A lightweight image sensitive information detection model based on yolov5s

Yueheng Mao[1,2,a,*], Bin Song[1,2], Zhiyong Zhang[1,2], Wenhou Yang[3], Yu Lan[3]

[1]Information Engineering College, Henan University of Science and Technology, Luoyang, 471023, Henan, China
[2]Henan International Joint Laboratory of Cyberspace Security Applications, Luoyang, 471023, Henan, China
[3] Sunnetech Ltd., Quzhou, 324003, Zhejiang, China
[a]myhtel@163.com
**Corresponding author*

Abstract: *Current sensitive information detection methods are prone to problems such as low detection accuracy, long training time, and slow detection speed, resulting in models that are usually not suitable for practical deployment. To solve this problem, a lightweight image sensitive information detection model based on yolov5s is proposed in this paper. First, this paper designs an efficient attention module GPSA module based on PSA module in the feature extraction part, which enables the network model to learn richer multi-scale feature representations and improve the detection accuracy of the model for sensitive information. In the feature fusion part, this paper adopts the BiFPN structure instead of the PAN structure of the original model, so that the feature fusion ability of the model can be improved. After experimental comparison, the results show that the detection accuracy and speed of the proposed method in this paper on the homemade sensitive image dataset are better than the current mainstream methods. The experimental results show that the final mAP of this model on the self-made sensitive image data set can reach 71%, and the detection time of a single image is 2.8ms, which can meet the requirements of network platform deployment in practical application.*

Keywords: *Sensitive Information, yolov5s, PSA Module, Attention Module, BiFPN*

1. Introduction

With the rapidly growing Internet, a great number of videos and pictures are uploaded to the Internet for free by individual users every day. In these public messages, images containing harmful or illegal information not only endanger the psychological health of individuals, but also threaten social security and stability. This sensitive information mainly includes three categories: pornography, politics, and violence. This information is very important for the auditors of the network platform, and once it lacks supervision, it will have very adverse effects.

During recent years, a large number of experts and scholars have conducted a lot of researches on the recognition of sensitive images, and these researches mainly focus on the detection and recognition of image classification methods. However, these methods perform multi-layer convolutional operations on images, and the final output image classification can only identify the main categories in the images, and it is difficult to classify once this sensitive information is slightly smaller or the scene is more complicated, which is extremely unfavorable to the review of images.

Therefore, this paper uses target detection and proposes a lightweight image sensitive information detection model based on yolov5s. The model adds an improved PSA module [1] (named GPSA module) before the C3 module of the original backbone network, thus increasing the extraction capability of the model in terms of multi-scale features of images. Finally, the BiFPN [2] structure is used to replace the FPN for image feature extraction and enhancement, which further improves the detection precision of the model.

2. Related work

2.1. Research status of sensitive information detection of images

Before the advent of deep learning, traditional methods for sensitive image recognition often used various low-level image features to classify them. H. Yin [3] et al. proposed a geometric filter based on fractal dimension and a coarse text filter to detect skin in images. JORGE [4] et al. proposed a YCbCr color pattern to detect detection tasks that implement pornographic images. Bermejo et al. [5] created an action recognition word packet framework and two action descriptors. STIP and MoSIFT to detect violent images.

Although an image can be represented using low-level features of the image, its detection capability is weak, and such features cannot understand the basic content of the image and do not meet the requirements of practical applications, so they often remain at the research stage only.

With the advent of deep learning and convolutional neural networks (CNNs), several studies have demonstrated their powerful recognition capabilities in image classification, and thus they are widely used for the recognition of sensitive images. This approach was used by Moustafa et al. [6], who combined GoogleNet and AlextNet and obtained a model capable of recognizing sensitive images. In their paper, they showed that the recognition precision of their model exceeded that of any other model available at that time. In the context of violent image detection, Mark Marsden [7] et al. in 2017 proposed the ResnetCrowd residual network structure, which uses a multitasking approach to solve crowd counting as well as violence detection and crowd density classification problems.

Deep learning-based methods can effectively distinguish sensitive images from normal images by combining partial features and global features. However, these methods are almost always based on image classification methods to distinguish whether an image is sensitive or not, and these methods are easily influenced by the background as well as other objects, leading to misclassification and omission. Therefore, it is very easy to lead to the wrong detection and omission of sensitive images in practical applications.

2.2. Research on target detection methods

Since target classification methods do not fully satisfy the detection task of sensitive images, this paper uses target detection.

In 2014, R-CNN [8], Fast R-CNN [9], and Faster R-CNN [10] were successively proposed and tested on VOC datasets, and the detection effect was gradually improved. These models use the RPN (Region Proposal Network) approach instead of the traditional manual feature extraction method.

In 2015, yolo [11] was first proposed by Redmon et al. This algorithm changes the detection task into an end-to-end regression problem, thus surpassing the R-CNN family of models consisting of two-stage detection in terms of detection speed. Later, Liu et al. [12] introduced the multiscale detection method to the target detection task and proposed the SSD algorithm model. Then, yolov2 [13], yolov3 [14],yolov4 [15] were proposed one after another. In 2020, Ultralytics proposed yolov5, which achieved a new benchmark for the best balance of speed and accuracy.

2.3. Bidirectional Feature Pyramid Networ

The processing of multiscale features is often a major difficulty in target detection tasks. First, the Feature Pyramid Network (FPN) proposed a top-down approach to combine multiscale features. Later, PANet [16] added an additional bottom-up path aggregation network on its basis.

In 2020, BiFPN [2] is proposed in EfficientDet, which makes multi-scale feature fusion much easier and more efficient. This strategy first removes some nodes on the PAN without feature fusion and adds a path between the input and output nodes of the original, thus fusing more feature information in a repetitive stacking manner while saving resources. Figure 1 illustrates the network architecture of BiFPN.

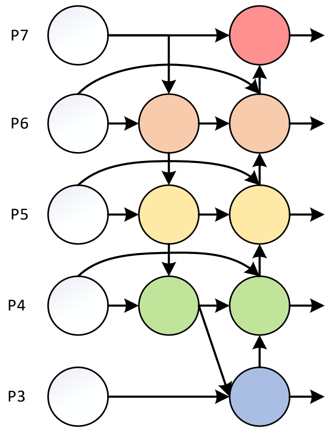
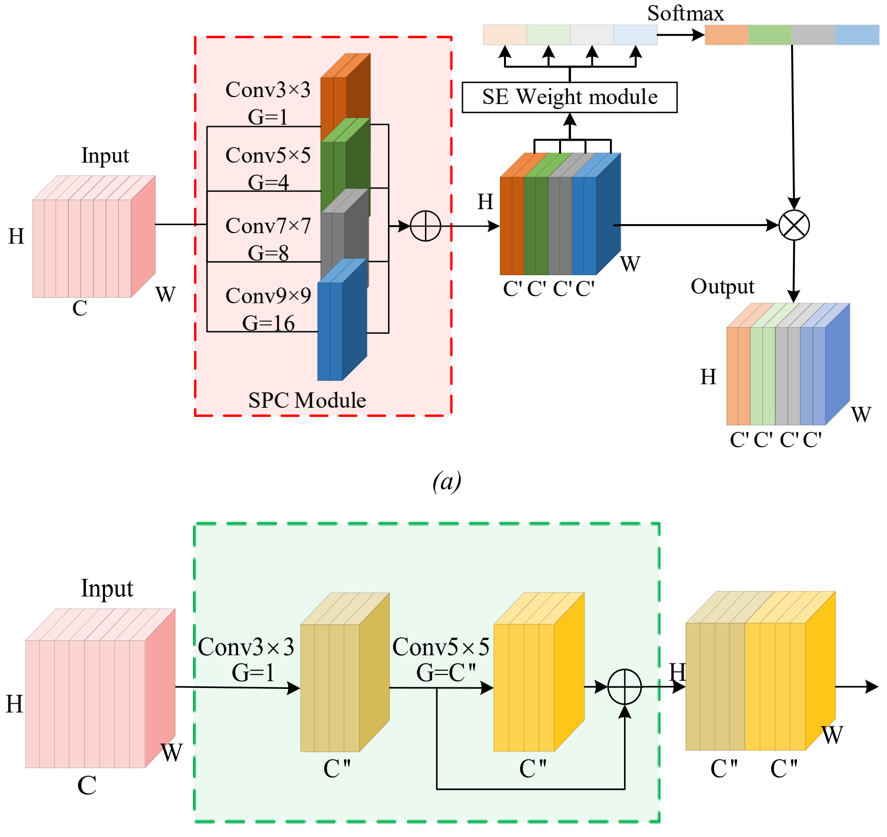


Figure 1: The architectural diagram of BiFPN. P3 to P7 represent the recursive level of the backbone convolutional network

2.4. Pyramid split attention module (PSA module)



(a)

(b)

Figure 2: (a): Efficient Pyramid Compression Attention Block PSA Module; (b): Ghost Module structure diagram. $C'=\frac{C}{4}$.

$\oplus$ *represents the channel concat operation.* $\otimes$ *represents the channel multiplication operation.SE Weight module indicates that the SE channel attention mechanism is calculated for each group of features separately.*

The most common approach for channel attention is the SE module [17], which can improve the performance of model detection significantly with a very small number of parameters and computational effort. But it has the disadvantage of ignoring the importance of spatial information. Therefore, Woo S et al. [18] proposed the CBAM, which is aimed at enhancing the attention graph by combining spatial and channel attention effectively.

To be able to effectively enrich the feature space, spatial information from different scale feature maps is captured and utilized to establish long-term channel dependencies. PyConv [19], Res2Net [20], and HS-ResNet [21] have been proposed successively, but the drawback of all these methods is that they require a large amount of computation, thus leading to a long training time.

PSA module [1] has shown excellent performance since it was proposed. This module can extract more multi-scale spatial information at a smaller pixel level. As shown in Figure 2(a), SPC first splits the input features into S groups (S=4), and the convolution kernel size of each group is increased in order, such as k=3, 5, 7, 9. Larger convolution kernel often brings more computation, so in SPC Module of PSA Module, each group is convolved in groups except the first group, and the specific number of groups is $G = 2^{\frac{k-1}{2}}$. When K = 3,5,7,9, G=1, 4, 8, 16.

Through the SE Weight Module, it has better fine-grained attention because it combines the feature information of different scales. Finally, the attention weights of each group of channels are stitched together, and softmax normalization is performed to weight the output of the SPC module. However, due to the characteristics of convolutional networks, the deeper the network is, the more channels the feature map will have. Therefore, even though the original network is designed with the number of grouped convolutions of 4, 8, and 16, respectively, it still requires a large amount of computation.

In this paper, a more efficient GPSA module is proposed to solve these problems. The inspiration for this improvement is Ghost module [22], a module that obtains more feature maps through linear changes, as shown in Figure 2(b). This paper reduces the part of the box in the figure to two groups, while the feature map of the second group is obtained by layer-by-layer convolution of the features from the first group. Thus reducing the number of parameters of the model.

3. A lightweight image sensitive information detection model based on yolov5s

yolov5 is divided into five versions. The model uses different coefficients to control the size of the model. According to the adjustment of the network depth and width coefficients, the depth and width of the model of yolov5 will also be adjusted. The larger the coefficient, the larger the volume of the model, and the more accurate the accuracy, but the speed will also slow down. Considering the number of images to be detected in the network, the yolov5s model with a depth factor of 0.33 and a width factor of 0.5 is chosen in this paper.

Figure 3 shows the structure of the improved model based on yolov5 in this paper.

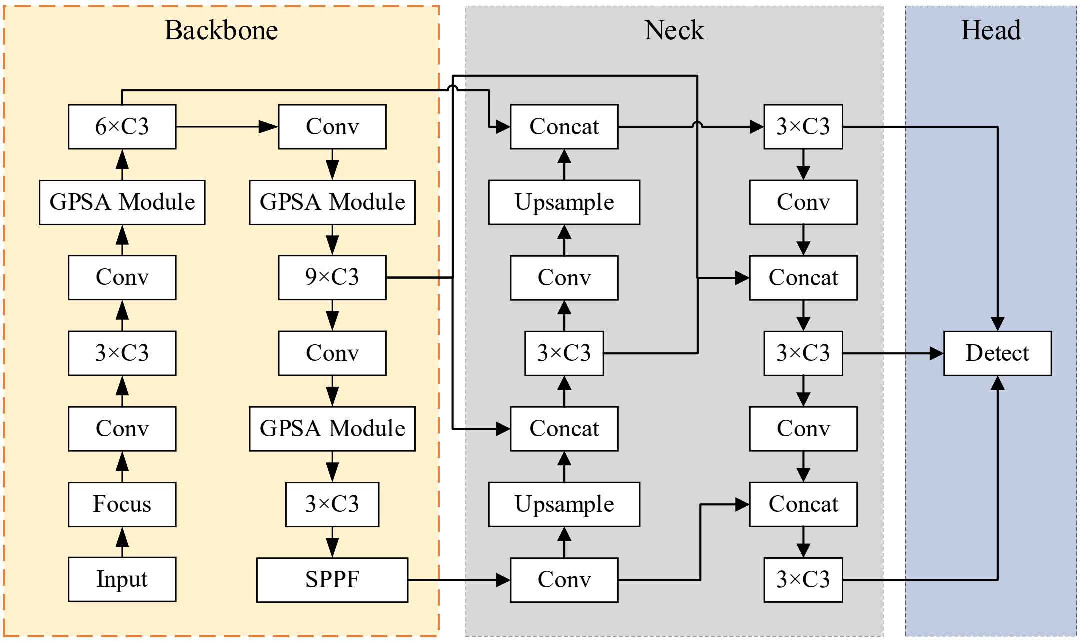


Figure 3: The structure of the improved yolov5s network schematic

The improved network model in this paper is mainly reflected in:

(1) The GPSA module is added in front of each C3 module that needs to build a pyramid network, thus enhancing the capability of extracting the model for multi-scale features. Adding the GPSA module to the Ghost model can extract the fine-grained spatial information of sensitive images more effectively and improve the detection precision of the model in sensitive information detection while having higher detection efficiency compared with the PSA module.

(2) There are differences between shapes, textures, and colors for the same type of sensitive information, and excessive attention to model details often leads to overfitting and reduces the generalization performance of the model. Therefore, the BiFPN structure is introduced in this paper, and the method improves the robustness of the model by fusing feature information across different scales.

The Backbone in the model contains Conv, GPSA Module, C3 and SPPF (spatial pyramid pooling-fast) structures for extracting image features. Conv is used for image feature extraction, the GPSA module is used to improve the network feature extraction capability, followed by C3 to reduce the computation and memory, and the SPPF network is used to integrate the features and increase the perceptual field of the model so that the model can perform better feature extraction.

The neck module adopts BiFPN structure, which integrates the image processed by backbone network with different level of feature information through pyramid network, so that the model can extract the multi-scale features of the image more accurately and efficiently.

The Head uses GIOU loss, while the non-maximum suppression method is used to filter and create the target box.

4. Experiment

4.1. Experimental preparation

Experimental environment: Python version is 3.6, using the open source Python machine learning library Pytorch 1.10.0 with Cuda 11.3 for training.
Training environment: GPU is RTX 3080 with 10GB of video memory; CPU is 12-core Intel(R) Xeon(R) Platinum 8255.

4.2. Dataset Collection and Enhancement

Due to the specificity of sensitive image detection, there is no uniform public dataset available for training in this study, so the datasets in this paper were collected from the Internet. The collected image data is then initially filtered to get the images that meet the requirements as the annotated dataset.

Then pre-process the data, convert the compliant images to jpg format, and manually label each image using the Label-Img annotation tool. Total 3681 pictures. In order to identify sensitive information in images more accurately, this paper makes a specific division of sensitive information. Pornographic images include naked breasts, large naked hips and sexual organs; political images include national leaders, national flags and national emblems; terrorist images include explosions, gunmen, arms and bloody parts. Considering the size of the dataset, all the experiments were trained using the mosaic data enhancement method.

4.3. Model evaluation indicators

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{1}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{2}$$

$$\mathrm{AP} = \int_0^1 \mathrm{P(Recall)dR} \tag{3}$$

$$\mathrm{mAP} = \frac{1}{\mathrm{c}}\sum_{\mathrm{i=1}}^{\mathrm{c}} \mathrm{AP_i} \tag{4}$$

Where TP represents the number of correct predictions identified by the model among the predictions, while FP represents the number of samples with incorrect predictions, FN represents the number of other correct classes identified by the model from the incorrect data.

AP represents the average precision of the model on one class of the dataset, and mAP is the average of AP across all classes. The single image detection time T was used as an evaluation criterion for the model speed.

4.4. Results and Analysis

4.4.1. Ablation experiment

To demonstrate the performance of the proposed method in sensitive information detection and recognition more intuitively, ablation experiments are conducted in this paper under the same training environment (No pre-trained model is used in the ablation experimental part, which is to prevent the influence of pre-trained model parameters).

As seen from the results in Table 1, the first group represents the baseline model of yolov5s. The second group represents the GPSA module added to the first group. The third group of models represents a separate replacement of the BiFPN module on top of the baseline model to enable more effective fusion of more scale features of the model. In order to represent the innovative improvement of the GPSA module to the PSA module.

Table 1: Results of ablation experiments.

No.	GPSA Module	BiFPN	Precision	Recall	mAP50	map@.5:.95	T(ms)
1	×	×	76.7%	59%	63.6%	36.3%	**2.5**
2	√	×	78.8%	57.8%	64.5%	38.2%	2.6
3	×	√	80.7%	59.4%	64.7%	38.3%	2.6
4	(PSA Module)	√	78.3%	**60.8%**	66.4%	39.2%	4.7
Ours	√	√	**81.4%**	59.7%	**66.4%**	**39.4%**	2.8

As shown in Table 1, the first and second groups demonstrate that this efficient pyramidal segmentation of attention blocks can significantly improve the feature extraction ability of the model. And by comparing the baseline model with the third group, it can be seen that the addition of BiPFN enables the model to detect sensitive images more accurately. Meanwhile, the results of the fourth and fifth groups demonstrate that the improvement of GPSA Module greatly increased the detection speed of PSA module and obtains the improvement of accuracy through the intergroup correlation of GhostModule. The final Precision of the model reached 81.4% and the recall rate reached 59.7%, while the mAP value of reached 66.4%. The results demonstrate the effectiveness of the BiFPN module and the improved GPSA module added in this paper.

4.4.2. Analysis of model comparison results

To prove the detection effects of the model proposed in this paper, this study used target detection models including Faster R-CNN [10], SSD [12] and retinanet_r50 [23] and trained them using the same test set under the same configuration conditions. The results of the experiments were evaluated by five outcome metrics, that is, AP50, mAP, Flops, Params, and single image detection time T.

As it can be seen from Table 2, the model in this paper has higher precision in sensitive information detection with the mAP value of 71% compared with other classical algorithms such as SSD and Faster-RCNN.

Table 2: Results of each model on the data set in this paper.

Module Name	Ap50			mAP.5	mAP@.5:.95	Flops (G)	Params (M)	T (ms)
	pornography	politics	Violence					
Faster-RCNN	54.3%	66.7%	78.42%	67.7%	34.6%	206.71	41.17	45.0
SSD512	55.6%	67.4%	**82.8%**	70.4%	34.4%	347.61	25.71	17.2
centernet_resnet18	38.2%	30.9%	74.9%	50.8%	29.0%	51.05	14.21	11.0
yolov3	51.0%	59.6%	79.6%	65.0%	31.7%	194.04	61.57	18.3
Retinanet_r50	58.1%	71.6%	78.9%	70.3%	39.2%	208.34	36.29	20.3
Ours	**68.8%**	**83.5%**	65.85%	**71%**	**42.8%**	**19.1**	**8.68**	**2.8**

SSD and retinanet_r50 perform better in the detection of targets involved in Violence, However, the detection performance of sensitive pornographic information is poor, and the overall detection accuracy is not as good as the model in this paper, while the detection speed is only 17.2 ms, which is much lower than the detection speed of the model in this paper.

For centernet_resnet18, although it can achieve a single detection speed of 11 ms, which is faster than any other target detection models, its detection accuracy is only 50.8%, which is 20.2% worse than the accuracy of this paper's model, and it is not as fast as this paper's model in terms of speed.

mAP value of yolov3 on the test set of this paper is 65.0%, which is lower in precision, and the model weight is larger and not easy to deploy. Due to the problem of data volume, Faster-RCNN cannot fully exploit the advantages of its two-stage algorithm on the dataset of this paper, and it also has the problem of having too many parameters and longer training and detection times. The low detection accuracy of the model in this paper is due to its poor performance in detecting bloody information, and it is easier to discriminate that part of the information as background, which is also a problem we have to solve subsequently. Figure 4 shows the detection effect of the model.

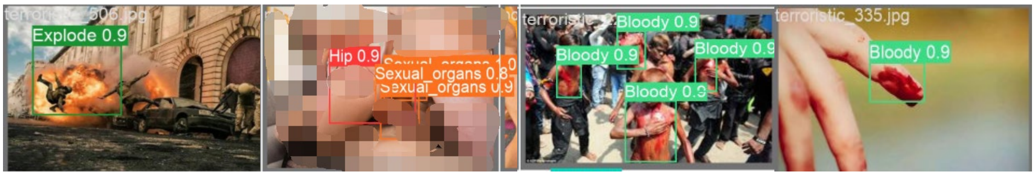


Figure 4: The detection results of the model in this paper

In summary, the lightweight image sensitive information detection method based on yolov5s is better than classical algorithms including Faster-RCNN and SSD, and has better performance and smaller model weights in the detection accuracy and speed, which is more conducive to carry out practical application deployment.

5. Conclusions

In order to detect sensitive information in images accurately and quickly, a lightweight image sensitive information detection model based on yolov5s is proposed in this paper.

Firstly, the PSA module is improved to obtain the GPSA module with higher efficiency, and For improving the detection accuracy and detection speed, this paper incorporates the GPSA module in the feature extraction network of the model;

Secondly, this paper uses the BiFPN network instead of the FPN network in the original model, thus enabling the network to learn feature information at more scales and avoiding wrong and missed detections of sensitive images. By comparing the proposed model with the mainstream target detection algorithm models. The final results show that the model incorporating the GPSA module and BiFPN is able to achieve 71% of the mAP accuracy in the test set, while the model detects an image in 2.8 ms. These data show that the model in this paper can offer a better balance between detection precision and detection speed, which is important in the case of sensitive image information detection tasks.

We will further increase the number of datasets in future work and try to apply the model specifically to practical work to assist in cleaning up the information security problem in cyberspace.

Acknowledgements

The work was sponsored by National Natural Science Foundation of China Grant No.61972133, Project of Leading Talents in Science and Technology Innovation for Thousands of People Plan in Henan Province Grant No.204200510021, Henan Province Key Scientific and Technological Project Grant No.222102210177, No.232102211060, Henan Province University Key Scientific Research Project Grant No.23A520008.

References

[1] Zhang H, Zu K, Lu J, et al. EPSANet: An efficient pyramid squeeze attention block on convolutional neural network[C]//Proceedings of the Asian Conference on Computer Vision. 2022: 1161-1177

[2] Tan M, Pang R, Le Q V. Efficientdet: Scalable and efficient object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10781-10790.

[3] Haiming Yin, Xiaodong Xu, Lihua Ye. Big Skin Regions Detection for Adult Image dentification[C]// Workshop on Digital Media & Digital Content Management. IEEE, 2011.

[4] Basilio J A M, Torres G A, Gabriel Sánchez Pérez, et al. Explicit image detection using YCbCr space color model as skin detection[C]// Proceedings of the 2011 American conference on applied mathematics and the 5th WSEAS international conference on Computer engineering and applications. World Scientific and Engineering Academy and Society (WSEAS), 2011.

[5] Bermejo Nievas E, Deniz Suarez O, Bueno García G, et al. Violence detection in video using computer vision techniques[C]//Computer Analysis of Images and Patterns: 14th International Conference, CAIP 2011, Seville, Spain, August 29-31, 2011, Proceedings, Part II 14. Springer Berlin Heidelberg, 2011: 332-339.

[6] Moustafa M. Applying deep learning to classify pornographic images and videos [J]. arXiv preprint arXiv:1511.08899, 2015.

[7] Mark Marsden, Kevin McGuinness, Suzanne Little, et al. Resnet Crowd: a residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification [C].Proceedings of 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance, 2017:1-7.

[8] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.

[9] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[10] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks [J]. Advances in neural information processing systems, 2015, 28.

[11] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[12] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 21-37.

[13] Redmon J, Farhadi A. yolo9000: better, faster, stronger[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.

[14] Redmon J, Farhadi A. yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.

[15] Bochkovskiy A, Wang C Y, Liao H Y M. yolov4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.

[16] Liu S, Qi L, Qin H, et al. Path aggregation network for instance segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8759-8768.

[17] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[18] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[19] Duta I C, Liu L, Zhu F, et al. Pyramidal convolution: Rethinking convolutional neural networks for visual recognition[J]. arXiv preprint arXiv:2006.11538, 2020.

[20] Gao S H, Cheng M M, Zhao K, et al. Res2net: A new multi-scale backbone architecture[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 43(2): 652-662.

[21] Yuan P, Lin S, Cui C, et al. HS-ResNet: Hierarchical-split block on convolutional neural network[J]. arXiv preprint arXiv:2010.07621, 2020.

[22] Han K, Wang Y, Tian Q, et al. Ghostnet: More features from cheap operations[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 1580-1589.

[23] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.