Research on Object Detection of High Resolution Remote Sensing Image based on Improved YOLOV4 Algorithm

Zairui Li
Shandong University of Science and Technology
Yongguo Zheng (skd991317@sdust.edu.cn)
Shandong University of Science and Technology https://orcid.org/0000-0002-8859-5920
Changlei Dong-ye
Shandong University of Science and Technology
Ping Wang
Shandong University of Science and Technology
Muhammad Yasir
China University of Petroleum Huadong - Qingdao Campus

Research Article

Keywords: remote sensing image, small object, object detection, shallow features

Posted Date: April 19th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1516274/v1

Abstract

Object detection is one of the fundamental tasks in computer vision. Although excellent progress has been made, there still exist challenges for objects with dense distribution, fuzzy feature, and small size. This paper presents a detector for object detection of small size in High-Resolution remote sensing images, namely SF-YOLOV4. The proposed SF-YOLOV4 exploits information for shallow layers and contextual information along with spatial attention to address the above challenges. Specifically, a shallow semantic information extraction network(SFN) is designed which introduces low-level semantic information into the backbone, to alleviate the loss of the small object features. Meanwhile, we replace the original Nick with multi-scale context feature pyramid (MSC-FPN) to improve the utilization of lower layers information and integrate the context information, we also add spatial attention module to find attention region at different scales. Experiments on two remote sensings public datasets DIOR and RSOD show the good detection performance of SF-YOLOV4.

1. Introduction

The fast development of deep learning, especially neural networks, has increasingly applied to remote sensing images[1][2]. First, compared with natural images, small objects in remote sensing images are densely distributed and fuzzy[3][4]. With the continuous deepening of the backbone, small objects such as densely distributed ships and airplanes, vehicles with fuzzy features in cities lost a lot of semantic information[5], which makes the model easy to ignore small objects, thus affecting the detection accuracy. Secondly, the background of remote sensing images is relatively complex, the object features are not obvious, and it is easy to be confused with surrounding objects, which also brings great challenges to detection[6].

At present, many scholars have made a research on the above problems [7]. YANG et al. [8] improved the detection accuracy of small objects in remote sensing images by adding shallow feature pyramids [9] and dense connections [10]. R3DET [11] proposed a feature fine-tuning module to solve the problem of misalignment between features and objects, thereby improving the detection performance of dense objects. Wang et al. [12] improved the loss function of the module [13–15] to highlight the weight of small objects while better combining shallow information.

In this article, we propose SF-YOLOV4 based on YOLOV4, which is used to solve the problem of detecting small objects with dense distribution and fuzzy features in remote sensing images. we first design a shallow feature extraction network(SFN), and use a feature extraction module to obtain small object information, thereby supplementing the semantic information lost by the deep network. In addition, our model consists of a multi-scale context feature pyramid (MSC-FPN) that fuse multi-scale feature map and tune their receptive fields. Furthermore, a spatial attention module is used to suppress background information and make the network pay more attention to the object area.

2. Related Work

With the introduction of the YOLO [16–19] series of networks, it has show excellent performance in computer vision. Among them, YOLOV4 improves the performance of extracting features by improving the backbone[20], and combines the structure of SPP [21] and PLANET [22] in NECK. Many scholars have also proposed different detectors based on YOLO[23]. SCALED-YOLOV4 [24] developed a model scaling technology, which can not only

modify the depth, width, resoluion, but also the network structure can be modified to achieve the best trade-off of speed accuracy. YOLOX [25] integrathe tes anchorless detection method [26] into the YOLO series of detectors, and decouples the prediction branches, which greatly improves the convergence speed and performance of the model. SCR++[27] improves the detection performance of the model by improving instance-level denoising, class scores, and predicting bounding boxes. TPH-YOLOv5[28] combines the Transformer with the network to enhance the ability of the model to extract features. In order to further enhance the ability of feature expression, some scholars carry out the regression of horizontal box and rotating box to improve the accuracy[29, 30].

As shown in Fig. 1, we choose CSPDARKNET53 combined with shallow feature network (SFN) backbone, multi-scale context feature pyramid network (MSC-FPN) path-aggregation neck, and final YOLOv4 head as the architecture of SF-YOLOV4.

3. Improved Yolov4 Detection Framework

3.1 Shallow Feature Network:

In order to solve the problem that small objects such as vehicles, ships, and airplanes lose most of semantic information in deep convolutional network [31], we design a shallow feature network (SFN) to supplement semantic features in backbone .

As shown in Fig. 1, the shallow feature network (SFN) is composed of four shallow feature extraction block(SFBLOCK), which are used to adjust the size of featuremap and produce shallow features. Shallow feature network(SFN) fuses low-level semantic information with backbone to improve network performance. It can be expressed as:

$$P_i = \left(S_i \oplus C_i\right) \oplus \left(C_{i+1} f_i\right) \#(1)$$

i represents the different levels of backbone, S and C represent the features of the output of SFBlock and CSPDarknet, respectively. $\oplus$ is the element-by-element addition operation, and $f$is the feature extraction operation of each layer in CSPDarknet.

The shallow feature extraction block (SFBLOCK) is shown in Fig. 2, it is composed of down-sampling and a residual block. First, image is downsampled by convolution operation, and then it is passed through residual block including convolution, batch-norm and ReLU layers.

3.2 Multi-scale context feature pyramid

Large-scale changes in the detected object may cause the problem of inconsistency between the deep-level feature map and the real object [32]. In the process of a continuous deepening of the network, feature map will only focus on a small part of the image, while the complex background of remote sensing images brings great challenges to object detection.

Based on this, we adopt the multi-scale context feature pyramid (MSC-FPN) (shown in Fig. 3), including a top-down, bottom-up bidirectional branch similar to PANET and a channel splicing that can integrate contextual information[33], and the spatial attention module for feature maps of different scales.

3.2.1 Context Fusion Module:

In the fusion between low-resolution features and high-resolution features of the feature pyramid, design a contextual information fusion module (CFM) with three parallel branches (shown in Fig. 4), after feature fuses with local and global contexts, it can enrich its expressive ability[34].

Specifically, high-level semantic information increases the receptive field through three mapconvolutions with different dilation rate [35], and adds residual connections on each parallel expansion branch to supplement semantic information, feature is then stitched with the low-level feature on the number of channels. It can be expressed as:

$$F_i = \phi\left\{Concat\left(F_{i+1}\left(D_1, D_2, D_3\right) \oplus F_{i+1}, F_i\right)\right\}\#(2)$$

F_{i+1} □ $i \in (1,2)$ rep, esents a feature in the deep layer of the pyramid network, F_i is the feature of the current layer, $D_j, j \in (1,2,3)$ represents the dilated convolution of three different dilation rate the in parallel branch. The deep feature is convoluted with different dilation rates respectively and then added with the low-level feature. Finally, a 1 * 1 convolution merging channel is represented by $\phi\{\}$, and $\oplus$ denotes element-wise addition.

3.2.2 Attention mechanism:

The attention mechanism can guide the network to focus on more prominent information in remote sensing images[36], so as to achieve the effect of enhancing features and suppressing the background. Among them, the spatial attention module [37] (as shown in Fig. 5) helps to enhance the object features with sparse texture and mixed background. The corresponding semantic information can help the network deal with different proportions of objects.

3.3 SF-YOLOV4 algorithm steps

SF-YOLOV4 algorithm steps are shown in Fig. 6. The first is the data preprocessing stage. Images need to be preprocessed such as cropping and flipping. Then enter the model training stage, load the model, training data, and save the weight in turn. The testing phase is similar to training. Network architecture is analyzed according to training weights, and then images are read in to make predictions. Finally, prediction boxes are drawn to the pictures for display, and related indicators such as MAP are calculated.

4. Experiments

4.1 Dataset

SF-YOLOV4 is evaluated through the remote sensing public data sets DIOR and RSOD.

DIOR: The DIOR data set used in this chapter is the largest and most category high-resolution remote sensing image dataset proposed by the Northwestern Polytechnical University research team in 2019[38].

RSOD: RSOD is a public data set for object detection in remote sensing images, including four types of objects: airplanes, playgrounds, overpasses, and oil drums[39].

4.2 Experimental environment and evaluation indicators:

In the experiment, the average precision mean (MAP) Eq. (3)is used as the evaluation index[40].

$$mAP = \frac{1}{N}\sum_i AP_i \#(3)$$

Among them, N represents the total number of detection object categories, where the average precision AP (AVERAGE PRECISION) represents the result of calculating the area under the P-R curve of a single object.

4.3 Analysis:

Table 1 shows the experimental results on the test set of DIOR data set. SF-YOLOV4 is realized for small objects and fuzzy features, and compared with FASTER-RCNN[41], YOLOV4, SSD[42], RETINANET[43], YOLOX in the same environment. Among them, FASTER-RCNN is a representative work of two-stage detector, SSD and RETINANET are both popular general one-stage detector, and YOLOX is another anchorless improved version of YOLO.

Table 1
comparative test under dior data set.

METHOD	Faster-RCNN	SSD	RetinaNet	Yolov4	Yolox	SF-yolov4
BACKBONE	ResNet50	VGG16	ResNet50	CSPdarknet53	darknet53	**CSPdarknet53**
Expresswayservicearea	65%	64%	90%	89%	80%	**91%**
Basketball court	71%	76%	90%	87%	89%	**90%**
Tennis court	77%	76%	87%	88%	90%	**91%**
golffield	70%	65%	85%	74%	72%	**80%**
Groundtrackfield	62%	69%	83%	82%	81%	**83%**
Stadium	94%	61%	81%	70%	74%	**75%**
Chimney	89%	66%	81%	80%	76%	**80%**
Airport	68%	72%	79%	80%	71%	**85%**
Dam	59%	57%	75%	70%	61%	**72%**
Baseball field	92%	72%	74%	85%	84%	**86%**
Wind mill	44%	66%	70%	83%	89%	**86%**
Airplane	91%	60%	68%	73%	85%	**82%**
Trainstation	40%	55%	61%	63%	48%	**70%**
Expresswaytollstation	55%	53%	59%	71%	71%	**78%**
Harbor	54%	49%	59%	63%	52%	**66%**
Overpass	51%	48%	57%	62%	61%	**64%**
Ship	21%	59%	47%	85%	88%	**89%**
bridge	22%	30%	37%	44%	44%	**51%**
Storagetank	73%	47%	34%	63%	70%	**68%**
Vehicle	30%	27%	21%	44%	49%	**52%**
MAP	61.58%	58%	66.92%	72.69%	71.7%	**77.07%**

As shown in Table 1, SF-YOLOV4 outperforms YOLOV4, and the MAP is increased by 4.38–77.07%. Specifically, SF-YOLOV4 has a greater improvement in the detection accuracy of dense small object categories such as vehicles, ships, windmills, airplanes, and oil storage tanks. In addition, it also improved the detection of large objects with fuzzy and easy to be confused with the background, such as railway station, golffield.

Figure 7 shows a few sample images from the DIOR data set and the corresponding detection by using the model— YOLOX (in the second columeature), RETINANET (in the third column), YOLOV4(in the fifth column) and the proposed SF-YOLOV4 (in the last column). The leftmost column is GROUND TRUE. (A) Image is more

complicated. Small objects such as vehicles are blurry and difficult to detect. In this case, SF-YOLOV4 can greatly improve the detection of vehicles compared with YOLOV4. (B) Featrues and edge information of a large subject such as a trainstation are not obvious in the city. It is easy to be confused with other objects such as railroad tracks and houses. SF-YOLOV4 can detect well. (C) Airplanes are densely distributed and small in size. Our model significantly improves the detection ability of densely distributed small objects.

In addition, we evaluated SF-YOLOV4 on the RSOD data set released by Wuhan University, and also compared several better detect to verify the effect of our model.

Table 2
comparative test under rsod data set

method	SF-YOLOV4	SSD	RetinaNet	Faster-rcnn	Yolox	Efficientdet[44]
playground	**99%**	98%	99%	99%	99%	99%
oiltank	**98%**	91%	97%	93%	99%	91%
aircraft	**96%**	72%	77%	66%	95%	89%
overpass	**83%**	90%	78%	88%	78%	83%
map	**93.72%**	87.8%	87.7%	86.51%	93.02%	90.69%

5. Conclusion

This paper presents a novel SF-YOLOV4, a high-resolution remote sensing image detection network for small objects, which designs a shallow information extraction network to supplement small object feature, provides a multi-scale context feature pyramid network to adjust the receptive field of the feature map and guide the model to pay attention to more effective semantic information. Experiments on two popular remote sensing public data sets DIOR and RSOD, it is shown that SF-YOLOV4 has a good effect on the detection of objects in high-resolution remote sensing images.

Declarations

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Author Contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Zairui li. The first draft of the manuscript was written by Zairui li and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Data Availability

The RSOD datasets generated during and/or analysed during the current study are available in GitHub - RSIA-LIESMARS-WHU/RSOD-Dataset-: An open dataset for object detection in remote sensing images.

The DIOR datasets during and/or analysed during the current study are not publicly available due to the link failure but are available from the corresponding author on reasonable request.

References

1. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 25:1097–1105
2. Chen J, Yue A, Wang C et al (2018) Wind turbine extraction from high spatial resolution remote sensing images based on saliency detection. J Appl Remote Sens 12(1):016041. https://doi.org/10.1117/1.JRS.12.016041
3. Wang T, Anwer RM, Cholakkal H et al (2019) Learning rich features at high-speed for single-shot object detection. Proceedings of the IEEE/CVF International Conference on Computer Vision: 1971–1980. https://doi.org/10.1109/iccv.2019.00206
4. Felzenszwalb PF, Girshick RB, Mcallester DA et al (2010) Object Detection with DiscriminativelyTrained Part-Based Models. IEEE Trans Pattern Anal MachineIntelligence 32(9):1627–1645. https://doi.org/10.1109/mc.2014.42
5. He K, Zhang X, Ren S et al (2016) Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition: 770–778. https://doi.org/10.1109/cvpr.2016.90
6. Li Z, Peng C, Yu G et al (2018) Detnet: A backbone network for object detection[J]. arXiv preprint. https://doi.org/10.1007/978-3-030-01240-3_21. arXiv:1804.06215
7. Cheng G, Han JW (2016) A survey on object detection in optical remote sensing images. ISPRS J Photogrammetry Remote Sens 117:11–28. https://doi.org/10.1016/j.isprsjprs.2016.03.014
8. Yang X, Sun H, Sun X et al (2018) Position detection and direction prediction for arbitrary-oriented ships via multitask rotation region convolutional neural network. IEEE Access 6:50839–50849. https://doi.org/10.1109/access.2018.2869884
9. Lin TY, Dollár P, Girshick R et al (2017) Feature pyramid networks for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117–2125. https://doi.org/10.1109/cvpr.2017.106
10. Huang G, Liu Z, Van Der Maaten L et al (2017) Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700–4708. https://doi.org/10.1109/cvpr.2017.243
11. Yang X, Liu Q, Yan JC, Li A (2021) R3Det: Refined single-stage detector with feature refinement for rotating object. In: Proceedings of the 35th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2021
12. Wang PJ, Sun X, Diao WH, Fu K (2019) FMSSD: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery. IEEE Trans Image Process 28(1):256–178. https://doi.org/10.1109/tgrs.2019.2954328

13. Yu J, Jiang Y, Wang Z et al (2016) Unitbox: An advanced object detection network. Proceedings of the 24th ACM international conference on Multimedia: 516–520. https://doi.org/10.1145/2964284.2967274
14. Rezatofighi H, Tsoi N, Gwak JY et al (2019) Generalized intersection over union: A metric and a loss for bounding box regression. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 658–666. https://doi.org/10.1109/cvpr.2019.00075
15. Zheng Z, Wang P, Liu W et al (2020) Distance-IoU loss: Faster and better learning for bounding box regression. Proceedings of the AAAI Conference on Artificial Intelligence, 34(07): 12993–13000. https://doi.org/10.1609/aaai.v34i07.6999
16. Redmon J, Divvala S, Girshick R, Farhadi A, Recognition P, Vegas L (2016) NV, 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91
17. Redmon J, Farhadi A, Recognition P (2017) (CVPR), Honolulu, HI, pp. 6517–6525, doi: 10.1109/CVPR.2017.690
18. Redmon J (2018) and Ali Farhadi "YOLOv3: An Incremental Improvement.". ArXiv abs/1804.02767(2018): n.pag
19. Bochkovskiy A et al (2020) "YOLOv4: Optimal Speed and Accuracy of Object Detection.". ArXiv abs/2004.10934: n. https://doi.org/10.48550/arXiv.2004.10934. pag
20. Wang CY, Liao HYM, Wu YH et al (2020) CSPNet: A new backbone that can enhance learning capability of CNN. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops: 390–391. https://doi.org/10.1109/cvprw50498.2020.00203
21. 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
22. Liu S, Qi L, Qin H et al (2018) Path aggregation network for instance segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition: 8759–8768. https://doi.org/10.1109/cvpr.2018.00913
23. Yang Y, Liao Y, Cheng L, Zhang K, Wang H, Chen S (2021) "Remote Sensing Image Aircraft Target Detection Based on GIoU-YOLO v3," 2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP), pp. 474–478, doi: 10.1109/ICSP51882.2021.9408837
24. Wang CY, Bochkovskiy A, Liao HYM (2021) Scaled-yolov4: Scaling cross stage partial network. Proceedings of the IEEE/cvf conference on computer vision and pattern recognition. 2021: 13029–13038. https://doi.org/10.1109/cvpr46437.2021.01283
25. Ge Z, Liu S, Wang F et al (2021) Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430, 2021
26. Law H, Deng J (2018) Cornernet: Detecting objects as paired keypoints[C]//Proceedings of the European conference on computer vision (ECCV): 734–750. https://doi.org/10.1007/s11263-019-01204-1
27. Yang X, Yan J, Yang X et al (2020) Scrdet++: Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing. arXiv preprint arXiv:2004.13316, 2020. https://doi.org/10.48550/arXiv.2004.13316
28. Zhu X, Lyu S, Wang X et al (2021) TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 2778–2788. https://doi.org/10.1109/iccvw54120.2021.00312

29. Cao L, Zhang X, Wang Z et al (2021) Multi angle rotation object detection for remote sensing image based on modified feature pyramid networks. Int J Remote Sens 42(14):5253–5276. https://doi.org/10.1080/01431161.2021.1910371
30. Xu YC, Fu MT, Wang QM, Wang YK, Chen K, Xia GS et al (2021) Gliding vertex on the horizontal bounding box for multioriented object detection. IEEE Trans Pattern Anal Mach Intell 43(4):1452–1459. https://doi.org/10.1109/tpami.2020.2974745
31. Wang T, Anwer RM, Cholakkal H et al (2019) Learning rich features at high-speed for single-shot object detection. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 1971–1980. https://doi.org/10.1109/iccv.2019.00206
32. Li Y, Chen Y, Wang N et al (2019) Scale-aware trident networks for object detection. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 6054–6063. https://doi.org/10.1109/iccv.2019.00615
33. Zhang G, Lu S, Zhang W (2019) CAD-Net: A context-aware detection network for objects in remote sensing imagery. IEEE Trans Geosci Remote Sens 57(12):10015–10024. https://doi.org/10.1109/tgrs.2019.2930982
34. Yu F, Koltun V (2015) Multi-scale context aggregation by dilated convolutions[J]. arXiv preprint arXiv:1511.07122, 2015
35. Yang X, Yang J, Yan J et al (2019) Scrdet: Towards more robust detection for small, cluttered and rotated objects. Proceedings of the IEEE/CVF International Conference on Computer Vision: 8232–8241. https://doi.org/10.1109/iccv.2019.00832
36. Zhu X, Cheng D, Zhang Z et al (2019) An empirical study of spatial attention mechanisms in deep networks. Proceedings of the IEEE/CVF International Conference on Computer Vision: 6688–6697. https://doi.org/10.1109/iccv.2019.00679
37. Li K, Wan G, Cheng G et al (2020) Object detection in optical remote sensing images: A survey and a new benchmark[J]. ISPRS J Photogrammetry Remote Sens 159:296–307. https://doi.org/10.1016/j.isprsjprs.2019.11.023
38. Xiao Z, Liu Q, Tang G et al (2015) Elliptic Fourier transformation-based histograms of oriented gradients for rotationally invariant object detection in remote-sensing images[J]. Int J Remote Sens 36(2):618–644. https://doi.org/10.1080/01431161.2014.999881
39. Zheng L, Shen L, Tian L et al (2015) Scalable person re-identification: A benchmark[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1116–1124. https://doi.org/10.1109/iccv.2015.133
40. Ren S, He K, Girshick R et al (2015) Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Adv Neural Inf Process Syst 28:91–99. https://doi.org/10.1109/tpami.2016.2577031
41. Liu W, Anguelov D, Erhan D et al (2016) SSD: Single Shot MultiBox Detector[J]. 2016
42. Lin T-Y, Goyal P, Girshick R, He K, Dollar P (2017) “Focal loss for dense object detection,”. in Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988. https://doi.org/10.1109/iccv.2017.324
43. Tan M, Pang R, Le QV (2020) Efficientdet: Scalable and efficient object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition: 10781–10790. https://doi.org/10.1109/cvpr42600.2020.01079

Figures

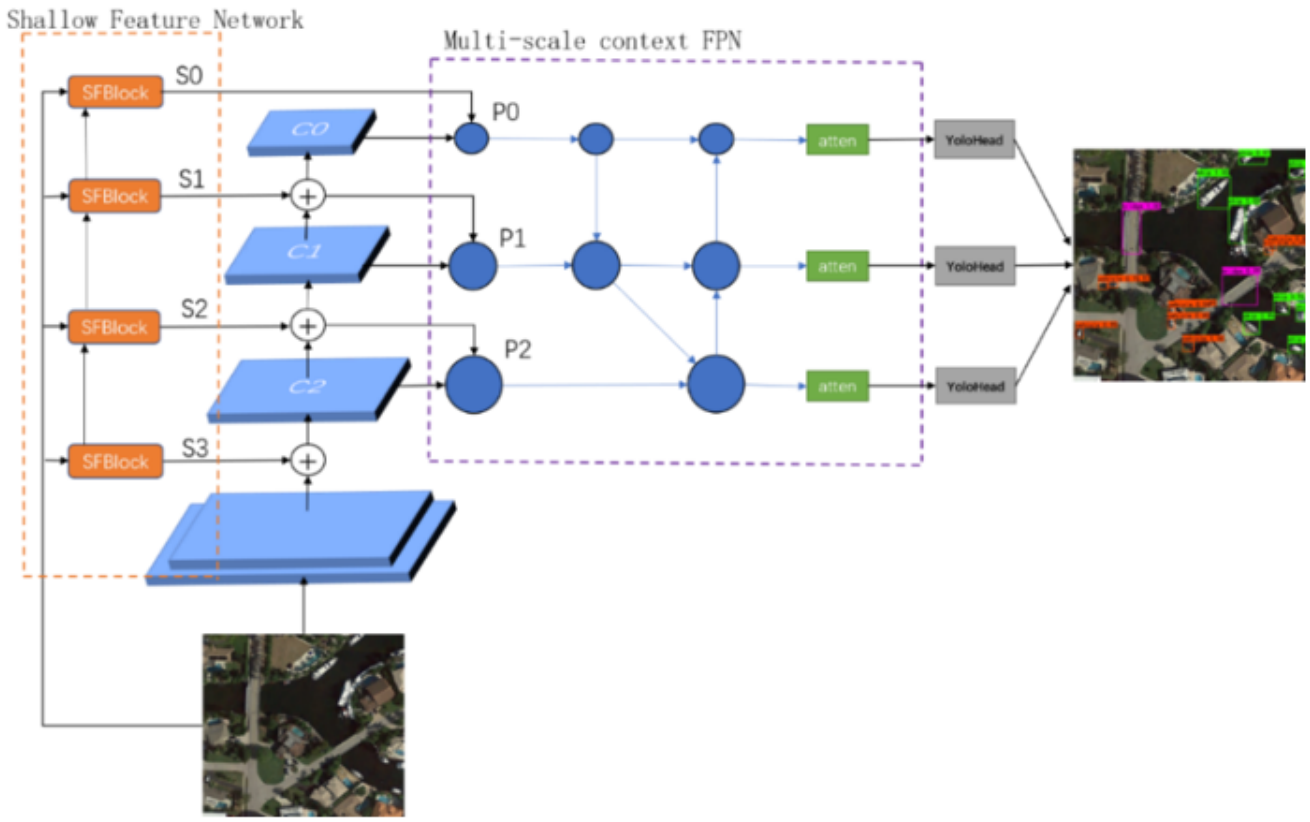

Figure 1

The overall structure of the sf-yolov4 algorithm.

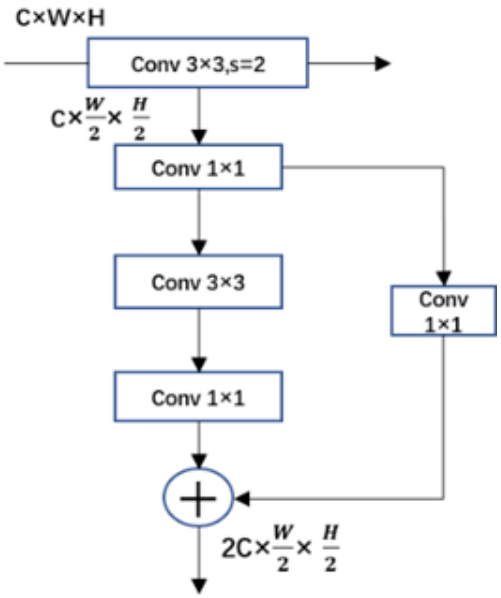

Figure 2

Sfblock.

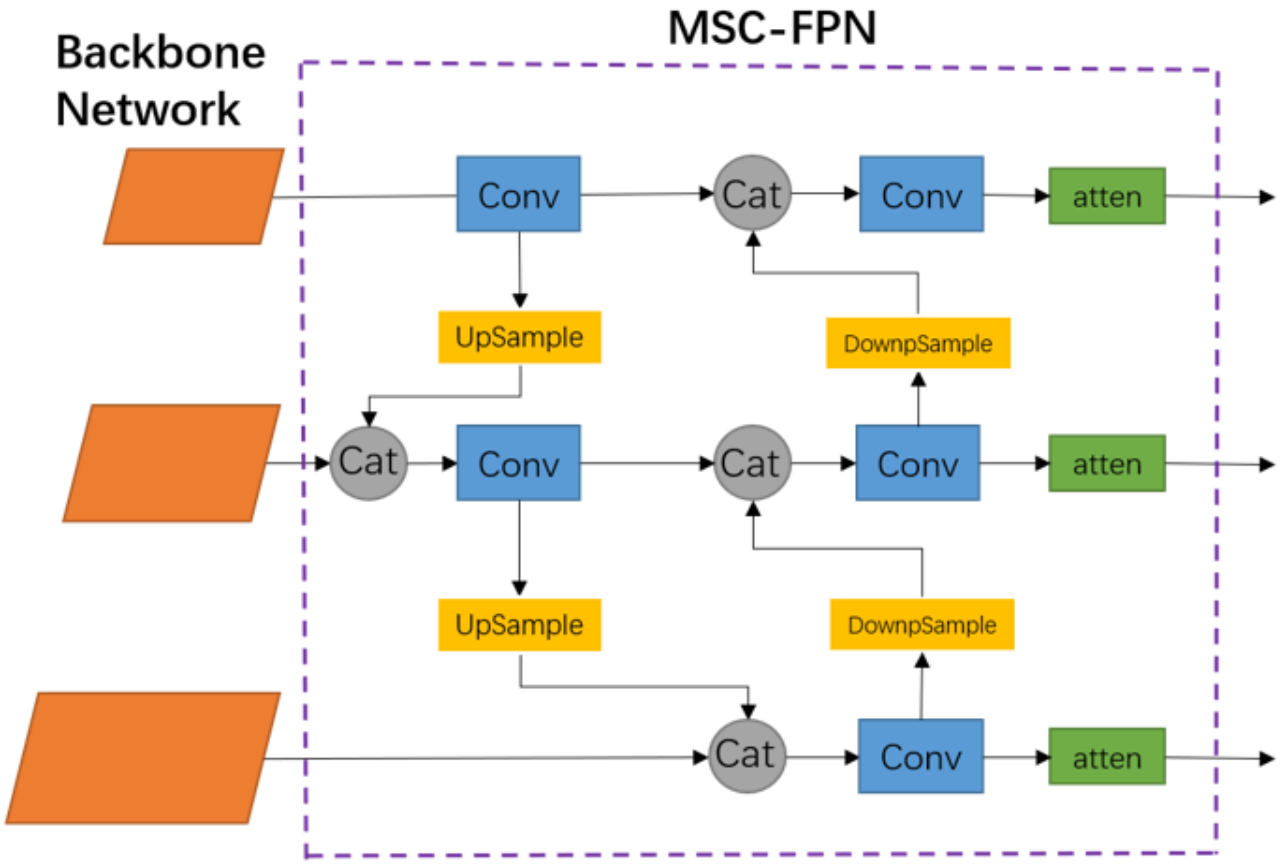

Figure 3

Multi-scale context fpn.

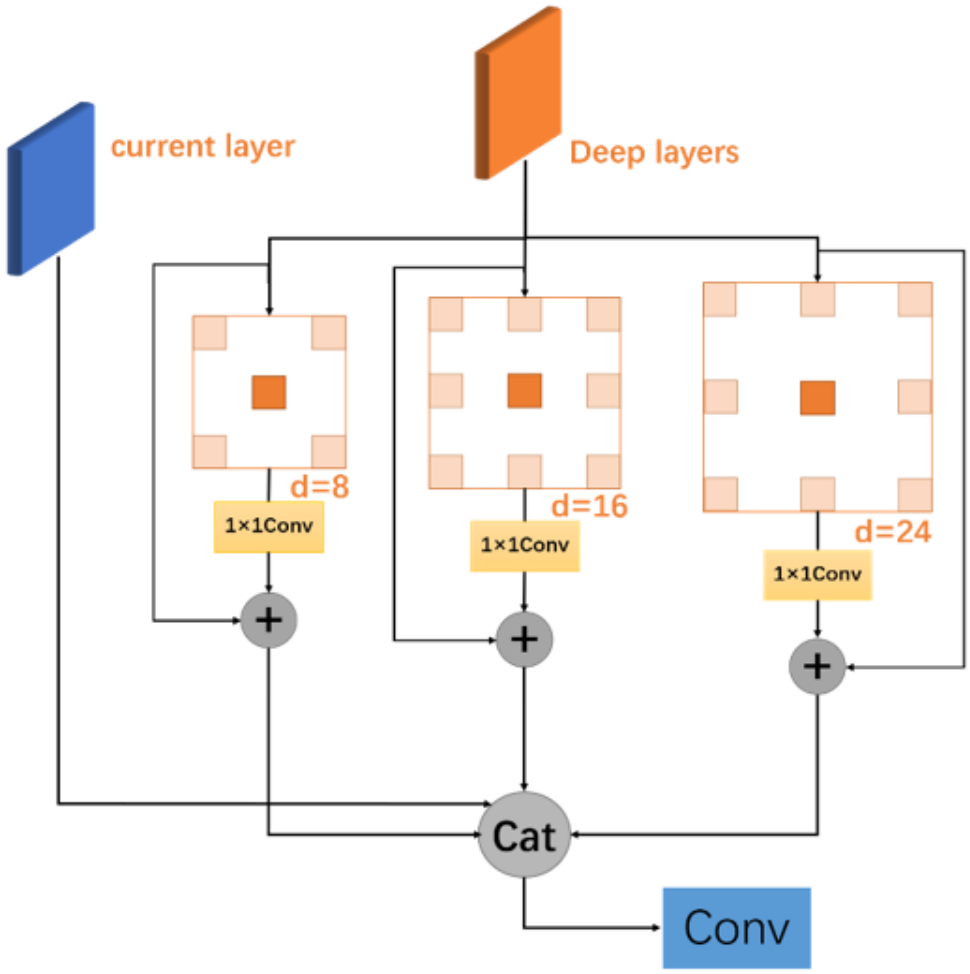

Figure 4

contextual information fusion module.

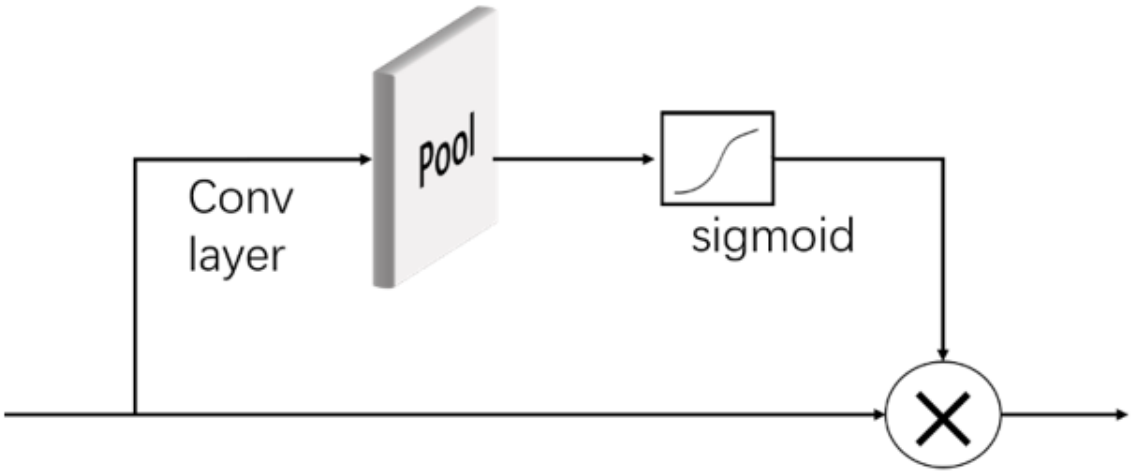

Figure 5

spatial attention module

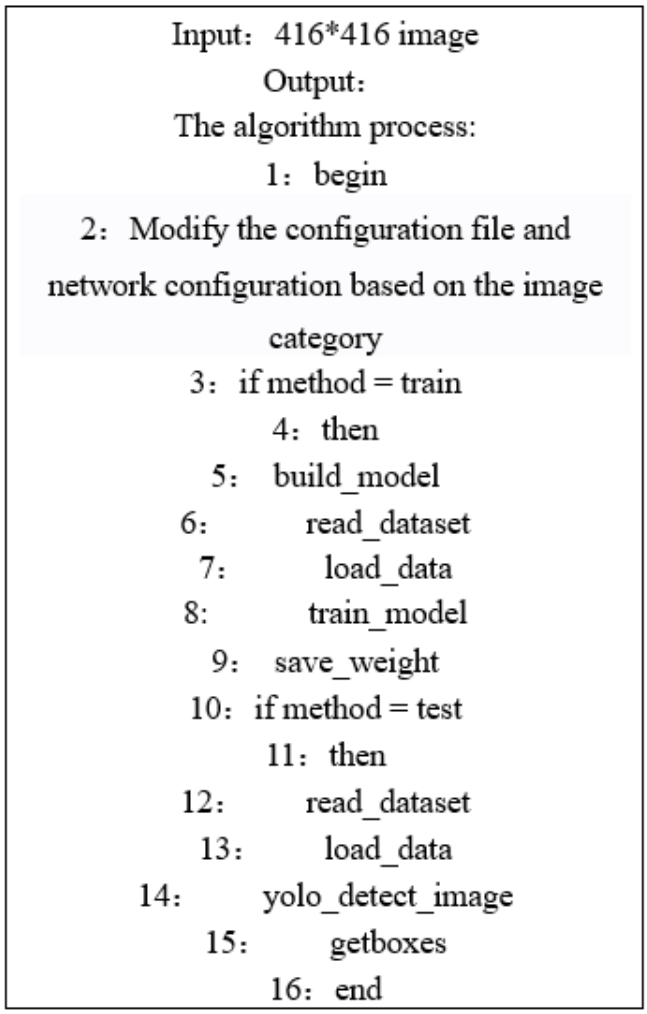

Figure 6

sf-yolov4 algorithm steps.

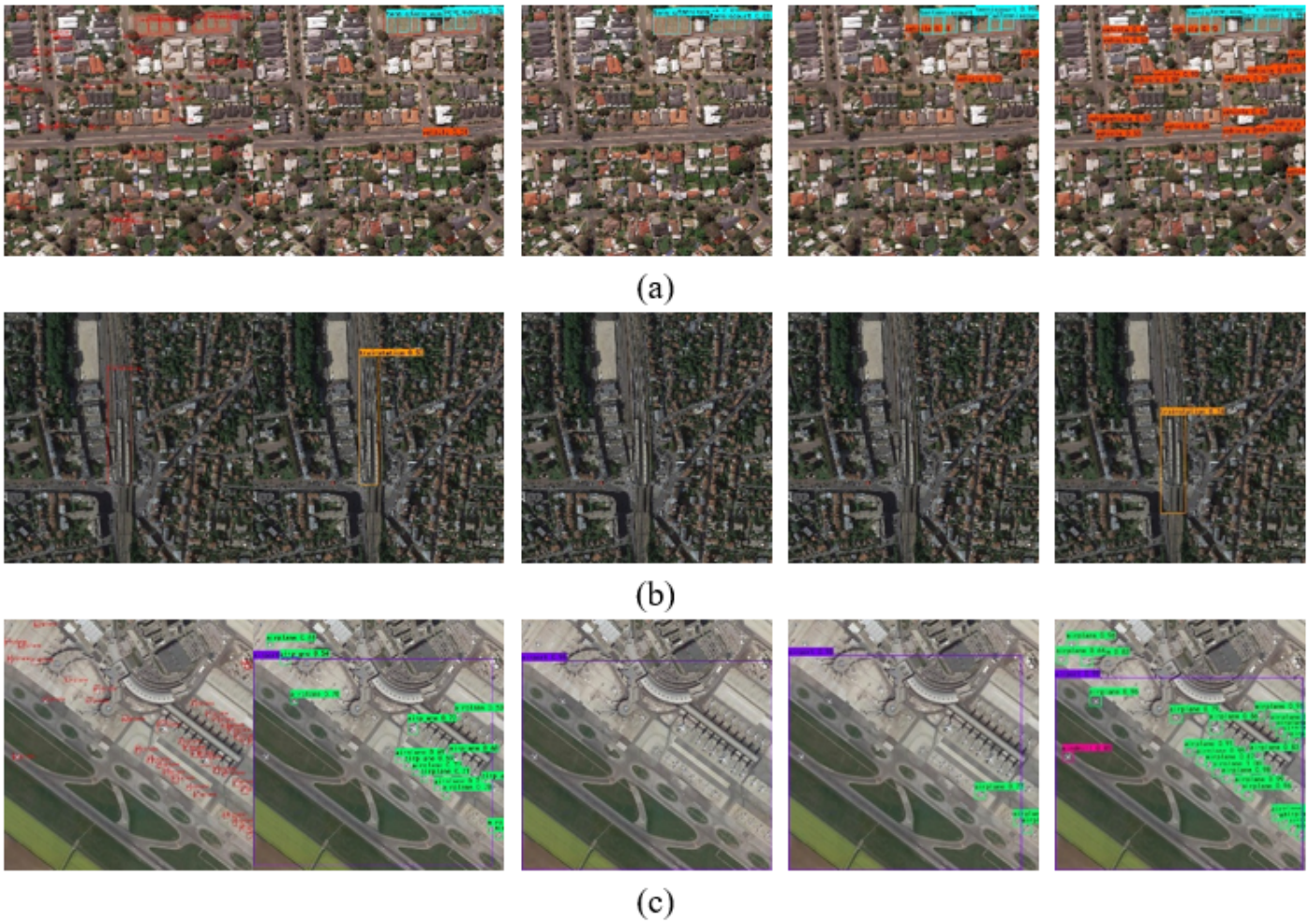

Figure 7

effect drawing of three target types detection