Improved YOLOv5 Network with Attention and Context for Small Object Detection

Tian-Yu Zhang, Zhong-Qiu Zhao and Wei-Dong Tian

School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, 230009, China
Intelligent Manufacturing Institute of HFUT
Guangxi Academy of Sciences
Intelligent Interconnected Systems Laboratory of Anhui Province (Hefei University of Technology)
tianyu_zhang@mail.hfut.edu.cn

Abstract. Object detection is one of the most important and challenging branches in computer vision. Although impressive progress have been achieved on large or medium scale objects, detecting small objects from images is still difficult due to the limited image size and feature information. To deal with the small object detection problem, we explore how the popular YOLOv5 object detector can be modified to improve its performance on detecting small objects. To achieve this, we integrate Coordinate Attention (CA) and Context Feature Enhancement Module (CFEM) in YOLOv5 network. Coordinate Attention is based on attention mechanism and it embeds positional information into channel attention, which enables deep neural network to augment the representations of the objects of interest. Context Feature Enhancement Module explores rich context information from multiple receptive fields and only contains several additional layers. Extensive experimental results on the VisDrone-Detection dataset demonstrate that our approach can improved the performance of the proposed method for small object detection.

Keywords: Small Object Detection, Attention Mechanism, Context Information.

1 Introduction

Object detection is fundamental task of many advanced computer vision problem, such as instance segmentation [1] and image caption [2]. Over the past few years, the emergence of deep convolutional neural network [3, 4] has boosted the performance of object detection, which mainly include two-stage object detection [5, 6, 7] and one-stage object detection [9, 10, 11]. Although these general object detection methods have improve accuracy and efficiency, the limited resolution and context information are not enough to a model, detecting small objects in images can be still difficult.

Efforts have been made to improve the small object detection. Feature pyramid network (FPN) [8] is the first method to enhance features by fusing features from

different levels and constructing feature pyramids. Another approach to small object detection is to generate high-resolution feature to the detection model. Li et al. [12] propose Perceptual GAN to enhance features of small objects with the characteristics of large objects. Leveraging the relationship between an object and its coexisting environment in the real world, context information is another novel method to improve small object detection. Many methods [13, 14, 15] employ additional layers to build context information from multiple layers. Augmented RCNN [16] proposes a novel region proposal network (RPN) to encode the context information around a small object proposal. A context module consisting of three sub-networks is designed to obtain the context information around the proposal network.

As a one-stage object detector, YOLOv5 network [17] is widely used in academia and industry with its excellent detection accuracy and speed. Compared with other one-stage object detectors, YOLOv5 network has a lightweight model size and is easier to train, so many systems are built on it and further improved. However, it is designed to be a general-purpose object detector and is not optimized to detect small objects.

In this paper, an improved YOLOv5 network is proposed. We take YOLOv5 as the main network of our model, then we improve the YOLOv5 network with an attention module to capture key visual information and add a context feature enhancement module. The main contribution of this paper are summarized as follows:

(1) In order to capture more visual information for the detection model, we use Coordinate Attention (CA) to capture key visual information. Coordinate attention can embed positional information into channel attention to enable deep neural network to augment the representations of the objects of interest.

(2) A Context Feature Enhancement Module (CFEM) is proposed, which can capture rich context information from different receptive fields by using multi-path dilated convolutional layers. Furthermore, it uses concatenation to merge the layers with different receptive fields for fusing the coarse-and-fine-grained features in CFEM.

(3) We evaluate our method on VisDrone-Detection dataset. The results demonstrate that the improved YOLOv5 network can get better performance than the baseline method (YOLOv5).

2 Related Work

2.1 CNN-based Object Detection

CNN-based object detection can be mainly divided into two categories: 1) two-stage detectors and 2) one-stage detectors, where the former generate a lot of regions proposals and then classifies each proposal into different object categories. And the later regard object detection as a regression or classification problem, using a unified network to achieve final detection results directly.

Two-stage object detection: In 2014, R.Girshick et al. [18] proposed the Regions with CNN features (RCNN) for object detection. It generates 2000 candidate proposals by Selective Search [19]. These proposals are fed into a CNN model to extract features. Finally the presence of objects and object categories are predicted by linear

SVM classifiers. One of the major issues with RCNN was the need to train multiple systems separately. Fast-RCNN [5] solved this problem by creating a single end-to-end trainable system. Moreover, for Faster-RCNN [6], Region Proposal Network (RPN) integrates proposal generation with the classifier into a single convolutional network. Besides, a lot of two-stage object detection methods have been proposed, such as FPN [8], R-FCN [7], Mask-RCNN [1], and Cascade RCNN [20].

One-stage object detection: In 2015, R.Joseph et al. proposed YOLO [9], which is the first one-stage object detector in computer vision era. The core idea of YOLO is to use the whole feature map to directly predict the location and category of the bounding box. Then, SSD [10] was proposed by Liu et al. in 2015. The main contribution of SSD is the introduction of the multi-reference and multi-resolution detection techniques, which significantly improves the detection accuracy of a one-stage detector, especially for some small objects. There are also extensive other one-stage object detection methods enhancing the detection process in the prediction objectives or the network architectures, such as YOLOv4 [30], RetinaNet [11], and CenterNet [21].

2.2 Attention mechanism

Attention mechanism is a data processing method in machine learning. The basic idea of attention mechanism in computer vision is to let the model learn to focus on key information and ignore unimportant information. SENet [22] is the first to use channel attention. The core of SENet is a squeeze-and-Excitation block which is used to collect global information, capture channel-wise relationships and improve representation ability. In 2018, Woo et al. [23] proposed the convolutional block attention module (CBAM) which stacks channel attention and spatial attention. It decouples the channel attention map and spatial attention map for computational efficiency, and leverages spatial global information by introducing global pooling. For object detection, Cao et al. [24] use an attention-guided module to adaptively extract the useful information around the salient object through the attention mechanism. Moreover, in recent years, Non-local neural network [25] and Self-attention [26] have become very popular due to their capability of building spatial or channel-wise attention. However, due to the large amount of computation inside the self-attention modules, they are often adopted in large models but not suitable for mobile networks.

2.3 Context Information

Many Studies have proved that the context information can improve the performance of object detection and image classification. The feature from the top layers in generic object detectors are enough to capture large objects but the information is greatly limited for small objects. While the feature from the bottom layers contain too specific information which is not useful for detecting large objects but useful for small objects. Then, some detection methods based on context information were proposed to use the relationship between small objects and other objects or background. Oliva et al. [27] illustrate that the around region of small object could provide useful context information to help detect small object. Moreover, the experimental result in [28] also

demonstrate that adding a special context module can significantly improve the detection accuracy. Some studies [28, 29] also propose to use dilated convolution layer to better segment small objects because dilated convolution layer convers larger receptive fields without losing resolution. In this paper, we also use dilated convolution layer to extract features with different receptive fields.

3 Proposed Method

3.1 Network Architecture

Generally, YOLOv5 network can be divided into three parts: the architecture of CSPDarknet53 as backbone, SPP layer and PANet as Neck and YOLO detection head [9]. To further optimize the whole architecture, bag of freebies and specials [30] are provided. In order to balance the influence of detection performance and computing resources, we select YOLOv5-l as our baseline model.

In this paper, the Coordinate Attention and the Context Feature Enhancement Module are introduced to improve the performance of small object detection on the original YOLOv5 network. In order to extract more feature for small objects, we add one more prediction head for small object detection. The structure of four prediction heads can ease the negative influence caused by violent object scale variance and can be more helpful to detect small objects. The improved YOLOv5 network structure is shown in Fig. 1.

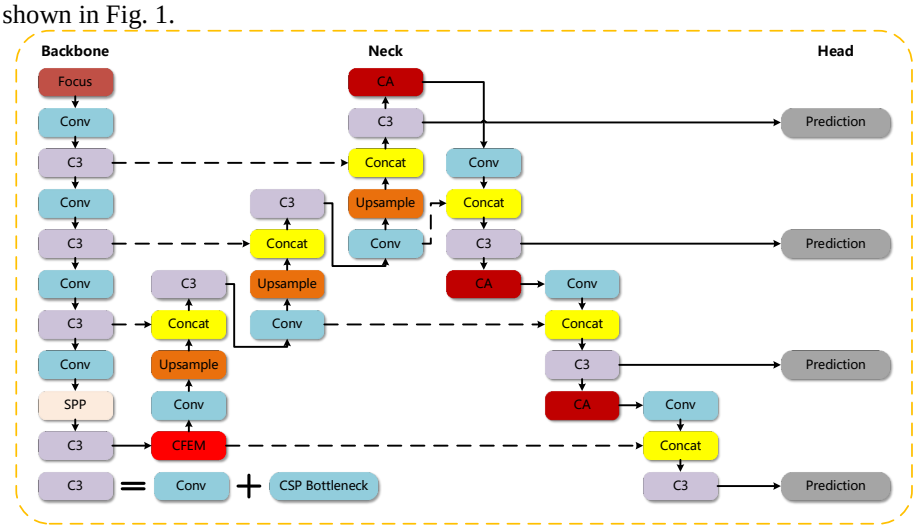

Fig. 1. The complete structure of the improved YOLOv5 network.

3.2 Coordinate Attention

Coordinate Attention [31] can be viewed as a computational unit that aim to enhance the expressive power of the learned features for CNN-based network. The structure of Coordinate Attention module is shown in Fig. 2. Coordinate attention is based on the

Squeeze-and-Excitation (SE) network [20], which can be divided into two steps: squeeze and excitation. The squeeze step is designed for global information embedding and excitation step is used for adaptive recalibration of channel relationships. The squeeze step can be formulated as follows:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i, j) \tag{1}$$

where x_c is the c-th channel for the input $\mathbf{X}$ and z_c is the output related to the c-th channel. The excitation step aims to fully capture channel-wise dependencies, which can be formulated as

$$\hat{X} = X \cdot \sigma\left(T_2\left(\mathrm{ReLU}\left(T_1(z)\right)\right)\right) \tag{2}$$

where $\cdot$ refers to channel-wise multiplication, σ is the sigmoid function and T_1 and T_2 are two linear transformations which can be learned to capture the importance of each channel.

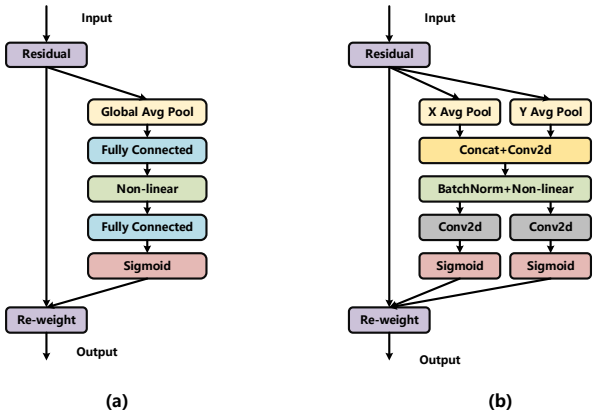

Fig. 2. (a) The architecture of SE network. (b) The architecture of Coordinate Attention module

Compared to SE network, Coordinate Attention takes into account both inter-channel relationships and positional information. It can be decomposed into two steps: coordinate information embedding and coordinate attention generation.

First of all, in order to encourage attention blocks to capture long-range interactions spatially with precise positional information, coordinate attention factorizes the global pooling into a pair of 1D feature encoding operations, which encode each channel along the horizontal coordinate and the vertical coordinate through two spatial extents of pooling kernels $(H,1)$ and $(1,W)$, respectively. The coordinate information embedding step can be formulated as follows:

$$z_c^h(h) = \frac{1}{W} \sum_{0 \le i < W} x_c(h, i) \tag{3}$$

$$z_c^w(w) = \frac{1}{H} \sum_{0 \le j < H} x_c(j, w) \tag{4}$$

where $z_c^h(h)$ is the output of the c-th channel at height h and $z_c^w(w)$ is the output of the c-th channel at width w. These two transformations aggregate features along the two spatial directions respectively and can also allow our attention block to capture long-range dependencies along one spatial direction and preserve precise positional information along the other spatial information. This step can help the networks locate the objects of interest more accurately.

The second step, coordinate attention generation, aims to take advantage of resulting expressive representations from Eq. (3) and Eq. (4), which enable a global receptive field and encode precise positional information. Specifically, coordinate attention generation step first concatenate the aggregated feature maps produced by Eq. (3) and Eq. (4), and then send them to a shared 1×1 convolutional transformation function F_1, which can be formulated as:

$$f = \delta\left(F_1\left(\left[z^h, z^w\right]\right)\right) \tag{5}$$

where $[\cdot,\cdot]$ means the concatenation operation along the spatial dimension, δ is a non-linear activation function and f is the intermediate feature map that encodes spatial information in both the horizontal and vertical direction. Then f will be split into two separate tensors f^h and f^w along the spatial dimension. Two 1×1 convolutional transformations F_h and F_w will be used for separately transforming f^h and f^w to tensors with the same channel number to the input **X**, which can be formulated as:

$$g^h = \sigma\left(F_h\left(f^h\right)\right) \tag{6}$$

$$g^w = \sigma\left(F_w\left(f^w\right)\right) \tag{7}$$

where σ is the sigmoid function. The outputs g^h and g^w are expended and used as attention weights. Finally, the output of coordinate attention module **Y** can be written as:

$$y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(j) \tag{8}$$

3.3 Context Feature Enhancement Module

The proposed Context Feature Enhancement Module (CFEM) is a multi-branch convolution module. CFEN can integrate the context information from different receptive fields and only contains several additional layers. The structure of the CFEM is shown in Fig. 3.

The operation of the CFEM can be performed in two steps. First of all, CFEM consists of multi-path dilated convolutional layers, and each branch uses dilated convolution with different dilation rates, e.g., rate = 1,3,5,7. These separated convolutional layers can harvest multiple feature maps in various receptive fields. The concept of dilated convolutional layer is originally introduced in Deeplab. Dilated convolutional layer can capture information at a larger receptive fields with more context while

keeping the same number of parameters. Besides, to enhance the capacity of modeling geometric transformations, this module also introduces deformable convolutional layers in each path. It ensures CFEM can learn transformation-invariant features from the given feature. Secondly, to maintain the coarse-grained information of the initial inputs, CFEM employs concatenation to merge the outputs of the dilated convolutional layers and feed them into a 1×1 convolutional layer to fuse the coarse-and-fine-grained features.

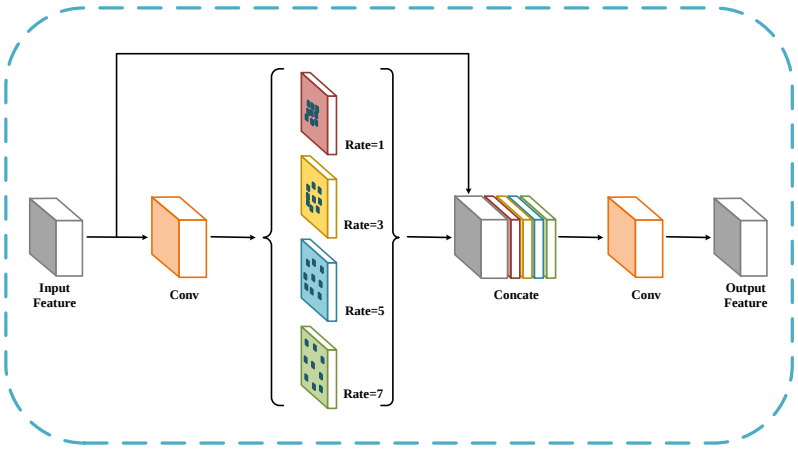

Fig. 3. The overall architecture of Context Feature Enhancement Module (CFEM).

4 Experiments

4.1 Datasets and Evaluation Metrics

VisDrone-Detection: VisDrone-Detection [32] is dataset which has 10209 images which is captured by drone platforms at different locations and at different heights. It includes 6471 images in the training subset, 548 in the validation subset, 1580 in the test-challenge subset, and 1610 in the test-dev subset. About 60.5% of objects in this dataset are small objects. The dominant majority of small objects in VisDrone make it an excellent benchmark for small object detection.

Evaluation Metrics: In this paper, we use the VOC/COCO-based mean Average Precision (mAP) as evaluation metrics to measure the performance of the detectors. Average Precision is originally introduced in VOC2007 dataset. The VOC2007 dataset usually uses a fixed IoU threshold of 0.5 to calculate the mean Average Precision value. The VOC dataset uses a fixed IOU threshold of 0.5 to calculate the AP value. However, after 2014, MS-COCO (Microsoft common objects) dataset gradually emerged. In the COCO dataset, more attention is paid to the accuracy of the location of the prediction bounding box. The AP value is the average AP value of multiple IoU thresholds, specifically taking 10 IOU thresholds (0.5, 0.55, 0.6 ... 0.9, 0.95) between 0.5 and 0.95. Therefore, mAP in VOC dataset is usually marked as mAP@IoU

= 0.5, mAP@0.5 or mAP@.5. In the COCO dataset, it is marked as mAP@IOU = 0.5:0.05:0.95, mAP@IoU = 0.5:0.95 or mAP@.5:.95.

4.2 Implementation Details

We implement Coordinate Attention and Context Feature Enhancement Module with a YOLOv5 detector. All of our models are implement on Pytorch 1.7.1 and use two NVIDIA RTX2080 GPUs for training and testing. In VisDrone-Detection experiment, we train the model on VisDrone-Detection trainset for 200 epochs, and the first 3 epochs are used for warm-up. We use adam optimizer for training, and use 0.01 as the initial learning rate with the cosine annealing strategy. Due to the fixed size of input images needed by the YOLOv5 network, we also adjusted the input images to the uniform size of 640×640. Finally, we evaluate our model on the VisDrone-Detection test-dev set.

4.3 Evaluation on VisDrone-Detection Datasets

To demonstrate the advantages of our proposed method in small object detection, we evaluated our model on VisDrone-Detection test-dev set and compared it with SSD512 [10], FPN [8], RetinaNet [11], YOLOv3 [33], YOLOX [34], and the original YOLOv5. We evaluated performance using metrics including model precision, recall and mean average precision (mAP). The specific results are shown in Table 1.

Table 1. Detection performance comparison with original YOLOv5 network on VisDrone-Detection test-dev set.

Model	Backbone	Precision	Recall	mAP@.5	mAP@.5:.95
SSD512	VGG16	0.11	0.405	0.239	-
FPN	ResNet50	0.273	0.397	0.292	-
RetinaNet	ResNet50	0.138	0.299	0.212	-
YOLOX-s	Darknet53	0.246	0.446	0.338	0.202
YOLOX-l	Darknet53	0.354	0.444	0.371	0.211
YOLOv3	Darknet53	0.459	0.348	0.323	0.183
YOLOv3-spp	Darknet53	**0.494**	0.337	0.324	0.181
YOLOv5	CSPDarknet53	0.314	0.462	0.339	0.192
Ours	CSPDarknet53	0.369	**0.496**	**0.385**	**0.218**

4.4 Ablation Studies

Effect of different attention mechanism. To demonstrate the performance of the coordinate attention, we use our improved YOLOv5 network as baseline to see the performance of the coordinate attention compared to the SE attention and CBAM. The corresponding results of which are all list in Table 2.

Table 2. Comparisons of different attention methods

Model	Precision	Recall	mAP@.5	mAP@.5:.95
YOLOv5	0.299	0.456	0.332	0.189
Improved YOLOv5+SE	0.347	0.493	0.378	0.214
Improved YOLOv5+CBAM	0.346	0.491	0.375	0.214
Improved YOLOv5+CA	**0.369**	**0.496**	**0.385**	**0.218**

Some detection results on VisDrone-detection test-dev set. We have selected some representative images as the display of the test result. Fig. 5 shows the result of small objects, dense objects and the image covering a large area.

Fig. 4. Some visualization results on VisDrone-detection test-dev set.

5 Conclusion

In this paper, we propose an improved YOLOv5 network with attention mechanism and context information to remedy the problem of small object detection. The Coordinate Attention embeds positional information into channel attention to enable neural network to attend over large regions while avoiding significant computation cost. Additionally, we propose a Context Feature Enhancement Module to capture rich context information by using multi-branch dilated convolutional layers. Experiments show that our proposed method can improve the performance of YOLOv5 network for small objects.

6 Acknowlegment

This work was supported in part by the National Natural Science Foundation of China under Grant 61976079, in part by Guangxi Key Research and Development Program under Grant 2021AB20147, and in part by Anhui Key Research and Development Pro-gram under Grant 202004a05020039.

References

1. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
2. Wu, Q., Shen, C., Wang, P., Dick, A., Van Den Hengel, A.: Image captioning and visual question answering based on attributes and external knowledge. IEEE transactions on pattern analysis and machine intelligence 40(6), 1367–1381 (2017)
3. Zhao, Z.Q., Gao, J., Glotin, H., Wu, X.: A matrix modular neural network based on task decomposition with subspace division by adaptive affinity propagation clustering. Applied mathematical modelling 34(12), 3884–3895 (2010)
4. Zhao, Z., Wu, X., Lu, C., Glotin, H., Gao, J.: Optimizing widths with pso for center selection of gaussian radial basis function networks. Science China Information Sciences 57(5), 1–17 (2014)
5. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 1440–1448 (2015)
6. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems 28 (2015).
7. Dai, J., Li, Y., He, K., Sun, J.: R-fcn: Object detection via region-based fully convolutional networks. Advances in neural information processing systems 29 (2016)
8. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
9. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
10. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision. pp. 21–37. Springer (2016)
11. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
12. Li, J., Liang, X., Wei, Y., Xu, T., Feng, J., Yan, S.: Perceptual generative adversarial networks for small object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1222–1230 (2017)
13. Bell, S., Zitnick, C.L., Bala, K., Girshick, R.: Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2874–2883 (2016)
14. Shrivastava, A., Sukthankar, R., Malik, J., Gupta, A.: Beyond skip connections: Top-down modulation for object detection. arXiv preprint arXiv:1612.06851 (2016)
15. Fu, C.Y., Liu, W., Ranga, A., Tyagi, A., Berg, A.C.: Dssd: Deconvolutional single shot detector. arXiv preprint arXiv:1701.06659
16. Chen, C., Liu, M.Y., Tuzel, O., Xiao, J.: R-cnn for small object detection. In: Asian conference on computer vision. pp. 214–230. Springer (2016)
17. Ultralytics. YOLOv5 2020 Available from: https://github.com/ultralytics/yolov5
18. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 580–587 (2014)
19. Uijlings, J.R., Van De Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. International journal of computer vision 104(2), 154–171 (2013)

20. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6154–6162 (2018)
21. Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., Tian, Q.: Centernet: Keypoint triplets for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6569–6578 (2019)
22. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
23. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
24. Cao, J., Chen, Q., Guo, J., Shi, R.: Attention-guided context feature pyramid network for object detection. arXiv preprint arXiv:2005.11475 (2020)
25. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7794–7803 (2018)
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems 30 (2017)
27. Oliva, A., Torralba, A.: The role of context in object recognition. Trends in cognitive sciences 11(12), 520–527 (2007)
28. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
29. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062 (2014)
30. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
31. Hou, Q., Zhou, D., Feng, J.: Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13713–13722 (2021)
32. Zhu, P., Wen, L., Bian, X., Ling, H., Hu, Q.: Vision meets drones: A challenge. arXiv preprint arXiv:1804.07437 (2018)
33. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
34. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021)