AgBFPN: Attention Guided Bidirectional Feature Pyramid Network for Object Detection

Lanjie Jiang[1,2], Xiang Zhang[1,2](✉), Ruijing Yang[1,2], and Yudie Liu[1,2]

[1] University of Electronic Science and Technology of China, Chengdu 611731, Sichuan, China
{202052012112,202022012111,202152011924}@std.uestc.edu.cn
[2] Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou 324000, Zhejiang, China
uestchero@uestc.edu.cn

Abstract. Object detection is increasingly in demand in IoT service applications. Deep learning based object detection algorithms are now in fashion. As the most popular multi-scale object detection network at present, Feature Pyramid Network achieves feature augmentation by fusing features of neighboring layers. It is widely used in the most advanced object detectors to detect objects of different scales. In this paper, we propose a new attention mechanism guided bidirectional feature pyramid architecture named AgBFPN to enhance the transfer of semantic and spatial information between each feature map. We design Channel Attention Guided Fusion(CAGF) Module and Spatial Attention Guided Fusion(SAGF) Module to enhance feature fusion. The CAGF mitigates the loss of information induced by channel reduction and better transfers the semantic information from high-level to low-level features. The SAGF passes the rich spatial information of shallow features into deep features. Our experiments show that AgBFPN achieves higher Average Precision for multi-scale object detection.

Keywords: Deep learning · Object detection · Feature pyramid network · Attentional mechanisms

1 Introduction

With the rapid expansion of IoT, there is an increasing demand for object detection in IoT application scenarios such as intelligent transportation and public

Supported by National Science Foundation of China (U19A2052, U1733111), National Key R&D Program of China (2021YFB1600500), Chengdu Science and Technology Project (2021-JB00-00025-GX), Key R&D Program of Sichuan Province (2020YFG0478), the Municipal Government of Quzhou under Grant Number 2021D012.

Y. Cao and X. Shao (Eds.): MONAMI 2022, LNICST 474, pp. 386–397, 2023.
https://doi.org/10.1007/978-3-031-32443-7_28

safety. Object detection algorithms based on the deep convolutional network have already achieved significant advancements in recent years. The object scale is the important factor related to the performance of object detection. Some detailed information about small objects is contained in shallow features. With deeper layers, the geometric details may vanish entirely (oversized receptive field), making it hard to detect small objects using deep features. Deeper feature maps can provide semantic information about large objects. Thus, object detection with a wide range of object scale changes is still a challenging problem [1].

Deep features in convolutional neural networks have a large receptive field and rich semantic information but lose geometric detail information. In contrast, shallow features have rich detail information with small receptive fields, but lack of semantic information. Multi-scale learning combines deep semantic information and shallow representation information, which is an effective strategy to improve the performance of object detection [2–4]. FPN [4] is the frequently utilized multi-scale object detection network at present. It passes down the high-level feature information and supplements the low-level semantics to solve the multi-scale problem in object detection.

We think about two issues that may exist in feature pyramid network. The first is before feature fusion, different level features will go through a convolutional layer with a convolution kernel of size 1×1 to reduce feature channels, and excessive channel attenuation will bring about unavoidable information loss. In addition, in the top-down pathway, the top-level pyramid does not get supplementary information, so the reduction of channels will lose information instead.

Based on these issues, we design the Channel Attention Guided Fusion (CAGF) Module, which introduced the attention mechanism. The features of high-level layers with sufficient classification details can be applied as attention to guide the low-level features. It transfers different scale semantic features from top to bottom, so that can obtain high-resolution and strong semantic features, which is beneficial to the detection of multi-scale objects. Furthermore, we add a new bottom-up spatial perception pathway by Spatial Attention Guided Fusion (SAGF) Module to pass the rich spatial information of shallow features into deep features. Combined Channel Attention Guided Fusion Module and Spatial Attention Guided Fusion, our AgBFPN architecture archives effective accuracy improvements on PASCAL VOC2007 [5] and MS COCO [6].

Based on these issues, the main contributions of our paper are as follows:

- Firstly, we design the Channel Attention Guided Fusion (CAGF) Module, which introduces the attention mechanism. The deep features have sufficient classification information to guide the shallow features. It conveys semantic feature information from top to bottom at various scales, which helps multi-scale object detection.
- Furthermore, we add a new bottom-up spatial perception pathway by Spatial Attention Guided Fusion (SAGF) Module to pass the rich spatial information of shallow features into deep features.

- Combined Channel Attention Guided Fusion Module and Spatial Attention Guided Fusion, our AgBFPN architecture archives effective accuracy improvements on PASCAL VOC2007 [5] and MS COCO [6].

2 Related Work

Early multi-scale detection has two ideas. One is to utilize different convolution kernel sizes to acquire various scale information through different sizes of the receptive field, and the other is to use image pyramids to detect different scale objects by inputting images at various scales. However, these two methods are computationally expensive and suffer from a limited range of receptive fields. Later, multi-scale detection is gradually developed to execute object detection based on the feature pyramid, using feature maps of various stages to build a feature pyramid network to detect multi-scale objects. Since FPN [4] was proposed, multiple versions have been iterated successively [9–12], from no fusion to top-down unidirectional fusion, and then gradually to bidirectional fusion as in Fig. 1.

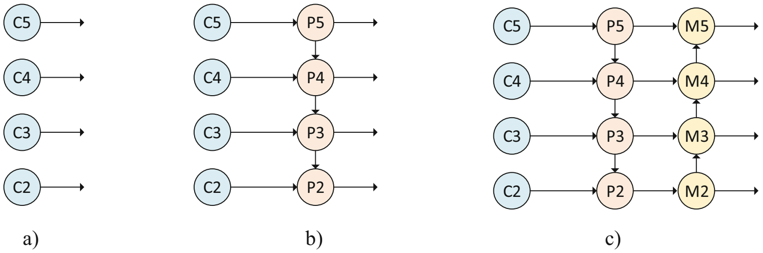

Fig. 1. Evolution of Feature Pyramid Networks: a) No fusion; (b) Top-down unidirectional fusion; and (c) Simple bidirectional fusion

2.1 No Fusion

Most classical object detection networks use the last layer of deep neural networks to make predictions. However, it is going to be hard to detect small objects in the last feature map due to the loss of spatial and detailed feature information. SSD [2] is one of the typical representatives of no fusion using multi-scale features. It uses shallower feature maps to detect smaller objects and Deeper feature maps to detect larger objects.

2.2 Top-Down Unidirectional Fusion

The current object detection model's main fusion mode is top-down unidirectional fusion FPN [4]. It introduces a top-down network architecture to enhance features with feature fusion from neighboring layers. Based on FPN [4], Liang [8] proposes a deep feature pyramid network, which enhances the semantic features of small objects by using feature pyramids with lateral connections. Libra R-CNN [9] fuses and refines multi-scale feature elements with a balanced feature pyramid. AugFPN [10] proposes a series of FPN enhancement methods.

2.3 Bidirectional Fusion

Only top-to-bottom feature maps are fused by FPN [4]. Secondary fusion from bottom to top has been proposed for the first time by PANet [13]. Based on traditional feature pyramid networks, PANet [13] increases the shallow information to the deep layer just by adding a bottom-up fusion pathway. Since the proposal of PANet [13] proves the effectiveness of bidirectional fusion, several relevant researches try more complex bidirectional fusion, such as NAS-FPN [14], ASFF [15] and BiFPN [16]. NAS-FPN [14] employs neural architecture search to learn all cross-scale connections for better fusion. For simple and fast feature fusion, BiFPN [16] proposes a weighted bidirectional feature pyramid network.

3 Proposed Methods

We describe our attention guided bidirectional Feature Pyramid Network architecture in this section. By introducing an attention mechanism, it fully utilizes semantic information from deep features and spatial information from shallow features to optimize the fusion of feature information at different scales. In AgBFPN, two main components are proposed: a Channel Attention Guided Fusion (CAGF) Module and a Spatial Attention Guided Fusion (SAGF) Module. We will describe them in detail below.

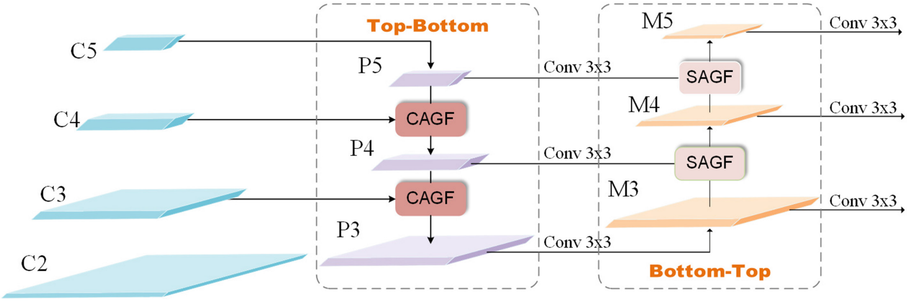

Fig. 2. An overview of our Attention guided Bidirectional Feature Pyramid Network(AgBFPN)

3.1 Overall

Figure 2 depicts the overall framework of AgBFPN. Following the configuration of FPN [4], the outputs of the backbone features are indicated as $\{C2, C3, C4, C5\}$ to build a feature pyramid, which corresponds to the $\{4, 8, 16, 32\}$ strides. We separate $C2$ from the four-level input features entering the feature pyramid network because the $C2$ would take up more computational resources. We keep $\{C3, C4, C5\}$ to build the feature pyramid. In FPN, horizontal connections are required to reduce the number of channels of each feature layer to the same 256. Different from this, we retain the number of input channels and complete the top-down semantic information transfer between different features through the CAGF. $\{P3, P4, P5\}$ are the features generated by the top-down path of the feature pyramid. We build a spatial perception bottom-top pathway with the SAGF that successively transfers the spatial information from low-level features to high-level features. $\{M3, M4, M5\}$ are the features generated by the spatial perception bottom-top pathway.

3.2 Channel Attention Guided Fusion Module

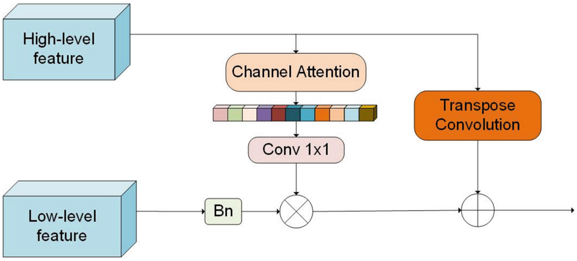

Fig. 3. The structure of Channel Attention Guided Fusion Module (CAGF)

With output channels of $\{256, 512, 1024, 2048\}$, residual network [17] is frequently applied as backbone network, where low-level feature maps include rich spatial information and high-level feature maps include rich semantic information.

In the top-down pathway, FPN [4] firstly uses a convolutional layer with a convolution kernel of size 1×1 to decrease the channel dimension of C_i to 256. On the basis that each feature map has the same number of channels, FPN [4] uses nearest neighbor interpolation to upsample and then fuse them by adding to transfer the features from the upper layer to the bottom. This approach reduces the number of channels of the top-level feature $C5$ from 2048 to 256 before fusion, which will result in severe loss of channel information. For this purpose, we introduce a method to fuse the features of neighboring layers without changing the number of channels.

We design Channel Attention Guided Fusion Module (CAGF) inspired by PAN [18]. The channel attention mechanism is introduced in CAGF, as illustrated in Fig. 3. Each channel mapping of high-level layer features can be seen as the response to a specific class. Obtaining the interdependence between different channel mappings can effectively enhance the characterization of feature maps for specific semantics. High-level layer features have adequate classifieds, that can be directed as the attention to direct the low-level.

The basic idea is that high-level features are weighted by predicting a channel weight mask and then weighting the low-level features. In specific, high-level feature map channel weight masks are predicted using a channel attention module [19].

This mask is then multiplied by the low-level feature map after the batch normalization layer to obtain a weighted feature map. At last, the high-level feature maps upsampled by transposed convolution are fused to the weighted low-level feature map and passed layer by layer.

3.3 Spatial Attention Guided Fusion Module

The top-down pathway complements the semantics from high-level features for high-resolution low-level features. But the features of the top-level pyramid lose information due to the reduction of channels in the top-level feature map. So we build a bottom-up spatial awareness path, aiming to supplement high-level features with spatial and detailed information from low-level to help multi-scale object detection.

After the top-down pathway, the result passes through a 3×3 convolution to mitigate the upsample aliasing effect. At this point, each layer of the feature map has the same number of 256 channels. For the deepest feature $C5$, the number of channels is reduced from 2048 to 256 without additional information, so there is a loss of information instead.

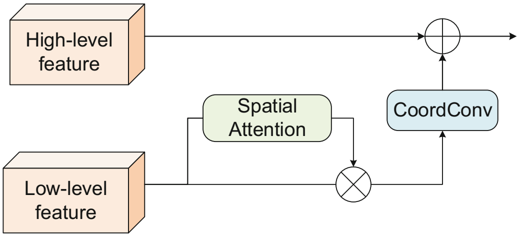

Fig. 4. The structure of Spatial Attention Guided Fusion Module (SAGF)

As Fig. 4 shows, we construct a spatial attention guided fusion module (SAGF) in the spatial perception pathway, which introduces a spatial attention mechanism in SAGF. The high-resolution low-level features have enough detailed spatial information to complement the high-level features.

To obtain the spatial attention map, low-level features are first passed through the spatial attention module. Then, applying the spatial attention map to the original feature map completes the spatial information calibration. After that, we downsample the low-level features using CoordConv [20], which adds two coordinate channels to enable the convolutional downsampling process to sense the feature map's spatial information. The downsampled low-level features are additively fused with the high-level features so that the high-level features fuse the spatial information from the low-level features.

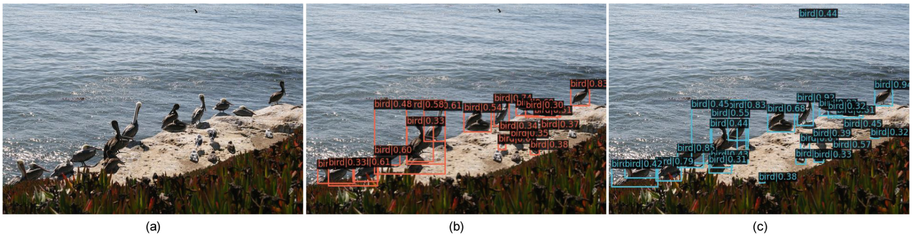

Fig. 5. Result comparison: (a) is the original image; (b) is the result of RetinaNet with FPN; (c) is the result of RetinaNet with AgBFPN (ours).

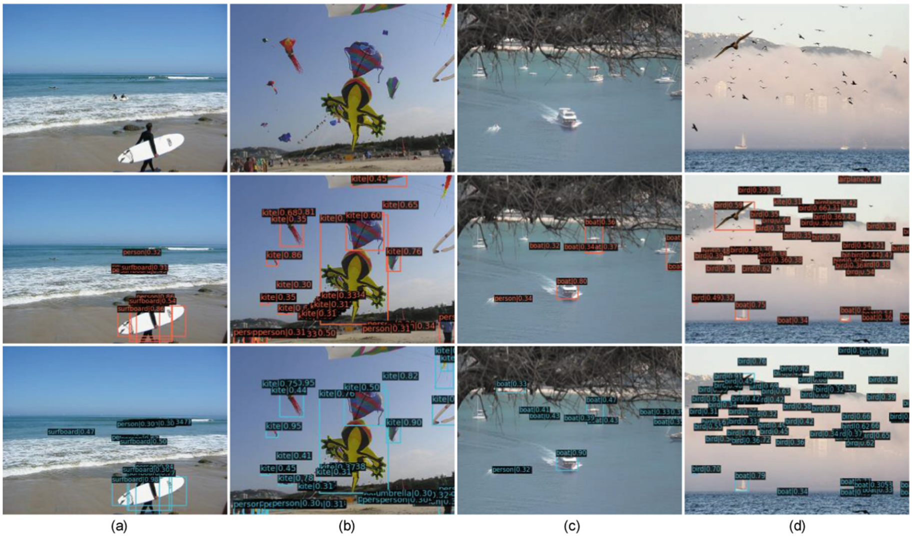

Fig. 6. Qualitative result comparison. The first row is the original image, the second row is the result of RetinaNet with FPN and the third row is the result of RetinaNet with AgBFPN (ours).

4 Experiments

4.1 Dataset and Evaluation Metric

We conduct experiments on the PASCAL VOC2007 [5] and MS COCO2017 [6] detection datasets. PASCAL VOC2007 [5] has 9,963 images with 20 classes, 50% of which are used for training/validation and 50% for testing. MS COCO2017 [6] has 80 classes and provides train2017 containing 115k images, val2017 containing 5k images, and test2017 containing 20k images.

For PASCAL VOC2007 [5], we report Mean Average Precision(mAP) on the basis that the IOU threshold is selected as 0.5. For MS COCO2017 [6], all reported results adhere to the standard COCO-style Mean Average Precision (mAP) metrics at multiple IoU thresholds from 0.5 to 0.95 with a 0.05 interval.

Table 1. Comparison of object detection performance on COCO test-dev. The asterisk (*) indicates that the results were re-implemented with MMDetection v2.0.

Baseline	Neck	Dataset	Schedule	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
RetinaNet* [21]	FPN	COCO	1x	34.6	52.7	36.7	19.3	37.8	45.3
RetinaNet* [21]	PAFPN [13]	COCO	1x	36.0	55.5	38.4	20.1	39.9	47.0
RetinaNet* [21]	FPN	mini	2x	24.6	39.9	25.8	12.6	28.4	33.9
Faster RCNN* [7]	FPN	mini	1x	24.2	45.4	23.8	12.5	30.5	32.6
FCOS* [23]	FPN	mini	1x	18.3	31.3	18.8	12.5	20.2	23.8
RetinaNet	**AgBFPN**	COCO	1x	37.1	56.6	39.2	22.2	40.9	47.5
RetinaNet	**AgBFPN**	mini	1x	21.4	36.0	22.9	12.6	25.7	30.5
RetinaNet	**AgBFPN**	mini	2x	26.4	42.7	26.7	15.9	31.7	36.1
FCOS	**AgBFPN**	mini	1x	20.2	33.7	21.5	14.2	24.6	27.5
Faster RCNN	**AgBFPN**	mini	1x	27.0	48.6	27.3	15.5	31.7	34.9

Table 2. Comparison of object detection performance on VOC test-dev. The asterisk (*) indicates that the results were re-implemented with MMDetection v2.0.

Baseline	Neck	Dataset	Schedule	AP
RetinaNet* [21]	FPN	VOC	1x	72.4
RetinaNet* [21]	PAFPN [13]	VOC	1x	72.7
RetinaNet* [21]	NASFPN [14]	VOC	1x	73.1
RetinaNet	**AgBFPN**	VOC	1x	74.7

4.2 Implementation Details and Main Results

Each of our experiments is based on MMDetection v2.0 [22]. By default, we train the networks for 12 epochs using NVIDIA 3060 TI (2 images per GPU). For the training process, the 1x schedule represents 12 epochs and the 2x schedule represents 24 epochs. The initial learning rate is 0.001. It respectively decreases by 0.1 at 9 and 12 epochs in the 1x schedule, corresponding to the 17 and 23 epochs in the 2x schedule.

Figure 5 compares the outcomes between the FPN and our AgBFPN. As can be observed, our AgBFPN is more sensitive to multi-scale object detection. More contrast can be seen in Fig. 6.

We assess AgBFPN on the COCO test-dev subset to validate the effectiveness of our approach for performance enhancement. To facilitate the verification, we randomly extracted part of the data of the MS COCO2017 detection dataset as miniCOCO (the same ratio of train/val/test to COCO) for part of the experiments.

To guarantee the designed network's generalization capabilities, we train the model on training data, validate on validation data, and lastly test with the optimal parameters on test data.

We exhibit re-implemented results of the corresponding baselines for fair comparisons. By swapping out FPN for AgBFPN, RetinaNet using ResNet-50 as the backbone achieves 37.1 AP on COCO test-dev, 2.5 points above the baseline, as demonstrated in Table 1. The same network achieves 74.7 AP on VOC test-dev, 2.3 points above the baseline, as demonstrated in Table 2.

4.3 Ablation Experiments

we also test the impact of each proposed AgBFPN component with RetinaNet [21] baseline on PASCAL VOC2007 [5] in Table 3. The training procedure runs on 1x schedule (12 epochs). For fair comparisons, ablation experiments are conducted under the same conditions.

Table 3. Effect of each component on VOC test-dev.

baseline	CAGF	SAGF	AP
✓			72.4
✓	✓		73.4
✓		✓	73.3
✓	✓	✓	**74.7**

Table 4. Ablation studies of Channel Attention Guided Fusion Module on VOC test-dev.

baseline	SE	CAM	BN	AP
✓				72.4
✓	✓			72.7
✓	✓		✓	73.2
✓		✓		72.6
✓		✓	✓	**73.4**

Channel Attention Guided Fusion Module. CAGF introduces channel attention to optimize feature fusion of adjacent feature layers in the top-down pathway and CAGF boosts performance by 1.0 AP.

We conduct ablation experiments in this module to examine the impact of different attentional mechanisms. In addition, to better integrate semantic information from higher levels, we verified the effectiveness of adding a Bn layer before the fusion of low-level features shown in Table 4.

We speed up network convergence by adding a Batch Normalization [24] layer. To avoid the gradient from vanishing or exploding and speed up training, Batch Normalization can address the issue that the middle layer's data distribution changes during the training process.

Spatial Attention Guided Fusion Module. Then we add SAGF on RetinaNet with the CAGF. According to Table 3, the combined module increases AP by 2.3 points above the corresponding baseline. Adding the SAGF module raises the AP by 1.3 points above the CAGF-only baseline. We also do ablation tests to assess CoordConv's effect compared with the traditional 3×3 convolution downsample layer in Table 5.

Table 5. Ablation studies of Spatial Attention Guided Fusion Module on VOC test-dev.

baseline w/ CAGF	Conv	CoordConv	AP
✓			73.4
✓	✓		73.3
✓		✓	**74.7**

5 Conclusion

In this paper, we propose a novel Attention guided Bidirectional Feature Pyramid Network (AgBFPN) for object detection to further improve the performance of multi-scale learning. To better convey the semantic information in the deep feature maps, we design a Channel Attention Guided Fusion Module. The module uses the higher-level feature maps to guide the lower-level feature maps during the top-down pathway in the feature pyramid network. Moreover, we build a Spatial Perception Bottom-up Pathway with Spatial Attention Guided Fusion Module to effectively transfer the spatial information in the underlying feature maps. According to the results of our experiment, the proposed methods can improve the performance of object detection algorithms based on the FPN framework on MS COCO2017 and PASCAL VOC2007 object detection benchmark. AgBFPN improves RetinaNet by 2.3 points AP on PASCAL VOC2007 and 2.5 points AP on MS COCO2017 when using ResNet50 as the backbone.

References

1. Zou, Z., Shi, Z., Guo, Y., Ye, J.: Object detection in 20 years: a survey. arXiv preprint arXiv:1905.05055 (2019)
2. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
3. Cao, G., Xie, X., Yang, W., Liao, Q., Shi, G., Wu, J.: Feature-fused SSD: fast detection for small objects. In: Ninth International Conference on Graphic and Image Processing, vol. 10615, pp. 381–388 (2018)
4. 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)
5. Everingham, M., Gool, L.V., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results (2007). https://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html
6. Lin, T.-Y.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
7. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
8. Liang, Z., Shao, J., Zhang, D., Gao, L.: Small object detection using deep feature pyramid networks. In: Hong, R., Cheng, W.-H., Yamasaki, T., Wang, M., Ngo, C.-W. (eds.) PCM 2018. LNCS, vol. 11166, pp. 554–564. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00764-5_51
9. Pang, J., Chen, K., Shi, J., Feng, H., Ouyang, W., Lin, D.: Libra R-CNN: towards balanced learning for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 821–830 (2019)
10. Guo, C., Fan, B., Zhang, Q., Xiang, S., Pan, C.: Augfpn: improving multi-scale feature learning for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12595–12604 (2020)
11. Luo, Y., et al.: CE-FPN: enhancing channel information for object detection. Multimed. Tools Appl. 1–20 (2022). https://doi.org/10.1007/s11042-022-11940-1
12. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890 (2017)

13. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8759–8768 (2018)
14. Ghiasi, G., Lin, T.Y., Le, Q.V. : NAS-FPN: learning scalable feature pyramid architecture for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7036–7045 (2019)
15. Liu, S., Huang, D., Wang, Y.: Learning spatial fusion for single-shot object detection. arXiv preprint arXiv:1911.09516 (2019)
16. Tan, M., Pang, R., Le, Q.V.: Efficientdet: scalable and efficient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10781–10790 (2020)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
18. Li, H., Xiong, P., An, J., Wang, L.: Pyramid attention network for semantic segmentation. arXiv preprint arXiv:1805.10180 (2018)
19. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: CBAM: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision, pp. 3–19 (2018)
20. Liu, R., et al.: An intriguing failing of convolutional neural networks and the coordconv solution. In: Advances in Neural Information Processing Systems, vol. 31 (2018)
21. 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)
22. Chen, K., et al.: MMDetection: open MMlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
23. Tian, Z., Shen, C., Chen, H., He, T.: FCOS: fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9627–9636 (2019)
24. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning, pp. 448–456 (2015)