and YOLOv4*

| Models | PR | AP | MAE |
|---|---|---|---|
| YOLO v3 | 89.3% | 87.1% | 87.7% |
| YOLOv4 | 91.2% | 89.6% | 88.4% |
| Our | 93.7% | 93.2% | 90.6% |

As can be seen from the above table, compared with model 1, mAP of model 2 increases by 2.6%, P value by 1.7%, and R value by 2.2%. In model 3, mAP increases by 0.6% on the basis of model 2, and P and R values have little change. The mAP of Model 4 is as high as 94.9%, which is increased by 0.7% on the basis of model 3 and 3.9% on the basis of model 1, which has a good detection effect.

The table shows that our model has the highest PR, AP, and MAE among the three models. Compared to YOLO v3, our model has a 4.4% increase in precision, a 6.1% increase in AP, and a 3.2% decrease in MAE. Compared to YOLOv4, our model has a 2.5% increase in precision, a 3.6% increase in AP, and a 2.2% decrease in MAE.

As can be seen from the figure 6, YOLOv3 has the lowest mAP value, the lowest detection accuracy for the three types of traffic signs in the dataset, and the detection effect is not good. YOLOv4 has the highest mAP value and the best detection effect. Compared with our model the mAP value of our model has decreased, but it is still higher than that of YOLOv3, and the detection accuracy is higher. At the same time, it can be seen that YOLOv4 and our model can achieve an ideal result in fewer rounds, and have a faster convergence rate than the benchmark model.

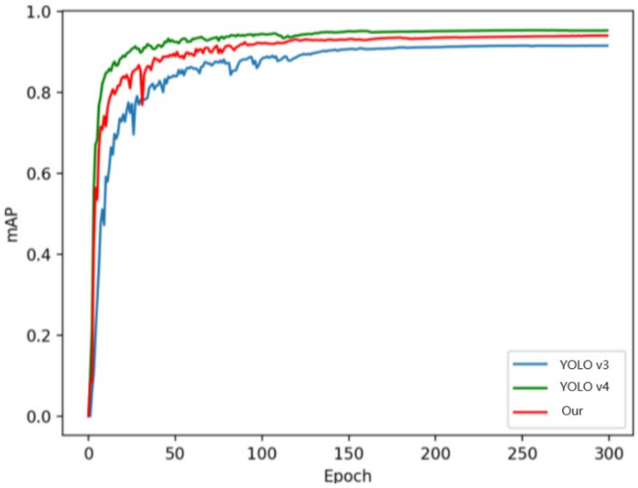

*Figure 6: mAP Curve comparison*

## 5. Conclusion

In summary, the object detection algorithm based on neural network has a wide application prospect in the field of computer vision. This paper reveals the potential of object detection through neural network technology. Compared with traditional methods, the object detection algorithm based on neural network automatically learns image features through convolutional neural network, which has stronger generalization ability, and provides an effective means to improve detection accuracy and reduce false detection rate. The paper also highlights the performance differences of neural network architectures in solving target detection problems, and points out the advantages and disadvantages of different approaches in different application scenarios. By delving into datasets and evaluation metrics, we provide critical support for the training and evaluation of algorithms, helping researchers compare the performance of different algorithms.

## References

*[1] Sherstinsky, Alex. "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network." Physica D: Nonlinear Phenomena 404 (2020): 132306.*

*[2] Graves, Alex, and Jürgen Schmidhuber. "Framewise phoneme classification with bidirectional LSTM and other neural network architectures." Neural networks 18.5-6 (2005): 602-610.*

*[3] Chang, Yue-Shan, et al. "An LSTM-based aggregated model for air pollution forecasting." Atmospheric Pollution Research 11.8 (2020): 1451-1463.*

*[4] Lee I, Kim D, Kang S, et al. Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks[C]//Proceedings of the IEEE international conference on computer vision. 2017: 1012-1020.*

*[5] Hou, Jialu, Hang Wei, and Bin Liu. "iPiDA-GCN: Identification of piRNA-disease associations based on Graph Convolutional Network." PLOS Computational Biology 18.10 (2022): e1010671.*

*[6] Eliasof, Moshe, Eldad Haber, and Eran Treister. "Pde-gcn: Novel architectures for graph neural networks motivated by partial differential equations." Advances in neural information processing systems 34 (2021): 3836-3849.*

*[7] Peng, Shaowen, Kazunari Sugiyama, and Tsunenori Mine. "SVD-GCN: A simplified graph convolution paradigm for recommendation." Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 2022.*

*[8] Wu, Lingfei, et al. "Graph neural networks: foundation, frontiers and applications." Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2022.*

*[9] Liu, Zhiyuan, and Jie Zhou. Introduction to graph neural networks. Springer Nature, 2022.*

*[10] Wang, Xiyuan, and Muhan Zhang. "How powerful are spectral graph neural networks." International Conference on Machine Learning. PMLR, 2022.*

*[11] Zeng, D., Liu, W., Chen, W., Zhou, L., Zhang, M., & Qu, H. (2023, June). Substructure aware graph neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 9, pp. 11129-11137).*

*[12] Zhang, Chenxiao, et al. "A domain adaptation neural network for change detection with heterogeneous optical and SAR remote sensing images." International Journal of Applied Earth Observation and Geoinformation 109 (2022): 102769.*

*[13] Gianola, Alessandro. "DABs: a Theoretical Framework for Data-Aware BPMN." Verification of Data-Aware Processes via Satisfiability Modulo Theories. Cham: Springer Nature Switzerland, 2023. 213-238.*

*[14] Kinoshita, M., Komasaka, M., & Tanaka, K. (2023). ChatGPT's performance on JSA-certified anesthesiologist exam. Journal of Anesthesia, 1-2.*