AAAS
Journal of Remote Sensing
Volume 2021, Article ID 9805389, 14 pages
https://doi.org/10.34133/2021/9805389

Journal of Remote Sensing
A SCIENCE PARTNER JOURNAL

*Research Article*

# Feature Enhancement Network for Object Detection in Optical Remote Sensing Images

**Gong Cheng, Chunbo Lang, Maoxiong Wu, Xingxing Xie, Xiwen Yao, and Junwei Han**

*School of Automation, Northwestern Polytechnical University, Xi'an 710129, China*

Correspondence should be addressed to Junwei Han; jhan@nwpu.edu.cn

Received 22 February 2021; Accepted 6 June 2021; Published 8 July 2021

Automatic and robust object detection in remote sensing images is of vital significance in real-world applications such as land resource management and disaster rescue. However, poor performance arises when the state-of-the-art natural image detection algorithms are directly applied to remote sensing images, which largely results from the variations in object scale, aspect ratio, indistinguishable object appearances, and complex background scenario. In this paper, we propose a novel Feature Enhancement Network (FENet) for object detection in optical remote sensing images, which consists of a Dual Attention Feature Enhancement (DAFE) module and a Context Feature Enhancement (CFE) module. Specifically, the DAFE module is introduced to highlight the network to focus on the distinctive features of the objects of interest and suppress useless ones by jointly recalibrating the spatial and channel feature responses. The CFE module is designed to capture global context cues and selectively strengthen class-aware features by leveraging image-level contextual information that indicates the presence or absence of the object classes. To this end, we employ a context encoding loss to regularize the model training which promotes the object detector to understand the scene better and narrows the probable object categories in prediction. We achieve our proposed FENet by unifying DAFE and CFE into the framework of Faster R-CNN. In the experiments, we evaluate our proposed method on two large-scale remote sensing image object detection datasets including DIOR and DOTA and demonstrate its effectiveness compared with the baseline methods.

## 1. Introduction

Object detection has always been a popular and important task in computer vision [1]. In recent years, the volume of remote sensing data is exploding with the development of earth observation technologies. Faced with the need of automatic and intelligent understanding of remote sensing big data, multiclass object detection is becoming a key issue in remote sensing data analysis [2, 3]. More recently, deep learning methods have achieved promising results on natural images, which resulted from the powerful ability of exploiting high-level feature representations, thus offering an opportunity in the interpretation applications of satellite images including urban planning, land resource management, and rescue missions.

However, object detection in optical remote sensing images still remains as a tough challenge due to the particular characteristics of the data, as shown in Figure 1. Firstly, compared with natural scene images that are usually captured by the ground-level cameras with horizontal perspectives, remote sensing images are obtained in the bird's-eye view perspective with a wide range of imaging area. Secondly, remote sensing images vary largely in object scale and aspect ratios. This is not only due to the difference of the Ground Sampling Distance (GSD) of aerial and satellite sensors but also as a result of intraclass variations. Thirdly, the objects in remote sensing images often present different visual appearances and optical properties due to diverse imaging conditions such as viewpoints, illumination, and occlusion [3, 4]. Last but not least, there exists unbalanced distribution of foreground objects and complex background information, especially in intricate landforms and urban scenarios. All of these issues pose great challenges for current state-of-the-art natural image detection algorithms.

Aiming at addressing these challenges to some extent, we propose a novel Feature Enhancement Network (FENet) for

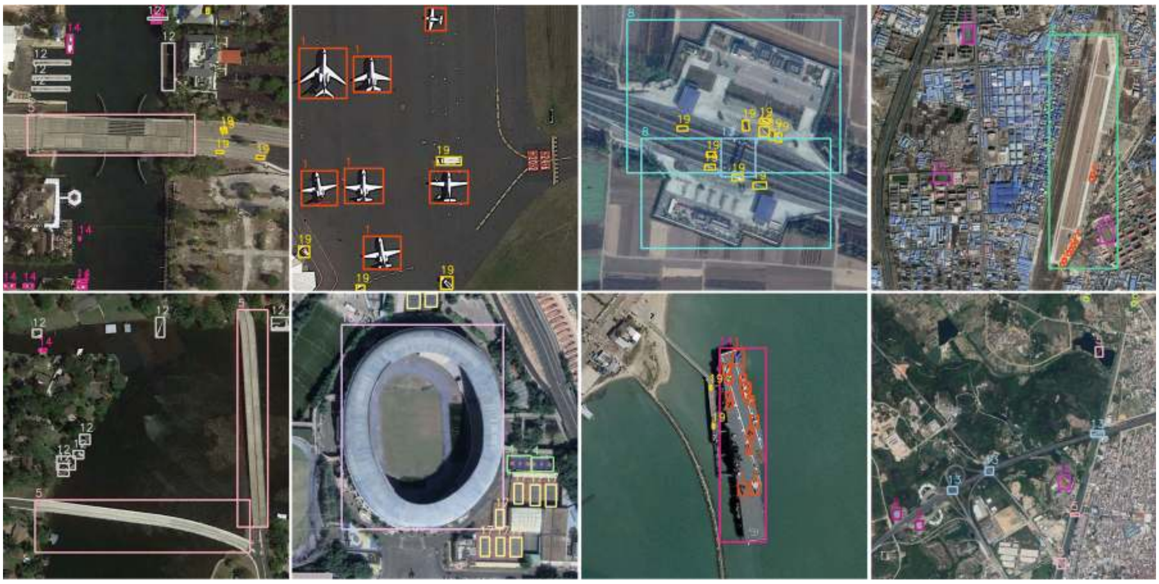

FIGURE 1: Some example images of the DIOR dataset [3] used in our experiments, where the numbers above the bounding boxes indicate the object classes as follows: 1, airplane; 2, airport; 3, baseball field; 4, basketball court; 5, bridge; 6, chimney; 7, dam; 8, expressway service area; 9, expressway toll station; 10, golf field; 11, ground track field; 12, harbor; 13, overpass; 14, ship; 15, stadium; 16, storage tank; 17, tennis court; 18, train station; 19, vehicle; 20, wind mill.

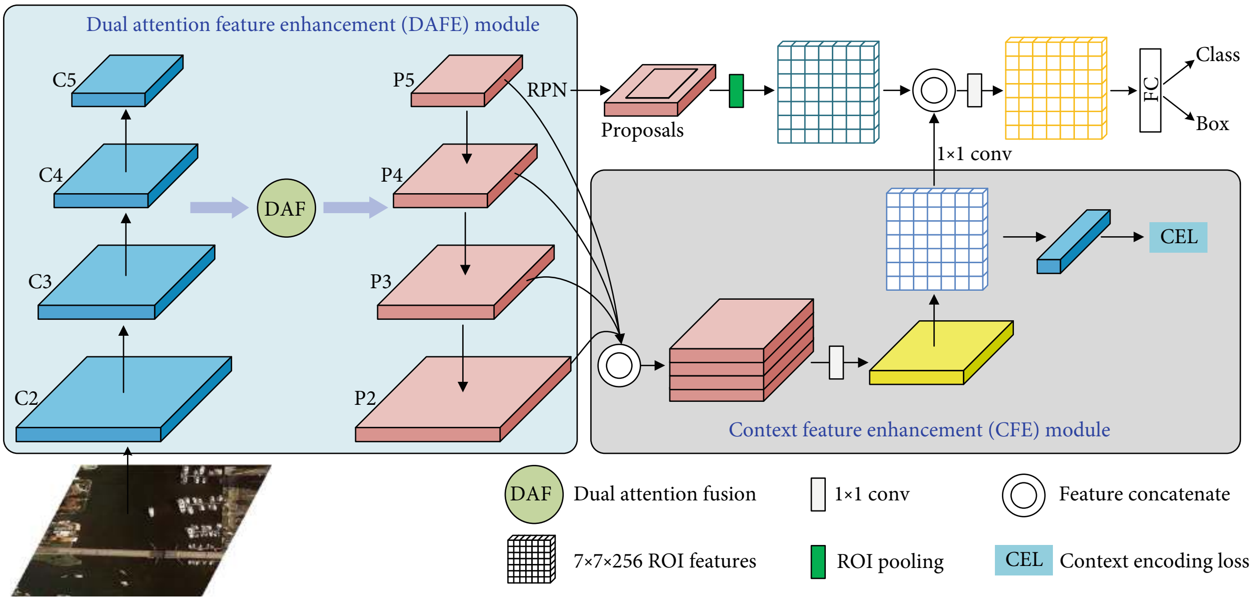
Dual attention feature enhancement (DAFE) module
C5
C4
C3
C2
DAF
P5
P4
P3
P2
RPN
Proposals
1×1 conv
FC
Class
Box
CEL
Context feature enhancement (CFE) module
DAF Dual attention fusion
1×1 conv
Feature concatenate
7×7×256 ROI features
ROI pooling
CEL Context encoding loss

FIGURE 2: Framework of our proposed Feature Enhancement Network (FENet) for object detection in remote sensing images. Building on the popular Faster R-CNN with FPN and adopting the backbone of ResNet-101, our FENet mainly consists of a Dual Attention Feature Enhancement (DAFE) module and a Context Feature Enhancement (CFE) module. The DAFE module is used to strengthen the feature representations of FPN by using the Dual Attention Fusion (DAF) of spatial attention and channel attention. The CFE module is used for capturing global semantic information for better classification and bounding box regression by using a context encoding loss.

robust object detection in remote sensing images. Figure 2 shows the overview of our proposed network. On the one hand, remote sensing images often contain rich spatial and texture cues as well as complex background environment information, which is a collection of both useful and useless information. Therefore, there is a need to guide the network to focus on the features that are more distinguishable for the current object detection task. To this end, we design a Dual Attention Feature Enhancement (DAFE) module to explore discriminative feature representations in both spatial and channel dimensions. On the other hand, there usually exist highly rich ground object categories in remote sensing images and a dataset cannot hold up all the appearances of the objects of interest, which makes it hard for the object detector to infer the object categories we are concerned. However, there are both advantages and disadvantages in view of this trait. The exposure of the ground objects or spatial patterns of the scenes provides useful context clues [5–7] on object classification and localization to some extent. For each object in the training procedure, an object label determines what category the object belongs to and a ground-truth box describes where the object locates, in which the contextual information is not fully utilized. It is noticeable that the scene-level context information like correlative objects and surroundings plays a nonnegligible role in object category and location reasoning. Our inspiration is based on the observation that the contextual information in remote sensing images is of great complementation for object classification and localization. For example, airplanes often appear in airports rather than lakes or residential areas and cars would be more likely appear in bridges, overpasses, or expressway service areas rather than rivers or harbors. This motivates us to design a Context Feature Enhancement (CFE) module to leverage global contextual information to extract more semantic features.

In summary, the main contributions of our work are as follows. First, we present a Dual Attention Feature Enhancement (DAFE) module to highlight the network to focus on the distinctive features of the objects of interest and suppress useless ones by reweighting the spatial and channel feature responses. Second, we design a Context Feature Enhancement (CFE) module to exploit global context cues and selectively strengthen class-aware features by leveraging image-level contextual information that indicates the presence or absence of an object class. Besides, we employ a context encoding loss to regularize the model training which promotes the object detector to understand the scene better and narrows the probable object categories in prediction. Third, our DAFE and CFE modules are generic and thus can be easily applied to existing object detection methods. In this work, we propose a new Feature Enhancement Network (FENet) by unifying DAFE and CFE into the famous object detection framework of Faster R-CNN. Fourth, we comprehensively evaluate our proposed method on two large-scale remote sensing image object detection datasets, namely, DIOR [3] and DOTA [8], and demonstrate its effectiveness compared with the baseline methods.

## 2. Related Work

*2.1. Object Detection in Natural Images.* Feature extraction plays an important role in object detection since it maps raw input data to high-level feature representations. Traditional methods like Histogram of Oriented Gradients (HOG) [9, 10] and Scale Invariant Feature Transform (SIFT) [11] require careful manual engineering and a large amount of time when faced with considerable data examples.

On the contrary, deep learning-based methods can learn powerful feature representations directly and automatically from the raw input data. Therefore, deep learning architecture releases the heavy burden of traditional feature modeling and engineering and thus achieves superior results over traditional feature extraction methods. In recent years, the milestone frameworks of generic object detection can be broadly organized into two mainstream approaches: two-stage detection framework and one-stage detection framework [1, 3].

Two-stage methods refer to region proposal generation at the first stage and the following evaluation of the region proposals. R-CNN [12] generates candidate proposals by selective search and becomes one of the pioneers in generic object detection. Fast R-CNN [13] outperforms R-CNN in both detection speed and accuracy with the idea of sharing feature extraction network for all region proposals. Then, an internal region proposal generation framework based on shared deep CNN arises, which shares the convolutional feature maps for region proposal generator and object detector. Typically, Faster R-CNN [14] proposed by Ren et al. designs a Region Proposal Network (RPN) for region proposal generation, encapsulating the task of proposal generation and the detection task in a single network with many shared convolution layers.

One-stage framework directly predicts class probabilities and bounding box offsets in a unified manner. For example, YOLO [15] integrates category classification and bounding box regression into a unified network, which can reach faster detection speed but usually trailed detection accuracy, especially faced with a successive appearance of small object instances. SSD [16] detects multiscale bounding boxes from multilevel feature maps with fully convolutional neural networks. RetinaNet [17] downweights the loss of numerous well-classified examples by reshaping the crossentropy loss and surpasses two-stage methods without compromising detection speed.

Besides, detecting objects with multiscale CNN layers also promotes detection accuracy, since it is clear that the prediction of objects of different scales is suboptimal with the features from a single layer. An alternative way is to use feature pyramids [18]. FPN [19] achieves a top-down architecture to learn features with hierarchical convolution layers and variant scales, which has shown remarkable improvement as a generic feature extractor in several computer vision tasks including object detection.

Since remote sensing images can be obtained with a wide range of ground sample distance, the object size can be varied from tens to thousands of pixels with dramatic aspect ratios [3, 20]. Compared with one-stage detection methods, most two-stage methods build proposal generation network firstly, which eliminates most of the easy negative examples and reaches a balanced trade-off in the training procedure. Consequently, we adopt the widely used two-stage detector Faster R-CNN with FPN [19] as our backbone in this paper for accurate detection performance.

*2.2. Object Detection in Remote Sensing Images.* Deep learning methods for object detection in remote sensing images have been investigated for years and have achieved promising results [20–31]. A detailed survey on object detection in optical remote sensing images can be found in [2, 3].

In recent years, comprehensive studies have been made to exploit different solutions to the problems of object detection in remote sensing images. For example, for the problem of rotation variations of objects in remote sensing images, [20] designed a rotation-invariant layer to extract robust feature representations. References [24, 32] proposed an effective rotation-invariant and Fisher discriminative CNN (RIFD-CNN) model to improve detection accuracy. Reference [25] presented a rotation-insensitive and context-augmented object detection method. Aiming at multiscale object detection problem, [26] introduced a crossscale feature fusion (CSFF) framework. Reference [27] developed an object detection method for remote sensing images by combining multilevel feature fusion and an improved bounding box regression scheme. Reference [33] designed a multiscale object proposal network (MS-OPN) for proposal generation and an accurate object detection network (AODN) for detecting objects of interest in remote sensing images with large-scale variability.

More recently, some literature began to pay attention to the research of oriented object detection in remote sensing images [34–44]. For example, [34] presented a region of interest (RoI) transformer through applying spatial transformations on RoIs and learning the parameters of transformation with the supervision of oriented annotations. Reference [35] proposed to describe an oriented object by gliding the vertexes of each horizontal bounding box on their corresponding sides, and an obliquity factor based on area ratio was further introduced to remedy the confusion issue. R3Det [37] encodes centers and corners information in the features to get a more accurate location. Reference [41] presented a dynamic refinement network which enabled neurons to adjust receptive fields according to the shapes and orientations of target objects and refined the prediction dynamically in an object-aware manner. Reference [36] proposed a new rotation detector, named SCRDet, for detecting small, cluttered, and rotated objects in remote sensing images, which alleviated the influence of angle periodicity by designing a novel IoU-Smooth L1 Loss. Reference [39] used image cascade and feature pyramid jointly with multisize convolution kernels to extract multiscale strong and weak semantic features for oriented object detection. Yao et al. [44] proposed a Single-shot Alignment Network ($S^2$A-Net) to alleviate the inconsistency between classification score and localization accuracy, which achieved state-of-the-art performance on two aerial object datasets. To achieve better detection speed, [42] used a set of default boxes with various scales like SSD to predict oriented bounding boxes. Reference [43] defined a rotatable bounding box to predict the exact shape of objects for detecting vehicles, ships, and airplanes, showing superior capability of locating multiangle objects.

Also, some methods were proposed for weakly supervised object detection (WSOD) in remote sensing images [21, 23, 45–48]. For instance, [21] proposed a coupled weakly supervised learning framework for aircraft detection. Reference [45] proposed a WSOD framework based on dynamic curriculum learning to progressively train object detectors by feeding training images with ascending difficulty. Reference [46] proposed a new progressive contextual instance refinement (PCIR) method to perform WSOD in remote sensing images.

*2.3. Attention Mechanism.* Feature-based attention has proved its effectiveness in many computer vision tasks as a perception-adapted mechanism [49]. For instance, Squeeze-and-Excitation network (SENet) [50] proposed by Hu et al. adaptively recalibrates channel relationships by global information embedding and fully connected (FC) layers. Reference [51] computed weights from nonlocal and local pixels/features as the spatially refined representations. Reference [52] achieves domain attention by a series of universal adaptation layers, following the principle of squeeze and excitation. For the task of object detection in remote sensing images, [22] puts forward an inception fusion strategy as well as pixel-wise and channel-wise attention for small object detection in aerial images. Reference [26] inserted a SENet block into the top layer of FPN to model the relationship of different feature channels. Inspired by Mask R-CNN, [40] proposed a refine FPN and multilayer attention network for oriented object detection of remote sensing images.

## 3. Proposed Method

*3.1. Review of Faster R-CNN.* Faster R-CNN proposed by Ren et al. [14] is an efficient two-stage detection algorithm, which consists of two main branches, namely, RPN and Fast R-CNN. In the first stage, RPN generates a set of anchor boxes with predefined scales and aspect ratios at each feature map location, followed by two sibling fully connected layers, one for object classification and one for bounding box regression, respectively. In the second stage, a ROI pooling layer is employed to obtain fixed-size outputs for each region proposal before classification and bounding box refinement. The two stages are integrated by several shared convolution layers and can be trained and tested end to end.

*3.2. Overview of Feature Enhancement Network (FENet).* The architecture of our proposed Feature Enhancement Network (FENet) for object detection in remote sensing images is illustrated in Figure 2. Building on the popular Faster R-CNN with FPN and adopting the backbone of ResNet-101, our FENet mainly consists of a Dual Attention Feature Enhancement (DAFE) module and a Context Feature Enhancement (CFE) module. The DAFE module is used to highlight the FPN to focus on the distinctive features of the objects of interest and suppress useless ones by using the Dual Attention Fusion (DAF) to jointly reweight the spatial and channel feature responses. The CFE module is used to selectively strengthen class-aware features by leveraging image-level contextual information that indicates the presence or absence of the object classes. The feature representations of the CFE module are concatenated with each ROI feature to make per-proposal prediction. To this end, we employ a context encoding loss to regularize the model training, which could enforce the network to learn the global semantic information through predicting the presence of the object classes in the images, thus promoting the object detector to better understand the images for classification and bounding box regression.

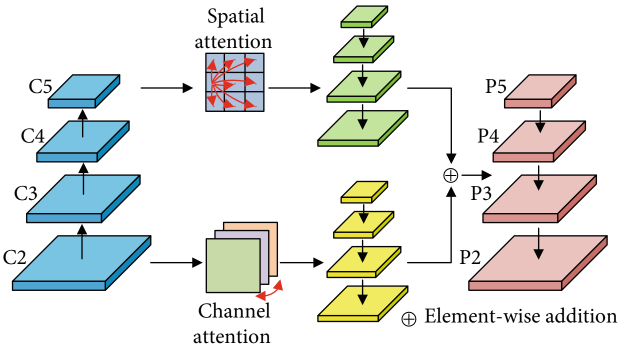

Figure 3: The proposed Dual Attention Feature Enhancement (DAFE) module. In brief, we use nonlocal block and SE block in parallel to jointly recalibrate the spatial and channel feature responses, respectively, and then fuse them for a better capability.

*3.3. Dual Attention Feature Enhancement (DAFE).* The CNN has shown powerful ability in feature extraction and representation with a large number of parameters. Low-level layers in the CNN architecture contain a large amount of detailed information such as edges and boundaries. As the network goes deeper, the high-level feature representations have diminished location information and specialized in semantic information. How to obtain and choose more discriminative features determines the detection performance. To this end, a Dual Attention Feature Enhancement (DAFE) module is constructed to prompt the network to focus on the distinctive features and suppress the redundant ones that are not useful for the current task by jointly recalibrating the spatial and channel feature responses, as shown in Figure 3. Specifically, in the spatial dimension, we use nonlocal building block [51] to acquire spatial dependencies in the whole feature map. As for channel dimension, SE block [50], which models the channel relationship explicitly from inherent feature maps and so can be directly applied to existing state-of-the-art CNN architectures, is selected for our implementation. These two kinds of attentions are carried out in parallel and then fused for a better capability. Next, we briefly introduce the nonlocal block [51] and SE block [50].

The nonlocal block was designed to capture long-range dependencies through nonlocal operation which calculates the new feature response of each position as a weighted sum of the original features of all positions [51]. Specifically, given an input feature $x$, its output feature $z$ of a nonlocal block is computed as follows:

$$z_i = W_z y_i + x_i, \tag{1}$$

where $W_z$ is the weight matrix that is implemented as 1 × 1 convolution, "$+x_i$ "represents a residual connection which makes it possible to insert a new nonlocal block into any pretrained CNN model without breaking its initial behavior, and $y_i$ is the output of the nonlocal operation of the same size as $x$, which is defined in the following equation:

$$y_i = \frac{1}{C(x)} \sum_{\forall j} f\left(x_i, x_j\right) g\left(x_j\right), \tag{2}$$

where $C(x)$ is the normalization factor set as $C(x) = \sum_{\forall j} f(x_i, x_j)$. $i$ is the index of an output position of the features, and $j$ is the index enumerating all possible positions. $f$ is a pair-wise function used to calculate a scalar to represent the relationship between $x_i$ and all $x_j$. The function $g$ is used to compute the embeddings of the input signal at the position $j$ by using $g(x_j) = W_g x_j$ with $W_g$ being a 1 × 1 convolutional operation. In this paper, we use the embedded Gaussian function as the pair-wise function as defined in Equation (3) for the computation of the relationship scalar.

$$f\left(x_i, x_j\right) = e^{\theta(x_i)^T \phi(x_j)}, \tag{3}$$

where $\theta(x_i) = W_\theta x_i$ and $\phi(x_j) = W_\phi x_j$ are two embeddings computed through the 1 × 1 convolutional filters $W_\theta$ and $W_\phi$.

The nonlocal module is inserted into the end of the convolutional stage of ResNet-101 in our experiments, and we investigate the results of different combinations of stages by using the nonlocal block in the experiments.

The SE block can be embedded into any regular CNN architectures with the operations of embedding global information and recalibrating channel-wise dependencies. First, a global average pooling is applied on the spatial dimensions and generate a $K \times 1 \times 1$ vector $z$, in which the $k$th element of $z$ is defined as

$$z_k = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_k(i, j), \tag{4}$$

where $K$ is the depth of the feature map and $x_k(i, j)$ is the value of the $k$th channel at position $(i, j)$ of the input feature map.

Then, two FC layers are followed to recalibrate the channel dependencies and a sigmoid activation function is employed to learn nonlinear relationships:

$$\text{Scale} = \sigma(W_2 \delta(W_1 z)), \tag{5}$$

where $\sigma(\cdot)$ denotes the ReLU function. Finally, the output feature map is obtained by the implementation of channel-wise multiplication.

Similar to the nonlocal module, the SE block is also added on the end of the convolutional stage to capture channel-wise responses and highlight discriminative features.

Nevertheless, what is the best arrangement for these two blocks in the network? Reference [50] also suggests that the importance of feature channels tends to share a similar weight when using SE block in low-level features, while in high-level features, the importance of each channel becomes more class-specific. To thoroughly investigate this problem, we deployed these two blocks in different residual stages of ResNet [53], respectively, and evaluated their performances by using different combinations, and the results of various combinations can be found in Section 4. Although there are small gaps between different results, we observe that the

Table 1: Object classes of DIOR and DOTA datasets. The short names are adopted for simplicity.

| | | | | |
|---|---|---|---|---|
| DIOR | C1<br>Airplane | C2<br>Airport | C3<br>Baseball field | C4<br>Basketball court |
| | C5<br>Bridge | C6<br>Chimney | C7<br>Dam | C8<br>Expressway service area |
| | C9<br>Expressway toll station | C10<br>Golf field | C11<br>Ground track field | C12<br>Harbor |
| | C13<br>Overpass | C14<br>Ship | C15<br>Stadium | C16<br>Storage tank |
| | C17<br>Tennis court | C18<br>Train station | C19<br>Vehicle | C20<br>Windmill |
| DOTA | PL<br>Plane | BD<br>Baseball diamond | BR<br>Bridge | GTF<br>Ground field track |
| | SV<br>Small vehicle | LV<br>Large vehicle | SH<br>Ship | TC<br>Tennis court |
| | BC<br>Basketball court | ST<br>Storage tank | SBF<br>Soccer-ball field | RA<br>Roundabout |
| | HA<br>Harbor | SP<br>Swimming pool | HC<br>Helicopter | |

entire framework is not sensitive to the implementation of the blocks that are used in different layers.

### *3.4. Context Feature Enhancement (CFE).*

Finally, we propose a novel Context Feature Enhancement (CFE) module that utilizes task-specific features and scene semantics generated from hierarchical feature layers. Since high-level features have more semantic information while low-level features contain specific geometric information such as context and edges, they are good complementation for each other in object detection task. In this model, we integrate the multilevel feature maps to obtain both high-level semantic features and low-level detailed features, which can guide the object category classification and location reasoning in a global manner.

More specifically, we empirically set the downsample rate of 16 to preserve some localization information. Max pooling is used for P2 and P3, and nearest upsampling operation is used for P5, ensuring the consistency of spatial scale. With the above approach, diverse feature representations from different levels can be aggregated. Then, two additional fully connected convolutional layers with sigmoid activation function are added on top of the fusion features to predict the confidence of object categories in the remote sensing scene, and the binary crossentropy loss is adopted for training. This auxiliary branch processes the multilabel classification task through intermediate feature map, thus providing the basic classifiers with global and local knowledge of contextual clues that are correlative to the region of interest. The object category prediction is typically achieved by computing softmax probabilities, which is not feasible for the object classification in such task. As a consequence, we adopt the sigmoid crossentropy loss to measure the probability error in which each class is independent and not mutually exclusive. Specifically, given an input image $X \in \mathbb{R}^{3\times H\times W}$, the ground-truth label can be denoted as a vector $y = [y_0, y_1, \cdots, y_C]^T$, where $C$ is the total number of object categories. $y_i$ is set to 1 if objects in image $X$ correspond to class $i$, otherwise it is set to 0, where $i \in \{1, \cdots, C\}$. We represent the predicted class score vector of image $X$ as $p = [p_0, p_1, \cdots, p_C]^T$, and for all the $j$ training images, the multilabel classification loss is calculated by

$$\mathscr{L}_{\mathrm{CFE}} = \sum_j \left[ y_j * \log\left(\frac{1}{1+\exp\left(-p_j\right)}\right) + \left(1-y_j\right) * \log\left(\frac{\exp\left(-p_j\right)}{1+\exp\left(-p_j\right)}\right)\right]. \tag{6}$$

What is more, [54] has demonstrated that the multilabel classification task based on CNN features retains coarse localization information of objects without using any bounding box annotations. Inspired by this, we aggregate the features obtained by CFE module with box prediction head, which provides not only global and local context information for object category reasoning but also localization information for bounding box regression. In our method, the context feature maps are downsampled to 7 × 7 to match the same resolution as region proposals after ROI pooling. Then, we concatenate the context features with ROI features and apply a 1 × 1 convolution operation to reduce channel dimensions while powering the informative representations, which can be seen as a complementation for region proposal detection task. Let $\mathscr{L}_{\mathrm{cls}}$ denote the object category classification loss and $\mathscr{L}_{\mathrm{reg}}$ denote the bounding box regression loss. Finally, the loss function can be defined as

$$\mathscr{L} = \mathscr{L}_{\mathrm{cls}} + \mathscr{L}_{\mathrm{reg}} + \lambda \mathscr{L}_{\mathrm{CFE}}, \tag{7}$$

where $\lambda$ is a hyperparameter that controls the factor of $\mathscr{L}_{\mathrm{CFE}}$. In Section 4, we discuss the choice of $\lambda$ in detail.

Table 2: Comparison of FENet and the state-of-the-art methods on the DIOR test set. FR and MR indicate the Faster R-CNN [14] and Mask R-CNN [59] methods, respectively.

| Method | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 | C9 | C10 | C11 | C12 | C13 | C14 | C15 | C16 | C17 | C18 | C19 | C20 | mAP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| FR + FPN [19] | 54.0 | 74.5 | 63.3 | 80.7 | 44.8 | 72.5 | 60.0 | 75.6 | 62.3 | 76.0 | 76.8 | 46.4 | 57.2 | 71.8 | 68.3 | 53.8 | 81.1 | 59.5 | 43.1 | 81.2 | 65.1 |
| FR + FPN + OHEM [60] | 54.2 | 78.0 | 63.2 | 81.0 | 41.2 | 72.6 | 60.7 | 79.1 | 62.9 | 78.3 | 77.0 | 56.9 | 59.3 | 71.4 | 62.3 | 53.5 | 81.1 | 56.3 | 43.4 | 81.3 | 65.7 |
| MR + FPN [59] | 53.9 | 76.6 | 63.2 | 80.9 | 40.2 | 72.5 | 60.4 | 76.3 | 62.5 | 76.0 | 75.9 | 46.5 | 57.4 | 71.8 | 68.3 | 53.7 | 81.0 | 62.3 | 43.0 | 81.0 | 65.2 |
| MR + FPN + OHEM | 54.2 | 78.3 | 63.3 | 81.0 | 46.7 | 72.6 | 61.6 | 80.1 | 66.4 | 78.4 | 76.7 | 57.2 | 59.6 | 71.6 | 65.5 | 53.8 | 81.2 | 58.8 | 43.3 | 81.2 | 66.5 |
| CornerNet [58] | 58.8 | 84.2 | 72.0 | 80.8 | 46.4 | 75.3 | 64.3 | 81.6 | 76.3 | 79.5 | 79.5 | 26.1 | 60.6 | 37.6 | 70.7 | 45.2 | 84.0 | 57.1 | 43.0 | 75.9 | 64.9 |
| Libra R-CNN [61] | 62.5 | 78.8 | 72.0 | 80.8 | 46.7 | 72.6 | 64.4 | 79.9 | 69.1 | 77.5 | 76.3 | 46.2 | 59.3 | 71.8 | 68.0 | 53.9 | 81.1 | 62.4 | 43.2 | 81.3 | 67.4 |
| FENet (ours) | 54.1 | 78.2 | 71.6 | 81.0 | 46.5 | 79.0 | 65.2 | 76.5 | 69.6 | 79.1 | 82.2 | 52.0 | 57.6 | 71.9 | 71.8 | 62.3 | 81.2 | 61.2 | 43.3 | 81.2 | 68.3 |

Table 3: Comparison of FENet and the state-of-the-art methods on the DOTA test set. FR and MR indicate the Faster R-CNN [14] and Mask R-CNN [59] methods, respectively.

| Method | PL | BD | BR | GTF | SV | LV | SH | TC | BC | ST | SBF | RA | HA | SP | HC | mAP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| FR + FPN [19] | 88.70 | 75.10 | 52.60 | 59.20 | 69.40 | 78.80 | 84.50 | 90.60 | 81.30 | 82.60 | 52.50 | 62.10 | 76.60 | 66.30 | 60.10 | 72.00 |
| EFR [62] | 88.36 | 83.90 | 45.78 | 67.24 | 76.80 | 77.15 | 85.35 | 90.77 | 85.55 | 75.77 | 54.64 | 60.76 | 71.40 | 77.90 | 60.94 | 73.49 |
| IoU-adaptive R-CNN [63] | 88.62 | 80.22 | 53.18 | 66.97 | 76.30 | 72.59 | 84.07 | 90.66 | 80.95 | 76.24 | 57.12 | 66.65 | 84.08 | 66.36 | 56.85 | 72.72 |
| FMSSD [64] | 89.11 | 81.51 | 48.22 | 67.94 | 69.23 | 73.56 | 76.87 | 90.71 | 82.67 | 73.33 | 52.65 | 67.52 | 72.37 | 80.57 | 60.15 | 72.43 |
| ICN [39] | 89.97 | 77.71 | 53.38 | 73.26 | 73.46 | 65.02 | 78.22 | 90.79 | 79.05 | 84.81 | 57.20 | 62.11 | 73.45 | 70.22 | 58.08 | 72.45 |
| ASBL-RetinaNet [65] | 89.51 | 74.07 | 46.91 | 55.54 | 73.78 | 66.87 | 78.48 | 90.86 | 70.09 | 73.20 | 46.71 | 61.34 | 70.50 | 72.17 | 32.84 | 66.86 |
| FENet (ours) | 88.47 | 80.54 | 54.65 | 71.70 | 78.09 | 80.65 | 87.36 | 90.81 | 84.53 | 84.74 | 53.23 | 64.14 | 76.87 | 69.94 | 57.66 | 74.89 |

To sum up, the DAFE and CFE modules complement each other well to some extent. The features with poor positioning ability and poor discrimination can be enhanced by contextual information, while features with good discrimination are guaranteed not to be significantly weakened.

Table 4: The running time of the proposed method on different datasets for a single test image of given size. The running time is tested on a NVIDIA Titan Xp GPU with the batch size of 1.

| Dataset | Image size | Running time (ms) |
|---|---|---|
| DIOR | 800 × 800 | 115 |
| DOTA | 1024 × 1024 | 134 |

## 4. Experiments

In the following section, we first present the implementation details of DAFE and CFE and conduct an ablation study on the newly published datasets DIOR [3] and DOTA [8].

*4.1. Datasets and Evaluation Metrics.* In this paper, we perform our experiments on two large-scale remote sensing datasets DIOR [3] and DOTA [8]. As for the former, it consists of 23463 optical remote sensing images and covers with 20 categories. 192472 manually labelled instances with axis-aligned boxes are involved, following similar annotation format as PASCAL VOC. The images in the DIOR dataset have the size of 800 × 800 and vary in spatial resolution from 0.5 m to 30 m. We take 11725 images from train and validation splits for training and the rest 11738 images for testing. As for the latter, it contains 2806 aerial images from various sensors and 15 common object categories. The fully annotated DOTA images consist of 188282 instances labeled by arbitrary quadrilaterals, and the image size of the DOTA dataset is large: from 800 × 800 to 4000 × 4000 pixels. We use training and validation sets for training and the rest for testing. The detection accuracy is obtained by submitting testing results to DOTA's evaluation server. All the object categories of these two datasets are reported in Table 1.

In our results, we follow the mean Average Precision (mAP) as the evaluation metric for our experiment and the evaluation of mAP is the same as the metric definition in PASCAL VOC 2007 object detection challenge.

*4.2. Implementation Details.* Our experiment is performed under the framework of PyTorch and based on the Faster R-CNN with FPN [55]. ResNet-101 is adopted as the backbone network. We run 12 epochs on a NVIDIA Titan Xp GPU with the batch size of 2. The initial learning rate is set to 0.0025 with a learning rate decay of 0.1 at the end of epoch 8 and epoch 11. The momentum is 0.9, and the weight decay is set to 0.0005. During the training process, a horizontal flip data augmentation method is used in the end-to-end procedure with stochastic gradient descent (SGD) optimizer. The parameter setting of SE block is the same as [50].

For the images in the DIOR dataset, we keep the original size of 800 × 800 for training and testing. With regard to the DOTA dataset, we crop the original images in the DOTA dataset into 1024 × 1024 patches. The stride of cropping is set to 824; that is, the pixel overlap between two adjacent patches is 200. As commonly used in object detection, ResNet-101 network is pretrained on the ImageNet [56] and fine-tuned on the aforementioned training set.

*4.3. Experimental Results.* We evaluate our model on the test set of DIOR and DOTA datasets and compare it with the state-of-the-art methods. The experiments are implemented on mmdetection [57] to make a fair comparison, except for CornerNet [58]. As shown in Table 2, on the DIOR dataset, our method achieves 68.3% mAP and outperforms the baseline Faster R-CNN with FPN by 3.2%, which demonstrates its effectiveness for object detection in remote sensing images. Our method shows competitive performance compared to state-of-the-art methods like Libra R-CNN and CornerNet. Moreover, CornerNet performs better results in large objects such as airport, expressway service area, and overpass while it struggles in small and crowed objects including ships and vehicles. As for individual class predictions, we notice that the AP values of the classes of airplane, basketball court, ship, tennis court, vehicle, and windmill only show little improvement. We analyze the reasons as follows. For the ship and vehicle categories, although there are many instances available, they account for a relatively small proportion of the entire images, leading to the information loss seriously after being sampled by the backbone network, which brings difficulty to feature extraction and further enhancement, so the improvement is not obvious. In contrast, for the golf field and ground track field categories with large object sizes,

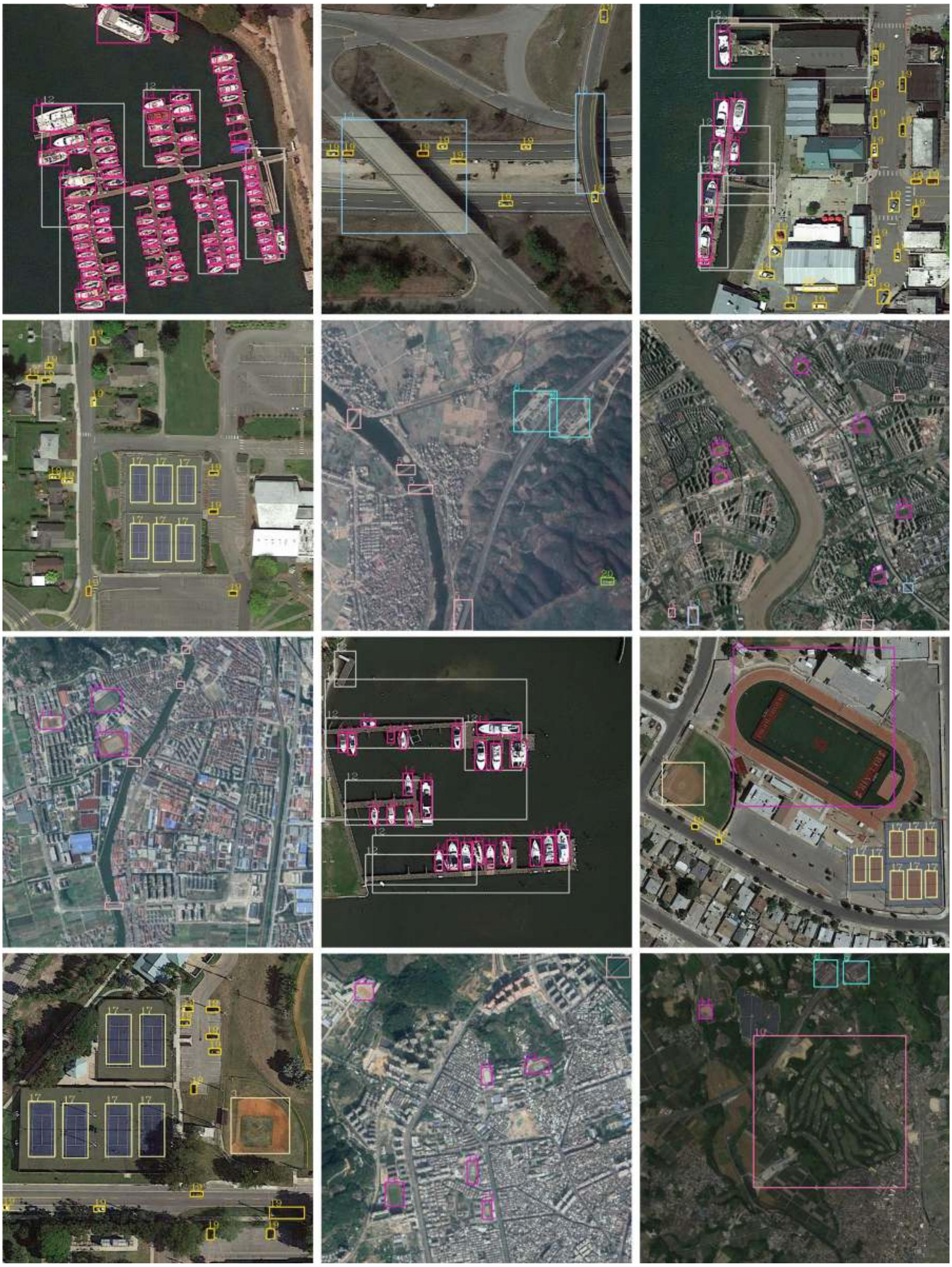

Figure 4: Visualization of the detection results. The detected boxes are shown with different colors according to different classes.

although the sample numbers of them are relatively small, they still have big accuracy gains (3.1% AP gain for the golf field class and 5.4% AP gain for the ground track field class). Besides, for the classes of airplane, basketball court, tennis court, and windmill, the experimental results are closely related to their characteristics. Specifically, the aircraft category has large-scale differences, the appearances of tennis courts and basketball courts are similar and easy to be confused, and the windmill category has shadow interference. These factors undoubtedly increase the difficulty of

Table 5: Comparison of different combinations of spatial and channel attention methods.

| Type | Nonlocal |  |  |  | SE |  |  |  | mAP |
|---|---|---|---|---|---|---|---|---|---|
|  | Stage2 | Stage3 | Stage4 | Stage5 | Stage2 | Stage3 | Stage4 | Stage5 |  |
| Spatial | ✓ |  |  |  |  |  |  |  | 66.6 |
|  |  | ✓ |  |  |  |  |  |  | 66.7 |
|  |  |  | ✓ |  |  |  |  |  | 66.8 |
|  |  |  |  | ✓ |  |  |  |  | 66.6 |
|  | ✓ | ✓ | ✓ | ✓ |  |  |  |  | 66.8 |
| Channel |  |  |  |  | ✓ |  |  |  | 66.6 |
|  |  |  |  |  |  | ✓ |  |  | 66.2 |
|  |  |  |  |  |  |  | ✓ |  | 66.5 |
|  |  |  |  |  |  |  |  | ✓ | 66.8 |
|  |  |  |  |  | ✓ | ✓ | ✓ | ✓ | 66.1 |
| Combination | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 67.0 |
|  |  |  | ✓ |  |  |  |  | ✓ | 67.7 |

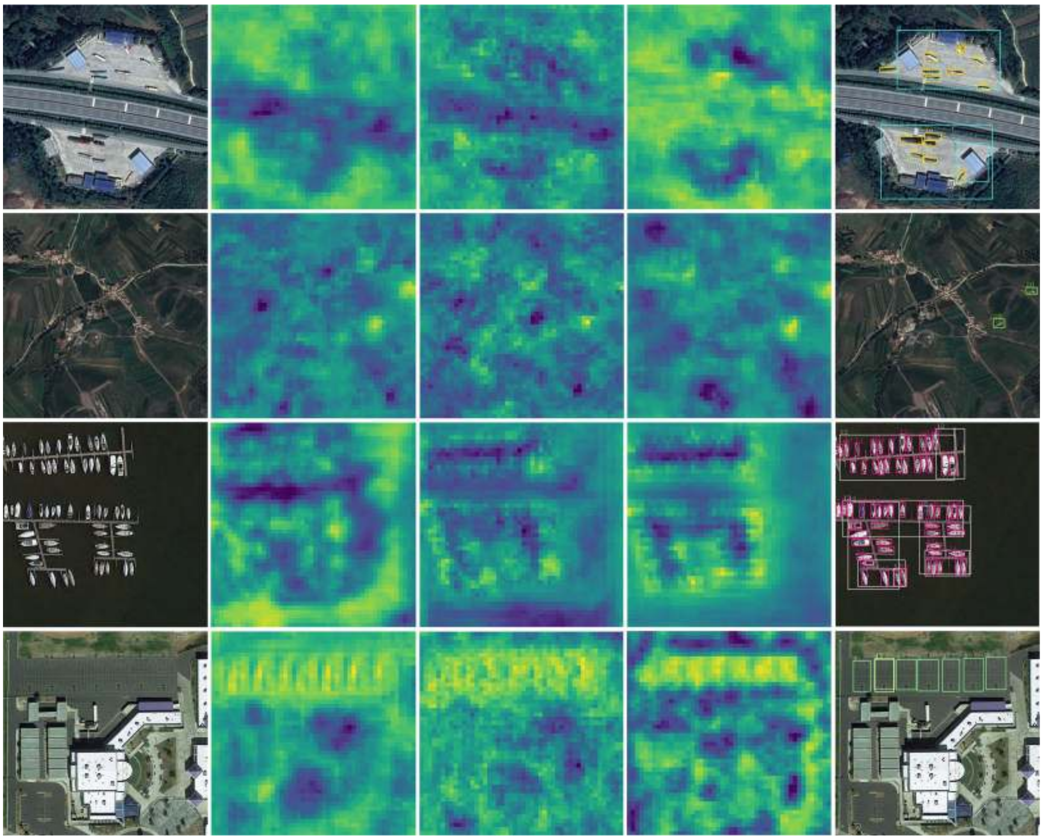

Figure 5: Visualization result of the DAFE module. The response value of the target region in the corresponding part of feature enhancement map is larger, indicating the successful acquisition of the task-related features.

detection, resulting in less performance improvement compared to the baseline method.

As for the detection results on the DOTA dataset (see Table 3), our proposed FENet once again achieves the highest accuracy, namely, 74.89% mAP, which outperforms the baseline FPN by 2.89%. The reason of the above promising results is closely related to the proposed DAFE and CFE modules, which enhance the capability of capturing task-related features and balancing global and local information and thus performing well in most object categories. Furthermore, Table 4 presents the running time of FENet on different datasets for a test image of given size. The running time is tested on a NVIDIA Titan Xp GPU with the batch size of 1. It can be seen that the proposed approach maintains a fast inference speed while achieving high detection accuracy.

Figure 4 illustrates some test samples and the corresponding detection results on the DIOR dataset. As can be found, our proposed method is suitable for some small-sized and medium-sized objects, such as vehicles, ships, and storage tanks, indicating the contributing guidance provided by contextual information. More specifically, these objects usually crowded together and cannot be easily distinguished. The low-level detailed features can provide some localization information, while the high-level semantics facilitate the object reasoning. In addition, the proposed FENet also achieves robust detection performance in the object categories with large scale variation compared with the state-of-the-art methods, such as baseball field, ground track field, harbor, and stadium. Although the objects in each of these classes present different visual appearances, they may share some common contextual clues to some extent, resulting in relatively stable detection performance.

*4.4. Ablation Study.* In order to evaluate the effectiveness of the proposed DAFE and CFE modules, we conduct a series of experiments on the DIOR dataset in this section. The impact of different components on detection performance is presented in Table 5. As can be found from the 3rd to 7th rows, the usage of nonlocal module shows no apparent difference in either separate stage or combined stages. The model with the spatial attention mechanism achieved the highest accuracy in two cases: all stages used and only the stage 4 used. According to the "channel" row of Table 5, the results fluctuate very little with diverse groups of stages that utilize SE blocks. Adding SE block to all the convolutional stages makes 1% improvement. In contrast, applying SE block to convolutional stage 5 achieves the highest performance with 1.7% increment compared to baseline result. It is worth noting that the table does not show the results of more combinations of attention blocks at different stages (i.e., double stages and triple stages), because it does not lead to significant performance improvement and sometimes even worse. One possible reason is that the emphasized features from different levels are not properly refined. As a consequence, the Context Feature Enhancement module is designed, where we associate multilevel features to accomplish this goal.

To further investigate how the different combinations of spatial and channel methods affect the final results, we make comparisons between the utilization of single stage and multiple stages in the last two rows of Table 5. It reveals that the DAFE achieves the best performance of 67.7% mAP when we use the nonlocal module in stage 4 and SE block in stage 5. However, the model with nonlocal module and SE block applied on all the stages only achieves 67.0% mAP. Figure 5 gives several visualization results of DAFE. The first column is the original images; the second column is the feature enhancement in the spatial dimension; the third column is the feature enhancement in the channel dimension; the fourth column corresponds to the total feature enhancement of DAFE; the last column is the corresponding detection result.

Table 6: Effect of Context Feature Enhancement by diverse values of $\lambda$. Note that the detection result of baseline method is 65.1% mAP.

| Parameter | Baseline + CFE | Baseline + DAFE + CFE |
|---|---|---|
| $\lambda = 1$ | 66.3 | 67.5 |
| $\lambda = 5$ | 66.6 | 68.3 |
| $\lambda = 10$ | 66.3 | 68.0 |

Furthermore, we also examine how the choice of $\lambda$ contributes to the detection results. The ablation study is mainly conducted from the following aspects:

(1) The individual impact of CFE on baseline. In Table 6, we compare the diverse values of $\lambda$ in a wide range from 1 to 10. The results suggest that the performance approximately grows 1% by average when the contextual information provided by CFE module is included. While $\lambda$ takes the value of 5, we obtain the highest performance, particularly up to 1.5% improvement compared to baseline

(2) The interaction between CFE and DAFE. From Table 6, we notice that the overall method shows no improvement when $\lambda = 1$. Then, we change the choice of $\lambda$ and find better results at 5 and 10. The method also shows little improvement when we further enlarge the hyperparameter $\lambda$. This indicates that there is imbalance between losses. When $\lambda$ is too small, CFE hardly contributes to the network with contextual information. While $\lambda$ is too large, $\mathscr{L}_{\text{cls}}$ and $\mathscr{L}_{\text{reg}}$ can be overwhelmed. We also find that CFE has significant effect on class 5, 6, 9, 12, 15, and 16 of the DIOR dataset, which are typical objects with great scale variations. This indicates that CFE can learn common contextual clues of certain object categories and guide the network to reason reliable possibilities. Besides, the experiments also demonstrate that these two components of the proposed network are complementary to each other

## 5. Conclusion

In this paper, we present a novel approach FENet for multiclass object detection in optical remote sensing images, which is aimed at addressing the complex background scenario and sparse object distribution problems. Firstly, the framework utilizes Dual Attention Feature Enhancement module to selectively emphasize informative features from multiple resolutions, thus guiding the network for robust object detection. In the next phase, a Context Feature Enhancement module is introduced to fully leverage the abundant information emerged in remote sensing objects. This branch explores both global and local contextual information like semantics and textures, which bridges the gap of multiscale feature maps. The experiments on DIOR and DOTA datasets verify its effectiveness and show that our proposed method achieves remarkable performance compared with the state-of-the-art algorithms. For future works, we plan to carry on our work in oriented bounding box detection and focus on unusual appearances of objects like exceptional aspect ratios and scales.

## Data Availability

The data of DIOR and DOTA used to support this study are publicly available. The DIOR data can be downloaded from the website https://gcheng-nwpu.github.io/datasets while the DOTA data can be downloaded from the website https://captain-whu.github.io/DOTA/index.html. The code is freely available upon request.

## Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

## Authors' Contributions

G. Cheng, C. Lang, M. Wu, and X. Xie proposed the method. M. Wu and X. Xie implemented the experiments. G. Cheng, C. Lang, M. Wu, X. Xie, and X. Yao participated in the analysis of experimental results. G. Cheng, C. Lang, and M. Wu wrote the manuscript. G. Cheng and J. Han provided supervisory support. All authors read and approved the final manuscript.

## Acknowledgments

This work was supported in part by the National Science Foundation of China under Grant 61772425 and in part by the Shaanxi Science Foundation for Distinguished Young Scholars under Grant 2021JC-16.

## References

[1] L. Liu, W. Ouyang, X. Wang et al., "Deep learning for generic object detection: a survey," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 261–318, 2020.

[2] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 117, pp. 11–28, 2016.

[3] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: a survey and a new benchmark," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 159, pp. 296–307, 2020.

[4] Y. Long, Y. Gong, Z. Xiao, and Q. Liu, "Accurate object localization in remote sensing images based on convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 5, pp. 2486–2498, 2017.

[5] C. Galleguillos and S. Belongie, "Context based object categorization: a critical survey," *Computer Vision and Image Understanding*, vol. 114, no. 6, pp. 712–722, 2010.

[6] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3588–3597, Salt Lake City, UT, USA, June 2018.

[7] X. Zhang, Y. H. Yang, Z. Han, H. Wang, and C. Gao, "Object class detection," *ACM Computing Surveys*, vol. 46, no. 1, pp. 1–53, 2013.

[8] G.-S. Xia, X. Bai, J. Ding et al., "Dota: a large-scale dataset for object detection in aerial images," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3974–3983, Salt Lake City, UT, USA, June 2018.

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 886–893, San Diego, CA, USA, 2005.

[10] G. Cheng, J. Han, P. Zhou, and L. Guo, "Multi-class geospatial object detection and geographic image classification based on collection of part detectors," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 98, pp. 119–132, 2014.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, Columbus, OH, USA, June 2014.

[13] R. Girshick, "Fast R-CNN," in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448, Santiago, Chile, December 2015.

[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[15] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, Las Vegas, NV, USA, June 2016.

[16] W. Liu, D. Anguelov, D. Erhan et al., "SSD: single shot multibox detector," in *European conference on computer vision*, pp. 21–37, Springer, 2016.

[17] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, Venice, Italy, October 2017.

[18] X. Yang, H. Sun, K. Fu et al., "Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks," *Remote Sensing*, vol. 10, no. 1, p. 132, 2018.

[19] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2117–2125, Honolulu, HI, USA, July 2017.

[20] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 12, pp. 7405–7415, 2016.

[21] F. Zhang, B. Du, L. Zhang, and M. Xu, "Weakly supervised learning based on coupled convolutional neural networks for aircraft detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 9, pp. 5553–5563, 2016.

[22] X. Yang, K. Fu, H. Sun et al., "R2CNN++: multidimensional attention based rotation invariant detector with robust anchor strategy," vol. 2, p. 7, 2018, https://arxiv.org/abs/1811.07126.

[23] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 6, pp. 3325–3337, 2014.

[24] G. Cheng, J. Han, P. Zhou, and D. Xu, "Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 265–278, 2018.

[25] K. Li, G. Cheng, S. Bu, and X. You, "Rotation-insensitive and context-augmented object detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2337–2348, 2017.

[26] G. Cheng, Y. Si, H. Hong, X. Yao, and L. Guo, "Cross-scale feature fusion for object detection in optical remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 3, pp. 431–435, 2021.

[27] X. Qian, S. Lin, G. Cheng, X. Yao, H. Ren, and W. Wang, "Object detection in remote sensing images based on improved bounding box regression and multi-level features fusion," *Remote Sensing*, vol. 12, no. 1, p. 143, 2020.

[28] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: remote sensing image scene classification via learning discriminative cnns," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 5, pp. 2811–2821, 2018.

[29] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1923–1938, 2018.

[30] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 5966–5978, 2020.

[31] D. Hong, L. Gao, N. Yokoya et al., "More diverse means better: multimodal deep learning meets remote-sensing imagery classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 5, pp. 4340–4354, 2020.

[32] G. Cheng, P. Zhou, and J. Han, "RIFD-CNN: rotation-invariant and fisher discriminative convolutional neural networks for object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2884–2893, Las Vegas, NV, USA, June 2016.

[33] Z. Deng, H. Sun, S. Zhou, J. Zhao, L. Lei, and H. Zou, "Multi-scale object detection in remote sensing imagery with convolutional neural networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 145, pp. 3–22, 2018.

[34] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning ROI transformer for oriented object detection in aerial images," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2849–2858, Long Beach, CA, USA, June 2019.

[35] Y. Xu, M. Fu, Q. Wang et al., "Gliding vertex on the horizontal bounding box for multi-oriented object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 4, pp. 1452–1459, 2020.

[36] X. Yang, J. Yang, J. Yan et al., "SCRDet: towards more robust detection for small, cluttered and rotated objects," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8232–8241, Seoul, Korea (South), October 2019.

[37] X. Yang, Q. Liu, J. Yan, A. Li, Z. Zhang, and G. Yu, "R3Det: refined single-stage detector with feature refinement for rotating object," 2019, https://arxiv.org/abs/1908.05612.

[38] C. Li, C. Xu, Z. Cui et al., "Learning object-wise semantic representation for detection in remote sensing imagery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 20–27, Long Beach, CA, USA, 2019.

[39] S. M. Azimi, E. Vig, R. Bahmanyar, M. Körner, and P. Reinartz, "Towards multi-class object detection in unconstrained remote sensing imagery," in *Asian Conference on Computer Vision*, pp. 150–165, Springer, 2018.

[40] Y. Li, Q. Huang, X. Pei, L. Jiao, and R. Shang, "RADet: refine feature pyramid network and multi-layer attention network for arbitrary-oriented object detection of remote sensing images," *Remote Sensing*, vol. 12, no. 3, p. 389, 2020.

[41] X. Pan, Y. Ren, K. Sheng et al., "Dynamic refinement network for oriented and densely packed object detection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11207–11216, Seattle, WA, USA, June 2020.

[42] T. Tang, S. Zhou, Z. Deng, L. Lei, and H. Zou, "Arbitrary-oriented vehicle detection in aerial imagery with single convolutional neural networks," *Remote Sensing*, vol. 9, no. 11, p. 1170, 2017.

[43] L. Liu, Z. Pan, and B. Lei, "Learning a rotation invariant detector with rotatable bounding box," 2017, https://arxiv.org/abs/1711.09405.

[44] J. Han, J. Ding, J. Li, and G. S. Xia, "Align deep features for oriented object detection," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–11, 2021.

[45] X. Yao, X. Feng, J. Han, G. Cheng, and L. Guo, "Automatic weakly supervised object detection from high spatial resolution remote sensing images via dynamic curriculum learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 1, pp. 675–685, 2020.

[46] X. Feng, J. Han, X. Yao, and G. Cheng, "Progressive contextual instance refinement for weakly supervised object detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 11, pp. 8002–8012, 2020.

[47] G. Cheng, J. Yang, D. Gao, L. Guo, and J. Han, "High-quality proposals for weakly supervised object detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 5794–5804, 2020.

[48] X. Feng, J. Han, X. Yao, and G. Cheng, "TCANet: triple context-aware network for weakly supervised object detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–10, 2020.

[49] G. Cheng, R. Li, C. Lang, and J. Han, "Task-wise attention guided part complementary learning for few-shot image classification," *SCIENCE CHINA Information Sciences*, vol. 64, no. 2, pp. 1–14, 2021.

[50] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, Salt Lake City, UT, USA, June 2018.

[51] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7794–7803, Salt Lake City, UT, USA, June 2018.

[52] X. Wang, Z. Cai, D. Gao, and N. Vasconcelos, "Towards universal object detection by domain attention," in *2019 IEEE/CVF Conference on Computer Vision and Pattern*

*Recognition (CVPR)*, pp. 7289–7298, Long Beach, CA, USA, June 2019.

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[54] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921–2929, Las Vegas, NV, USA, June 2016.

[55] A. Paszke, S. Gross, S. Chintala et al., *Automatic differentiation in PyTorch*, 2017.

[56] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: a large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, Miami, FL, USA, June 2009.

[57] K. Chen, J. Wang, J. Pang et al., "MMDetection: open MMLab detection toolbox and benchmark," 2019, https://arxiv.org/abs/1906.07155.

[58] H. Law and J. Deng, "CornerNet: detecting objects as paired keypoints," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 734–750, Munich, Germany, 2018.

[59] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2961–2969, Venice, Italy, October 2017.

[60] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 761–769, Las Vegas, NV, USA, June 2016.

[61] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra R-CNN: towards balanced learning for object detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 821–830, Long Beach, CA, USA, June 2019.

[62] K. Fu, Z. Chen, Y. Zhang, and X. Sun, "Enhanced feature representation in detection for optical remote sensing images," *Remote Sensing*, vol. 11, no. 18, p. 2095, 2019.

[63] J. Yan, H. Wang, M. Yan, W. Diao, X. Sun, and H. Li, "IoU-adaptive deformable R-CNN: make full use of IoU for multi-class object detection in remote sensing imagery," *Remote Sensing*, vol. 11, no. 3, p. 286, 2019.

[64] P. Wang, X. Sun, W. Diao, and K. Fu, "FMSSD: feature-mergedsingle-shot detection for multiscale objects in large-scale remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 5, pp. 3377–3390, 2019.

[65] P. Sun, G. Chen, and Y. Shang, "Adaptive saliency biased loss for object detection in aerial images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 10, pp. 7154–7165, 2020.