Hierarchical Focused Feature Pyramid Network for Small Object Detection

Siwei Wang, Zhiwei Chen, Haoyang Ding, and Liujuan Cao(✉)

Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, 361005 Xiamen, People's Republic of China

caoliujuan@xmu.edu.cn

Abstract. Small object detection has been a persistently practical and challenging task in the field of computer vision. Advanced detectors often utilize a feature pyramid network (FPN) to fuse the features generated from various receptive fields, which improve the detection ability of multi-scale objects, especially for small objects. However, existing FPNs typically employ a naive addition-based fusion strategy, which neglects crucial details that may exist only at specific levels. These details are vital for accurately detecting small objects. In this paper, we propose a novel Hierarchical Focused Feature Pyramid Network (HFFPN) to enhance these details while ensuring the detection performance for objects of other scales. HFFPN consists of two key components: Hierarchical Feature Subtraction Module (HFSM) and Feature Fusion Guidance Attention (FFGA). HFSM is first designed to selectively amplify the information important to small object detection. FFGA is devised to focus on effective features by utilizing global information and mining small objects' information from high-level features. Combining these two modules contributes greatly to the original FPN. In particular, the proposed HFFPN can be incorporated into most mainstream detectors, such as Faster RCNN, Retinanet, FCOS, *etc.* Extensive experiments on small object datasets demonstrate that HFFPN achieves consistent and significant improvements over the baseline algorithm while surpassing the state-of-the-art methods.

Keywords: Small object detection · Feature pyramid network · Self-attention

1 Introduction

Object detection is a widely studied task that aims to locate and classify the objects of interest. In recent years, object detection has achieved remarkable progress due to the powerful ability of Convolutional Neural Networks (CNNs) and the availability of an enormous amount of data [4]. However, as an important branch of object detection, small object detection has always been a bottleneck for detector performance. Small objects, typically refer to objects with a pixel

Q. Liu et al. (Eds.): PRCV 2023, LNCS 14436, pp. 432–444, 2024.
https://doi.org/10.1007/978-981-99-8555-5_34

size of less than 1024 (32×32) [18], have very important research significance in practical scenarios such as remote sensing detection [1,14], disaster rescue [22, 38], and intelligent transportation system [20,31]. Unfortunately, the features of small objects are extremely limited, making them susceptible to background and noise interference. Moreover, these weak features are likely to be lost during the feature extraction and downsampling process, leading to a noticeable drop in detection performance when dealing with small objects. For example, Faster R-CNN [24] achieves an mAP of 41.0% and 48.1% for medium and large objects on the COCO dataset [18], respectively, but the result for small objects drops significantly to only 21.2%. Therefore, as a task with both theoretical significance and practical demand, how to effectively enhance the detection performance on small objects is an urgent and important problem to be solved.

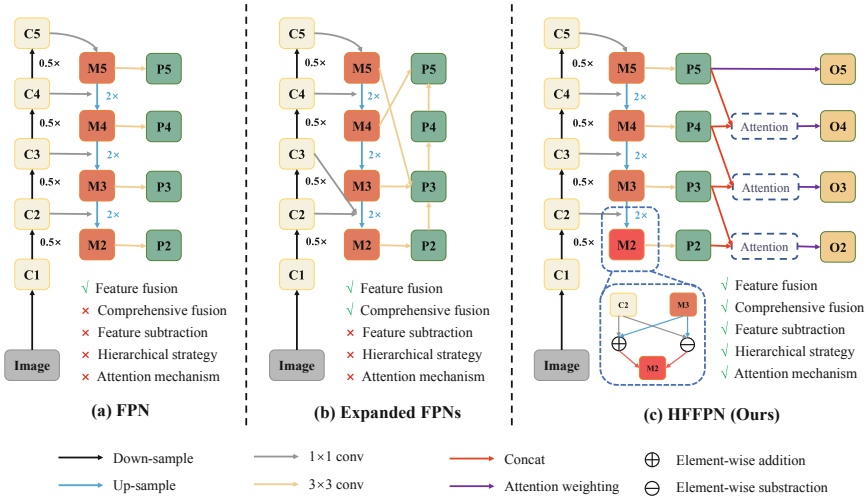

Fig. 1. Pictorial demonstrations of existing feature pyramid networks.

In order to detect objects of various sizes, advanced detectors often adopt a divide-and-conquer approach that utilizes larger receptive fields to detect large objects and smaller ones to detect small objects. This principle is usually reflected in the Feature Pyramid Network (FPN) [16]. As shown in Fig. 1, many studies have noticed the importance of FPN and attempted to fuse low-level and high-level features in a more effective manner to obtain better detection results. Consequently, many FPN variants have been devised to achieve more comprehensive feature fusion [7,19,27]. We collectively refer to them as the Expanded FPNs. However, the fusion strategies of expanded FPNs are generally accomplished by the element-wise addition operation, and the only difference between them is the level of the fused features. In contrast, to extract detailed features that are conducive to small object detection, the element-wise subtraction between the corresponding levels may better obtain edge information [28]. It should be noted

that in the high-level feature layers, the information of small objects is almost submerged in the frequent downsampling process, and subtracting such features cannot extract small object features. Instead, it may lead to the loss of main body features. Therefore, a hierarchical feature fusion strategy is necessary. On the other hand, we notice that the fused features have information of different scales. Using global information can help to guide the refinement of each level of features, thus improving detection performance [15].

Based on the above observations, we propose a novel approach called Hierarchical Focused Feature Pyramid Network (HFFPN). HFFPN mainly consists of two parts: Hierarchical Feature Subtraction Module (HFSM) and Feature Fusion Guidance Attention (FFGA). HFSM leverages the feature subtraction operation to obtain the edge information of objects. To avoid erasure effects on main body information caused by subtraction operations at higher semantic levels, HFSM adopts a hierarchical subtraction strategy. Besides, the proposed FFGA introduces a novel attention mechanism for small object detection by incorporating both self-features and higher-level features in the generation of attention weights. It deviates from the common self-attention methods [12,30], which solely relies on the self-features. The adjacent feature levels often contain richer interaction information, particularly with low-level features assisting high-level features in exploring potential information on small objects.

To sum up, our contributions are summarized as follows:

- We design a brand-new Hierarchical Feature Subtraction Module (HFSM). It fully utilizes the difference of information between feature layers and helps to improve the performance of small object detection. The hierarchical strategy employed in HFSM further enhances the robustness of the model.
- We introduce a Feature Fusion Guidance Attention (FFGA) to utilize the global fused information. The self-attention mechanism used highlights useful information and suppresses noise information by weighting the features of itself, helping to explore potential information of small objects.
- Extensive experiments on the DOTA and COCO datasets demonstrate that the proposed HFFPN significantly improves the performance of the baseline algorithm and surpasses the current state-of-the-art detectors.

2 Related Work

2.1 Small Object Detection

With the development of deep learning, extensive research has been carried out on small object detection. There have been numerous attempts to enhance the performance of small object detection from different perspectives, all with the common goal of increasing the exploitable features of small objects. SCRDet [37] achieves a more refined feature fusion network by introducing flexible downsampling strides, allowing for the detection of a broader spectrum of smaller objects with greater precision. R3Det [36] designs a feature refinement module to enhance the detection performance of small objects. Oriented RepPoints [13]

captures features from adjacent objects and background noise for adaptive point learning, which utilizes contextual information to discover small objects.

2.2 Feature Pyramid Network

It is a consensus that the shallow layers are usually rich in detailed information but lack abstract semantic information, while the deeper layers are on the contrary due to the downsampling. Smaller objects predominantly rely on shallow features and can be more effectively detected by detectors with smaller receptive fields. Feature Pyramid Network [16] combines the deep layer and shallow layer features by building a top-down pathway to form a feature pyramid. PAFPN [19] enriches the feature hierarchy by adding a bottom-up path, enhancing deeper features without losing information from the shallow layers. HRFPN [27] utilizes multiple cross-branch convolution to enhance feature expression. NAS-FPN [7] searches for the optimal combination method for feature fusion in each layer.

2.3 Self-Attention

The attention mechanism exhibits an impressive capability to quickly concentrate on and distinguish objects within a scene, while effectively ignoring irrelevant aspects. And self-attention is also a powerful technique in deep learning that allows a model to selectively focus on different parts of input, effectively capturing dependencies and relationships within it. Spatial self-attention and channel self-attention are two common kinds of self-attention. SENET [12] is the first proposed channel attention. It uses a SE block to gather global information through channel-wise relationships and enhance the representation capacity. CBAM [30] can sequentially generate attention feature maps in both channel and spatial dimensions for adaptive feature refinement, resulting in the final feature map. Self-attention mechanism has shown outstanding performance in handling small objects to some extent. SCRDet [37] utilizes pixel attention and channel attention to highlight small object regions while mitigating the impact of noise interference. CrossNet [14] develops a cross-layer attention module to enhance the detection of small objects by generating more pronounced responses.

3 Methodology

3.1 Overview

In order to fully utilize the information of small objects, we propose a novel feature pyramid network, named HFFPN, as shown in Fig. 2. The detector receives the input image I and sends it to the backbone network for feature extraction. The image feature C_i gradually becomes richer in semantic information during the subsampling process while losing detailed information. C_i is then passed through the proposed Hierarchical Feature Subtraction Module (HFSM) to obtain intermediate feature M_i in a top-down manner. Next, M_i is further

fused through convolution with a kernel size of 3 to obtain fused feature P_i. Finally, P_i is sent to the proposed Feature Fusion Guidance Attention (FFGA) to obtain focused feature O_i, which are particularly focused on effective information, especially small objects. The focused feature O_i will be used by the model to predict the category and location of objects.

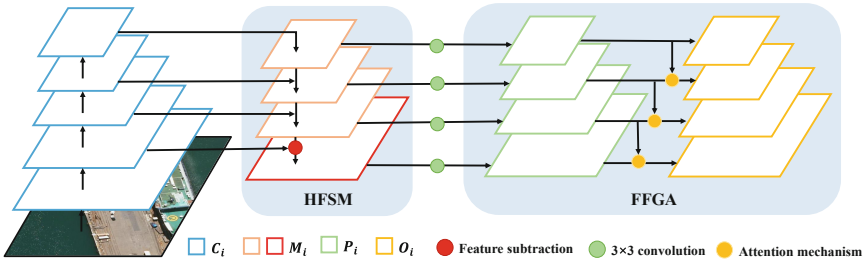

Fig. 2. Overview of the proposed HFFPN, which consists of HFSM and FFGA.

3.2 Hierarchical Feature Subtraction Module

The Hierarchical Feature Subtraction Module (HFSM) is designed to enhance the specific details of low-level features in the feature pyramid. Generally, features at the bottom of pyramid have higher resolution and smaller receptive fields, and contain local information such as edges, textures, and colors, which are crucial for detecting small objects. However, the widely used fusion strategy, *i.e.*, element-wise addition, fails to enhance the local information due to its uniqueness at each level. To cope with it, we propose HFSM that adopts the subtraction operation with hierarchy to highlight the local information, thereby alleviating the above-mentioned problem. The specific process of HFSM is as follows.

Firstly, the input image I passes through the backbone network to obtain the image feature C_i:

$$C_i = \begin{cases} I, & i = 0, \\ \mathcal{F}(C_{i-1}), & i = 1, \ldots, t \end{cases}, \tag{1}$$

where $\mathcal{F}(\cdot)$ denotes the convolution block in backbone and t is the number of feature layers.

Secondly, C_i is then processed by HFSM to obtain the intermediate feature M_i. The proposed HFSM aims to better extract detailed information from different feature levels. The subtraction operation can capture the differential information between two feature levels, which often includes fine-grained or edge information, crucial for detecting small objects. Afterwards, the intermediate

features are further fused through a 3×3 convolutional layer. These processes can be represented by the following equations:

$$M_i = \begin{cases} \sigma(C_i), & i = t, \\ \sigma(C_i) \oplus \texttt{UP}(M_{i+1}), & i = l+1, \ldots, t-1\ , \\ \frac{1}{2}(\sigma(C_i) \oplus \texttt{UP}(M_{i+1})) \oplus |\sigma(C_i) \ominus \texttt{UP}(M_{i+1})|\ , & i = 2, \ldots, l, \end{cases} \quad (2)$$

$$P_i = \texttt{conv}_{3\times3}(M_i). \quad (3)$$

where $\sigma(\cdot)$ denotes a 1×1 convolution, and $\texttt{UP}(\cdot)$ represents upsampling with ratio of 2. $\oplus$ and $\ominus$ denote element-wise addition and element-wise subtraction, respectively. $|\cdot|$ indicates the operation of taking absolute values. l is a hyperparameter for hierarchical strategy.

3.3 Feature Fusion Guidance Attention

Feature Fusion Guidance Attention (FFGA) is a generalized self-attention mechanism that can effectively focus on useful information, especially small object information. In the feature pyramid, the fused features contain multi-scale information from different levels, and adjacent levels have stronger complementary abilities in feature distribution due to their similar receptive fields. Based on the features between adjacent levels, self-attention is designed to guide the current level of features to focus on useful parts, which can effectively improve the quality of each feature layer and thus improve detection performance. Specifically, the process of FFGA guiding feature focusing can be expressed as Fig. 3.

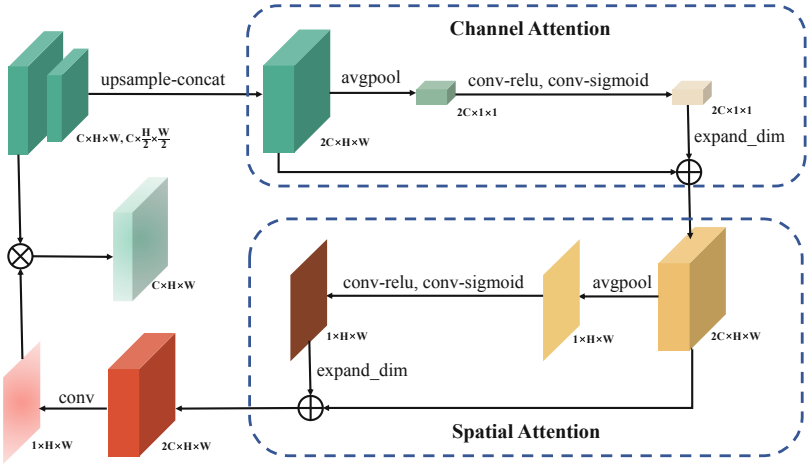

Fig. 3. Diagram of the FFGA.

Firstly, the input of FFGA are the current layer feature $P_i \in \mathbb{R}^{C\times H\times W}$ and the up one $P_{i+1} \in \mathbb{R}^{C\times \frac{H}{2}\times \frac{W}{2}}$. These two features are concatenated along the

channel dimension to obtain the guided feature $F_g \in \mathbb{R}^{2C\times H\times W}$. F_g is then sequentially fed into the channel attention (CA) and spatial attention (SA) modules, and we obtain the attention feature $F_a \in \mathbb{R}^{2C\times H\times W}$. Afterwards, F_a is passed through a 1×1 convolution to generate the attention map $W_a \in \mathbb{R}^{1\times H\times W}$. This map is multiplied as attention weight with the current layer feature P_i to obtain the focused feature $O_i \in \mathbb{R}^{C\times H\times W}$ after attention guidance. This process can be represented by the following formulas:

$$F_g = \texttt{concat}(P_i, \texttt{UP}(P_{i+1})), \tag{4}$$

$$F_a = SA(CA(F_g)), \tag{5}$$

$$W_a = \texttt{conv}_{1\times1}(F_a), \tag{6}$$

$$O_i = \begin{cases} P_i \otimes W_a, & i = 2, \ldots, t-1, \\ P_i, & i = t, \end{cases} \tag{7}$$

where the composition of channel attention and spatial attention has been detailed in Fig. 3. They have a similar structure that mainly consists of an average pooling layer, a 1×1 convolution layer followed by a ReLU activation, and a 1×1 convolution layer followed by a sigmoid activation. The input feature generates attention focusing on channel and spatial dimensions in the two modules respectively. After dimension expansion, they are element-wise added to the original feature, allowing the original feature to obtain a different degree of attentional gain in the channel and spatial dimensions.

4 Experiments

4.1 Datasets

DOTA [32] is a rotation-based small object dataset in the remote sensing field. It contains $2,806$ images with a total of $188,282$ instances. The detection targets in DOTA include 15 common categories in remote sensing images, namely Bridge (BR), Harbor (HA), Ship (SH), Plane (PL), Helicopter (HC), Small vehicle (SV), Large vehicle (LV), Baseball diamond (BD), Ground track field (GTF), Tennis court (TC), Basketball court (BC), Soccer-ball field (SBF), Roundabout (RA), Swimming pool (SP), and Storage tank (ST). **COCO** [18] is the most popular dataset for object detection. Due to its definition of small object and specialized evaluation metric mAP_s, COCO is commonly used as a well recognized benchmark for small object detection.

4.2 Experiment Settings

We employed Resnet50 and Resnet101 [11] pre-trained on ImageNet [25] as backbone networks. We utilized the SGD algorithm with a momentum of 0.9 and a weight decay of 0.0001 for network optimization. The initial learning rate warms up at a rate of 0.001 per iteration for the first 500 iterations. The training schedule for all experiments was consistent. We trained 12 epochs on the two datasets, and the learning rate decays at the epoch 8 and 11 with ratio of 0.1. The code for all experiments was built on the MMdetection [2] platform.

4.3 Comparison Results

Results on DOTA. We selected the RoI Transformer [5], a general method for aerial object detection, as the baseline algorithm. Table 1 reports the comparison result on DOTA test set. With Resnet50 as the backbone, our method obtains 76.64% mAP_{50}, improving the performance of baseline by approximately 1%, thereby surpassing the performance of the state-of-the-art algorithms. With Resnet101 as the backbone, HFFPN also increases the baseline's performance by 0.87% mAP_{50}, achieving the best result on the DOTA dataset. These results fully demonstrate HFFPN's advantages on small object detection and reflect its potential applications. Figure 4 provides a more intuitive visual comparison.

Table 1. Comparison with state-of-the-art methods on DOTA test set. The reported results come from AerialDetection [6] and OBBDetection [33]. ‡ indicates that it is the result of our re-implement. Note that we only list some classes for better display.

Methods	Backbone	GTF	SV	SH	SBF	HA	SP	mAP_{50}
Single-stage Methods								
RSDet [23]	R152-FPN	68.50	70.20	73.60	64.30	66.10	69.30	74.10
R^3Det [36]	R152-FPN	66.10	70.92	78.21	61.81	68.16	69.83	73.74
S^2A-Net [9]	R50-FPN	71.11	78.11	87.25	60.36	65.26	69.13	74.12
R^3Det-DCL [35]	R152-FPN	69.70	76.84	87.30	63.50	68.96	68.79	75.54
Two-stage Methods								
SCRDet [37]	R101-FPN	68.36	68.36	72.41	65.02	66.25	68.24	72.61
Gliding Vertex [34]	R101-FPN	77.34	73.01	86.82	59.55	72.94	70.86	75.02
ReDet [10]	ReR50-ReFPN	74.00	78.13	**88.04**	61.76	72.10	68.07	76.25
Oriented R-CNN [33]	R101-FPN	76.92	74.27	87.52	**65.51**	74.36	70.15	76.28
Anchor-free Methods								
PIoU‡ [3]	R50-FPN	68.90	77.58	81.57	60.47	57.68	65.12	69.68
O^2-DNet [29]	H-104	61.21	71.32	78.62	60.93	58.21	66.98	71.04
DRN [21]	H-104	64.10	76.22	85.84	57.65	69.30	69.63	73.23
CFA [8]	R101-FPN	67.17	**79.99**	84.46	54.86	73.04	70.24	75.05
Oriented RepPoints [13]	R101-FPN	71.76	**79.95**	87.33	59.15	75.23	**73.75**	76.52
RoI-Trans.‡ [5]	R50-FPN	76.65	78.40	87.55	60.12	74.89	69.70	75.70
RoI-Trans.‡	R101-FPN	75.75	78.11	87.46	63.80	**76.05**	71.35	76.02
RoI-Trans. (Ours)	R50-HFFPN	**78.11**	78.33	**87.71**	**65.24**	75.39	**72.99**	**76.64**
RoI-Trans. (Ours)	R101-HFFPN	**79.84**	77.97	87.68	65.00	**76.36**	71.67	**76.89**

Results on COCO. On the COCO dataset, we applied HFFPN to two-stage [24], one-stage [17], and anchor-free [26] detectors, respectively. Table 2 shows the performance gain brought by HFFPN. Although the overall mAP improvement is not significant due to the small proportion of small objects in the COCO dataset, the consistent and significant increase in the mAP_s metric indicates that HFFPN makes detectors more capable of detecting small objects while maintaining their detection capabilities of other scales of objects.

Table 2. Comparison experiment on COCO. The baseline results come from [2].

Methods	Backbone	mAP	mAP_{50}	mAP_{75}	mAP_s	mAP_m	mAP_l
Faster RCNN [24]	R50-FPN	37.4	58.1	40.4	21.2	41.0	48.1
Faster RCNN	R50-HFFPN	37.6	58.4	40.7	21.9 **(+0.7)**	40.9	48.3
RetinaNet [17]	R50-FPN	36.5	55.4	39.1	20.4	40.3	48.1
RetinaNet	R50-HFFPN	36.6	55.7	39.1	21.2 **(+0.8)**	40.3	48.0
FCOS [26]	R50-FPN	36.6	56.0	38.8	21.0	40.6	47.0
FCOS	R50-HFFPN	36.6	55.9	38.7	21.7 **(+0.7)**	40.2	47.2

4.4 Ablation Study

To further verify the advantages and effectiveness of the proposed method, we conduct a series of experiments on the DOTA dataset. The baseline algorithm is RoI Transformer with Resnet50.

Evaluation for Component Effectiveness. To evaluate the effects of HFSM and FFGA, we carry out several ablation experiments, and the experimental results are shown in the Table 3. Without any improvement schemes, the mAP_{50} detected by the baseline is 75.70%. The introduction of HFSM and FFGA gradually improves the detection accuracy to 76.24% and 76.64%. The results indicate that each combination in HFFPN brings improvement to the detector.

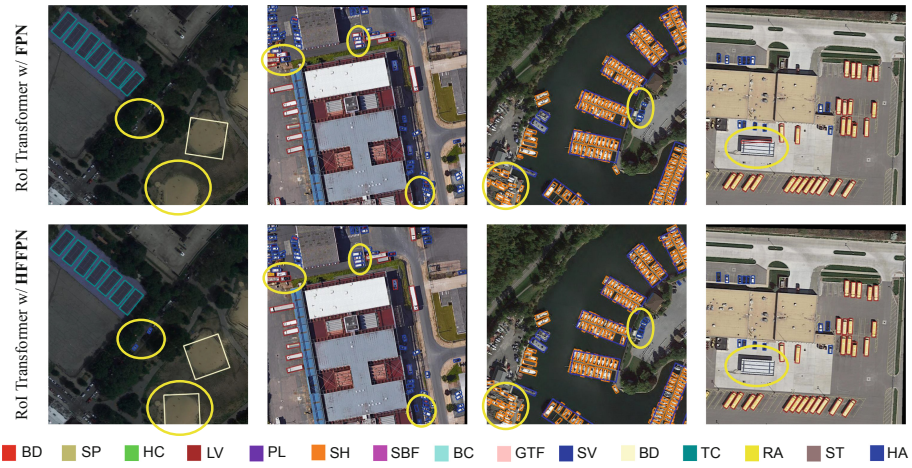

Fig. 4. Visualization on DOTA test set. The yellow circles highlight the difference of detection result. We can easily find that HFFPN (second row) can help detect more small objects and achieve higher accuracy in classification and regression. (Color figure online)

Evaluation on Different Settings of l in HFSM. Hierarchical level l, as a hyperparameter in HFSM, determines in which feature layers the operation of feature subtraction is performed. Specifically, the feature subtraction module will be introduced when the level lower than l. Table 4 shows the results under different values of l. When l is 2, the performance of baseline with HFFPN reaches the highest. Assuming we do not employ the hierarchical strategy by setting l equals to 5, where feature subtraction is performed between each level of features, we would observe a significant drop in results. The hierarchical strategy ensures that the subtraction is performed only on detailed features, making it applicable to a wide range of input images and thus enhancing the model's robustness.

Comparison with Other FPNs. Table 5 presents performance of the baseline algorithm with different FPNs. It can be observed that some expanded FPNs do enhance the detector's performance to some extent, but the improvements are not as significant as those of the proposed HFFPN.

Evaluation on Different Detectors. To verify that the proposed HFFPN is a common method for most detectors, experiments were conducted on several different detectors. Table 6 shows the comparison results of these detectors with or without using HFFPN. The experimental results show that the use of HFFPN has led to performance improvements for all detectors, strongly indicating the universality and effectiveness of the proposed method.

Table 3. Evaluation on the effectiveness of each component. FS, HS, CA and SA denote feature subtraction, hierarchical strategy, channel attention, and spatial attention, respectively.

Baseline	FS	HS	CA	SA	mAP_{50}
✓					75.70
✓	✓				76.13
✓	✓	✓			76.24
✓	✓	✓	✓		76.38
✓	✓	✓		✓	76.41
✓	✓	✓	✓	✓	**76.64**

Table 4. Results of differnent l.

l	2	3	4	5
mAP_{50}	**76.64**	76.33	76.07	75.59

Table 5. Comparison with other FPNs.

Backbone	mAP_{50}	mAP
R50-FPN [16]	75.70	46.27
R50-PAFPN [19]	76.26	46.56
R50-HRFPN [27]	76.33	46.85
R50-NASFPN [7]	73.71	45.03
R50-HFFPN	**76.64**	**47.07**

Table 6. Improvements on DOTA by applying HFFPN to different detectors.

Methods	Backbone	mAP_{50}	mAP
PIoU [3]	R-50-FPN	69.68	40.05
PIoU	R-50-HFFPN	70.26 **(+0.58)**	40.55 **(+0.50)**
Gliding Vertex [34]	R-50-FPN	72.65	40.93
Gliding Vertex	R-50-HFFPN	73.24 **(+0.59)**	41.42 **(+0.49)**
Oriented RCNN [33]	R-50-FPN	75.72	46.78
Oriented RCNN	R-50-HFFPN	76.14 **(+0.42)**	46.85 **(+0.07)**
RoI-Trans. [5]	R-50-FPN	75.70	46.27
RoI-Trans.	R-50-HFFPN	76.64 **(+0.94)**	47.07 **(+0.80)**

5 Conclusion

To better utilize the detailed information for small object detection, this paper proposes a hierarchical focused feature pyramid network. It mainly contains a hierarchical feature subtraction module and feature fusion guidance attention. This design overcomes the problem of neglecting edge information that exists in common FPN methods, thus improving the detection ability of small objects without affecting the detection performance of objects at other scales. Comparison and ablation experiments on multiple datasets demonstrate the excellent performance of the proposed method, fully verifying the effectiveness of HFFPN.

Acknowledgements. This work was supported by National Key R&D Program of China (No. 2022ZD0118201), the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No. U21B2037, No. U22B2051, No. 62176222, No. 62176223, No. 62176226, No. 62072386, No. 62072387, No. 62072389, No. 62002305 and No. 62272401), and the Natural Science Foundation of Fujian Province of China (No. 2021J01002, No. 2022J06001).

References

1. Bashir, S.M.A., Wang, Y.: Small object detection in remote sensing images with residual feature aggregation-based super-resolution and object detector network. Remote Sens. **13**(9), 1854 (2021)
2. Chen, K., et al.: Mmdetection: open MMLAB detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
3. Chen, Z., et al.: PIoU loss: towards accurate oriented object detection in complex environments. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12350, pp. 195–211. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58558-7_12
4. Cheng, G., Yuan, X., Yao, X., Yan, K., Zeng, Q., Han, J.: Towards large-scale small object detection: survey and benchmarks. arXiv preprint arXiv:2207.14096 (2022)

5. Ding, J., Xue, N., Long, Y., Xia, G.S., Lu, Q.: Learning ROI transformer for oriented object detection in aerial images. In: CVPR, pp. 2849–2858 (2019)
6. Ding, J., et al.: Object detection in aerial images: a large-scale benchmark and challenges. TPAMI **44**(11), 7778–7796 (2021)
7. Ghiasi, G., Lin, T.Y., Le, Q.V.: NAS-FPN: learning scalable feature pyramid architecture for object detection. In: CVPR, pp. 7036–7045 (2019)
8. Guo, Z., Liu, C., Zhang, X., Jiao, J., Ji, X., Ye, Q.: Beyond bounding-box: convex-hull feature adaptation for oriented and densely packed object detection. In: CVPR, pp. 8792–8801 (2021)
9. Han, J., Ding, J., Li, J., Xia, G.S.: Align deep features for oriented object detection. IEEE Trans. Geosci. Remote Sens. **60**, 1–11 (2021)
10. Han, J., Ding, J., Xue, N., Xia, G.S.: Redet: a rotation-equivariant detector for aerial object detection. In: CVPR, pp. 2786–2795 (2021)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
12. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR, pp. 7132–7141 (2018)
13. Li, W., Chen, Y., Hu, K., Zhu, J.: Oriented reppoints for aerial object detection. In: CVPR, pp. 1829–1838 (2022)
14. Li, Y., Huang, Q., Pei, X., Chen, Y., Jiao, L., Shang, R.: Cross-layer attention network for small object detection in remote sensing imagery. IEEE J. Select. Top. Appl. Earth Observ. Remote Sens. **14**, 2148–2161 (2020)
15. Liao, M., Zou, Z., Wan, Z., Yao, C., Bai, X.: Real-time scene text detection with differentiable binarization and adaptive scale fusion. TPAMI **45**(1), 919–931 (2022)
16. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR, pp. 2117–2125 (2017)
17. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV, pp. 2980–2988 (2017)
18. Lin, T.Y., et al.: 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
19. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: CVPR, pp. 8759–8768 (2018)
20. Liu, Y., Sun, P., Wergeles, N., Shang, Y.: A survey and performance evaluation of deep learning methods for small object detection. Expert Syst. Appl. **172**, 114602 (2021)
21. Pan, X., et al.: Dynamic refinement network for oriented and densely packed object detection. In: CVPR, pp. 11207–11216 (2020)
22. Pi, Y., Nath, N.D., Behzadan, A.H.: Convolutional neural networks for object detection in aerial imagery for disaster response and recovery. Adv. Eng. Inform. **43**, 101009 (2020)
23. Qian, W., Yang, X., Peng, S., Yan, J., Guo, Y.: Learning modulated loss for rotated object detection. Proc. AAAI Conf. Artif. Intell. **35**(3), 2458–2466 (2021)
24. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: towards real-time object detection with region proposal networks. NeurIPS **28** (2015)
25. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
26. Tian, Z., Shen, C., Chen, H., He, T.: FCOS: fully convolutional one-stage object detection. In: ICCV, pp. 9627–9636 (2019)
27. Wang, J., et al.: Deep high-resolution representation learning for visual recognition. TPAMI **43**(10), 3349–3364 (2020)

28. Wang, L., Tong, Z., Ji, B., Wu, G.: TDN: temporal difference networks for efficient action recognition. In: CVPR, pp. 1895–1904 (2021)
29. Wei, H., Zhang, Y., Chang, Z., Li, H., Wang, H., Sun, X.: Oriented objects as pairs of middle lines. ISPRS J. Photogramm. Remote. Sens. **169**, 268–279 (2020)
30. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: convolutional block attention module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 3–19. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_1
31. Wu, J., Zhou, C., Zhang, Q., Yang, M., Yuan, J.: Self-mimic learning for small-scale pedestrian detection. In: ACMMM, pp. 2012–2020 (2020)
32. Xia, G.S., et al.: Dota: a large-scale dataset for object detection in aerial images. In: CVPR, pp. 3974–3983 (2018)
33. Xie, X., Cheng, G., Wang, J., Yao, X., Han, J.: Oriented r-cnn for object detection. In: ICCV, pp. 3520–3529 (2021)
34. Xu, Y., et al.: Gliding vertex on the horizontal bounding box for multi-oriented object detection. TPAMI **43**(4), 1452–1459 (2020)
35. Yang, X., Hou, L., Zhou, Y., Wang, W., Yan, J.: Dense label encoding for boundary discontinuity free rotation detection. In: CVPR, pp. 15819–15829 (2021)
36. Yang, X., Yan, J., Feng, Z., He, T.: R3det: refined single-stage detector with feature refinement for rotating object. In: AAAI, vol. 35, pp. 3163–3171 (2021)
37. Yang, X., et al.: Scrdet: towards more robust detection for small, cluttered and rotated objects. In: ICCV, pp. 8232–8241 (2019)
38. Zhang, M., Yue, K., Zhang, J., Li, Y., Gao, X.: Exploring feature compensation and cross-level correlation for infrared small target detection. In: ACMMM, pp. 1857–1865 (2022)