Multimedia Tools and Applications (2024) 83:66517–66538
https://doi.org/10.1007/s11042-024-18219-7

PANetW: PANet with wider receptive fields for object detection

Ran Chen[1] · Dongjun Xin[1] · Chuanli Wang[1] · Peng Wang[1] · Junwen Tan[1] · Wenjie Kang[2,3]

Received: 12 September 2023 / Revised: 25 December 2023 / Accepted: 8 January 2024 /
Published online: 24 January 2024
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract
PANet is widely used in various object detection tasks due to its powerful feature expression ability. However, PANet's performance in complex scenarios is subpar, with frequent object omission or misidentification. We find that the reason for this phenomenon is that the receptive field of PANet can't cover sufficient feature information, to deal with drastic changes of source object size. In order to solve this problem, this paper adopts dilated convolution technology and applies it to each parallel branch directly following the PANet network. This method can effectively represent the feature information of objects at different scales by integrating the information from small and large receptive fields into a new feature output. We also introduce residual structure to circumvent the network degradation caused by excessive convolutions. By combining the above methods, we build a new module named PANetW (PANet with Wider Receptive Fields). Taking YOLOX-S as the baseline, we comprehensively evaluated the proposed module PANetW on two datasets, VOC2007 and MSCOCO2017. The test results show that our PANetW achieves a high level of mean average precision (AP). On the VOC2007 dataset, the AP of our PANetW improves by 4.9% to 43.0%; on the MS COCO2017 dataset, the AP of PANetW is as high as 44.3%, far exceeding the current mainstream modules. The experimental results fully demonstrate the effectiveness of our module.

Keywords Object detection · Receptive fields · Objects of different scales · Dilated convolution · Residual structure

1 Introduction

Object detection is an important and challenging task in computer vision. Its purpose is to detect as many target objects as possible, classify the target objects, and divide the target

✉ Dongjun Xin
xindj@126.com

[1] Central South University of Forestry and Technology, 498 Shaoshan South Road, Changsha 410004, Hunan, China

[2] Hunan Provincial Key Laboratory of Network Investigational Technology, Hunan Police Academy, Changsha, China

[3] College of Systems Engineering, National University of Defense Technology, Changsha, China

object domain frames. This task plays an important role in various industries, such as video surveillance, robotics, and drone scene analysis. With the help of deep convolutional neural network learning, researchers have proposed a variety of detection frameworks to accomplish this task, such as RCNN [9, 10, 26], SSD [22] and YOLO [1, 7, 13, 14, 16, 23–25, 31]series.

The YOLO series is a representative framework for pursuing the best trade-off between speed and accuracy. Among them, YOLOX [7] plays a pivotal role in the field of object detection because of its obvious advantages in speed and accuracy. According to the original intention of the YOLO series, YOLOX integrates many advanced technologies in the field of object detection, such as decoupled head, Anchor-Free, adaptive label assignment strategy, and provides models with different parameter scales for users to choose flexibly. As time goes by, more and more people use YOLOX, and further improvements have been made to YOLOX, such as PP-YOLOE [34], YOLOCa [6]. There is no doubt that YOLOX is one of the current excellent models.

The way of information propagation in a neural network is very important. YOLOX uses PANet [21], uses the weight sharing of its neural network to extract image features, and integrates image features at different scales to obtain multiple better features. PANet is an enhanced version of FPN [19], which shortens the information path between the bottom and top features by fusing the bottom-up and top-down paths, and successfully avoids rigid matching of target size and network depth. Thus, the representation capability of the backbone network is enhanced. Therefore, PANet has become one of the most popular feature fusion structures.

Although, YOLOX-L won the first place in the Streaming Perception Challenge on WAD 2021. However, when we use YOLOX-L, we find that the average inference time is as high as 15.18ms, and the number of parameters is large and the floating point operation is many. However, the limited memory and computing resources of mobile devices are often not available for large-scale network. And the datasets are often difficult to collect in practical applications. This limits its applicability to devices. Therefore, in most real-life applications, people tend to use YOLOX-S. However, the experiments in this paper found that in complex scenes, the object detection capability of YOLOX-S would be affected, and the accuracy of YOLOX-S would drop rapidly. The ultimate reason is that receptive fields obtained by YOLOX-S using PANet has limitations and the target scale range is not flexible enough to handle the object detection scenarios with drastic changes in target size. As mentioned by Yanghao Li [17], the network is sensitive to receptive fields when measuring targets of different sizes. When this paper tries to use YOLOX-S to process the PASCAL VOC2007 [5] dataset, it is realized that this problem urgently needs to be solved.

In order to solve the above problems, this paper proposes PANetW (PANet with Wider Receptive Fields). While further expanding the range of PANet's receptive fields, it also retains the advantages of PANet supporting feature fusion at various scales to generate enhanced multi scale features. The specific module is as follows: Firstly, this paper uses dilated convolution to expand receptive fields without losing resolution, so that receptive fields grow exponentially; Secondly, we use a residual structure to superimpose the computational results of the inflated convolution with the original output, thus solving the problem of network degradation caused by the increase in network depth. At the same time, the use of the residual structure can not lose the detection accuracy of small targets, and can also combine the better semantic information obtained by the large receptive fields with the precise positioning information obtained by the small receptive fields, thereby improving the semantics and positioning accuracy of the network. The results show that using PANetW instead of PANet can significantly improve the performance of YOLOX.

The main contributions of this paper are as follows:

- We propose an innovative module called PANetW by combining dilated convolution technology with PANet, which not only preserves the resolution but also expand the model's receptive field. This module can effectively address the issue of drastic changes in target size in complex scenes when it is applied to object detection.
- We also propose an optimization method that combines our PANetW module with residual structure, effectively addressing the problem of network degradation caused by deepening the network, and further improving the success rate of object detection.
- Based on YOLOX-S, we conducted comprehensive experiments on the PASCAL VOC2007 and MS COCO2017 datasets, verifying the effectiveness of PANetW.

This paper focuses on the lightweight model YOLOX-S. If there is no specific description of YOLOX in the follow-up paper, the uniform default YOLOX-S. Code is available at https://github.com/ChenRan2000/PANetW.

The subsequent organization of this paper is as follows: Section 2 provides background knowledge about receptive fields and network degradation. Section 3 describes the improved method PANetW in detail. Section 4 gives the experimental results of this paper. Section 5 concludes the paper.

2 Related work

Feature pyramid [19] is a staple in areas such as object detection and image segmentation. In the realm of artificially intelligent research, Feature Pyramid proposal has sparked lively discourse and garnered substantial interest, sparking renewed passion for exploration.Feature Pyramid technique skillfully merges backbone network derived features across varying scales, creating detailed and adaptive visual info representations. It remains an active area of inquiry within academia today. Academics have mostly focused on two areas when delving into Feature pyramids - modifying the network's feature fusion method [8, 21, 30, 36] or adjusting its module [20, 37]. The main obstacles they face here are a restricted receptive field and the potential for degradation within the network.

2.1 Receptive fields

According to Karan [15], each of the generative models must complete a task: to extract semantic information from the input and generate a meaningful output. Therefore, the importance of the receptive field that model features can perceive is self-evident. Most object detection tasks require larger receptive fields. For example, in image classification tasks, the final receptive fields must be larger than the target area in order to effectively detect the target. There are many ways to increase network receptive fields. SPP (Spatial Pyramid Pooling) [11], a milestone in the object detection field, compresses the features through the MaxPool layer, and then fuses features through the residual structure to increase its receptive fields and avoid repeated extraction of image features. So that it can save computing costs. Since the proposal of SPP, researchers have followed this direction all the way. On this basis, the author of YOLOv5 [13] changes the original parallel connection to serial connection by changing the MaxPool layer connection method, resulting in a faster SPPF. YOLOv6 [16] also continues to follow the steps of the predecessor, changing the original activation function from SiLU to ReLU, which makes the speed even faster. Li names it as SimSPPF. The recent popular YOLOv7 [31] also continues to use this framework, but adds two convolutional layers after the MaxPool layer to expand receptive fields, and names it as SPPCSPC. Although the number of parameters and calculation of SPPCSPC has been greatly increased, its performance is significantly better than SPPF. This further shows that receptive fields of the current network structure design may generally have a problem of insufficient range. At present, the most popular modules like SPP, SPPF, SimSPPF and SPPCSPC, all use the MaxPool layer to increase receptive fields. But this method would also severely reduce feature resolution and loss image detail information. If you want to expand the receptive fields of the models, the above methods have some limitations. Using dilated convolution is a good solution to these limitations. DeepLabv2 [3] proposes ASPP. ASPP uses four dilated convolutional layers with different dilation factors in parallel to capture image information at different resolutions, fuse multiscale information, and successfully obtain more robust results. Then, TridentNet [17] uses three branches of the dilated convolution with different dilation factors, and each branch creates a size-specific feature map to identify objects of different sizes. In addition to the above methods, there are other methods that can also increase the receptive field. For example, MiSoNet [35] used deep separable convolution in parallel through a multi-level feature integration module and a deep convolutional residual encoder, which further expands the scale range of the receptive field. Zhou [38] also solved this problem by improving the model label assignment strategy. He upgraded the model backbone and feature fusion method by using the shortest-longest gradient strategy and self-attention mechanism in the model respectively, and named the new model YOLO-NL.In the process of research, different researchers proposed a variety of solutions for different application scenarios. For example, in the process of studying the significant object detection of light fields, Wang et al. [33] creatively embedded the Transformer into the light-field network, so that each pixel can obtain a larger receptive field, thus improving the effect of feature representation.In the process of studying pedestrian perception and vehicle perception, Alan [29] conducted in-depth research on the dataset. In order to improve the accuracy of the model, he added a synthetic dataset he created to the training dataset, which could expand the scope of the dataset and obtain better results.

2.2 Network degradation

Each layer of the neural network corresponds to extracting different sizes of feature information. Under certain conditions, deeper networks generally work better than shallow networks. However, when the network is too deep, it is likely to encounter problems such as slow convergence, over-fitting, gradient disappearance, and gradient explosion. Take VGG [27] as an example, as the network depth increases, the accuracy may be saturated and then degrade rapidly. This is not caused by over-fitting. Because the over-fitting is manifested by a small error in the training set and a large error in the test set. But when this happens, both the training set and the test set perform worse than the shallow network. This is called network degradation. How can we deepen network layers without causing gradient disappearance? Highway Network [28] introduces a transform gate mechanism based on the traditional neural network. So that Highway Network can keep part of the input unchanged according to the data characteristics, and the other part of the input is processed in the same way as a general feed-forward neural network. It can skip useless layers, and speed up information transfer. Highway Network can be better iteratively optimized, and the convergence speed is faster. Since then, Kaiming He believes that Highway Network relies on data and uses parameters. When the selected data characteristics are not reasonable enough, the Highway Network will fail. Furthermore, there is no improvement in accuracy when the depth is greatly increased. Residual structure [12] comes into being. Residual structure is designed to enable the network to have the capability of identity mapping, while deepening the network, to ensure that the deeper networks and the shallower networks have the same impact on the output. Different from Highway Network, Residual structure implements identity mapping, so that data can flow across layers. It allows the model itself to have a more flexible structure, so that the model can choose whether to do more convolution and linear transformation, or more inclined to do nothing in each part during the training process. The model can adapt to its own structure during training. And the residual structure also achieves accuracy improvement after the network depth is extremely deepened. Residual structure has made great achievements in the neural network industry, and it can be seen in many subsequent models.

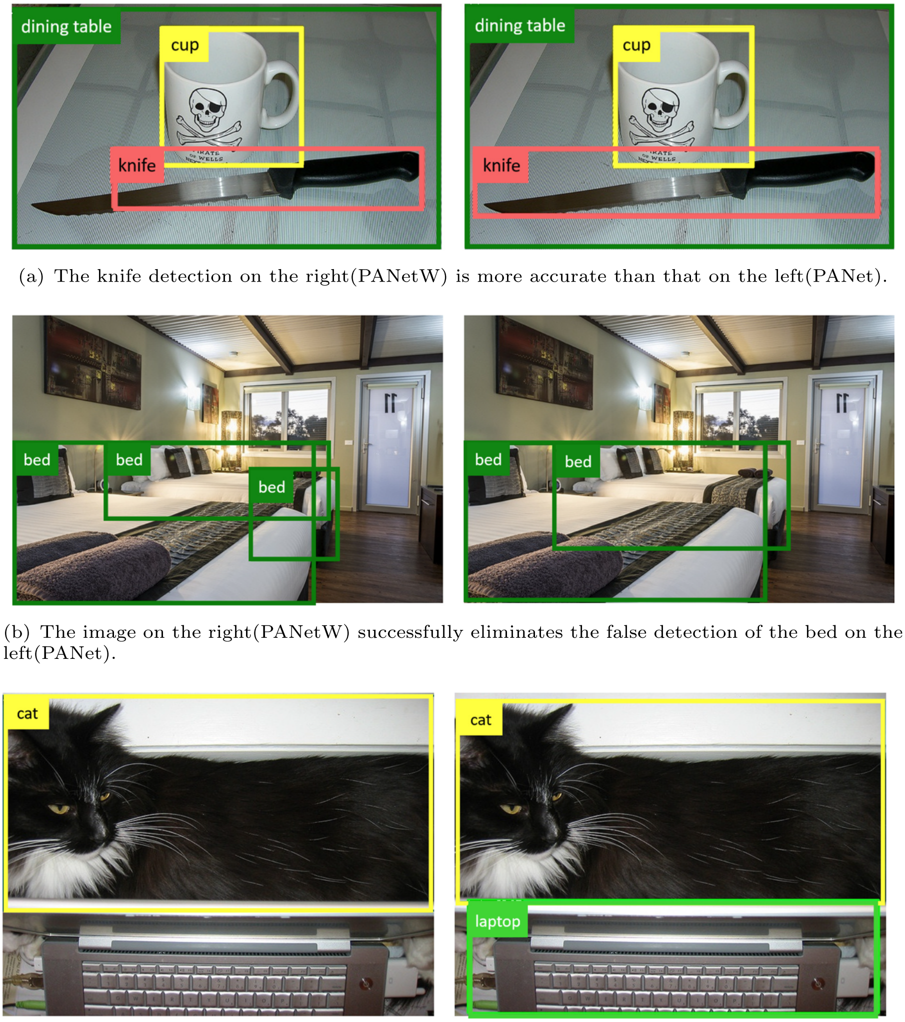
3 Methods

In object detection tasks, detecting objects of various scales is always challenging. Many aspects of network design, such as sampling rate, network depth, receptive field, etc., may affect the performance of object detection. This paper deeply studies the receptive field problem of PANet used in YOLOX. During the experiment, it is found that the receptive field extracted by PANet has a limited range of scales and cannot fully match the different receptive fields of various targets, which affects the performance of cross-scale detection objects. In complex scenarios, the detection effect is not good.

To further solve this problem, this paper proposes PANetW. As shown in Fig. 1, the left side of the comparison chart is YOLOX using PANet, and the right side is YOLOX using PANetW. It can be clearly seen that the precision of the detection frame using PANetW is generally higher than that using PANet. After observing the knife in Fig. 1a and comparing the detection effect, it is found that YOLOX using PANet can only detect a portion of the knife due to the insufficient receptive field, and cannot detect the overall structure of the knife. Figure 1b in the PANet detection effect drawing shows the repeated detection of the target bed. The main reason is that the features obtained by PANet, in which the small receptive field detection frame does not combine enough receptive field information. The network structure identifies the mattress of the bed while detecting the bed's overall structure. However, the small receptive field detection frame provides an additional detection frame for this bed, which results in redundant detection. The missed detection of the keyboard in the diagram on the right side of Fig. 1c is also one of the manifestations of insufficient PANet receptive field range. The designed framework for using PANetW on YOLOX is shown in Fig. 2. First, we perform 1×1 convolution operations on the branches of PANet with 512 channels, 256 channels, and 128 channels respectively. Then, we use the Batch Normalize and SiLU modules in sequence to perform fine semantic processing on the output features of PANet. Finally, we input the processed results into our dilated block. In this section, this paper will describe the PANetW structure in detail.

3.1 Dilated convolution

In this study, we found that the range of receptive fields extracted by PANet only covers a limited scale range and cannot flexibly cope with scenarios where the size of the target changes sharply. If the target object does not match the scale of the receptive field, the performance will be degraded. To make up for this shortcoming, we need to generate output features with multiple receptive field sizes. The traditional solution is to repeatedly use down-sampling and large-stride convolutional layers in network architectures, but this approach



(a) The knife detection on the right(PANetW) is more accurate than that on the left(PANet).

(b) The image on the right(PANetW) successfully eliminates the false detection of the bed on the left(PANet).

(c) The right figure(PANetW) successfully makes up for the missing detection of laptop in the left figure(PANet).

Fig. 1 Comparison of detection results between the module PANetW (right) and the original module PANet (left)

results in a significant reduction in the spatial resolution of feature maps. However, studies have shown that dilated convolution exhibits extremely high performance in reducing the loss of accuracy caused by the use of convolution [20, 22]. Compared with ordinary convolution, dilated convolution can expand the receptive field by increasing the size of the convolution kernel, so its number of parameters is less than that of ordinary convolution. However, dilated convolution adds holes in the convolution kernel, which increases the interval of the convolution kernel, so it can receive more context information, and realizes receptive field expansion. This is very beneficial for tasks that require a lot of context information, such as entity naming. It can increase the receptive field of the network structure well without

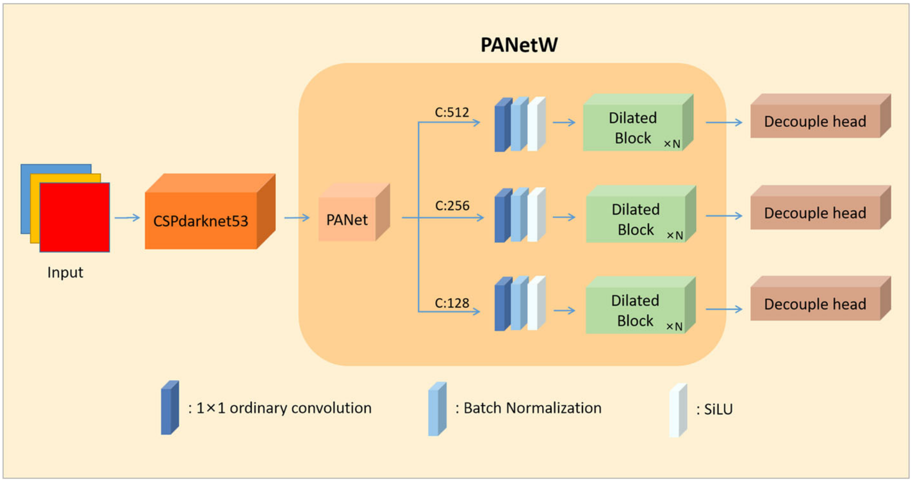


Fig. 2 Network structure figure of PANetW used in YOLOX. C indicates the number of channels. N indicates that N Dilated Block are used continuously

losing accuracy. Ordinary convolution usually requires operations such as pooling to reduce dimensionality, but this will lead to loss of information, while dilated convolution can avoid loss of information by reducing the number of holes. Therefore, we recommend using dilated convolution to solve the problem of insufficient receptive fields (Fig. 3).

Dilated convolution is an adjustable convolution kernel, and its hyper-parameter S controls the interval between elements to $S-1$. When S is equal to 1, dilated convolution degenerates into ordinary convolution; when S is equal to 2, the operation of dilated convolution is as shown in the Fig. 4.

It can be clearly seen that compared with ordinary convolution, dilated convolution adds holes in the convolution kernel, which makes it expand the field of view without adding parameters and complexity, while avoiding the boundary effect caused by the small convolution kernel. It can better preserve the image edges, so as to better expressive and utilize the

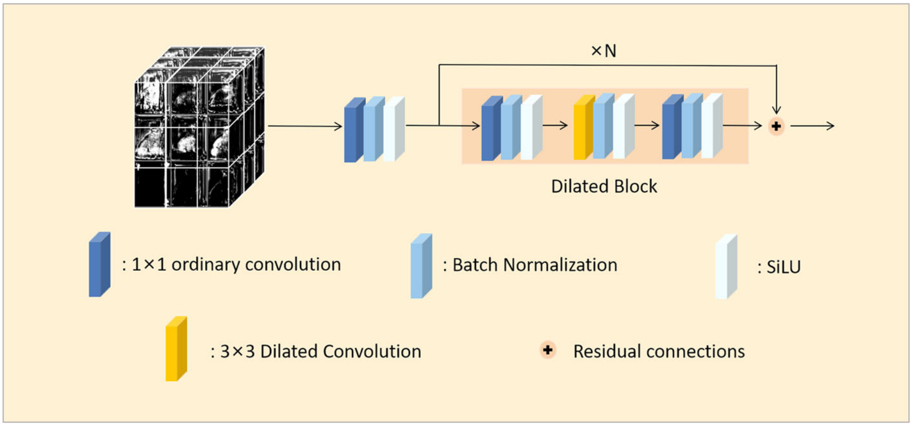
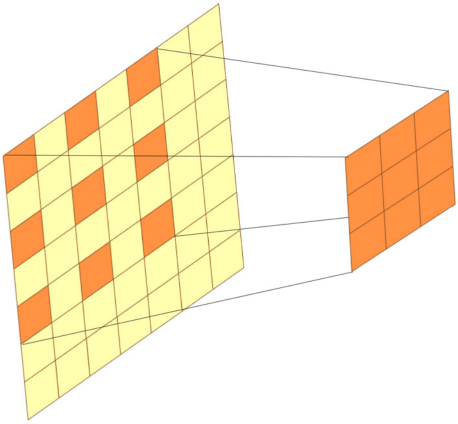


Fig. 3 Structure diagram of Dilated Blocks

Fig. 4 Visualization of Dilated Convolution



feature information in the image, and further improve the recognition ability of the model. Specifically, the receptive field relationship between ordinary convolution and dilated convolution using a uniform convolution kernel size can be expressed as:

$$K_c = H \times (K_o - 1) + 1 \tag{1}$$

$$S_c = H^2 S_o + 2H(1 - H)\sqrt{S_o} + (H - 1)^2 \tag{2}$$

Equation 1 reveals the relationship between the receptive field length of dilated convolution and ordinary convolution when the same convolution kernel is used, where K_c represents the receptive field length after dilated convolution processing, and K_o represents the receptive field length after ordinary convolution processing. Since both convolutions use convolution kernels of the same size, their parameter quantities are the same. Equation 2 illustrates the relationship between the receptive field area of dilated convolution and ordinary convolution when the same convolution kernel is used, where S_c represents the receptive field area after expansive convolution, and S_o represents the receptive field area after ordinary convolution. Equation 2 shows that the receptive field area of dilated convolution has been multiplied. At the same time, in order to retain the edge information of the image and increase the width of the receptive field without reducing the resolution, this paper uses padding to fill the edge information, so that the image information can be more fully utilized.

PANet outputs three parallel branches and inputs them to the Decouple head. In this paper, we extend the YOLOX receptive field by placing N Dilated Blocks on each of the three parallel output branches to generate output features with different size receptive field scales. In order to expand the PANet's receptive field, N Dilated Blocks are incorporated into each branch of the parallel output pathway. This configuration is designed to generate receptive field features with various scales. As specified in this paper, N is set to 8. As shown in Fig. 3, each Dilated Block consists of a dilated convolutional component and two ordinary convolutional components. An ordinary convolutional block is composed of an ordinary convolution, Batch Normalize, and SiLU in turn. This paper sets the convolution kernel size of ordinary convolution to 1 because a 1 × 1 convolution kernel can not only easily change dimensions, but also reduce the amount of computation. Therefore, the solution of this paper is to reduce the channel to X times in the previous ordinary convolutional component, and restore the channel to the original number of channels in the latter ordinary convolutional component, so that the subsequent output to the Decouple head can be kept unaffected. This paper sets different dilated factors for the dilated convolutional layer of the dilated convolutional component in each dilated Block, which are 1, 2, 3, 4, 5, 6, 7 and 8, respectively. The acceptance field for each extended convolution is shown in the Fig. 5. This paper assumes that the pixel value of the cat image is 19 × 19, and each square represents a pixel value. As can be seen from the figure, when the dilation factor increases, the receptive field range increases, and the network can more accurately identify that this is a cat. When the dilation factor is small, more edge details can be captured, making the detection frame more accurate. Referring to YOLOF [4], this paper set X to 0.25.

3.2 Residual structure

Although stacking Dilated Blocks can indeed expand the receptive field to a certain extent, it cannot be denied that in the process of stacking Dilated Blocks, we will inevitably add convolutional blocks. Adding convolutional blocks may lead to the degradation of the network or vanishing gradient problem. On the other hand, although stacking Dilated Blocks can expand the size of targets, it cannot completely solve all the problems of target sizes we encounter in daily applications. Because in the real world, we often need to deal with a variety of different sizes of targets, and these target sizes are usually continuously increasing, rather than a simple proportional relationship. Therefore, while stacking Dilated Blocks is an effective technique, it cannot cover all the problems of typical target sizes in daily applications.

In order to solve this problem, this paper adopts the residual structure, which combines the large target features acquired in each Dilated Block with the small target features generated by the original output, as shown in Fig. 3. The residual structure is an important structure in deep convolutional neural network, which can help improve the accuracy and performance of the model. This structure forms a so-called "skip connection" by connecting the input and output layers, so that the output of the network can better match the original input. It can not only improve the accuracy of the network, but also solve the problem of layer disappearing or exploding during the deepening process of the model, thus making learning easier. In addition, this structure can effectively suppress the overfitting phenomenon. This structure

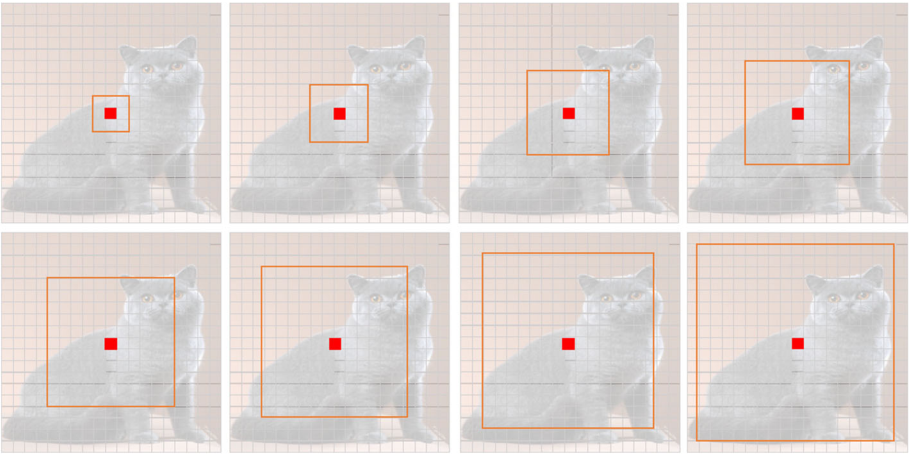


Fig. 5 Receptive fields size of the 8 dilated convolution used in PANetW. From left to right, from top to bottom are receptive fields effects when the dilation factors are 1, 2, 3, 4, 5, 6, 7, 8 respectively

can adjust the number and position according to the actual training needs to achieve better training results. The specific superposition formula is:

$$Y = F(x, \{x_i\}) + x \tag{3}$$

Where $x$represents the original output, $x_i$represents the current output after Dilated Block processing, and Frepresents the residual mapping to be learned. Through this superposition operation, the scale of the target object can be adapted, and the corresponding features can be output according to the information captured by the receptive field.

Finally, after the above experimental operation, we can obtain some feature maps of PANetW on three branches, as shown in Fig. 6. It can be seen that no matter which branch, PANetW can obtain better feature maps with semantic information and positioning information.

4 Experiments

4.1 Implementation details

This paper uses PyTorch to complete the implementation of PANetW. During the training process of the model, no pre-trained models are used, and all models are trained from scratch. To ensure a fair comparison, this paper keeps the original parameters of YOLOX unchanged, such as momentum set at 0.9, and trains for 300 epochs.

4.2 Datasets and indicators

In order to demonstrate the performance of PANetW, in the experimental testing process, this paper uses the PASCAL VOC2007 [5] and MS COCO2017 [18] datasets for training and testing. All the datasets in this paper are open-source datasets.

If not specified, the AP used for performance evaluation indicators in this experiment means: IoU is tested every 0.05 intervals from 0.5 to 0.95, and then the average of the measurement results is taken as the final AP. AP_{50} refers to the AP obtained by IoU setting 0.5, and AP_{75} refers to the AP obtained by IoU setting 0.75. $AP_s$is the detection accuracy for small targets (area $<$ 32 × 32), AP_m is the detection accuracy for medium targets (96 × 96 >area >32 × 32), and AP_l is the detection accuracy for large targets (area >96 × 96). AR is the average recall. The recall is the ratio of the sum of the correct result predicted by the positive sample to the correct result predicted by the positive sample and the wrong

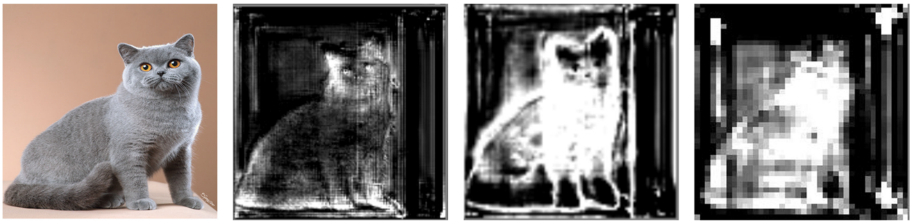


Fig. 6 From left to right is a pair of feature maps selected from the three branches output by PANetW respectively

result predicted by the positive sample, mainly reflecting the missed detection rate in the prediction results. AR_s is the average recall for small targets (area $< 32 \times 32$), AR_m is the average recall for medium targets ($96 \times 96 >$area $>32 \times 32$), AR_l is the average recall for large targets (area $>96 \times 96$). $AR_{max=1}$ is the average recall under the threshold of 1 number of predicted boxes per picture; $AR_{max=10}$ is the average recall under the threshold of 10 number of predicted boxes per picture; $AR_{max=100}$ is the average recall rate under the threshold of 100 number of predicted boxes per picture. $param$ is the abbreviation of parameter, and the unified unit in this paper is M. $FLOPs$ is the abbreviation of floating point operations, and the unified unit in this paper is G. It means floating point operations and is understood as the amount of computation. It can be used to measure the complexity of the model.

4.2.1 PASCAL VOC

The full name of PASCAL VOC is Pattern Analysis Statistical Modelling and Computational Learning Visual Object Classes. Everingham et al. [5] The PASCAL VOC dataset is provided by the PASCAL VOC project for object detection. This paper uses the VOC2007 dataset, hereinafter referred to as the VOC dataset. The VOC dataset includes 9963 labeled pictures, consisting of three parts: train/ verification /test, with a total of 24,640 marked objects. Among them, the training set includes 2501 pictures, the verification set contains 2510 pictures, and the test set contains 4952 pictures. The dataset contains a total of 20 classes, belonging to 4 categories (for example, buses and motorcycles belong to the same category). The data set can be downloaded through the following link: http://host.robots.ox.ac.uk/pascal/VOC/voc2007/index.html.

4.2.2 Microsoft common objects in context

The full name of MS COCO is Microsoft Common Objects in Context. Lin et al. [18] This paper mainly uses the MS COCO2017 dataset, hereinafter referred to as the COCO dataset. The COCO dataset is a large-scale object detection data set, which is divided into training set containing 118,287 images and test set containing 5,000 images. There are 123,287 pictures in total. Most of the pictures come from real life, with a total of 80 classes. The background is complex, and the number of instance targets on each picture is large. The COCO dataset contains an average of 3.5 categories and 7.7 instance targets per picture, less than 20% of pictures contain only one category, and only 10% of pictures contain one instance target. The dataset can be downloaded at the following link: https://cocodataset.org/

4.3 Comparison with PANet

Since YOLOX is one of the most widely used models in PANet, this paper uses YOLOX model as baseline. PANetW and PANet are used on YOLOX model for detailed comparative analysis. This paper selects a small VOC dataset and a large COCO dataset for experiments. The experimental results are shown in Tables 1 and 2. From Tables 1 and 2, it can be seen that compared with PANet, the use of PANetW on YOLOX has improved various indicators to varying degrees. Among them, AP on the VOC dataset increased by 4.9%. AP_l for large targets increased by 7.0%; AP on the COCO dataset increased by 4.0%. AP_l for large targets increased by 6.0%, indicating that PANetW has higher detection accuracy for large targets.

Table 1 Comparison table of PANetW and PANet accuracy (%) on VOC test set

Module	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
PANet	38.1	63.5	40.4	12.2	28.1	44.5
PANetW	**43.0**(+4.9)	**66.7**(+3.2)	**46.8**(+6.4)	**13.0**(+0.8)	**29.6**(+1.5)	**51.5**(+7.0)

Considering the actual situation of many large target objects in real life, this paper believes that PANetW can meet the needs of real life more effectively. At the same time, PANetW also significantly improved AP_{75}, indicating that when the threshold of IoU is increased, the effect of the model is small, and the model has strong robustness to IoU values. Experimental results show that PANetW has strong robustness to IoU values. The improvement of various indicators shows that the detection results of PANetW are more accurate, indicating that the improved PANetW.

In addition, PANetW has also achieved significant improvement in model convergence speed compared to PANet. As shown in the Fig. 7, this is the accuracy change curve of PANet and PANetW training on VOC data set (left) and COCO data set (right) respectively.By observing the curves in Fig. 7, it can be found that PANetW can significantly accelerate the convergence speed of the model when training with the VOC dataset and the COCO dataset, and the accuracy of the network will also improve rapidly. During the first 100 epochs of training, the network quickly begins to converge. During the entire training process, the accuracy of the network convergence AP exceeded that of the original module PANet, which also shows the effectiveness and practicality of the improved method in this paper.

Further, this paper analyses the finding effect of PANetW from AR metrics, and the experimental results are shown in Tables 3 and 4. It can be seen that all indicators of PANetW are higher than those of PANet. Experimental data proves that YOLOX using PANet has different degrees of missing targets for all scales of targets. After improving receptive fields of the network in this paper, the features cover wider receptive fields, and the omission phenomenon is significantly reduced. Especially for large targets, AR_l increases by 3.7% on the VOC dataset and AR_l increases by 5.5% on the COCO dataset. Whether the predicted frame number threshold per image is 1, 10 or 100, the AR of PANetW in this paper improves. It can be seen that the performance of the model is improved because PANetW can flexibly handle the sharp changes in target size in object detection scenarios.

At the same time, this paper also focuses on the Loss value using PANetW and PANet. In the field of object detection, conf Loss(confidence loss) is a commonly used evaluation metric to measure the detection accuracy of the predicted detection box and the real detection box and the confidence level of the model. If the conf Loss curve jitter is large, it will reduce the credibility of the model, and at the same time, it will also affect the prediction performance of the model, resulting in a large gap between the model and the real situation. In addition, the jitter of the cls loss(classification Loss) curve may suggest fluctuations in the performance of the model, resulting in the model showing instability during training. At the same time,

Table 2 Comparison table of PANetW and PANet accuracy (%) on COCO test set

Module	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
PANet	40.3	58.9	43.5	23.0	44.6	53.9
PANetW	**44.3**(+4.0)	**62.2**(+3.3)	**48.3**(+4.8)	**24.6**(+1.6)	**48.1**(+3.5)	**59.9**(+6.0)

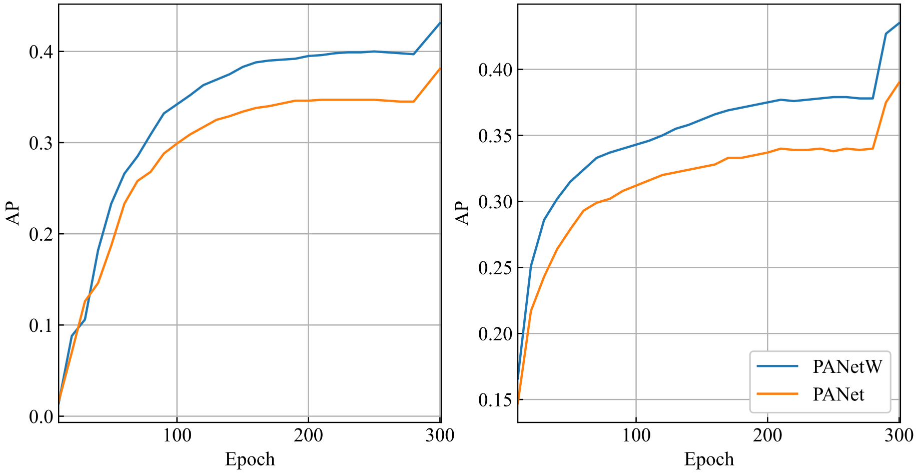


Fig. 7 The figure is the AP curve of PANet and PANetW based on YOLOX. The figure on the left shows the AP comparison curve evaluated on the VOC dataset, and the figure on the right shows the AP comparison curve evaluated on the COCO dataset. As shown in the figure, PANetW improved accuracy significantly faster and achieved better results on both VOC and COCO datasets

jitter may also make the training process of the model difficult to understand, because it may change the learning process of the model, thus making the training results of the model difficult to interpret. In the training of object detection models, IoU loss(Intersection over Union loss) is a common loss function used to measure the overlap rate between two labels. However, when the model needs to process different target sizes, this may cause the model to perform poorly when processing larger images, which in turn leads to the jitter of the IoU loss curve. The Fig. 8 consists of a total of 16 images, in which the Fig. 8 is from left to right: each loss curve of training PANet on VOC dataset, each curve of training PANetW on VOC dataset, each curve of training PANet on COCO dataset and each curve of training PANetW on COCO dataset. From top to bottom are Total Loss curve, IoU Loss curve, conf curve, and cls loss curve. It can be clearly observed from Fig. 8 that the training process using PANetW converges more rapidly, and the Loss value decreases more smoothly, mainly concentrated in the last 15 training cycles. In addition, the experimental results also clearly show that PANetW is more adaptable to YOLOX's requirement to cancel the data enhancement strategy within the last 15 rounds. This strategy requires the model to turn off the data enhancement strategy during the last 15 rounds of training, so that all training and test datasets are original real data. As can be seen from the Loss curve, the network can handle the real data distribution more accurately, which will have a significant positive impact on the practical application of the network.

Table 3 Comparison table of PANetW and PANet average recall(%) on VOC test set

Module	$AR_{max=1}$	$AR_{max=10}$	$AR_{max=100}$	AR_s	AR_m	AR_l
PANet	33.5	50.8	52.4	24.4	42.6	58.7
PANetW	**36.8**(+3.3)	**54.2**(+3.4)	**55.8**(+3.4)	**24.8**(+0.8)	**42.9**(+0.3)	**62.4**(+3.7)

Table 4 Comparison table of PANetW and PANet average recall(%) on COCO test set

Module	$AR_{max=1}$	$AR_{max=10}$	$AR_{max=100}$	AR_s	AR_m	AR_l
PANet	32.3	51.7	55.0	34.2	60.5	70.3
PANetW	**34.4**(+2.1)	**55.8**(+4.1)	**60.2**(+5.2)	**38.6**(+4.4)	**65.3**(+4.8)	**75.8**(+5.5)

4.4 Comparison with previous works

In this paper, PANetW is used on YOLOX-S to analyze and compare the mainstream models using FPN [19], NAS [39], SimSPPF [16], and One Feature fusion modules [2, 4] respectively. It can be known from the following analysis that the module PANetW has superior performance compared to the current mainstream module. The detailed analysis is as follows.

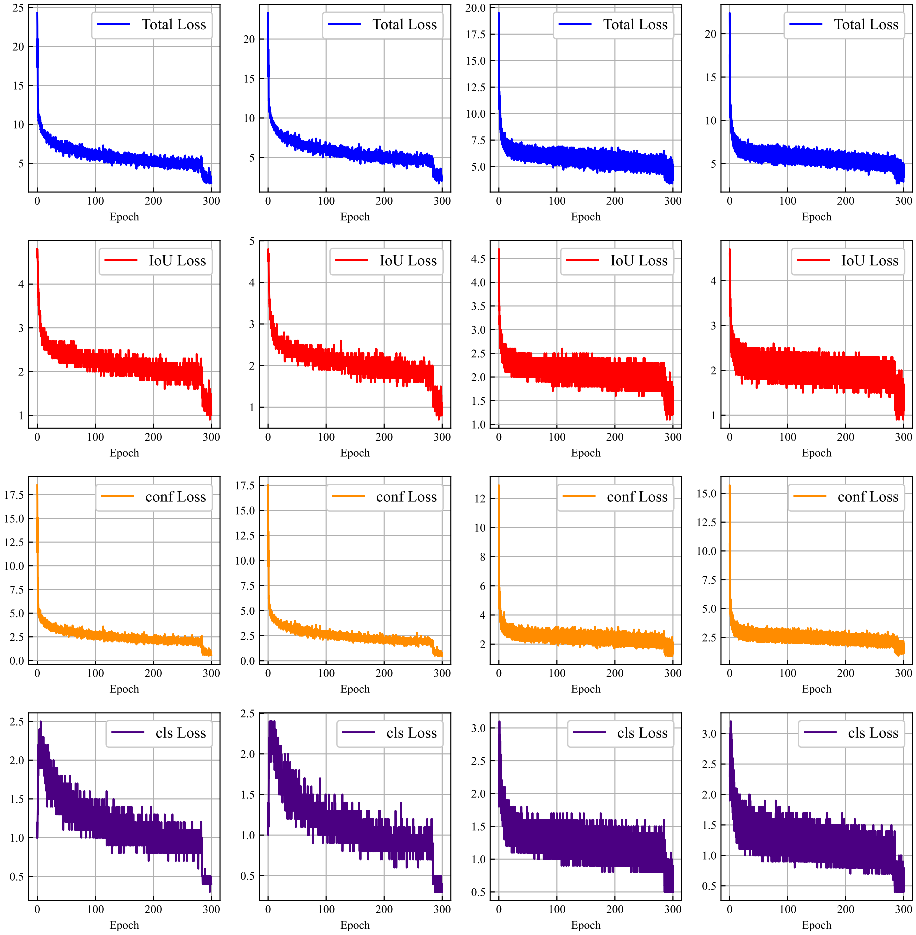


Fig. 8 The figure shows four loss curves for PANet(left 1 and left 3) and PANetW(left 2 and left 4) during iterative training on YOLOX using VOC and COCO datasets (Total loss, IoU loss, conf loss, and cls loss, respectively)

Since the following mainstream modules are not all for light-weight datasets, this paper will use more rigorous COCO dataset in the following experiments.

4.4.1 Comparison with FPN

FPN is undoubtedly still the most mainstream feature fusion module at present. This paper compares the effects of YOLOX using PANetW with Faster RCNN, RetinaNet, EfficientNet, and FCOS using FPN on the COCO dataset. The results are shown in Table 5. It can be seen that the use of FPN will generally lead to a substantial increase in the number of parameters, and the calculation is also more complicated. The module PANetW in this paper not only has fewer parameters than the above module (15.0M vs 21M), but also has a more accurate object detection effect (AP 44.3% vs AP 43.2%). Through in-depth analysis, we can conclude that PANetW has the following advantages over FPN: Firstly, it can effectively solve the problem of possible path length in FPN, so as to achieve high-precision and high-speed object detection; Secondly, it can effectively retain spatial information, make full use of semantic information and spatial relationships, so as to further improve the accuracy of object detection.

4.4.2 Comparison with FCOS-NAS

NAS has achieved great success in classification tasks, and it can automatically select the optimal neural network structure by searching in a large number of calculations. Under the research of Wang Ning et al., the NAS was successfully combined with the current hotspot module FCOS to obtain the FCOS-NAS [32] module. FCOS-NAS is compatible with major backbone networks and achieves excellent results. This paper compares it with the new module PANetW proposed in this paper on the COCO dataset, and the results are shown in the Table 6. It can be seen from this paper that compared with NAS searching in a large number of calculations, the computational cost of PANetW is lower, because it only needs part of the single-layer feature information for network fusion, which means it is more suitable for devices with limited computing resources. It can be seen from the table that PANetW outperforms FCOS-NAS in terms of accuracy (44.3% vs 43.4%) and number of parameters (15.0M vs 37.3M).

4.4.3 Comparison with SimSPPF

SimSPPF is proposed by YOLOv6. It is an improved version of SPPF in YOLOv5 and has been well adapted to YOLOv6. For a fair comparison, this paper chooses YOLOv6-S as the comparison model, because the two models have the same parameter level. The comparison data is shown in Table 7. Compared with SimSPPF, PANetW improved by 4.0% on AP and 1.8% on AP_{50}. In addition, not only does PANetW have fewer parameters (params 17.2M) than SimSPPF (params 15.0M), but its computation amount (FLOPs 44.2G) is also relatively low (SimSPPF 38.1G), which makes PANetW more hardware-friendly. This paper analyzes that PANetW can provide richer channels of detailed information and improve the regression accuracy of target boxes, so it can achieve better performance when dealing with complex object detection tasks.

Table 5 Comparison of various mainstream detectors using FPN and YOLOX using PANetW experiment results

Module	Decoder	$\#param(M)$	$AP(\%)$	$AP_{50}(\%)$	$AP_{75}(\%)$	$AP_s(\%)$	$AP_m(\%)$	$AP_l(\%)$
FPN	Faster RCNN	42	40.2	61	43.8	24.2	43.5	52
	RetinaNet-R50	33.6	36.1	-	-	-	-	-
	EffificientNet-B3	21	40.3					
	FCOS	89.6	43.2					
PANetW	YOLOX-S	**15.0**(-6.0)	**44.3**(+1.1)	**62.2**(+1.2)	**48.3**(+4.5)	**24.6**(+0.4)	**48.1**(+4.6)	**59.9**(+7.9)

Table 6 Comparison of the mainstream model using FCOS-NAS and YOLOX-S using PANetW

Module	Decoder	$\#param(M)$	$AP(\%)$
FCOS-NAS	ResNet50	37.3	40.3
	ResNet101	56.2	43.4
PANetW	YOLOX-S	**15.0**(-22.3)	**44.3**(+0.9)

4.4.4 Comparison with one feature

Using only one layer of features is a new idea recently proposed, such as DETR [2], which is very popular now. DETR is a recently proposed detector that introduces Transformer for object detection, and achieves surprising results on the COCO dataset. And it is proved that using only a single C5 feature can achieve comparable results to using a multi-level feature detector. Later, YOLOF [4] also uses this idea to bring considerable improvement to the model by using only one layer of C5 features on ResNet50, with its unique encoder and label assignment, making the network fast and accurate. This paper compares DETR, YOLOF with PANetW in Table 8. The results show that PANetW outperforms DETR in performance. And because the detection result is closer to the real result when the IoU threshold is set to 0.75, this paper finds that PANetW is significantly better than YOLOF (+3.3%) in IoU on AP_{75}, which just shows that the target frames detected correctly by PANetW is closer to the real target frame. At the same time, PANetW is not bad compared to DETR. PANetW only uses 36.59% of its parameters to achieve higher accuracy than it. And whether the IoU threshold is 0.5 or 0.75, PANetW can get better results. The analysis in this paper is that when the model can capture multi-scale features, it can model contextual information at different levels, which makes PANetW more robust and adaptable.

5 Conclusion

Through experimental observation, we found that PANet performs poor in certain scenarios, particularly when the target's scale varies significantly. Further analysis revealed that this is due to the limited receptive field of PANet, which results in limited perception of target scale information, rendering the object detection algorithm unable to deal with drastic changes in target size. To address this problem, we propose an innovative high-performance feature fusion module named PANetW (PANet with Wider Receptive Fields). The superiority of our module mainly lies in two aspects: Firstly, it can achieve larger receptive fields capturing in a low cost with the help of our meticulously designed dilated convolution networks; Secondly, by incorporating residual structure, we successfully avoid network degradation, and enhance the capability of PANetW to better represent the detailed features of targets. Through experimental analysis, our PANetW has achieved high levels in multiple metrics.

Table 7 Comparison of experimental results between YOLOv6-S using SimSPPF and YOLOX-S using PANetW

Module	Decoder	$\#param(M)$	$FLOPs(G)$	$AP(\%)$	$AP_{50}(\%)$
SimSPPF	YOLOv6-S	17.2	44.2	40.3	60.4
PANetW	YOLOX-S	**15.0**(-2.2)	**38.1**(-6.1)	**44.3**(+4.0)	**62.2**(+1.8)

Table 8 Comparison of experimental results between DETR, YOLOF using a single feature and YOLOX using PANetW on the COCO dataset

Module	Decoder	$\#param(M)$	$FLOPs(G)$	$AP(\%)$	$AP_{50}(\%)$	$AP_{75}(\%)$	$AP_s(\%)$	$AP_m(\%)$	$AP_l(\%)$
one feature	DETR	41.0	86.0	42.0	**62.4**	44.2	20.5	45.8	**61.1**
	YOLOF	44.0	86.0	41.6	60.5	45.0	22.4	46.2	57.6
PANetW	YOLOX-S	**15.0**(-26.0)	**38.1**(-47.9)	**44.3**(+2.3)	62.2(-0.2)	**48.3**(+3.3)	**24.6**(2.2)	**48.1**(+1.9)	59.9(-1.2)

Compared to the PANet method, the AP of PANetW improves by 4.9% on the VOC2007 dataset and by 4.0% on the MS COCO2017 dataset. Additionally, we have achieved a better trade off between speed and accuracy than mainstream modules. These experimental results fully demonstrate the effectiveness of our module. Therefore, by applying PANetW to object detection frameworks, the success rate of detection can be effectively improved.

While our work has been successful in optimizing object detection, there remains potential for enhancement: 1) The computational costs could be further optimized; 2) the effectiveness of PANetW was only analyzed based on YOLOX-S in this paper. In the future, we will continue to study how to effectively reduce the computational cost and model parameters, and plan to extend PANetW to more baseline models. We hope that this research report can provide some reference for developers and researchers.

Appendix A Abbreviations

We list the definitions for each abbreviation in Table 9.

Table 9 Abbreviation table

Abbreviations	Definitions
IoU	Intersection over Union
AP	Set the IoU to 0.5 to 0.95 and measure in steps of 0.05 to arrive at an average accuracy.
AP_{50}	Average precision at IoU = 0.5
AP_{75}	Average precision at IoU = 0.75
AP_s	Average precision for small objects: area $<$ 32
AP_m	Average precision for medium objects: 32 $<$ area $<$ 96
AP_l	Average precision for large objects: area $>$ 96
AR	Average Recall
$AR_{max=1}$	Average recall given 1 detection per image
$AR_{max=10}$	Average recall given 10 detection per image
$AR_{max=100}$	Average recall given 100 detection per image
AR_s	Average recall for small objects: area $<$ 32
AR_m	Average recall for medium objects: 32 $<$ area $<$ 96
AR_l	Average recall for large objects: area $>$ 96
IoU loss	Intersection over Union loss
conf loss	confidence loss
cls loss	classification Loss
#param	The total number of parameters in the network model that need to be trained.
GFLOPs	Giga Floating-point Operations Per Second
COCO	Microsoft Common Objects in Context 2017 dataset.
VOC	PASCAL VOC 2007 dataset

Acknowledgements The authors acknowledge funding from the Research and Application of Multi-core Extreme Learning Machine Based on Low Quality Samples, Research Project of Education Department, Hunan Province, China(20A511), Central South University of Forestry and Technology Degree and Post-graduate Education Teaching Reform Project(2022JG006), Hunan Provincial Natural Science Foundation of China(2023JJ40272) and Research Foundation of Education Bureau of Hunan Province, China(22B0938)

Availability of data and materials All data generated or analysed during this study are included in this published article

Code Availability Code is available at https://github.com/ChenRan2000/PANetW.

Declarations

Conflicts of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bochkovskiy A, Wang CY, Liao HYM (2020) Yolov4: Optimal speed and accuracy of object detection. arXiv:2004.10934 https://doi.org/10.48550/arxiv.2004.10934
2. Carion N, Massa F, Synnaeve G, et al (2020) End-to-end object detection with transformers. In: Computer Vision - ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I. Springer-Verlag, p 213-229. https://doi.org/10.1007/978-3-030-58452-8_13
3. Chen L, Papandreou G, Kokkinos I et al (2018) Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans Pattern Anal Mach Intell 40(4):834–848. https://doi.org/10.1109/TPAMI.2017.2699184
4. Chen Q, Wang Y, Yang T, et al (2021) You only look one-level feature. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 13034–13043. https://doi.org/10.1109/CVPR46437.2021.01284
5. Everingham M, Van Gool L, Williams C et al (2010) The pascal visual object classes (voc) challenge. Int J Comput Vis 88:303–338. https://doi.org/10.1007/s11263-009-0275-4
6. Gao Z (2023) Yoloca: Center aware yolo for dense object detection. In: Journal of Physics: Conference Series, IOP Publishing, p 012019. https://doi.org/10.1088/1742-6596/2425/1/012019
7. Ge Z, Liu S, Wang F, et al (2021) Yolox: Exceeding yolo series in 2021. arXiv:2107.08430 https://doi.org/10.48550/arXiv.2107.08430
8. Ghiasi G, Lin T, Le Q (2019) Nas-fpn: Learning scalable feature pyramid architecture for object detection. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 7029–7038. https://doi.org/10.1109/CVPR.2019.00720
9. Girshick R (2015) Fast r-cnn. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp 1440–1448. https://doi.org/10.1109/ICCV.2015.169
10. Girshick R, Donahue J, Darrell T, et al (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp 580–587. https://doi.org/10.1109/CVPR.2014.81
11. He K, Zhang X, Ren S et al (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell 37(9):1904–1916. https://doi.org/10.1109/TPAMI.2015.2389824
12. He K, Zhang X, Ren S, et al (2016) Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 770–778. https://doi.org/10.1109/CVPR.2016.90
13. Jocher G (2020) YOLOv5 by Ultralytics, 7.0. https://doi.org/10.5281/zenodo.3908559 https://github.com/ultralytics/yolov5
14. Jocher G, Chaurasia A, Qiu J (2023) YOLO by Ultralytics, 8.0.0. https://github.com/ultralytics/ultralytics
15. Karan A (2022) Has the future started? the current growth of artificial intelligence, machine learning, and deep learning. Iraqi J Comput Sci Math 3:115–123. https://doi.org/10.52866/IJCSM.2022.01.01.013
16. Li C, Li L, Jiang H, et al (2022) Yolov6: A single-stage object detection framework for industrial applications. arXiv:2209.02976 https://doi.org/10.48550/arXiv.2209.02976
17. Li Y, Chen Y, Wang N, et al (2019) Scale-aware trident networks for object detection. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp 6053–6062. https://doi.org/10.1109/ICCV.2019.00615
18. Lin T, Maire M, Belongie S, et al (2014) Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer, pp 740–755 https://doi.org/10.1007/978-3-319-10602-1_48

19. Lin T, Dollár P, Girshick R, et al (2017) Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 936–944. https://doi.org/10.1109/CVPR.2017.106
20. Liu S, Huang D, et al (2018a) Receptive field block net for accurate and fast object detection. In: Proceedings of the European conference on computer vision (ECCV), pp 385–400, https://doi.org/10.1007/978-3-030-01252-6_24
21. Liu S, Qi L, Qin H, et al (2018b) Path aggregation network for instance segmentation. In: 2018 IEEE/CVF conference on computer vision and pattern recognition, pp 8759–8768 https://doi.org/10.1109/CVPR.2018.00913
22. Liu W, Anguelov D, Erhan D, et al (2016) Ssd: Single shot multibox detector. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, Springer, pp 21–37. https://doi.org/10.1007/978-3-319-46448-0_2
23. Redmon J, Farhadi A (2017) Yolo9000: Better, faster, stronger. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 6517–6525. https://doi.org/10.1109/CVPR.2017.690
24. Redmon J, Farhadi A (2018) Yolov3: An incremental improvement. arXiv:1804.02767 https://doi.org/10.48550/arXiv.1804.02767
25. Redmon J, Divvala S, Girshick R, et al (2016) You only look once: Unified, real-time object detection. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR), pp 779–788. https://doi.org/10.1109/CVPR.2016.91
26. Ren S, He K, Girshick R 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. https://doi.org/10.1109/TPAMI.2016.2577031
27. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 https://doi.org/10.48550/arXiv.1409.1556
28. Srivastava R, Greff K, Schmidhuber J (2015) Highway networks. arXiv:1505.00387 https://doi.org/10.48550/arXiv.1505.00387
29. Tabata A, Zimmer A, dos Santos Coelho L et al (2023) Analyzing carla 's performance for 2d object detection and monocular depth estimation based on deep learning approaches. Expert Syst Appl 227:120200. https://doi.org/10.1016/j.eswa.2023.120200
30. Tan M, Pang R, Le Q (2020) Efficientdet: Scalable and efficient object detection. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 10778–10787. https://doi.org/10.1109/CVPR42600.2020.01079
31. Wang C, Bochkovskiy A, Liao H (2023a) Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: 2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 7464–7475. https://doi.org/10.1109/CVPR52729.2023.00721
32. Wang N, Gao Y, Chen H et al (2021) Nas-fcos: efficient search for object detection architectures. Int J Comput Vis 129:3299–3312. https://doi.org/10.1007/s11263-021-01523-2
33. Wang X, Chen S, Wei G et al (2023) Tenet: Accurate light-field salient object detection with a transformer embedding network. Image Vis Comput 129:104595. https://doi.org/10.1016/j.imavis.2022.104595
34. Xu S, Wang X, Lv W, et al (2022) Pp-yoloe: An evolved version of yolo. arXiv:2203.16250
35. Yang K, Li J, Dai S et al (2023) Multiscale features integration based multiple-in-single-out network for object detection. Image Vis Comput 135:104714. https://doi.org/10.1016/j.imavis.2023.104714
36. Zhang D, Zhang H, Tang J, et al (2020) Feature pyramid transformer. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16, Springer, pp 323–339.https://doi.org/10.1007/978-3-030-58604-1_20
37. Zhao G, Ge W, Yu Y (2021) Graphfpn: Graph feature pyramid network for object detection. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp 2743–2752. https://doi.org/10.1109/ICCV48922.2021.00276
38. Zhou Y (2024) A yolo-nl object detector for real-time detection. Expert Syst Appl 238:122256. https://doi.org/10.1016/j.eswa.2023.122256
39. Zoph B, Vasudevan V, Shlens J, et al (2018) Learning transferable architectures for scalable image recognition. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 8697–8710. https://doi.org/10.1109/CVPR.2018.00907

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.