ICIC Express Letters
Volume **18**, Number **2**, February **2024**
 ISSN 1881-803X
pp. **125–134**

IMPROVING SMALL OBJECT DETECTION IN REMOTE SENSING IMAGES USING EXTENDED FEATURE PYRAMID NETWORK

Hoanh Nguyen

Faculty of Electrical Engineering Technology
Industrial University of Ho Chi Minh City
12 Nguyen Van Bao, Ward 4, Go Vap District, Ho Chi Minh City 700000, Vietnam
nguyenhoanh@iuh.edu.vn

Received March 2023; accepted June 2023

Abstract. *Detecting small objects in remote sensing images poses a challenge because of the restricted resolution and the inconsistent sizes of the objects. This paper proposes an enhanced approach for detecting small objects in remote sensing images, utilizing an Extended Feature Pyramid Network (EFPN). By providing a high-resolution feature layer that incorporates multi-scale features and contextual information, EFPN combines the advantages of both Feature Pyramid Networks (FPN) and a top-down pathway, which enhances the representation of small objects and thus improving the accuracy of detecting tiny objects in remote sensing images. Additionally, this paper introduces an attention mechanism to further suppress noise from other layers and refine the feature representation of small instances. The proposed method is evaluated on a publicly remote sensing dataset. The results of the experiment demonstrate that EFPN considerably enhances the detection accuracy of small objects, attaining the best outcomes with an Average Precision (AP) of 64.4% on the test dataset. The proposed method can potentially benefit various remote sensing applications, such as intelligent monitoring, agriculture management, and urban planning.*
Keywords: Object detection, Deep learning, Remote sensing images, Feature pyramid network, Convolutional neural networks

1. **Introduction.** Object detection in images obtained through remote sensing has become a crucial research topic owing to its diverse applications in different domains such as agriculture, national security, urban planning, and environmental monitoring [1,2]. Remote sensing images provide a unique perspective for object detection by capturing large areas of the earth's surface at different spatial and spectral resolutions. However, detecting objects in images from remote sensing is a difficult task due to various factors such as the limited resolution of sensors, the varying size and shape of objects, and the presence of occlusions and clutter. Object detection in remote sensing imagery has been traditionally performed using manual or semi-manual methods, which are time-consuming and require expert knowledge. Therefore, there is a need for automated and accurate object detection methods in remote sensing imagery to enhance the efficiency and accuracy of various applications. The latest progressions in deep learning methodologies, such as Convolutional Neural Networks (CNNs), have shown promising results in improving the accuracy of generic object detection. Although generic object detection methods based on CNNs have achieved remarkable advancements in natural images, their direct application to remote sensing images yields unsatisfactory results. Remote sensing images are obtained from a top-down perspective using satellites or airborne cameras at high altitudes, which is different from ground-based cameras that capture natural images from a landscape perspective. The unique perspective of remote sensing imaging presents several challenges for

DOI: 10.24507/icicel.18.02.125

object detection, including significant variations in scale, intricate backgrounds, diminutive objects, and crowded distributions. To overcome these crucial challenges, numerous approaches have been proposed by researchers. In [3], the authors introduced a technique called RoI Transformer, which addresses the issue of inaccurate object classification and localization caused by misalignment. This approach utilizes spatial transformations to Regions of Interest (RoIs) and trains the parameters of the transformations through oriented box annotations. RoI Transformer can be readily integrated into detectors for detecting oriented objects. A self-reinforced network called remote sensing region-based CNN was introduced by Pang et al. [4]. The network comprises three components: a backbone network, an attention module, and a classifier and detector. The lightweight residual structure of the backbone network facilitates rapid and robust feature extraction from input images. The attention module, built on top of the backbone network, helps to minimize false positive candidates. The classifier is employed to anticipate the presence of the target in each image segment, followed by the detector to accurately detect them, if possible. An alternative technique for detecting aircraft in remote sensing imagery is to employ a bottom-up approach [5]. The method characterizes the aircraft detection challenge as a prediction and clustering task for intersecting line segments in pairs within each target. Detecting aircraft then becomes a matter of estimating the appearance-based line segments, without the need for classification of rectangular regions or implicit feature learning. To enhance the detection performance of small instances by acquiring more robust features, Li et al. [6] proposed a feature pyramid with both down-sampling and up-sampling capabilities. By integrating deep and shallow features in both directions, along with skip connections, this pyramid allows for a more comprehensive contextual understanding. In [21], the authors proposed an Extended Feature Pyramid Network (EFPN) to improve small object detection in existing feature pyramid networks. EFPN incorporates an extra high-resolution pyramid level and a novel Feature Texture Transfer (FTT) module to enhance feature resolution and extract regional details. It also utilizes a cross-resolution distillation mechanism and a foreground-background-balanced loss function to address scale-level and area imbalances. To achieve highly accurate detection of ships in remote sensing images, Yu et al. [7] introduced a cascaded rotating anchor-aided network based on RetinaNet. A hybrid method that performs aircraft detection in SAR images by utilizing information enhancement through scattering and an attention pyramid network was recently developed by Guo et al. [8]. To maintain semantic balance across multiple features on different layers, the Feature Balancing and Refinement Network (FBR-Net) [9] developed a pyramid with attention-guided balancing. To enhance the information flow, Squeeze Excitation Skip-Connection Path Networks (SESPNets) [10] employed a route-level skip-connection architecture and a channel feature calibration attention module. By modifying the input data distribution through a hard sample mining network, Zhang et al. [11] were able to effectively enable the network to learn the properties of difficult samples.

Despite some advancements in object detection in remote sensing imagery by existing techniques, they have overlooked several problems. One of these is the lack of use of global context, which has the potential to improve object feature representation, helping with object classification and localization. Additionally, these methods avoid overburdening the CNN backbone, but they only rely on the low-resolution feature map to fill in the missing details, disregarding the credible information stored in other features of the backbones. As a result, this leads to spurious textures and artifacts on the features of the CNN, which in turn leads to false positives.

Based on the above observations, this paper proposes the Extended Feature Pyramid Network (EFPN), which uses large-scale feature maps with rich location details to separate small object detection. EFPN expands on the original FPN with a high-resolution feature level designed specifically for detecting small objects. To enhance context-guided detection

of small objects, a global context enhancement block is designed to combine high-level and low-level contexts. Additionally, an attention mechanism is proposed to suppress noise from other layers and refine the feature information of small objects, enabling effective capture of small objects in complex scenarios. The main contributions of this paper can be summarized as the following.

- This paper introduces an extended feature layer that incorporates a high-resolution feature map to provide rich semantic information for small objects.
- This paper designs a global context enhancement module and an attention module, which capture strong semantic information and accurate location to effectively detect small instances in complex scenarios.
- The proposed designs are integrated into an anchor-free object detector that directly generates predictions of objects without relying on predefined anchor boxes as references. This eliminates the limitations of anchor-based object detection pipelines.
- The results of the experiments demonstrate that the proposed model improves the detection accuracy of small objects, which can potentially benefit various remote sensing applications, such as intelligent monitoring, agriculture management, and urban planning.

This paper follows the following structure: Section 2 presents a detailed explanation of the proposed model, Section 3 covers the experimental details and results, and lastly, Section 4 presents the conclusions that were drawn.

2. Methodology.

2.1. Overview of the proposed model. The proposed framework's overall architecture is depicted in Figure 1 and is based on the Fully Convolutional One-Stage object detector (FCOS) model [12]. It comprises three modules: a backbone model for extracting features from the input image, an extended feature pyramid architecture for feature fusion, and an anchor-free detection head for object predictions. To extract the features from the input image, this paper uses ResNet-50 model [13] that has been pre-trained on ImageNet [14]. For the feature pyramid subnet, the paper selects four output stages of the backbone network, namely $\{C_2, C_3, C_4, C_5\}$, instead of $\{C_3, C_4, C_5\}$ and two additional feature maps $\{P_6, P_7\}$ used in [12,15]. This choice is made to enable the model to

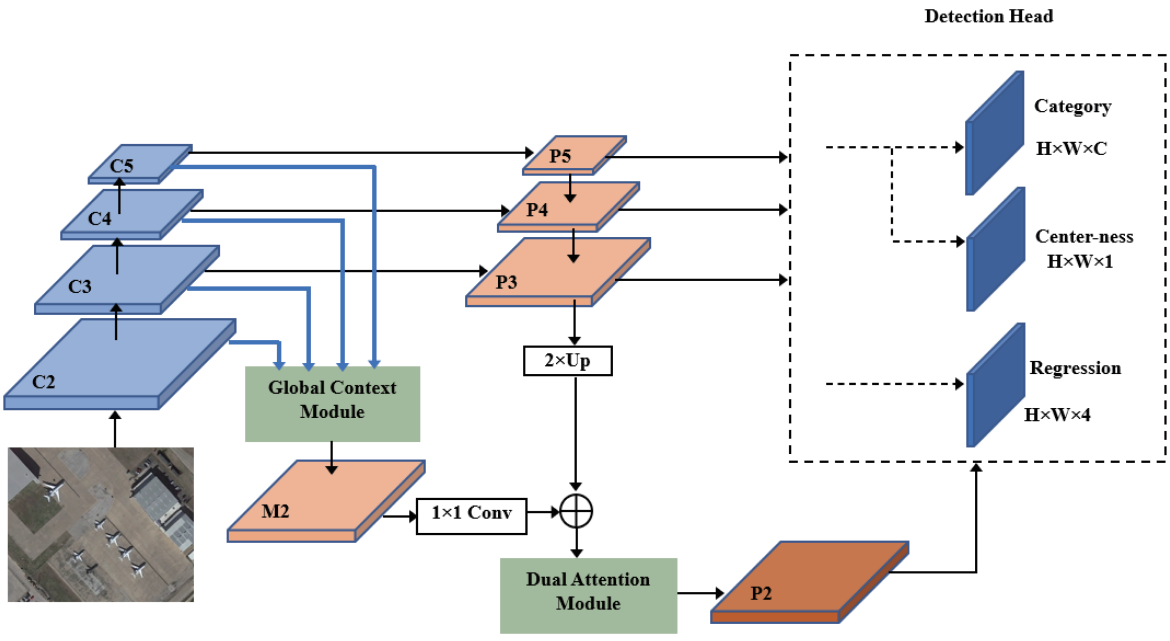

FIGURE 1. Overall flowchart of the proposed model

better detect smaller objects. Furthermore, the paper designs an extended feature layer that specializes in small objects and incorporates a high-resolution feature map to provide rich localization information. To improve the semantic information of the extended layer, the paper introduces a global context enhancement module and an attention module, which capture strong semantic information and accurate location to effectively detect small instances in complex scenarios. Finally, the paper employs a shared detection head for different feature layers, which consists of three branches: category, regression, and center-ness. These branches are used to predict object scores, object locations, and remove low-quality detected results, respectively.

In summary, the proposed model is an anchor-free object detector that directly generates predictions of objects without relying on predefined anchor boxes as references. This eliminates the limitations of anchor-based object detection pipelines. Additionally, the feature pyramid network in the proposed design separates the detection of instances with different sizes and assigns a more appropriate feature level for small instances. This improves the detection accuracy of object detection in images from remote sensing.

2.2. **Extended feature pyramid for detecting small objects.** Most contemporary methods for object detection use multi-level architecture with FPN [16] to fuse features of different semantic representations and resolutions through a top-down pathway and lateral connections. However, these methods mainly rely on high-level feature maps for object detection, which may cause spatial information degradation in low-resolution maps. The loss of location information and texture in high-level feature maps can result in poor performance in locating small instances in remote sensing images. To address this issue, this paper proposes an extension subnetwork to the vanilla feature pyramid network with an additional feature level for small object detection that contains more location information. First, a global context enhancement module is developed to generate a high-resolution feature map at the bottom of the top-down pathway, which extracts more useful location details of small objects. Additionally, an attention module, which consists of a channel attention block and a location attention block, is introduced to refine object information with strong semantics and accurate positioning while suppressing noise from other feature layers. The details of these modules will be explained in the next subsections.

2.2.1. *Global context enhancement module.* Figure 2 displays the structure of the global context enhancement module, which is designed to integrate semantic and detailed features from feature maps C_2 to C_5. The output layers of the ResNet-50 model consist of C_2, C_3, C_4, and C_5, where high-level feature maps (C_5 and C_4) are smaller in resolution but more beneficial for determining object categories due to their more semantic features. In contrast, low-level feature maps (C_3 and C_2) are reasonably large in resolution and more conducive to object localization due to their highly detailed features. The proposed method aggregates feature layers C_2 to C_5 to obtain global context information in a high-resolution feature layer. To fuse feature maps of different sizes, high-level feature maps are first processed using upsampling and a 1×1 convolution layer for reducing dimension. Specifically, deformable convolution with a 3×3 kernel [17] followed by a transposed convolution layer [18] is applied on C_5, C_4, and C_3 to upsampling the feature maps to the size of C_2. Deformable convolution is used to dynamically adjust the receptive field based on object scales and alleviate alignment issues. For C_2, one 3×3 deformable convolution layer is used since it is not subjected to upsampling operations. Finally, elementwise summation is used to fuse different feature layers, resulting in M_2 with dimension 256. Mathematically, operations of the global context enhancement module can be described as

$$M_2 = Conv(C_2) + Up(Conv(C_3)) + Up(Conv(C_4)) + Up(Conv(C_5)) \tag{1}$$

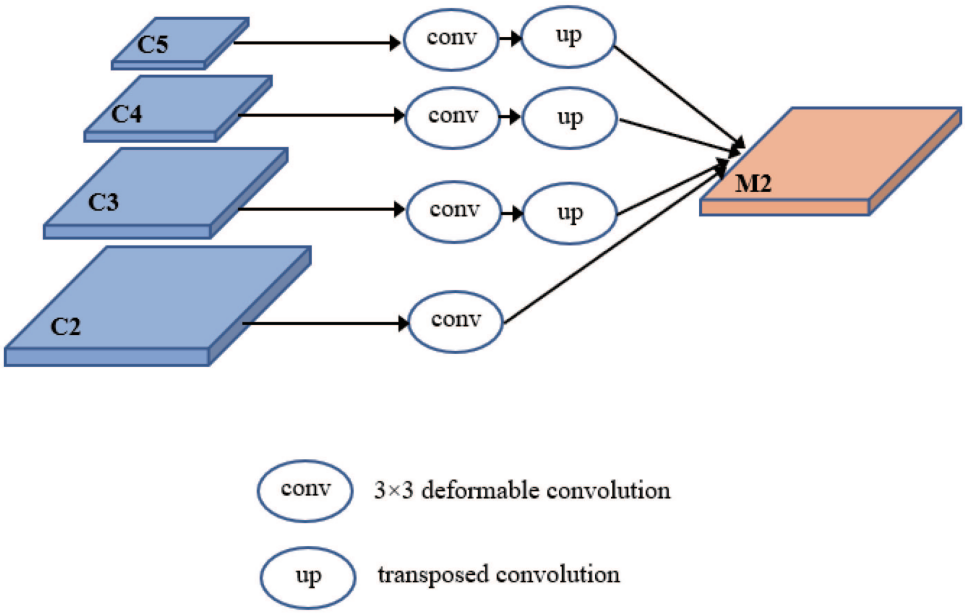

FIGURE 2. Architecture and details of the global context enhancement module

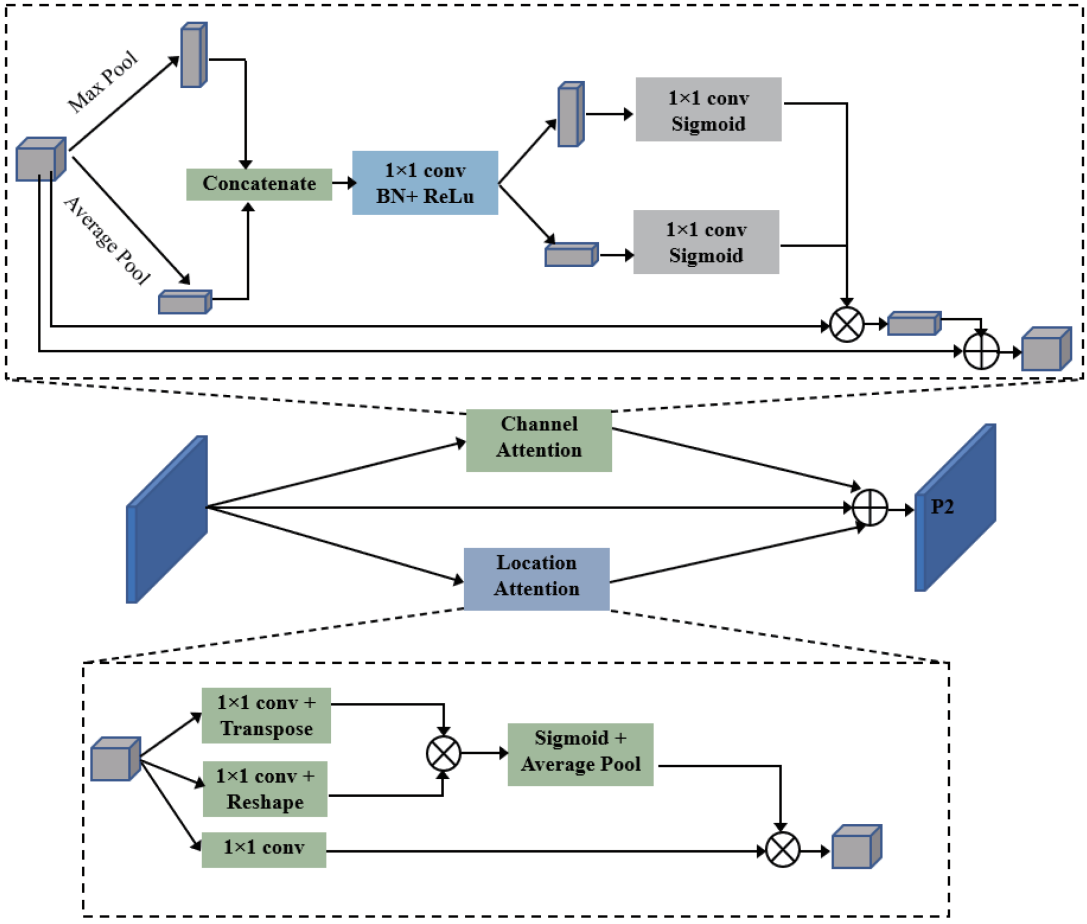

FIGURE 3. Network architecture of the attention module

2.2.2. *Attention module.* The architecture of the attention module is presented in Figure 3 and consists of two blocks: channel attention block and location attention block. In complex scenes, these branches effectively identify small objects by capturing precise positions and strong semantic dependencies. In the channel attention block, although

average pooling is typically utilized to incorporate global spatial information for channel attention, it fails to preserve spatial information, which is critical for detecting spatial patterns in object detection.

To overcome this constraint, this article proposes encoding operations that work along a single dimension to enable the attention subnetwork to capture distant spatial relationships with precise location details. Specifically, the channel attention block starts by combining spatial information from a feature map using average pooling and max pooling operations. This results in the creation of two distinct spatial context descriptors for each channel, with one descriptor capturing information along the vertical coordinate and the other along the horizontal coordinate. Both descriptors are then forwarded to a network consisting of convolutional layers followed by ReLU activation function, and a sigmoid function to produce pair of attention maps with the same channel numbers to capture channel correlations. By extending the output feature maps and leveraging them as attention weights, a channel attention map with greater prominence can be obtained. Finally, the paper combines the extracted feature maps with the attention weights through element-wise multiplication operations to generate output features.

To enhance representation capabilities, the location attention branch starts by converting the input features into latent spaces using 1×1 convolutions. To compute the location attention map, matrix multiplication is performed between the transformed feature maps to generate an efficient feature descriptor. The location attention map is calculated using average pooling and sigmoid activation function. Finally, the element-wise multiplication of the latent space and location attention map results in the final output. This branch is effective in modeling the context relationships of local features.

3. **Results and Discussion.**

3.1. **Dataset and metrics.** The DOTA [19] dataset is a comprehensive collection of aerial images that is ideal for object detection purposes. This dataset contains 2806 images that have been obtained from different devices and platforms. The images vary in size, with dimensions ranging from 800 × 800 to 20,000 × 20,000 pixels. The objects in the images have diverse scales, shapes, and orientations. Since this paper focuses on detecting small objects in remote sensing imagery, DOTA-v1.5 is used to evaluate the detection performance. DOTA-v1.5 contains 403,318 instances, which is 215,036 more than DOTA-v1.0. The majority of these extra examples are small in size, which makes them appropriate for testing the proposed model. It should be noted that the DOTA-v1.5 dataset contains 16 object classes.

For evaluation metrics, the Average Precision (AP) is used for evaluating detection performance. AP is defined as follows:

$$AP = \int_0^1 p(r)dr \tag{2}$$

where r represents the recall rate; $p(r)$ represents the corresponding precision value; r and $p(r)$ are calculated as follows:

$$r = \frac{TP}{TP + FN} \tag{3}$$

$$p(r) = \frac{TP}{TP + FP} \tag{4}$$

where quantities TP (True Positive samples), FP (False Positive samples), and FN (False Negative samples) are defined based on different IoU matching cases.

3.2. **Results and discussion.** In this study, the proposed model is evaluated against various existing object detection methods, including one-stage methods such as Single Shot MultiBox Detector (SSD) [20], RetinaNet [15], and YOLOv4 [25], and two-stage methods such as Faster R-CNN [22], FPN [16], PANet [23], and Context-Driven Detection Network (CDD-Net) [24]. Table 1 displays the detection results for the proposed model and all comparison models. Overall, the proposed model achieves a 64.4% mAP for all object classes in the DOTA dataset. The results in Table 1 indicate that the proposed model attains a substantial improvement of 30.9% and 28.1% mAP compared to RetinaNet and SSD, respectively. The proposed model surpasses both these methods in terms of detection accuracy across all classes. Moreover, the proposed model shows an 8.8% improvement in mAP compared to YOLOv4, highlighting its superior performance compared to one-stage object detection approaches. Regarding two-stage approaches, it is evident that most of them outperform one-stage methods significantly. PANet, for instance, boosts mAP by 5.6% compared to YOLOv4. This is because of the region proposal network's ability to eliminate most false positive candidates in two-stage approaches. When compared to two-stage methods, the proposed model performs better, achieving a 12.6% and 7.1% improvement compared to Faster R-CNN and FPN, respectively. Additionally, the proposed model enhances mAP over PANet and CDD-Net by 3.2% and 3.1%, respectively, and outperforms them in most of categories. This demonstrates the proposed model's suitability for remote sensing object detection. Furthermore, as observed from Table 1, the proposed model exhibits the best improvement in detecting ships, small vehicles, and

Table 1. Comparisons of the proposed model with previous approaches on the DOTA dataset (%)

	One-stage methods			Two-stage methods				Proposed model
	SSD	RetinaNet	YOLOv4	FPN	Faster R-CNN	PANet	CDD-Net	
plane	78.1	76.0	85.2	78.6	70.0	85.9	81.4	82.6
ship	30.6	33.4	79.5	40.3	32.5	58.4	49.2	79.8
storage tank	43.7	31.2	64.8	46.4	36.9	61.3	53.3	51.2
baseball diamond	38.6	44.5	61.7	70.1	63.4	74.1	74.7	74.1
tennis court	81.7	75.1	88.3	86.0	81.3	89.6	89.8	86.2
basketball court	42.0	30.8	55.6	69.4	68.4	67.0	71.4	69.4
ground track field	27.4	32.5	35.2	68.5	63.3	64.5	70.1	71.2
harbor	46.9	35.8	69.8	59.5	59.1	67.9	69.9	67.7
bridge	14.5	32.6	32.6	55.1	55.5	51.5	55.3	54.3
large vehicle	39.5	33.3	64.0	45.4	44.5	56.2	51.5	56.5
small vehicle	13.5	10.7	37.0	23.7	22.7	27.7	25.3	40.1
helicopter	30.4	0.2	67.6	68.3	51.5	71.3	71.3	72.2
roundabout	33.6	42.4	54.2	56.2	49.8	59.2	58.2	56.8
soccer ball field	20.7	13.0	34.4	61.1	57.6	63.4	65.6	64.0
swimming pool	39.3	43.9	58.5	64.5	55.1	73.4	60.4	68.2
container crane	0.0	0.0	0.7	24.4	16.9	7.6	32.7	35.4
mAP	36.3	33.5	55.6	57.3	51.8	61.2	61.3	64.4

other small instances. Using ship class as an example, the proposed model surpasses the best comparison model, YOLOv4, by 0.3%. This outcome indicates the effectiveness of the global context enhancement module in generating high-resolution feature maps to extract more useful location details of small objects, and the attention module in refining object information with strong semantics and accurate positioning. For large object instances, such as storage tanks and large vehicles, the proposed model achieves competitive results compared to other models. This result demonstrates the flexibility of the model in detecting objects of various sizes. The results of object detection produced by the proposed model on the DOTA-v1.5 dataset are visually presented in Figure 4. The figure indicates that the proposed method can detect multi-scale instances accurately in different scenarios with few false positive candidates. This suggests that the global context enhancement and attention features guide feature generation effectively and facilitate object detection, especially for small objects. However, occasional instances of undetected objects are observable in the proposed model due to the proximity between the objects, making it challenging to differentiate them from the background.

FIGURE 4. Visualization of detection results by the proposed method on the DOTA dataset

4. **Conclusions.** This study introduces an improved technique for detecting small instances in remote sensing imagery, which employs an Extended Feature Pyramid Network (EFPN) to separate small object detection using large-scale features with rich location details. EFPN expands on the original FPN by introducing a high-resolution level designed specifically for detecting small objects. To enhance context-guided detection of small objects, the proposed method includes a global context enhancement module that integrates low-level and high-level context information. Moreover, an attention mechanism is introduced to suppress noise from other layers and refine the feature representation of small objects, thus effectively capturing small objects in complex scenes. Results from experiments on the DOTA dataset demonstrate the outstanding performance and robustness of the proposed method for detecting small instances in remote sensing imagery. For future work, this paper will explore more efficient attention mechanisms to enable the model to capture both global and local contextual information effectively, thus improving the detection performance of objects with various sizes.

REFERENCES

[1] W. Han, J. Chen, L. Wang, R. Feng, F. Li, L. Wu, T. Tian and J. Yan, Methods for small, weak object detection in optical high-resolution remote sensing images: A survey of advances and challenges, *IEEE Geoscience and Remote Sensing Magazine*, vol.9, no.4, pp.8-34, 2021.

[2] A. A. S. Gunawan, I. Arifiany and E. Irwansyah, Semantic segmentation of aerial imagery for road and building extraction with deep learning, *ICIC Express Letters*, vol.14, no.1, pp.43-52, 2020.

[3] J. Ding, N. Xue, Y. Long, G.-S. Xia and Q. Lu, Learning RoI Transformer for oriented object detection in aerial images, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.2849-2858, 2019.

[4] J. Pang, C. Li, J. Shi, Z. Xu and H. Feng, R2-CNN: Fast tiny object detection in large-scale remote sensing images, *IEEE Transactions on Geoscience and Remote Sensing*, vol.57, no.8, pp.5512-5524, 2019.

[5] H. Wei, Y. Zhang, B. Wang, Y. Yang, H. Li and H. Wang, X-LineNet: Detecting aircraft in remote sensing images by a pair of intersecting line segments, *IEEE Transactions on Geoscience and Remote Sensing*, vol.59, no.2, pp.1645-1659, 2020.

[6] Y. Li, Q. Huang, X. Pei, Y. Chen, L. Jiao and R. Shang, Cross-layer attention network for small object detection in remote sensing imagery, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol.14, pp.2148-2161, 2020.

[7] Y. Yu, X. Yang, J. Li and X. Gao, A cascade rotated anchor-aided detector for ship detection in remote sensing images, *IEEE Transactions on Geoscience and Remote Sensing*, vol.60, pp.1-14, 2022.

[8] Q. Guo, H. Wang and F. Xu, Scattering enhanced attention pyramid network for aircraft detection in SAR images, *IEEE Transactions on Geoscience and Remote Sensing*, vol.59, no.9, pp.7570-7587, 2021.

[9] J. Fu, X. Sun, Z. Wang and K. Fu, An anchor-free method based on feature balancing and refinement network for multiscale ship detection in SAR images, *IEEE Transactions on Geoscience and Remote Sensing*, vol.59, no.2, pp.1331-1344, 2020.

[10] G. Huang, Z. Wan, X. Liu, J. Hui, Z. Wang and Z. Zhang, Ship detection based on squeeze excitation skip-connection path networks for optical remote sensing images, *Neurocomputing*, vol.332, pp.215-223, 2019.

[11] L. Zhang, Y. Wang and Y. Huo, Object detection in high-resolution remote sensing images based on a hard-example-mining network, *IEEE Transactions on Geoscience and Remote Sensing*, vol.59, no.10, pp.8768-8780, 2020.

[12] Z. Tian, C. Shen, H. Chen and T. He, FCOS: Fully convolutional one-stage object detection, *Proc. of the IEEE/CVF International Conference on Computer Vision*, pp.9627-9636, 2019.

[13] K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.770-778, 2016.

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and F.-F. Li, ImageNet: A large-scale hierarchical image database, *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp.248-255, 2009.

[15] T.-Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, Focal loss for dense object detection, *Proc. of the IEEE International Conference on Computer Vision*, pp.2980-2988, 2017.

[16] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, Feature pyramid networks for object detection, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.2117-2125, 2017.

[17] X. Zhu, H. Hu, S. Lin and J. Dai, Deformable ConvNets v2: More deformable, better results, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.9308-9316, 2019.

[18] B. Xiao, H. Wu and Y. Wei, Simple baselines for human pose estimation and tracking, *Proc. of the European Conference on Computer Vision (ECCV)*, pp.466-481, 2018.

[19] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo and L. Zhang, DOTA: A large-scale dataset for object detection in aerial images, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.3974-3983, 2018.

[20] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu and A. C. Berg, SSD: Single shot multibox detector, *Proc. of the European Conference on Computer Vision (ECCV)*, pp.21-37, 2016.

[21] C. Deng, M. Wang, L. Liu, Y. Liu and Y. Jiang, Extended feature pyramid network for small object detection, *IEEE Transactions on Multimedia*, vol.24, pp.1968-1979, 2021.

[22] S. Ren, K. He, R. Girshick and J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, *Advances in Neural Information Processing Systems*, vol.28, 2015.

[23] S. Liu, L. Qi, H. Qin, J. Shi and J. Jia, Path aggregation network for instance segmentation, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.8759-8768, 2018.

[24] Y. Wu, K. Zhang, J. Wang, Y. Wang, Q. Wang and Q. Li, CDD-Net: A context-driven detection network for multiclass object detection, *IEEE Geoscience and Remote Sensing Letters*, vol.19, pp.1-5, 2020.

[25] A. Bochkovskiy, C.-Y. Wang and H.-Y. M. Liao, YOLOv4: Optimal speed and accuracy of object detection, *arXiv.org*, arXiv: 2004.10934, 2020.