CoF-Net: A Progressive Coarse-to-Fine Framework for Object Detection in Remote-Sensing Imagery

Cong Zhang, *Graduate Student Member, IEEE*, Kin-Man Lam, *Senior Member, IEEE*, and Qi Wang, *Senior Member, IEEE*

***Abstract*— Object detection in remote-sensing images is a crucial task in the fields of Earth observation and computer vision. Despite impressive progress in modern remote-sensing object detectors, there are still three challenges to overcome: 1) complex background interference; 2) dense and cluttered arrangement of instances; and 3) large-scale variations. These challenges lead to two key deficiencies, namely, coarse features and coarse samples, which limit the performance of existing object detectors. To address these issues, in this article, a novel coarse-to-fine framework (CoF-Net) is proposed for object detection in remote-sensing imagery. CoF-Net mainly consists of two parallel branches, namely, coarse-to-fine feature adaptation (CoF-FA) and coarse-to-fine sample assignment (CoF-SA), which aim to progressively enhance feature representation and select stronger training samples, respectively. Specifically, CoF-FA smoothly refines the original coarse features into multispectral nonlocal fine features with discriminative spatial–spectral details and semantic relations. Meanwhile, CoF-SA dynamically considers samples from coarse to fine by progressively introducing geometric and classification constraints for sample assignment during training. Comprehensive experiments on three public datasets demonstrate the effectiveness and superiority of the proposed method.**

***Index Terms*— Coarse-to-fine paradigms, geometric constraints, object detection, remote-sensing imagery, spatial–spectral nonlocal features.**

I. Introduction

OBJECT detection has a wide range of real-world applications, including vehicle and people detection, and construction-site object detection. It also plays a vital role in processing large-scale optical remote-sensing imagery, which is one of the most fundamental tasks in civil and military intelligence systems and has applications in emergency rescue, environmental monitoring, resource exploration, and urban planning [1], [2], [3], [4], [5]. This topic has been studied in the Earth observation community for several decades, due to its promising practical value [6], [7], [8]. In general, remote-sensing object detection aims to accurately identify the locations and categories of specific geospatial objects, such as airplanes and vehicles [9], [10], [11]. With the vigorous development of deep convolutional neural networks (CNNs), the performance of object detection in natural images has been impressively improved in recent years, further promoting the development of object detection in remote-sensing images [12], [13], [14], [15].

In order to effectively facilitate the deep learning-based geospatial object detection methods, various large-scale datasets have been publicly released, such as DOTA [16], DIOR [17], NWPU VHR-10 [6], [18], HRRSD [19], and HRSC2016 [20]. Compared with common natural scenes, remote-sensing and aerial images are usually captured from a bird's-eye view of arbitrary horizontal or oriented objects in unforeseen circumstances, making object detection much more challenging. Some algorithms have been specifically explored for rotated object detection [21], [22], [23], [24], [25], [26] and achieved improved performance in different scenarios. However, more generally and conclusively, there are still three prevailing major challenges to be tackled in remote-sensing images.

1) *Highly Complex Background Interference:* The foreground geospatial regions are easily interfered with, or even overwhelmed by, the complicated surrounding scenes, such as buildings, vegetation, and other background noises, which coarsens both the spatial details and semantic relations.
2) *Densely Packed and Cluttered Instances:* Geospatial objects with specific categories, such as vehicles, ships, and storage tanks, appear inclined in densely arranged and cluttered forms, leading to instance-level coupling and ambiguity.
3) *Large-Scale Variations:* Since remote-sensing images are often taken from various ground sampling distances with respect to diverse capture devices, the object scale varies dramatically with image resolution. Such tremendous scale variations hinder accurate object detection.

To address these issues, most state-of-the-art geospatial object detectors are constructed for better performance, based on a sophisticated two-stage R-CNN framework [27], [28], [29], which is composed of a detection stem and an additional region proposal network (RPN). Furthermore, modern mainstream methods [30], [31], [32], [33], [34], [35] are devoted to introducing external auxiliary branches or modules for adaptation to challenging remote-sensing scenes. However, these enhanced detectors suffer from a problem with the

Manuscript received 2 August 2022; revised 18 October 2022 and 7 December 2022; accepted 28 December 2022. Date of publication 3 January 2023; date of current version 11 January 2023. This work was supported by the Key-Area Research and Development Program of Guangdong Province under Grant 2020B090928001. *(Corresponding author: Cong Zhang.)*

Cong Zhang and Kin-Man Lam are with the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong (e-mail: cong-clarence.zhang@connect.polyu.hk).

Qi Wang is with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China.

Digital Object Identifier 10.1109/TGRS.2022.3233881

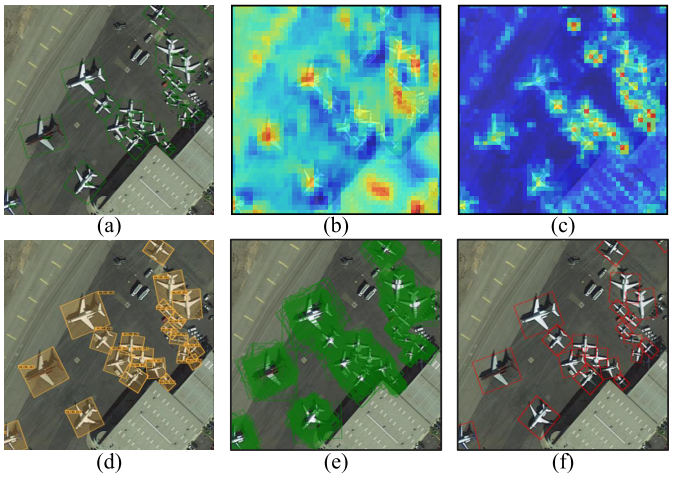

Fig. 1. Illustration of coarse features and coarse training samples that degrade object detection performance in remote-sensing images. (a) Input image with the ground truths. (b) Coarse features. (c) Fine features with noises suppressed and semantic relations enhanced. (d) Detection results based on the fine features and samples. (e) Coarse training samples. (f) Fine training samples that are accurate and explicit.

two-stage frameworks, which have high computational complexity. In contrast, with the removal of the region proposal stage, single-stage detectors directly regress the final detection results from deep features and dense anchors, achieving higher computational efficiency and fewer memory footprints. However, these methods usually struggle with reduced accuracy and robustness, especially in complicated scenarios. As shown in Fig. 1(b) and (e), the inferiority of single-stage detectors can be attributed to two critical aspects: coarse features and coarse samples. On the one hand, in view of the first challenge mentioned above, complex background interference in remote-sensing images causes difficulties in differentiating foreground geospatial regions with various appearances at the feature level. The backbone features are interfered with and misaligned by background noises, as shown in Fig. 1(b). This will result in inappropriate results, without using any effective refinement. In other words, with the degradation of spatially explicit details and semantically implicit relations, e.g., reduced subtle boundary information and global location constraints, only coarse features are generated for object localization and classification in the single-stage framework. On the other hand, referring to Fig. 1(e), redundant and near-duplicate anchors are selected during training, without using advanced assignment rules. Due to the dense and cluttered arrangements, as described in the second challenge, coarse samples result in inaccurate and confusing supervision at the instance level. These challenges make training less effective, and the detector has high computational requirements.

Concerning the two issues, i.e., coarse features and coarse samples, in this article, we explore a single-stage detection framework that performs a favorable tradeoff between performance and efficiency. The goal is to refine the coarse features and samples so as to enhance the features and the training process, making the object detector fine to conquer the aforementioned three challenges in remote-sensing scenarios. Fine features and samples are shown in Fig. 1(c) and (f), respectively. To this end, a novel progressive coarse-to-fine framework (CoF-Net) is proposed, which consists of two pivotal stages: coarse-to-fine feature adaptation (CoF-FA) and coarse-to-fine sample assignment (CoF-SA). Specifically, as shown in Fig. 2, CoF-FA and CoF-SA are designed separately as two parallel branches for model flexibility. The former forwardly refines coarse features to become finer against complex background interference (the first challenge), to prevent detail distortion and contextual confusion. The latter constrains the selection of samples from coarse to fine in stages under sparse and explicit supervision for the dense and cluttered arrangements of instances (the second challenge). Both CoF-FA and CoF-SA follow the progressive "coarse-to-fine" strategy proposed in this work, which guarantees their effectiveness and efficiency. Meanwhile, the overall feature extraction network adopts a feature pyramid network (FPN)-like architecture [36] to produce fine hierarchical features, with the capability of dealing with large-scale variations (the third challenge). Finally, extensive experiments are conducted on different challenging datasets, DOTA [16], DIOR [17], and NWPU VHR-10 [18], for horizontal and oriented object detection in remote-sensing images, which demonstrate the superiority of CoF-Net. Moreover, due to its high flexibility and efficiency, the proposed coarse-to-fine paradigm can be adapted to various real-world single-stage object detectors to improve their detection performance in complex scenarios. The main contributions of this article can be summarized as follows.

1) A novel end-to-end single-stage framework, called CoF-Net, whose most notable property is its progressive coarse-to-fine paradigm, is proposed for remote-sensing object detection. With this advanced paradigm that overcomes the crucial challenges, CoF-Net can achieve high detection accuracy and robustness.
2) The proposed CoF-FA is devised in a coarse-to-fine manner for multispectral nonlocal feature adaptation. CoF-FA distinctively enriches spatial–spectral feature details in the frequency domain and boosts the implicit semantic discrimination for fine-grained alignment.
3) The proposed CoF-SA introduces geometric and classification-aware constraints to progressively and dynamically assign samples during training, avoiding supervision disturbance. Complementary to CoF-FA, CoF-SA considerably contributes to accurate and robust localization of dense and cluttered objects in remote-sensing images.

The rest of this article is organized as follows. Section II briefly reviews the related work. The proposed CoF-Net is described in Section III. Section IV reports and analyzes the experimental results on three challenging datasets. Finally, the conclusion is drawn in Section V.

II. Related Work

Object detection has been extensively investigated in both computer vision and Earth observation communities due to its important role in various applications. In this section, first, we briefly review geospatial object detection in remote-sensing

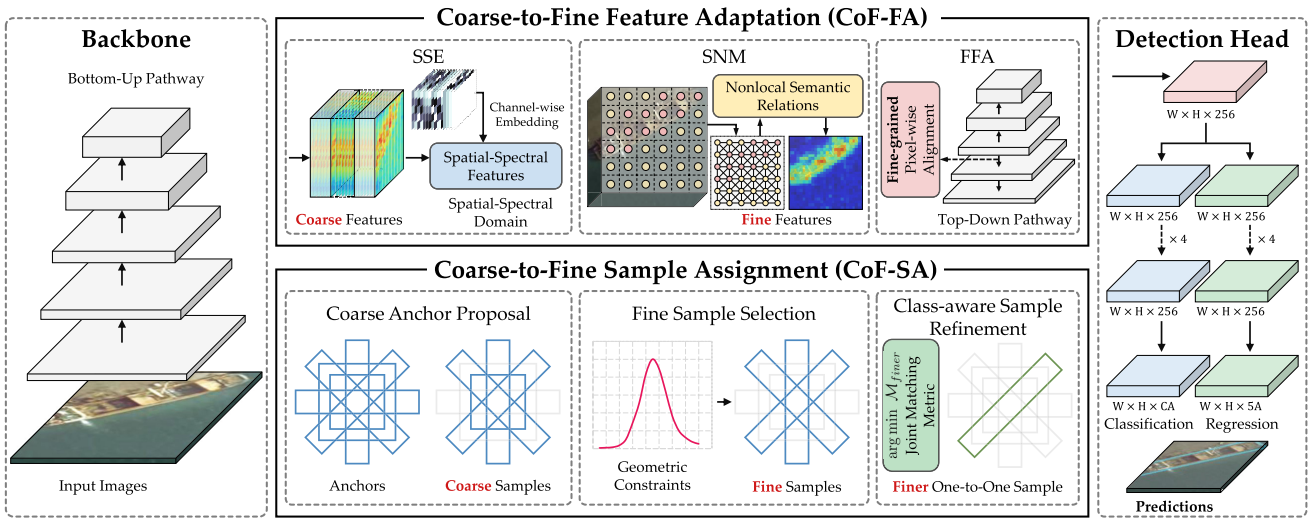

Fig. 2. Overview of the proposed CoF-Net. It consists of a backbone, a detection head, and two critical parallel branches, namely, CoF-FA and CoF-SA.

images and, then, discuss two important techniques, the visual attention mechanism and the anchor-based/anchor-free strategies.

A. Geospatial Object Detection

Geospatial object detection has received considerable attention over the past two decades since it forms the basis for high-level remote-sensing tasks such as urban planning. Different from objects in natural images, geospatial objects in remote-sensing images (such as planes, ships, and storage tanks) are usually cluttered and densely arranged with large variations in aspect ratio (AR), scale, and orientation, leading to the ineffectiveness of many generic object detection methods. In general, existing geospatial object detection methods can be divided into two categories: traditional algorithms and deep learning-based algorithms. In traditional methods, handcrafted features are designed to represent and localize specific objects, based on shape and texture information [37], saliency features [38], [39], [40], and scale-invariant features [41], [42]. For example, in [43], the histogram of oriented gradients (HOG) was employed as the feature descriptor to detect airborne vehicles in dense urban regions. Han et al. [41] proposed to integrate visual saliency and sparse coding for geospatial object localization. However, these traditional handcrafted feature-based methods are always sensitive to complex background interference, due to a lack of discriminative semantics, which limits their performance in real-world applications.

Due to the rapid development of deep learning, numerous object detection algorithms have been studied and proposed during the past few years. Compared to traditional algorithms, deep CNNs can extract powerful high-level feature representations with rich semantic information, yielding promising and robust classification and localization performance. Motivated by generic object detection, deep learning-based remote-sensing object detection methods can be categorized into two- and single-stage frameworks, according to whether or not region proposals are generated. With regard to the former, the representative Faster R-CNN architecture [29] has been widely exploited and improved for geospatial object detection [30], [32], [44], [45], [46], [47]. For instance, in [47], additional multiangle anchors were introduced into the RPN stage of Faster R-CNN, in view of the arbitrary orientations of geospatial objects. Following the pipeline of Faster R-CNN, Ye et al. [44] proposed a feature fusion and filtration network, namely, $\mathcal{F}^3$-Net, which leveraged a feature fusion module to extract multiscale contextual information and a feature filtration module to suppress the background interference, resulting in competitive detection performance. Based on FPN [36] and Faster R-CNN, Qin et al. [32] decoupled detection into several subtasks and proposed multiple heads for detecting objects with different scales. Recently, Wu et al. [45] devised a global context-weaving network to facilitate dense object detection in remote-sensing images. In [46], two transformer-based modules were designated to enhance the pixelwise representation capability and multiobject dependencies. Although two-stage detectors achieve leading detection accuracy, they suffer from slow inference speed, significantly hindering their real-world applications. On the other hand, for the purpose of real-time object detection, some researchers have explored more efficient single-stage detectors [4], [14], [16], [48], [49], [50]. Xia et al. [16] adopted YOLOv2 [51] to detect geospatial objects in remote-sensing images. A refined single-stage detector for rotated objects [50] was proposed by utilizing an advanced regression strategy. Recently, Liu et al. [48] introduced scene-contextual information related to remote-sensing objects of interest and a scene auxiliary detection head for more accurate detection. Although single-stage methods have achieved a certain degree of success, there is still a gap, in terms of accuracy, for remote-sensing object detection, compared with two-stage methods. The main reason is that these models are usually built on coarse features and samples, with limited semantic relation refinement and alignment as in two-stage detectors. Distinct from previous methods, in this work, we propose to adapt features to complicated cluttered

backgrounds from a novel spatial–spectral perspective and refine coarse features and samples following an efficient coarse-to-fine paradigm.

B. Visual Attention Mechanism

Extensive research on visual attention mechanisms has demonstrated its effectiveness in highlighting potential regions of interest, which has been widely introduced to deep learning-based vision tasks, including object detection [52]. The attention mechanism aims to reweight different image areas in terms of their importance or influence for specific visual tasks. In this way, the attention-based object detectors can better model the spatial and semantic relationships of both intraobjects and interobjects and, hence, boost the detection performance [34], [53]. Moreover, attention mechanisms can be applied to different dimensions of deep features, such as the spatial dimension [54], [55], the channel dimension [56], [57], [58], and both together [59].

Recently, some attention-based methods have also been proposed in the field of remote-sensing object detection, which achieved enhanced performance [35], [60], [61]. For instance, SCRDet [60] introduced a jointly supervised pixel and channel attention network for cluttered and small geospatial object detection. A multisource region attention network was constructed in [62] to simultaneously detect and recognize fine-grained objects by leveraging the attention-driven feature representation. Liu et al. [61] proposed the center-boundary dual attention mechanism by extracting features in the center and boundary regions of objects in remote-sensing images. Nevertheless, these aforementioned methods developed simple and suboptimal attention modules, by only introducing locally neighboring semantics, due to the limited receptive field of convolutional kernels, resulting in limited improvements. Considering the large-scale variations and complicated scenes in remote-sensing images, it is necessary to capture long-range dependencies and facilitate interactions between both global and local positions [63], [64], [65], [66], which are of vital importance for downstream vision tasks such as object detection.

C. Anchor-Based and Anchor-Free Strategies

Developed in Faster R-CNN [29], anchors are a set of predefined bounding boxes, treated as potential region candidates for classification and regression. They have played an essential role in modern object detectors and have recently attracted considerable interest. Previous Faster R-CNN-like methods [67], [68] applied the region proposal mechanism to filter out invalid anchors, and then, they concentrated only on positive samples during training. Unlike these two-stage detectors, single-stage methods usually suffer from the imbalance issue of foreground–background or positive–negative anchors. Thus, in [69], the focal loss was designed to alleviate this problem, especially for dense prediction based on single-stage detectors. After that, some anchor generation and selection mechanisms were explored successively [1], [21], [70], [71], [72], [73], [74], [75]. FreeAnchor [71] formulated object-anchor matching as the maximum likelihood estimation, which is conceptually complicated. Regarding remote-sensing object detection, in [1], with two parallel branches corresponding to classification and regression, an extra anchor refinement network was proposed to generate high-quality anchors. Ming et al. [21], [72] introduced a dynamic anchor learning strategy to adaptively evaluate the quality of predefined anchors and, then, select appropriate ones. In order to get rid of the dilemma of manually predefined anchors and mismatches, some anchor-free assignment strategies have been studied recently. For example, FCOS [76] and FoveaBox [77] predicted objects pixelwise and regressed the offset from the point to the four sides of the bounding box of each object. Regardless of anchor-based and anchor-free methods, most of these existing algorithms neglect the geometrical characteristics of specific objects, such as shape and scale, when assigning training samples. The geometrical information, however, will benefit object detection if it is taken as prior knowledge.

III. Proposed Method

Geospatial objects in remote-sensing images are characterized by large variations in scale and densely arranged distribution, leading to damaged or obscure boundaries, misaligned features, and ambiguous bounding-box regression targets. Previous object detectors, especially single-stage frameworks, fail to yield satisfactory detection performance in such cases. In order to address this issue, some approaches have been proposed toward highly complicated remote-sensing scenarios, but they still only operate on coarsely defined features or samples and directly force the deep networks to perform learning and prediction. Consequently, the methods may suffer from semantic feature degradation and result in target misclassification, further hindering the detection accuracy and robustness. Therefore, in this article, we propose CoF-Net to detect geospatial objects from remote-sensing images in a novel coarse-to-fine manner, which can enhance discriminative features and dynamically determine high-quality training samples. In this section, the proposed framework will be first overviewed, and then, the two crucial components of CoF-Net, i.e., CoF-FA and CoF-SA, will be presented in detail.

A. Framework Overview

The overall pipeline of the proposed CoF-Net is shown in Fig. 2. Built on RetinaNet [69], CoF-Net is a single-stage framework with fully convolutional layers [78], which can achieve fast inference speed, with a small number of parameters, for multiscale geospatial object detection in real-world scenarios. CoF-Net mainly consists of four components: 1) ResNet [79] as the CNN backbone for coarse feature extraction; 2) CoF-FA based on the FPN structure [36], for enhancing the coarse and misaligned features to adaptively generate fine-grained nonlocal features, containing both spectral and spatial information; 3) CoF-SA for dynamically generating and nominating finer, high-quality samples during training, instead of simply using massive coarse anchors, by introducing geometric prior and classification constraints as selection criteria; and 4) the detection head, composed of two compact and parallel subnetworks for classification and regression to produce the final detection results.

With the proposed coarse-to-fine strategy, both the deep features and training samples are refined to become more reliable and robust against complicated interference in remote-sensing images, thereby boosting the detection performance of the single-stage framework. Moreover, in order to adapt CoF-Net to detect rotated objects in remote-sensing images, following [1], [21], [32], and [50], we use five parameters (x, y, w, h, θ) for arbitrary-oriented rectangle representation, where $\theta \in [-\pi/2, 0)$ denotes the acute angle between an object and the x-axis. In addition, for those datasets that only involve horizontal objects, such as the DIOR dataset [17] and the NWPU VHR-10 dataset [6], the parameter θ is simply fixed to an arbitrary value without updating, to form the typical four-parameter representation (x, y, w, h) [69], which represents the center coordinates, width, and height, respectively.

B. Coarse-to-Fine Spectral Feature Adaptation

Aimed at extracting more robust and discriminative features with global semantic relations for foreground targets with background interference, spectral nonlocal feature adaptation is developed, based on the proposed coarse-to-fine approach. Our method considers the spectral property in remote-sensing imagery and adapts deep features to the spatial and frequency domains. As shown in Fig. 2, CoF-FA consists of three modules, spatial–spectral embedding (SSE), spectral nonlocal modulation (SNM), and fine-grained feature alignment (FFA), which will be explained in detail in this section.

1) Spatial–Spectral Embedding: As shown in Fig. 2, given an input image $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$, the feature extractor will produce feature maps for the subsequent feature-level operations. Specifically, by feeding $\mathcal{I}$ to the ResNet backbone [79], multiscale feature maps are extracted in the classic bottom-up pathway [36], denoted as $\mathcal{F}^{(i)} \in \mathbb{R}^{M_i\times N_i\times C_i}$, where $i = 3, 4$, and 5 for multiscale representations. However, limited by the receptive fields of CNN, the extracted features are susceptible to deformation and interference in remote-sensing images. The coarse granularity can be caused by two reasons, using the spatial domain representation only and global semantic deficiency.

The former motivates the SSE module to introduce spectral information to the spatial features, which can enhance feature discriminability and diversity. Discrete cosine transform (DCT) [80] is applied to convert the coarse spatial features into the frequency domain to form spectral features, which are then compressed and embedded in the coarse spatial features. Generally, given the input coarse spatial features $\mathcal{F}$ (i is removed for brevity), the DCT of $\mathcal{F}$ can be written as follows:

$$\begin{aligned}\tilde{\mathbf{F}}(k_x, k_y) &= \frac{1}{MN}\sum_{u=0}^{M-1}\sum_{v=0}^{N-1}\mathcal{F}(u, v)\odot G_{k_x,k_y}(u, v)\\ &\text{where } G_{k_x,k_y}(u, v)\\ &= \cos\left(k_x\frac{\pi\left(u+\frac{1}{2}\right)}{M}\right)\cos\left(k_y\frac{\pi\left(v+\frac{1}{2}\right)}{N}\right)\\ &\text{and } k_x \in \{0, 1, \ldots, M-1\},\quad k_y \in \{0, 1, \ldots, N-1\}\end{aligned} \tag{1}$$

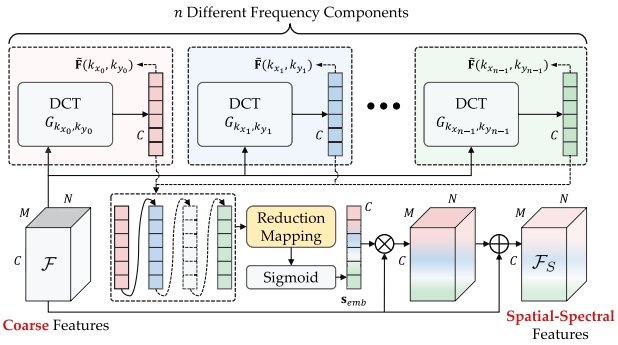

Fig. 3. Illustration of the proposed SSE module, in which the original coarse features are promoted to spatial–spectral features by embedding multiple frequency components in a channelwise manner.

$\tilde{\mathbf{F}}(k_x, k_y)$ is the spectral representation of $\mathcal{F}$ and $G_{k_x,k_y}(u, v)$ is the basis function of 2-D DCT. M and N represent the height and width of $\mathcal{F}$, respectively, while $\odot$ denotes the elementwise multiplication along the channel dimension. To boost object boundary response and discriminability, multispectral information is embedded into the original coarse features by using the basis function $G_{k_x,k_y}(\cdot)$ with different values of k_x and k_y. As shown in Fig. 3, n different frequency components, from low to high frequencies, are generated and fused into the original coarse features, based on channelwise embedding [56]. Specifically, given n different DCT bases G_{k_x,k_y}, n different spectral representations $\tilde{\mathbf{F}}(k_x, k_y)$ are first derived by (1), and then, the multispectral embedding vector can be computed as follows:

$$\begin{aligned}\mathbf{s}_{\text{emb}} &= \text{Sigmoid}\big(C\big(\big[\tilde{\mathbf{F}}(k_{x_0}, k_{y_0}); \ldots; \tilde{\mathbf{F}}(k_{x_{n-1}}, k_{y_{n-1}})\big]\big)\big)\\ &\text{s.t.}(k_x, k_y) \in S = \left\{\left(k_{x_0}, k_{y_0}\right), \ldots, \left(k_{x_{n-1}}, k_{y_{n-1}}\right)\right\}\end{aligned} \tag{2}$$

$C([;])$ is a learnable reduction mapping function after channelwise feature concatenation, which is based on 1-D convolutions, while S represents the superset of the specific combinations (k_x, k_y) with a cardinality of $|S| = n$.

To further clarify the operation of the SSE module and its advantages over other methods, suppose that $(k_x, k_y) = (0, 0)$ and $\mathbf{1}$ is an all-ones vector. Then, $G_{0,0} = \mathbf{1}\mathbf{1}'$ represents the lowest frequency basis, i.e., the zero frequency, and we have

$$\tilde{\mathbf{F}}(0, 0) = \frac{1}{MN}\sum_{u=0}^{M-1}\sum_{v=0}^{N-1}\mathcal{F}(u, v) \tag{3}$$

which can actually be regarded as the global average-pooling operation for $\mathcal{F}$, commonly utilized in [56], [57]. However, only the zero frequency or dc component is used in (3). This descriptor potentially eliminates or weakens the boundary information and retains only the coarse global distribution. To remedy this issue, as shown in Figs. 3 and 4(a), more spectral components are utilized in our method for multiple frequency embedding. Different spectra are generated based on the corresponding basis frequency functions, as visualized in Fig. 4(b), where the cosine bases are orthogonal to each other. Consequently, with the channelwise multispectral embedding vector $\mathbf{s}_{\text{emb}} \in \mathbb{R}^C$, the output features of the SSE module can

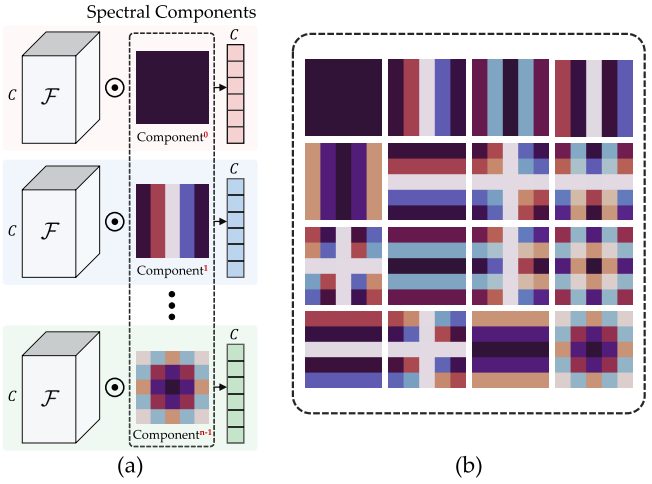

Fig. 4. (a) Based on the coarse features and different spectral components, multiple spectral vectors are generated by elementwise multiplication. (b) Visualization of 16 different spectral components, from low to high frequencies.

be formulated as follows:

$$\mathcal{F}_S = \lambda \mathbf{s}_{\text{emb}} \otimes \mathcal{F} + (1-\lambda)\mathcal{F} \tag{4}$$

where $\lambda \in (0, 1)$ controls the relative importance of the two terms and $\otimes$ denotes the tensor product. In this way, with negligible extra computation, all the details and key frequency information about object boundaries are aggregated into the spatial–spectral features.

2) Spectral Nonlocal Modulation: The SSE module converts coarse features from the spatial domain to the spatial–spectral domain. However, as mentioned above, there is still another challenge, i.e., deficiency in nonlocal dependencies. Previous methods [35], [60], [61] applied simple attention mechanisms to enhance the discrimination of features, but their performance is limited, due to the local receptive fields of the convolutional kernels used. Local attention is usually suboptimal, especially for detecting rotated and scale-varying geospatial objects in remote-sensing images, as modeling long-range dependencies in the feature space has been proven to be critical for many computer vision tasks [63]. Therefore, we introduce the SNM module to further refine the spatial–spectral features by incorporating both short- and long-range visual dependencies to flexibly construct nonlocal semantic relations. Unlike previous spatial-nonlocal blocks, which inevitably aggregate background noise [63], the proposed SNM is developed and analyzed theoretically from the spectral perspective. More importantly, following [66] and [81], the affinity matrix is designed to be symmetric to enhance its stability for suppressing background noise and boosting the response of the foreground regions.

For clarity, the regular nonlocal operator can be unified into matrix form, as in the following [65], [66]:

$$\mathcal{Y} = \mathcal{T}\left(\mathbf{X}_{\mathcal{F}}; \mathbf{W}_\theta, \mathbf{W}_\phi, \mathbf{W}_g\right) + \mathcal{F} \tag{5}$$

where $\mathbf{X}_{\mathcal{F}} \in \mathbb{R}^{MN\times C}$ is the spatially collapsed input feature matrix from $\mathcal{F}$ and $\mathbf{W}_{\theta,\phi,g} \in \mathbb{R}^{C\times C}$ are defined as the transformation weight matrices. Specifically, in Fig. 5, the proposed SNM module takes the spatial–spectral features $\mathcal{F}_S \in \mathbb{R}^{M\times N\times C}$ as input, which is fed into three 1×1 convolution blocks for channel reduction and feature transformation to generate three collapsed outputs $\mathbf{X}_{S;\theta,\phi,g} \in \mathbb{R}^{MN\times C_s}$. Then, $\mathbf{X}_{S;\theta}$ and $\mathbf{X}_{S;\phi}$ are exploited to generate the symmetric affinity matrix $\mathbf{A}$, while $\mathbf{X}_{S;g}$ maintains the spectral structure context for subsequent modulation. In the proposed SNM module, (5) can be realized and rewritten as follows:

$$\begin{aligned}\mathcal{F}_N &= \mathcal{T}_S\left(\mathbf{X}_{S;g}, \mathbf{A}\right) + \mathcal{F}_S \\ &= \mathbf{X}_{S;g}\mathbf{W}_\alpha + \mathbf{A}\mathbf{X}_{S;g}\mathbf{W}_\beta + \mathcal{F}_S \\ \text{s.t.}\mathbf{A} &= \mathbf{D}^{-\frac{1}{2}}\tilde{\mathbf{K}}\mathbf{D}^{-\frac{1}{2}}, \quad \tilde{\mathbf{K}} = \frac{\mathbf{K}+\mathbf{K}'}{2} \\ \mathbf{K} &= \mathbf{X}_{S;\theta}\mathbf{X}_{S;\phi}\end{aligned} \tag{6}$$

where $\mathcal{T}_S$ represents the spectral nonlocal modulator and $\mathbf{W}_\alpha, \mathbf{W}_\beta \in \mathbb{R}^{C_s\times C}$ are the transformation matrices used for feature restoration. The symmetric and normalized affinity matrix $\mathbf{A}$ is computed in (6), where $\mathbf{K}$ and $\mathbf{K}'$ are the pairwise pixel-similarity matrix and its transpose, respectively, and $\mathbf{D}$ denotes the diagonal degree matrix of $\tilde{\mathbf{K}}$. Therefore, with the SNM module, spectral properties can be stably preserved in the affinity matrix $\mathbf{A}$, while the nonlocal semantics can be effectively modulated in the output $\mathcal{F}_N$ to strengthen its discrimination.

3) Fine-Grained Feature Alignment: Based on the proposed SSE and SNM, the two aforementioned crucial problems caused by coarse granularity, which adversely affect feature robustness and discrimination, are mitigated effectively. With the proposed progressive feature adaptation strategy, the coarse features from the backbone become finer, in terms of both spatial–spectral details and semantics, suitable for potentially pixelwise operations, such as alignment. In Fig. 6, following the typical bottom-up top-down scheme to construct a hierarchical feature pyramid [36], the high-level features are upsampled for aggregation with the corresponding lower level features in the top-down pathway. However, the feature misalignment issue between the improved fine features and the coarse upsampled features may occur if the vanilla FPN is simply adopted as the neck [82]. Feature misalignment or aliasing is mainly caused by the cumulatively coarse and non-learned upsampling operations, such as bilinear interpolation, without accurate correspondence, which in turn destroys the refined features generated from the proposed SSE and SNM modules. To tackle this problem, the FFA module is devised to adaptively learn a fine-grained pixelwise transformation. Conceptually, supposing that the hierarchical features, after aggregation in the top-down pathway, are denoted as $\mathcal{P}^{(i)} \in \mathbb{R}^{M_i\times N_i\times \hat{C}_i}$, as shown in Fig. 6, which is generally produced from its higher level features, $\mathcal{P}^{(i+1)} \in \mathbb{R}^{(M_i/2)\times(N_i/2)\times \hat{C}_{i+1}}$ by upsampling, i.e., $\mathcal{P}^{(i)} = U(\mathcal{P}^{(i+1)}) + \mathcal{F}^{(i)}$, where $U(\cdot)$ denotes the regular upsampling operation. In contrast, in view of its misalignment, before feature merging, the FFA module adaptively adjusts and aligns $\mathcal{P}^{(i+1)}$ with reference to the accurate spatial distribution and semantics in the robust fine features $\mathcal{F}_N^{(i)}$. The operations in FFA for feature alignment can

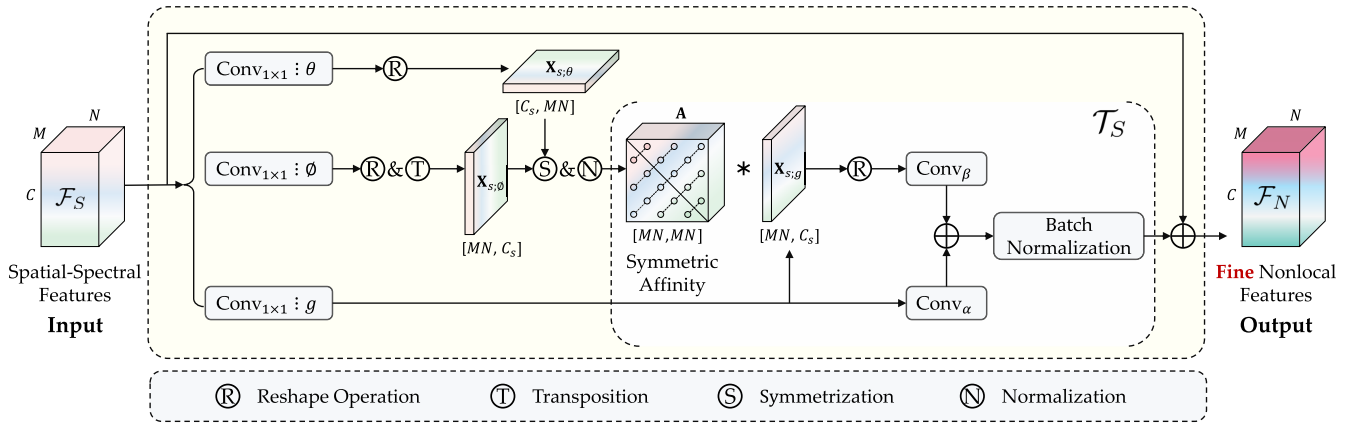

Fig. 5. Detailed structure of the proposed SNM module, which strengthens the spatial–spectral features by introducing nonlocal contextual dependencies and still retaining the spectral property.

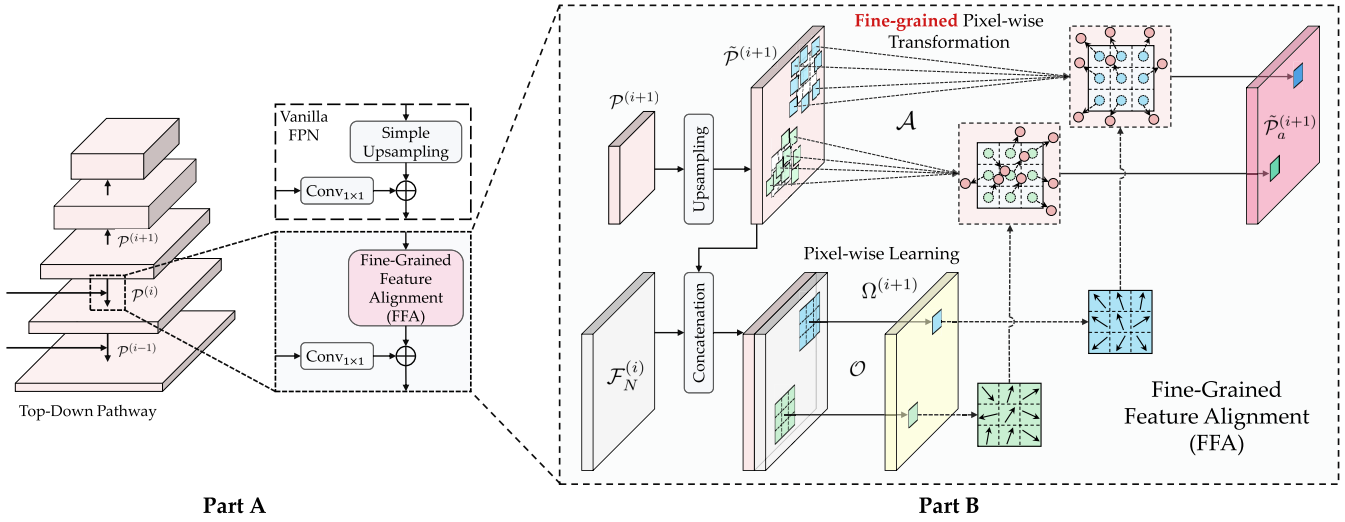

Fig. 6. Illustration of the proposed FFA module. Part A shows the difference between the FFA-based and vanilla FPN architectures, while Part B depicts the detailed procedure of aligning coarse features and fine features at the fine-grained pixel level.

be expressed as follows:

$$\mathcal{P}^{(i)} = \eta \tilde{\mathcal{P}}_a^{(i+1)} + (1-\eta)\mathcal{F}_N^{(i)} \tag{7}$$

where $\tilde{\mathcal{P}}_a^{(i+1)}$ represents the aligned upsampled features and $\eta \in (0, 1)$ is the normalization coefficient. Specifically, as shown in Fig. 6, with the learnable pixelwise offsets $\Omega^{(i+1)}$, we have

$$\tilde{\mathcal{P}}_a^{(i+1)} = \mathcal{A}\left(\tilde{\mathcal{P}}^{(i+1)}, \Omega^{(i+1)}\right) \tag{8}$$

$$\text{s.t. } \Omega^{(i+1)} = \mathcal{O}\left(\left[\mathcal{F}_N^{(i)}; \tilde{\mathcal{P}}^{(i+1)}\right]\right), \quad \tilde{\mathcal{P}}^{(i+1)} = U\left(\mathcal{P}^{(i+1)}\right) \tag{9}$$

$\mathcal{A}(\cdot)$ and $\mathcal{O}(\cdot)$ denote the offset-based fine-grained aligning function and the pixelwise difference learning function, respectively, implemented by deformable convolution [83], [84], and [;] denotes the concatenation operation to yield spatial differences between the fine $\mathcal{F}_N^{(i)}$ and coarse $\tilde{\mathcal{P}}^{(i+1)}$.

Thus, with the three proposed modules, the original coarse features $\mathcal{F}^{(i)}$ are progressively refined and adapted to the final finer features $\mathcal{P}^{(i)}$ for a detection head. These features become more discriminative and robust for remote-sensing object detection with complex background interference and scale variations.

C. Coarse-to-Fine Sample Assignment Strategy

Although the CoF-FA stage successfully constructs feature adaptation in a coarse-to-fine manner, there is still an issue to be solved for coarse training samples, parallel to the coarse feature representations, as shown in Figs. 1 and 2. To this end, this section explains the proposed CoF-SA strategy, referred to as CoF-SA, which dynamically selects higher quality samples during different training phases to improve the model's learning capability in complicated scenarios. As shown in Fig. 7, following the coarse–fine–finer scheme and the corresponding matching metrics, the number of training samples changes from dense to sparse, but with more precise supervision. This process is detailed in Algorithm 1. CoF-SA can be divided into three sequential training stages, namely, coarse anchor proposal, fine sample selection via geometric constraints, and finer class-aware sample refinement.

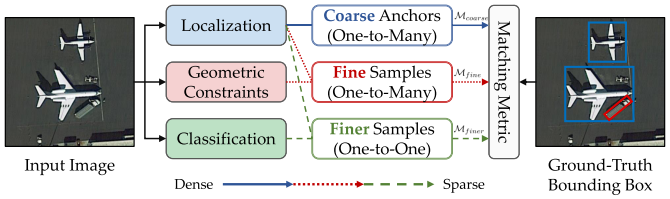

Fig. 7. Concept of the proposed CoF-SA strategy, which progressively assigns more precise samples for training in the coarse–fine–finer scheme. $\mathcal{M}_{\text{coarse}}$, $\mathcal{M}_{\text{fine}}$, and $\mathcal{M}_{\text{finer}}$ denote three progressive matching metrics.

Algorithm 1 Dynamic Coarse-to-Fine Sample Assignment

Input: The training image $\mathcal{I}$
Hierarchical features $\mathcal{P}^{(3)}, \mathcal{P}^{(4)}, \mathcal{P}^{(5)}, \mathcal{P}^{(6)}, \mathcal{P}^{(7)}$
The set of ground-truth bounding boxes $\mathbf{T}^*$
Output: The set of positive samples $\mathbf{A}$ for training
Initialization: The set of candidate anchor boxes $\mathbf{S}_a$ and three iteration (or epoch) thresholds $\mathcal{N}_{\text{coarse}}$, $\mathcal{N}_{\text{fine}}$, and $\mathcal{N}_{\text{finer}}$ for the coarse-fine-finer training strategy

for $i = 1$ **to** *max_iteration* **do**
  Coarse Anchor Proposal:
  if $i \leq \mathcal{N}_{\text{coarse}}$ **then**
    Calculate the location distance $\mathcal{D}_{\text{loc}}$
    Obtain $\mathcal{M}_{\text{coarse}}$ according to Eq. (10)
    Calculate *coarse* positive samples $\mathbf{A}_{\text{co}}$ by Eq. (11)
    $\mathbf{A} \leftarrow \mathbf{A}_{\text{co}}$
  Fine Sample Selection:
  else if $i \leq \mathcal{N}_{\text{coarse}} + \mathcal{N}_{\text{fine}}$ **then**
    Initialize the category number $\mathcal{C}$
    Obtain the prior probability $p^{(c)}_{\mathbf{a},\mathbf{t}}$
    Calculate geometric constraints $\mathcal{D}_{\text{geo}}$ by Eq. (12)
    Calculate $\mathcal{M}_{\text{fine}}$ according to Eq. (13)
    Calculate *fine* positive samples $\mathbf{A}_{\text{ge}}$ by Eq. (14)
    $\mathbf{A} \leftarrow \mathbf{A}_{\text{ge}}$
  Finer Class-aware Sample Refinement:
  else if $i \leq \mathcal{N}_{\text{coarse}} + \mathcal{N}_{\text{fine}} + \mathcal{N}_{\text{finer}}$ **then**
    Calculate classification distance $\mathcal{D}_{\text{cls}}$
    Calculate $\mathcal{M}_{\text{finer}}$ according to Eq. (15)
    Calculate *finer* positive samples $\mathbf{A}_{\text{fi}}$ by Eq. (16)
    $\mathbf{A} \leftarrow \mathbf{A}_{\text{fi}}$
  end if
end for
return $\mathbf{A}$, $\mathbf{T}^*$

1) Coarse Anchor Proposal: For most anchor-based object detectors, dense anchors are predefined and determined as positive–negative training samples. CoF-SA generates category-agnostic coarse anchors at the beginning of the training stage, e.g., in the first $\mathcal{N}_{\text{coarse}}$ epochs, when the framework usually cannot yield reliable classification results and requires a large number of preset anchors as learning candidates. Specifically, based on the coarse-to-fine hierarchical features $\mathcal{P}^{(i)}$ from CoF-FA, a number of anchors are predefined with fixed initial areas and scales at each level. Then, in the preliminary training stage, positive anchors are coarsely collected from all the initial areas and assigned to each target in a one-to-many manner, considering only the location-matching distances. Given the bounding box of the target $\mathbf{t} \in \mathbf{T}^*$ and the candidate anchor $\mathbf{a} \in \mathbf{S}_a$, the matching metric can be computed as follows:

$$\mathcal{M}_{\text{coarse}} = \mathcal{D}_{\text{loc}}(\mathbf{a}, \mathbf{t}). \tag{10}$$

In practice, the location distance $\mathcal{D}_{\text{loc}}(\cdot)$ is defined as the intersection-over-union (IoU) between the target and all anchors, which is only based on the spatial location and bounding-box intersection, without considering the category knowledge.

Furthermore, based on the whole anchor set $\mathbf{S}_a$ and a threshold ε_{co}, positive samples can be determined as follows:

$$\mathbf{A}_{\text{co}} = \{\mathbf{a}_{\text{co}} | \mathcal{M}_{\text{coarse}} < \varepsilon_{\text{co}}, \mathbf{a}_{\text{co}} \in \mathbf{S}_a\} \tag{11}$$

where $\mathbf{a}_{\text{co}}$ represents an anchor selected from $\mathbf{S}_a$. Consequently, all positive samples, as in (10) and (11), are exploited during training without any semantic constraints, such as object categories. This is a coarse and unrobust sample assignment strategy, considering only densely packed and small geospatial objects in remote-sensing images. In this way, a large number of anchors survive, particularly suitable for the initial unsophisticated training phase.

2) Fine Sample Selection via Geometric Constraints: Most existing remote-sensing object detectors [14], [30], [85] simply adopt a similar coarse sample assignment strategy throughout

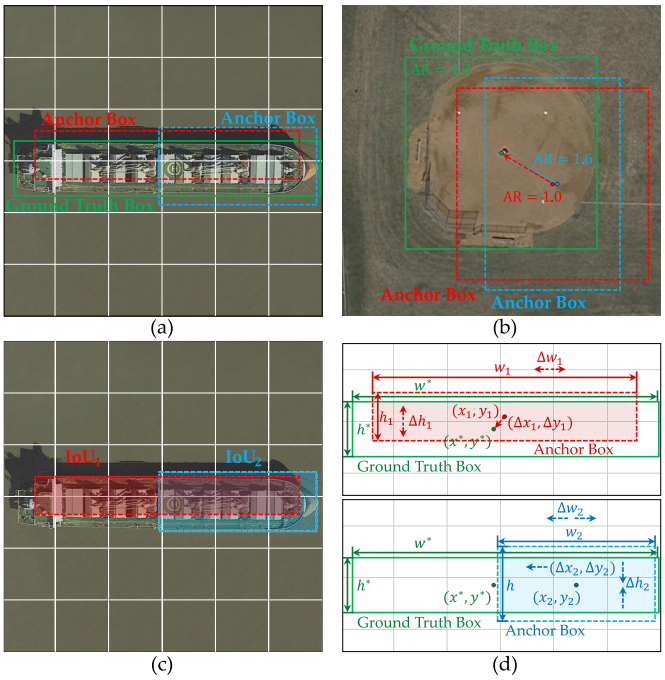

Fig. 8. Effect of selecting samples with different ARs for training. (a) Two anchors with the same IoU but different ARs are shown. The green solid box represents the ground truth, while the red and blue dashed boxes are two anchors with different ARs. (b) Another sample with an AR of 1.0 is presented. It can be seen that the same IoU cannot indicate that the two anchors are equally accurate, while AR can imply how well an anchor spatially matches its ground truth. (c) Despite $\text{IoU}_1 = \text{IoU}_2$, the red anchor in (a) with an approximate AR to the ground truth can capture more discriminative features, such as the hull and stern, which facilitate classification and localization. (d) Procedure and difficulty of regressing two anchors with different ARs to the ground truth.

the whole training process. However, such a static sample assignment method is suboptimal. On the one hand, with the progress of training, tiling numerous training samples cannot contribute to detection accuracy but leads to ineffective redundant computation. It is reasonable to dynamically reduce the number of positive samples so that training is focused on competitive sample candidates with critical discrimination. On the other hand, empirically limiting the number of samples, e.g., by increasing the threshold ε_{co}, cannot enhance the detection performance. This actually implies the bottleneck and deficiency of the category-agnostic IoU-guided sample assignment strategies introduced in Section III-C1. As shown in Figs. 1 and 12, during the later phase of training, there are still plenty of imprecise anchors treated as positives, which may adversely affect the accuracy of detecting densely arranged geospatial objects.

For common nonrigid objects in natural scenes, such as cats and dogs, the geometric characteristics of instances in the same category are often different, due to diverse shooting angles, postures, and object deformation. This leads to the domination of the IoU-based metrics over sample assignment for generic object detection, which is simply adopted in remote-sensing object detection. Different from generic visual objects in heads-up views, geospatial targets are always rigid objects captured by the top-down perspective, whose geometric properties can be explored as strong prior knowledge yet ignored in previous detectors. Relatively insensitive to rotation and scale variations, AR is the most representative and robust geometric attribute for geospatial objects, which usually conforms to a specific statistical distribution. Fig. 8 presents two selected anchors with the same IoU but different ARs, where AR is defined as the ratio of the long side to the short side. Compared to the blue anchor, the red anchor is more similar to the inherent AR of the target under consideration, thereby capturing more discriminative feature regions. Leveraging AR as a geometric constraint can effectively reduce the number of positives and make them more precise. More importantly, due to the rigidity of geospatial objects, ARs normally vary with their categories instead of orientations, which explicitly indicates the statistical prior knowledge. Therefore, we naturally introduce this geometric constraint for fine sample selection in the middle of the training phase to select more accurate samples. Specifically, considering the similar ARs of different object categories in remote sensing and the limited classification performance during intermediate training, it is unnecessary to clarify the specific category of each sample. Alternatively, the positives can be grouped into several clusters according to their different prior AR distributions. As shown in Fig. 9, there are five different clusters whose ARs are subject to their respective statistical distributions. In Fig. 9, the possible ARs are sliced into discrete intervals to form histograms, which are then normalized to form the distribution of each cluster. Given the anchor–target pair $(\mathbf{a}, \mathbf{t})$ and the discretized AR $\widehat{\text{AR}}_{\mathbf{a},\mathbf{t}} \sim N_c$ for cluster c, the prior probability $p^{(c)}_{\mathbf{a},\mathbf{t}}$ can be simply derived from Fig. 9(b). Then, the geometric constraint is defined as the cross-entropy-like cost as follows:

$$\mathcal{D}_{\text{geo}}(\mathbf{a}, \mathbf{t}) = \mathcal{D}_{\text{geo}}\left(p^{(c)}_{\mathbf{a},\mathbf{t}}\right) = -\left(1 - p^{(c)}_{\mathbf{a},\mathbf{t}}\right) \log p^{(c)}_{\mathbf{a},\mathbf{t}}. \quad (12)$$

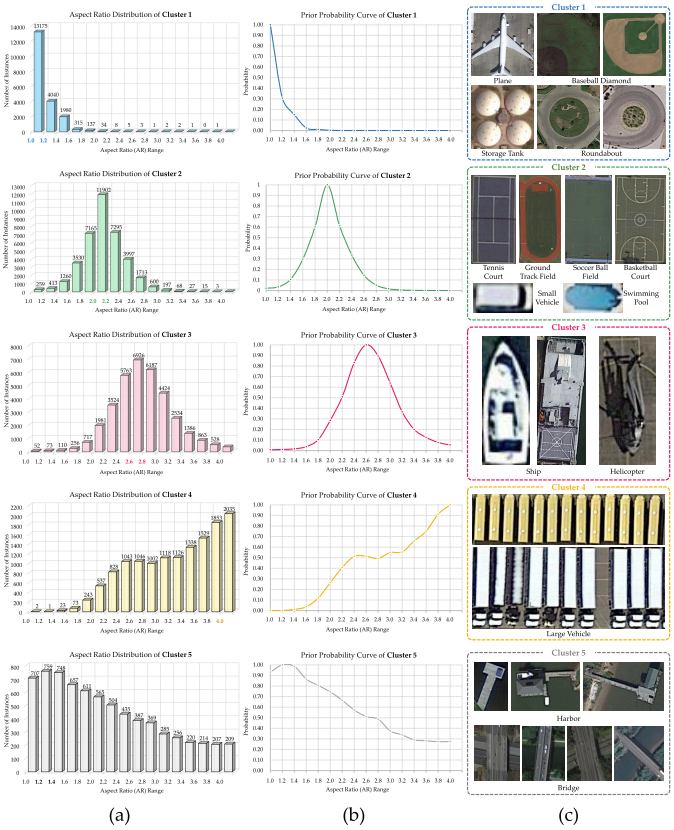

Fig. 9. AR-based prior probability distributions utilized for geometric constraints. (a) Clusterwise histograms of the number of instances against ARs. (b) Normalized prior probability functions. (c) Illustration of five different clusters in the DOTA dataset [16].

Thus, after training with coarse samples for a certain period, the sample selection criteria in (10) and (11) will be updated to

$$\mathcal{M}_{\text{fine}} = \mathcal{D}_{\text{loc}}(\mathbf{a}, \mathbf{t}) + \mathcal{D}_{\text{geo}}(\mathbf{a}, \mathbf{t}) \quad (13)$$

$$\mathbf{A}_{\text{ge}} = \{\mathbf{a}_{\text{ge}} | \mathcal{M}_{\text{fine}} < \varepsilon_{\text{ge}}, \mathbf{a}_{\text{ge}} \in \mathbf{S}_a\} \quad (14)$$

where $\mathcal{M}_{\text{fine}}$ and $\mathbf{a}_{\text{ge}}$ denote the updated matching metric and the fine samples constrained by both IoU and the geometric prior knowledge, respectively. Theoretically, in (13), such geometric constraints are added to the sample assignment metric as a prior regularizer, effectively refining the training samples and avoiding the struggle of the ongoing training on spatially and semantically inaccurate samples.

3) Finer Class-Aware Sample Refinement: As introduced above, sample assignment is dynamically divided into three stages in a progressive coarse–fine–finer manner throughout the training. To bridge the gap from coarse to fine granularity, the second stage exploits geometric prior constraints to refine coarse samples. However, there are still two significant problems to be alleviated. First, the cluster-based probabilities in (12) merely provide coarse-grained category instruction for sample selection, while category distances between samples and targets remain undiscovered. Second, in both the first and second stages, redundant and dense foreground samples should be calculated for each target, but it is ineffective to tile these duplicate candidates as one-to-many assignments in the

latter training phase [75]. To address these two issues, a finer class-aware one-to-one assignment rule is proposed to select the most appropriate sample for each potential candidate. Specifically, the matching metric in (13) is updated to

$$\mathcal{M}_{\text{finer}} = \mathcal{D}_{\text{loc}}(\mathbf{a}, \mathbf{t}) + \mathcal{D}_{\text{cls}}(\mathbf{a}, \mathbf{t}) \tag{15}$$

where $\mathcal{D}_{\text{cls}}(\cdot)$ represents the class-aware distance for finer sample assignment. In this work, it is naturally defined as the classification cost between the predicted category of the sample $\mathbf{a}$ and the ground-truth category of $\mathbf{t}$. In the latter training stage, the network discriminability is dramatically improved compared to the earlier training stages, making the classification more robust and reliable. This reveals why we adopt the "clusters" in (12), but finer "classes" here. More importantly, once the class-aware matching metric $\mathcal{M}_{\text{finer}}$ is computed, a finer one-to-one assignment rule can be applied by searching for the minimum distance positive samples as follows:

$$\begin{aligned} \mathbf{A}_{\text{fi}} &= \{\mathbf{a}_{\text{fi}}\} \\ \text{s.t.}\mathbf{a}_{\text{fi}} &= \underset{\mathbf{a}\in\mathbf{S}_a}{\arg\min}\ \mathcal{M}_{\text{finer}}(\mathbf{a}, \mathbf{t}). \end{aligned} \tag{16}$$

In this way, only sparse and finer samples are assigned, while their neighboring candidates are eliminated. Since CoF-SA dynamically and progressively introduces both spatial location and semantic category distances in a coarse-to-fine manner, it alleviates the inconsistency between the matching metric and loss function, thereby facilitating optimization. Moreover, it focuses on the most discriminative regions during training, primarily benefiting densely arranged and cluttered object detection in remote-sensing images.

IV. Experiments and Analysis

In this section, comprehensive experiments on both horizontal and oriented geospatial object detection were conducted to demonstrate the effectiveness and superiority of the proposed CoF-Net. We first introduce the datasets and evaluation metrics, and then, the components of CoF-Net, including CoF-FA and CoF-SA, are evaluated in ablation studies. Finally, the quantitative and qualitative results of our method on three public datasets are shown and analyzed, and compared to other state-of-the-art methods.

A. Dataset Description

To extensively evaluate the proposed framework, three representative and public datasets, namely, DIOR [17], NWPU VHR-10 [6], [18], and DOTA [16], are employed in our experiments.

1) DIOR Dataset: DIOR [17] is currently the largest public dataset for horizontal object detection in optical remote-sensing imagery, which contains 23 463 images. We divide it into three subsets, training set, validation set, and test set, with a ratio of 1:1:2, respectively. This dataset covers 20 different categories of geospatial objects, denoted as c1–c20 in the experiments: airplane (c1), airport (c2), baseball field (c3), basketball court (c4), bridge (c5), chimney (c6), dam (c7), expressway service area (c8), expressway toll station (c9), golf course (c10), ground track field (c11), harbor (c12), overpass (c13), ship (c14), stadium (c15), storage tank (c16), tennis court (c17), train station (c18), vehicle (c19), and windmill (c20).

2) NWPU VHR-10 Dataset: NWPU VHR-10 is another publicly available dataset for horizontal remote-sensing object detection, which was released by Cheng et al. [18]. This dataset consists of 800 very-high-resolution (VHR) remote-sensing images involving ten categories, including airplanes, ships, storage tanks, baseball diamonds, tennis courts, basketball courts, ground track fields, harbors, bridges, and vehicles. For a fair comparison in the experiments, following [6] and [18], 75% of the images are randomly selected as the training set, and the rest are used for testing.

3) DOTA Dataset: DOTA [16] is a large-scale dataset mainly for oriented object detection, which consists of 2806 aerial images, categorized into 15 geospatial classes. $1/2$, $1/6$, and $1/3$ of the original images are randomly selected for training, validation, and testing, respectively. The categories in the DOTA dataset are defined as plane (PL), baseball diamond (BD), bridge (BR), ground field track (GFT), small vehicle (SV), large vehicle (LV), ship (SH), tennis court (TC), basketball court (BC), storage tank (ST), soccer ball field (SBF), roundabout (RA), harbor (HA), swimming pool (SP), and helicopter (HC). Moreover, with a wide range of object scale and shape variation, spatial resolution, random arrangement, arbitrary orientations, and imaging conditions, this dataset is considerably challenging, which properly meets the goal to validate the effectiveness of our proposed coarse-to-fine strategy in CoF-Net.

B. Evaluation Metrics and Implementation Setup

To quantitatively evaluate the performance of different detectors, we adopt the widely used average precision (AP) as the evaluation metric. Specifically, all detection results can be categorized into four cases, true positive (TP), false positive (FP), true negative (TN), and false negative (FN). By calculating the number of samples for each case, the precision rate and recall rate can be expressed as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{17}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{18}$$

Precision represents the proportion of correctly predicted positives in all predicted positives, while recall measures the proportion of TPs to the total positives. Then, AP can be defined as the integral area of precision values over all recall values from 0 to 1. For multiclass object detection, the mAP represents the mean of APs of all categories, which is calculated as follows:

$$\text{mAP} = \frac{1}{\mathcal{C}} \sum_{c=1}^{\mathcal{C}} \text{AP}_c \tag{19}$$

where $\mathcal{C}$ represents the number of object categories. A higher mAP generally indicates better detection performance.

In the experiments, unless otherwise specified, ResNet50 is utilized as the backbone of the proposed framework, whose

TABLE I
PERFORMANCE EVALUATION AND COMPARISON OF THE PROPOSED COF-FA AND COF-SA

Method		PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP (Δ)	Param (M)	Speed (FPS)
CoF-FA	CoF-SA																		
×	×	89.26	69.66	34.20	66.34	65.95	70.84	79.45	87.99	55.65	81.26	62.76	65.32	37.65	59.80	56.54	65.51	37.56	**21**
✓	×	89.15	70.29	46.13	74.46	**66.52**	75.80	83.75	88.87	71.91	87.38	75.85	70.52	43.81	65.01	65.83	71.69 **(+6.18)**	39.43	18
×	✓	89.69	72.94	41.60	69.35	65.17	77.28	82.18	**89.69**	55.47	83.90	72.80	66.05	54.53	58.83	58.95	69.23 **(+3.72)**	37.56	**21**
✓	✓	**89.77**	**73.87**	**53.55**	**76.92**	65.22	**81.25**	**85.43**	89.67	**72.69**	**89.26**	**81.07**	**69.92**	**56.59**	**67.73**	**67.28**	**74.68 (+9.17)**	39.43	18

initial weights have been pretrained on the ImageNet dataset [86]. We train the models for a total of 24 epochs. The momentum stochastic gradient descent (SGD) optimizer is adopted during training, with an initial learning rate of 0.0001, which is decreased by 0.1 on the 12th and 18th epochs, and the warm-up strategy is also implemented for all models. The momentum is 0.9 and the weight decay is 0.0001. The confidence score threshold is set to 0.05, and the nonmaximum suppression (NMS) threshold is 0.3. We train the models on four NVIDIA Tesla V100 GPUs with a batch size of 8.

C. Ablation Studies

To verify the effectiveness of the proposed method, comprehensive ablation studies are carried out on the oriented object detection-specific DOTA validation set, and the details are shown in this section.

1) Effect of CoF-FA and CoF-SA: As CoF-FA and CoF-SA are parallel branches against coarse features and coarse training samples, respectively, their performances are evaluated separately in ablation experiments. As shown in Table I, without CoF-FA and CoF-SA, the baseline model achieves only 65.51% mAP. Only using CoF-FA to refine coarse features into fine features, it achieves 71.69% mAP, leading to an improvement of 6.18%. This demonstrates that CoF-FA can alleviate the negative effects of background noise and enrich semantic knowledge in the generated fine features. When only CoF-SA is employed without CoF-FA, the detection result is 69.23%, representing an improvement of 3.72%, compared to the baseline. This gain validates the contribution of coarse-to-fine samples to the overall model performance during training. Moreover, CoF-FA and CoF-SA can work well cooperatively (i.e., CoF-Net) to achieve the best result in detecting oriented objects, i.e., 74.68% mAP, which is about 9.17% higher than the baseline model. As reported in the last row of Table I, CoF-Net consistently improves the accuracy in almost all categories.

2) Component Analysis for CoF-FA: As described above, the three key components, SSE, SNM, and FFA, constitute CoF-FA, and their effectiveness is quantitatively and individually evaluated in Tables II–IV and Fig. 10.

a) Effect of each component of CoF-FA: Table II first reports the results of four experiments by ablating the three components. It can be seen that simply adding SSE can facilitate the baseline model to achieve a gain of about 4.10% mAP. When further utilizing "SSE + SNM" and "SSE + FFA," the improvement increases to 5.75% and 5.23%, respectively. If all three proposed modules are activated in CoF-FA, the best detection performance of 71.69% mAP can be obtained, with a total of improvement of 6.18%, as shown in the last row of

TABLE II
ABLATION STUDIES FOR EACH PROPOSED COMPONENT IN COF-FA

Model	SSE	SNM	FFA	mAP (Δ)	Param (M)	Speed (FPS)
Baseline	–	–	–	65.51	37.56	**21**
CoF-FA	✓	×	×	69.61 (+4.10)	**38.25 (+0.69)**	20
	✓	✓	×	71.26 (+5.75)	38.90 (+1.34)	19
	✓	×	✓	70.74 (+5.23)	38.77 (+1.21)	19
	✓	✓	✓	**71.69 (+6.18)**	39.43 (+1.87)	18

TABLE III
ABLATION STUDIES WITH DIFFERENT SETTINGS OF THE PROPOSED SNM

Model	Stage 1	Stage 2	Stage 3	mAP (Δ)	Param (M)	Speed (FPS)
SNM	×	×	×	69.61	38.25	**20**
	✓	×	×	71.26 (+1.65)	**38.90 (+0.65)**	19
	×	✓	×	70.63 (+1.02)	40.87 (+2.62)	17
	×	×	✓	69.89 (+0.28)	48.73 (+10.48)	15
	✓	✓	✓	**71.88 (+2.27)**	52.01 (+13.76)	14

TABLE IV
COMPARISON OF THE PROPOSED FFA AND VANILLA FPN

Vanilla FPN	FFA	mAP (Δ)	Param (M)
✓	×	65.51	37.56
×	×	61.47 (-4.04)	33.69 (-3.77)
×	✓	**66.58 (+1.07)**	38.08 (+0.52)

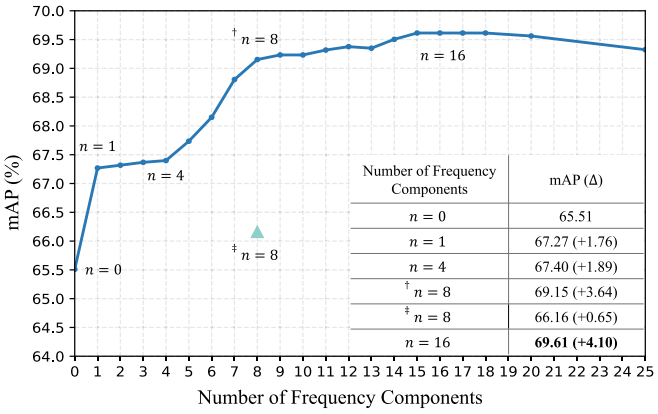

Fig. 10. Ablation experiments of the number of frequency components in SSE. † means that the first eight spectra are utilized, while ‡ means that the last eight spectra are employed.

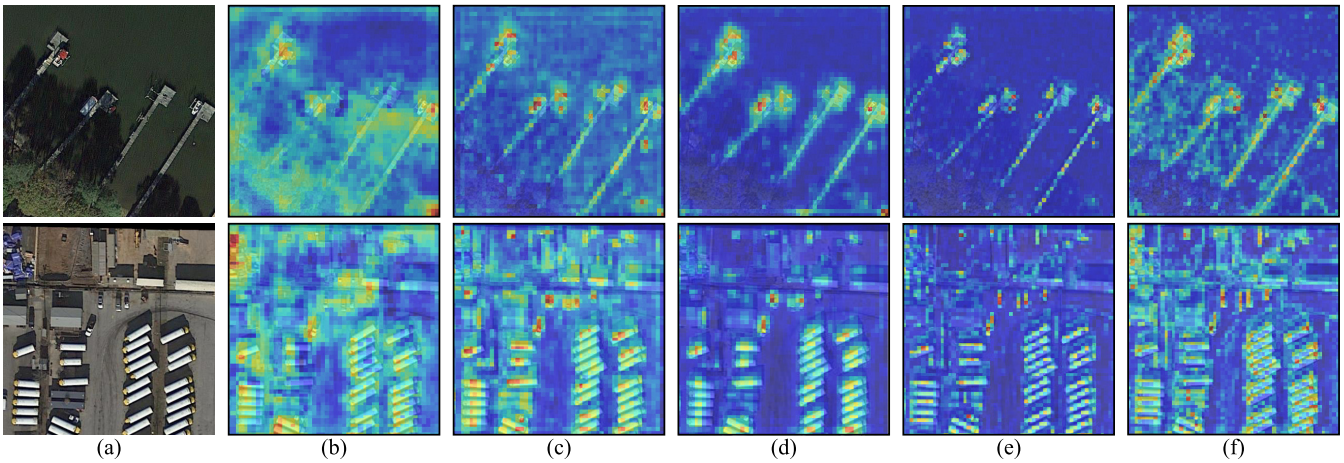

Fig. 11. Visualization of coarse-to-fine features enhanced progressively by the proposed components in CoF-FA. (a) Input images. (b) Coarse features without CoF-FA. (c) Features refined by SSE only. (d) Fine features refined by SSE and SNM. (e) Features refined by SSE, SNM, and FFA (the whole CoF-FA). (f) Features refined by vanilla FPN instead of the proposed FFA.

Table II. Some coarse-to-fine features are visualized in Fig. 11. In order to interpret how these modules can work effectively, more detailed ablation experiments were carried out and the results are tabulated in Tables III and IV and Fig. 10.

b) Effect of the number of frequency components in SSE: In Fig. 10, n represents the number of frequency components used in SSE. As the smallest dimension of the coarse features is 5×5, the number of frequency components is 25. When $n = 0$, it means the baseline model without using any frequency components. While $n = 1$, the lowest component, i.e., the zero frequency, is fused with the coarse features (the first component in Fig. 4(b) and $n = 1$ in Fig. 10), an improvement of around 1.76% mAP can be obtained. When the number of frequency components n increases, the performance in terms of mAP increases steadily to 69.61%, until $n = 16$. When n is more than 16, the performance will drop. This is due to the fact that the high-frequency components represent noise in the shape and appearance of objects. In our experiments, we set the number of frequency components to 16.

c) Effect of SNM and its position: The goal of SNM is to further refine the spatial–spectral features by incorporating global long-range dependencies. Actually, SNM can be inserted into different stages of the backbone with minor modifications. To evaluate its effectiveness at different positions of the backbone, several ablation experiments were conducted, and the results are tabulated in Table III. Adding SNM to the early stage (i.e., Stage 1) achieves a better performance of 71.26% mAP with a small increase in model parameters, compared to the latter two stages (Stages 2 and 3). The best detection accuracy can be achieved if SNM is inserted into all three stages, achieving 71.88% mAP, but causing a much larger increase in model parameters. Considering the speed–accuracy tradeoff, we only insert SNM into Stage 1 by default.

d) Effect of FFA: FFA operates on pixelwise features to mitigate misalignment and serves as an FPN-like structure to generate the final hierarchical fine features for the detection head. Table IV compares the performance of the proposed FFA and the vanilla FPN [36]. As shown in the second row, removing the FPN structure from the baseline model causes a dramatic performance drop of 4.04%, which demonstrates the importance of the multiscale hierarchical features for remote-sensing object detection, whereas replacing the vanilla FPN with FFA can obtain a 1.07% mAP improvement. Some visualized examples are shown in Fig. 11. This ablation result validates that FFA can inherit the feature pyramid structure and smooth the fine features simultaneously.

TABLE V

ABLATION STUDIES ON THE PROPOSED DIFFERENT MATCHING METRICS FOR COF-SA. ALL MODELS FOLLOW THE SAME TOTAL NUMBER OF TRAINING EPOCHS

Model	$\mathcal{M}_{coarse}$	$\mathcal{M}_{fine}$	$\mathcal{M}_{finer}$	mAP_{50} (Δ)	mAP_{75}	Param (M)
CoF-SA	✓	×	×	65.51	27.60	37.56
	✓	✓	×	66.84 (+1.33)	31.24	37.56
	✓	×	✓	67.27 (+1.76)	33.53	37.56
	✓	✓	✓	**67.87 (+2.36)**	**35.22**	37.56

TABLE VI

COMPARISON OF DIFFERENT RATIOS OF THREE SAMPLE ASSIGNMENT STRATEGIES DURING TRAINING. $\mathcal{N}_{\text{coarse}}$, $\mathcal{N}_{\text{fine}}$, AND $\mathcal{N}_{\text{finer}}$ REPRESENT THE NUMBER OF TRAINING EPOCHS USING THE CORRESPONDING MATCHING METRICS

$\mathcal{N}_{coarse} : \mathcal{N}_{fine} : \mathcal{N}_{finer}$	mAP (Δ)
1 : 1 : 1	67.87
2 : 1 : 1	67.45 (-0.42)
1 : 2 : 1	68.56 (+0.69)
1 : 1 : 2	**69.23 (+1.36)**

3) Component Analysis for CoF-SA: In parallel to CoF-FA, several ablation experiments were also carried out for CoF-SA, to verify its effectiveness in producing finer training samples and benefit to the overall model performance.

a) Effect of coarse-to-fine matching metrics: Table V shows the effect of the three proposed coarse-to-fine matching distance metrics on the model performance. It can be observed that, if only coarse samples are selected throughout

TABLE VII
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE DIOR DATASET

Method	Backbone	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10	c11	c12	c13	c14	c15	c16	c17	c18	c19	c20	mAP
Two-Stage																						
Faster R-CNN [29]	VGG16	53.6	49.3	**78.8**	66.2	28.0	70.9	62.3	69.0	55.2	68.0	56.9	50.2	50.1	27.7	73.0	39.8	75.2	38.6	23.6	45.4	54.1
CornerNet [87]	Hourglass104	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
Mask R-CNN [88]	ResNet101	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
PANet [89]	ResNet101	60.2	72.0	70.6	80.5	43.6	72.3	61.4	72.1	66.7	72.0	73.4	45.3	56.9	71.7	70.4	**62.0**	80.9	57.0	**47.2**	84.5	66.1
CSFF [90]	ResNet101	57.2	79.6	70.1	**87.4**	46.1	76.6	62.7	82.6	73.2	78.2	81.6	50.7	59.5	**73.3**	63.4	58.5	**85.9**	61.9	42.9	86.9	68.0
GLNet [91]	ResNet101	**62.9**	**83.2**	72.0	81.1	**50.5**	**79.3**	**67.4**	**86.2**	70.9	**81.8**	**83.0**	**51.8**	**62.6**	72.0	**75.3**	53.7	81.3	**65.5**	43.4	**89.2**	**70.7**
Single-Stage																						
YOLOv3 [92]	DarkNet53	72.2	29.2	74.0	78.6	31.2	69.7	26.9	48.6	54.4	31.1	61.1	44.9	49.7	87.4	70.6	68.7	87.3	29.4	48.3	78.7	57.1
RetinaNet [69]	ResNet101	53.3	77.0	69.3	85.0	44.1	73.2	62.4	78.6	62.8	78.6	76.6	49.9	59.6	71.1	68.4	45.8	81.3	55.2	44.4	**85.5**	66.1
MSFC-Net [4]	ResNeSt101	**85.8**	76.2	74.4	**90.1**	44.2	78.1	55.5	60.9	59.5	76.9	73.7	49.6	57.2	**89.6**	69.2	**76.5**	86.7	51.8	**55.2**	84.3	70.1
ASSD [93]	VGG16	85.6	82.4	75.8	89.5	40.7	77.6	64.7	67.1	61.7	80.8	78.6	**62.0**	58.0	84.9	76.7	65.3	87.9	62.4	44.5	76.3	71.1
CoF-Net (Ours)	ResNet50	84.0	**85.3**	**82.6**	90.0	**47.1**	**80.7**	**73.3**	**89.3**	**74.0**	**84.5**	**83.2**	57.4	**62.2**	82.9	**77.6**	68.2	**89.9**	**68.7**	49.3	85.2	**75.8**

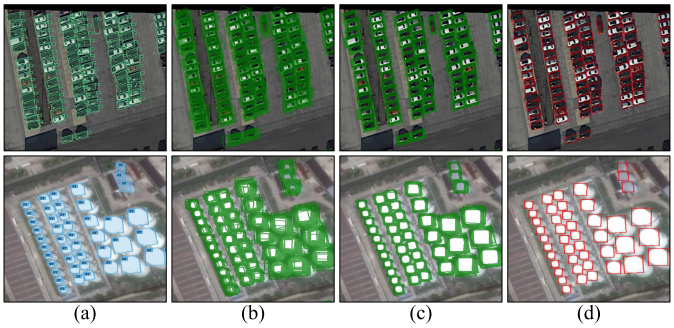

Fig. 12. Illustration of the proposed CoF-SA. (a) Predictions. (b) Coarse one-to-many samples. (c) Fine one-to-many samples. (d) Finer one-to-one samples.

the training phase, the baseline model achieves 65.51% mAP_{50} (IoU threshold is 0.50) and 27.60% mAP_{75} (IoU threshold is 0.75). Replacing $\mathcal{M}_{\text{coarse}}$ with a combination of "$\mathcal{M}_{\text{coarse}}$ and $\mathcal{M}_{\text{fine}}$" or "$\mathcal{M}_{\text{coarse}}$ and $\mathcal{M}_{\text{finer}}$" can improve the performance by 1.33% or 1.76% mAP_{50}, respectively. In terms of mAP_{75}, the corresponding gains further increase to about 3.64% and 5.93%. If the proposed progressive combination "$\mathcal{M}_{\text{coarse}}$ and $\mathcal{M}_{\text{fine}}$ and $\mathcal{M}_{\text{finer}}$" is employed, the final detection accuracy can reach 67.87% mAP_{50}, surpassing the other models in Table V. Some visualized results are presented in Fig. 12. The above ablation results notably validate the positive impact of the proposed CoF-SA strategy on object detection.

b) Effect of different ratios of the coarse–fine–finer assignment: In Table V, the coarse–fine–finer training strategy achieves the best performance, but the number of epochs is the same for the three training stages, i.e., the ratio $\mathcal{N}_{\text{coarse}}$:$\mathcal{N}_{\text{fine}}$:$\mathcal{N}_{\text{finer}}$ = 1:1:1. To further explore the optimal ratio $\mathcal{N}_{\text{coarse}}$:$\mathcal{N}_{\text{fine}}$:$\mathcal{N}_{\text{finer}}$, the training scheme is evaluated with different ratios, while the total number of epochs is still fixed at 24. The results are shown in Table VI. We observe that increasing the number of training epochs for coarse-sample assignment, i.e. $\mathcal{N}_{\text{coarse}}$, marginally reduces the performance by 0.42% mAP. Conversely, increasing the ratio for fine-sample training and finer-sample training yields better results of 68.56% and 69.23% mAP, respectively, as shown in the last two rows of Table VI. Furthermore, the setting of $\mathcal{N}_{\text{coarse}}$:$\mathcal{N}_{\text{fine}}$:$\mathcal{N}_{\text{finer}}$ = 1:1:2 results in the most significant mAP improvement, about 1.36%. Only the finest sample surviving as supervision for each potential instance marginally requires more training epochs or iterations. This phenomenon also consistently demonstrates the effectiveness of the proposed CoF-SA.

4) Analysis of Model Parameters and Speed: In addition to detection accuracy, the number of parameters of the models and the speed are evaluated and reported in detail in Tables I–V. As shown in Tables I and II, CoF-FA only brings an increase in the number of parameters by about 1.87 M and can achieve competitive inference speed. In Tables I and V, CoF-SA introduces no additional parameters and focuses on efficiently nominating fine samples from coarse samples, instead of using any heavy computational modules.

D. Comparative Experiments With State-of-the-Art Methods

In this section, we compare the proposed CoF-Net with other state-of-the-art methods on three popular aerial object detection datasets: DIOR, NWPU VHR-10, and DOTA.

1) Results on DIOR: The proposed CoF-Net is evaluated on DIOR and compared to other representative methods in Table VII, all of which are CNN-based horizontal detectors, including two- and single-stage methods. Recently, many studies have been devoted to promoting the performance of single-stage detectors to make them competitive with two-stage detectors, while our method successfully fills this gap by refining coarse features and coarse samples into fine features and samples, respectively. Specifically, our model obtains 75.8% mAP across all categories and achieves the best performance in 13 out of 20 categories. Compared to all the advanced object detectors in Table VII, the proposed CoF-Net shows the state-of-the-art performance with competitive model size and speed. The latest single-stage algorithms, MSFC-Net [4] and ASSD [93], also achieve promising overall accuracy, comparable to two-stage detectors, but our CoF-Net dramatically outperforms them by 5.7% and 4.7% mAP, respectively. Some detection results are visualized in Fig. 13.

2) Results on NWPU VHR-10: Table VIII compares the performance of the proposed method with other state-of-the-art horizontal detectors on the NWPU VHR-10 dataset. It can be seen that CoF-Net attains 94.5% mAP, achieving the best accuracy. Most two-stage methods rely on large backbones for higher mAP, which limits their model efficiency

TABLE VIII
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE NWPU VHR-10 DATASET

Method	Backbone	Airplane	Ship	Storage Tank	Baseball Diamond	Tennis Court	Basketball Court	Ground Track Field	Harbor	Bridge	Vehicle	mAP
Two-Stage												
RICNN [6]	AlexNet	88.4	77.3	85.3	88.1	40.8	58.5	86.7	68.6	61.5	71.1	72.6
Faster R-CNN [29]	ResNet50	94.6	82.3	65.3	95.5	81.9	89.7	92.4	72.4	57.5	77.8	80.9
CAD-Net [35]	ResNet101	97.0	77.9	95.6	93.6	87.6	87.1	99.6	**100.0**	86.2	89.9	91.5
MEDNet [30]	ResNet101	99.2	94.4	82.2	**98.5**	95.4	95.2	98.3	88.1	75.1	89.3	91.6
GLNet [91]	ResNet101	**100**	88.7	84.4	**98.5**	81.6	88.2	**100**	97.2	**88.4**	**90.9**	91.8
EGAT-LSTM [34]	GCN+LSTM	97.3	**96.7**	**97.2**	96.5	86.6	94.5	94.2	86.2	80.1	90.8	92.0
NL-LFPN-MR101 [9]	ResNet101	**100**	89.5	90.9	96.8	**96.7**	**99.2**	**100**	90.1	79.1	90.2	**93.2**
Single-Stage												
YOLOv2 [51]	DarkNet19	83.0	84.4	81.9	84.3	85.0	53.5	62.8	78.7	85.0	70.0	76.8
SCRDet [60]	ResNet101	**100**	89.4	97.2	97.0	83.2	87.5	99.2	**99.4**	74.5	90.1	91.8
FMSSD [94]	VGG16	99.7	89.9	90.3	98.2	86.0	**96.8**	99.6	75.6	80.1	88.2	90.4
CANet [58]	ResNet101	99.9	85.9	**99.3**	97.3	**97.8**	84.8	98.4	90.4	89.2	90.3	93.3
CoF-Net (Ours)	ResNet50	**100**	**90.9**	96.1	**98.8**	91.1	95.8	**100**	91.4	**89.7**	**90.8**	**94.5**

TABLE IX
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE DOTA DATASET

Method	Backbone	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
Two-Stage																	
FR-O [16]	ResNet101	79.4	77.1	17.7	64.1	35.3	38.0	37.2	89.4	69.6	59.3	50.3	52.9	47.9	47.4	46.3	54.1
RoI Trans. [67]	ResNet101	88.6	78.5	43.4	75.9	68.8	73.7	**83.6**	90.7	77.3	81.5	58.4	53.5	62.8	58.9	47.7	69.6
CAD-Net [35]	ResNet101	87.8	82.4	49.4	73.5	71.1	63.5	76.6	**90.9**	79.2	73.3	48.4	60.9	62.0	67.0	62.2	69.9
SCRDet [60]	ResNet101	90.0	80.7	52.1	68.4	68.4	60.3	72.4	**90.9**	**87.9**	**86.9**	65.0	66.7	66.3	68.2	65.2	72.6
APE [95]	ResNeXt101	90.0	83.6	53.4	**76.0**	**74.0**	**77.2**	79.5	90.8	87.2	84.5	67.7	60.3	**74.6**	**71.8**	65.6	75.8
CSL [22]	ResNet152	**90.3**	**85.5**	**54.6**	75.3	70.4	73.5	77.6	90.8	86.2	86.7	**69.6**	**68.0**	73.8	71.1	**68.9**	**76.2**
Single-Stage																	
O^2DNet [96]	Hourglass104	89.3	82.1	47.3	61.2	71.3	74.0	78.6	90.8	82.2	81.4	60.9	60.2	58.2	67.0	61.0	71.0
CFC-Net [21]	ResNet50	89.1	80.4	**52.4**	70.0	76.3	78.1	87.2	**90.9**	84.5	85.6	60.5	61.5	67.8	68.0	50.1	73.5
R^3Det [50]	ResNet152	89.5	81.2	50.5	66.1	70.9	78.7	78.2	90.8	85.3	84.2	61.8	63.8	68.2	69.8	67.2	73.7
CBDA-Net [61]	DLANet34	89.2	**85.9**	50.3	65.0	77.7	82.3	**87.9**	90.5	**86.5**	85.9	66.9	**66.5**	67.4	**71.3**	62.9	75.7
CoF-Net (Ours)	ResNet50	**89.6**	83.1	48.3	**73.6**	**78.2**	**83.0**	86.7	90.2	82.3	**86.6**	**67.6**	64.6	**74.7**	**71.3**	**78.4**	**77.2**

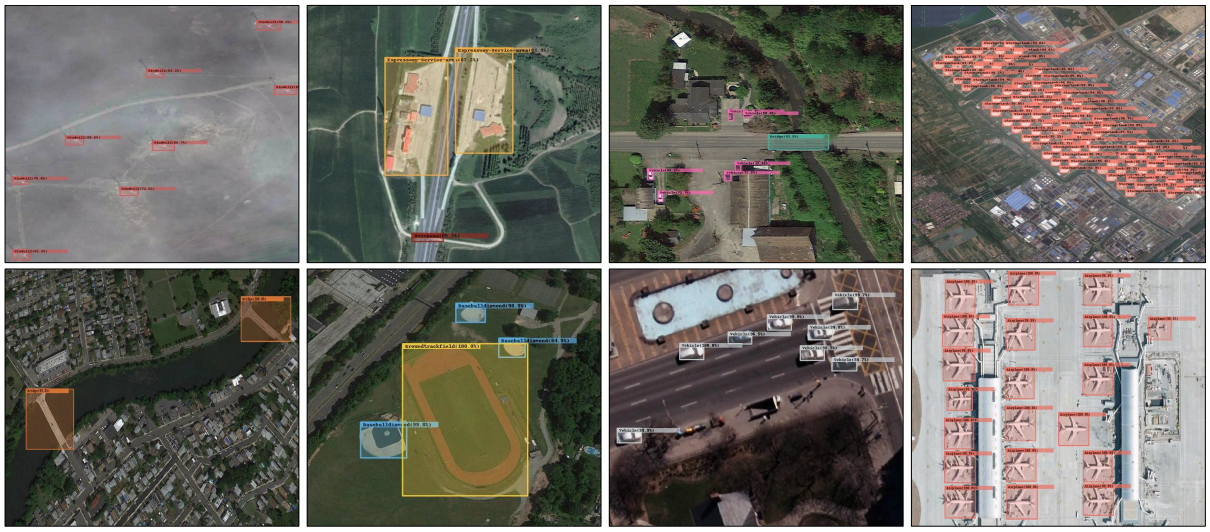

Fig. 13. Some qualitative detection results of the proposed CoF-Net on the DIOR dataset (the first row) and the NWPU VHR-10 dataset (the second row).

and flexibility, while CoF-Net can substantially outperform them with a more lightweight backbone. In terms of categorywise performance, our method achieves the best results on six categories and is also competitive in the remaining four categories. Some qualitative results are visualized in Fig. 13.

3) Results on DOTA: Different from the DIOR and NWPU VHR-10 datasets for horizontal bounding-box object detection, the DOTA dataset is widely used for oriented object detection. In Table IX, it can be observed that CoF-Net still outperforms other state-of-the-art methods, achieving 77.2% mAP. Furthermore, in all categories, the proposed framework

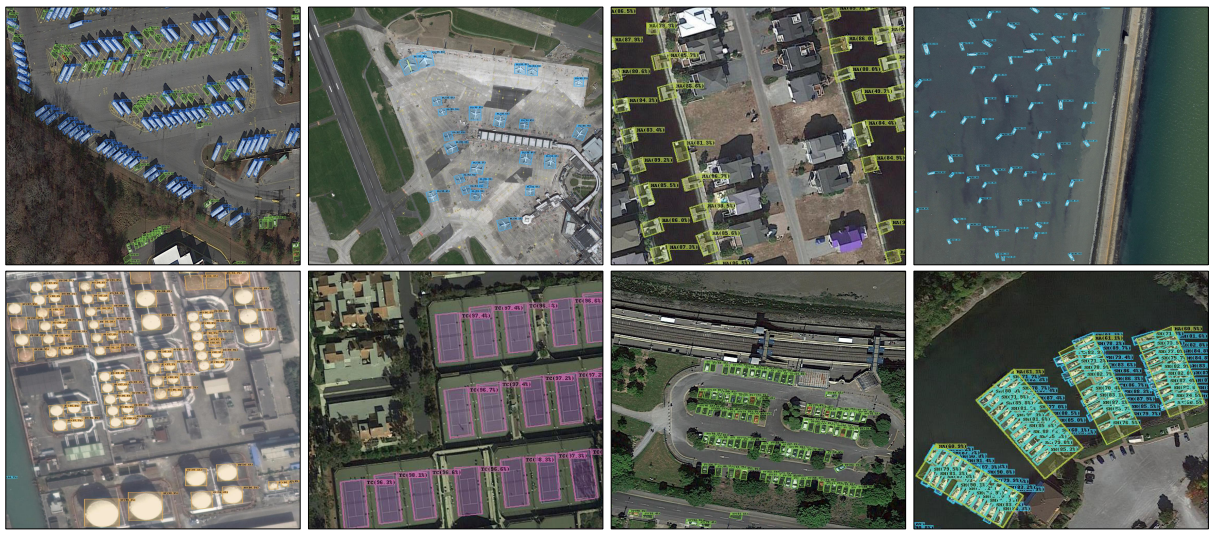

Fig. 14. Some qualitative detection results of the proposed CoF-Net on the DOTA dataset.

achieves the best or favorable detection results, beating most single-stage detectors and being comparable to two-stage methods. Some qualitative detection results of the proposed CoF-Net are presented in Fig. 14. It can also be seen that, benefiting from the coarse-to-fine features and samples, our method accurately localizes densely packed and cluttered instances with various scale and orientation variations, against complicated background interference.

V. Conclusion

This article proposes a novel single-stage framework, named CoF-Net, for object detection in remote-sensing imagery, which achieves high accuracy and low complexity. The deficiencies of previous geospatial object detectors, especially single-stage methods, which cause performance degradation, are elucidated as coarse features and coarse training samples. To pave the way, CoF-Net presents a friendly progressive coarse-to-fine architecture, in which CoF-FA and CoF-SA are placed in parallel. The former aims to progressively enrich spectral details, semantic dependencies, and alignment at the feature level, while the latter dynamically assigns fine and finer samples during training for accurate regression. Extensive experiments were conducted on three datasets, DIOR, NWPU VHR-10, and DOTA. The results have demonstrated the effectiveness and efficiency of the proposed method, as well as its superiority over other state-of-the-art methods.

References

[1] J. Han, J. Ding, J. Li, and G.-S. Xia, "Align deep features for oriented object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Mar. 2022, Art. no. 5602511.

[2] Q. Wang, W. Huang, X. Zhang, and X. Li, "Word–sentence framework for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10532–10543, Dec. 2020.

[3] Y. Liu, Q. Li, Y. Yuan, Q. Du, and Q. Wang, "ABNet: Adaptive balanced network for multiscale object detection in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Dec. 2022, Art. no. 5614914.

[4] T. Zhang, Y. Zhuang, G. Wang, S. Dong, H. Chen, and L. Li, "Multiscale semantic fusion-guided fractal convolutional object detection network for optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 2022, Art. no. 5608720.

[5] Y.-Y. Ma, Z.-L. Sun, Z. Zeng, and K.-M. Lam, "Corn-plant counting using scare-aware feature and channel interdependence," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, Jan. 2022.

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

[7] X. Lu, Y. Zhang, Y. Yuan, and Y. Feng, "Gated and axis-concentrated localization network for remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 179–192, Jan. 2019.

[8] X. Yang et al., "Automatic ship detection in remote sensing images from Google earth of complex scenes based on multiscale rotation dense feature pyramid networks," *Remote Sens.*, vol. 10, no. 1, p. 132, 2018.

[9] W. Zhang, L. Jiao, Y. Li, Z. Huang, and H. Wang, "Laplacian feature pyramid network for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 2022, Art. no. 5604114.

[10] Z. Shao, J. Han, D. Marnerides, and K. Debattista, "Region-object relation-aware dense captioning via transformer," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 11, 2022, doi: 10.1109/TNNLS.2022.3152990.

[11] Y. Xu et al., "Gliding vertex on the horizontal bounding box for multi-oriented object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1452–1459, Feb. 2020.

[12] G. Cheng et al., "Dual-aligned oriented detector," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Feb. 2022, Art. no. 5618111.

[13] J. Wang, W. Yang, H.-C. Li, H. Zhang, and G.-S. Xia, "Learning center probability map for detecting objects in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4307–4323, May 2021.

[14] J. Li, H. Zhang, R. Song, W. Xie, Y. Li, and Q. Du, "Structure-guided feature transform hybrid residual network for remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Aug. 2022, Art. no. 5610713.

[15] T. Zhang et al., "Foreground refinement network for rotated object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 2021, Art. no. 5610013.

[16] G.-S. Xia et al., "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3974–3983.

[17] 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 J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020.

[18] 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 J. Photogramm. Remote Sens.*, vol. 98, pp. 119–132, Dec. 2014.

[19] Y. Zhang, Y. Yuan, Y. Feng, and X. Lu, "Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5535–5548, Aug. 2019.

[20] Z. Liu, L. Yuan, L. Weng, and Y. Yang, "A high resolution optical satellite image dataset for ship recognition and some new baselines," in *Proc. Int. Conf. Pattern Recognit. Appl. Methods*, vol. 2, 2017, pp. 324–331.

[21] Q. Ming, L. Miao, Z. Zhou, and Y. Dong, "CFC-Net: A critical feature capturing network for arbitrary-oriented object detection in remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jul. 2022, Art. no. 5605814.

[22] X. Yang and J. Yan, "Arbitrary-oriented object detection with circular smooth label," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 677–694.

[23] X. Yang, L. Hou, Y. Zhou, W. Wang, and J. Yan, "Dense label encoding for boundary discontinuity free rotation detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 15819–15829.

[24] W. Qian, X. Yang, S. Peng, J. Yan, and Y. Guo, "Learning modulated loss for rotated object detection," in *Proc. AAAI Conf. Arti. Intell.*, vol. 35, no. 3, May 2021, pp. 2458–2466.

[25] X. Yang, J. Yan, Q. Ming, W. Wang, X. Zhang, and Q. Tian, "Rethinking rotated object detection with Gaussian Wasserstein distance loss," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 11830–11841.

[26] Q. Ming, L. Miao, Z. Zhou, X. Yang, and Y. Dong, "Optimization for arbitrary-oriented object detection via representation invariance loss," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, Oct. 2021.

[27] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[28] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[29] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[30] Q. Lin, J. Zhao, B. Du, G. Fu, and Z. Yuan, "MEDNet: Multiexpert detection network with unsupervised clustering of training samples," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jul. 2022, Art. no. 4703114.

[31] C. Xu, C. Li, Z. Cui, T. Zhang, and J. Yang, "Hierarchical semantic propagation for object detection in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 4353–4364, Jun. 2020.

[32] R. Qin, Q. Liu, G. Gao, D. Huang, and Y. Wang, "MRDet: A multihead network for accurate rotated object detection in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Oct. 2022, Art. no. 5608412.

[33] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented R-CNN for object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2021, pp. 3520–3529.

[34] S. Tian, L. Kang, X. Xing, J. Tian, C. Fan, and Y. Zhang, "A relation-augmented embedded graph attention network for remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, May 2022, Art. no. 1000718, doi: 10.1109/TGRS.2021.3073269.

[35] G. Zhang, S. Lu, and W. Zhang, "CAD-Net: A context-aware detection network for objects in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 10015–10024, Aug. 2019.

[36] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[37] C. Zhu, H. Zhou, R. Wang, and J. Guo, "A novel hierarchical method of ship detection from spaceborne optical image based on shape and texture features," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 9, pp. 3446–3456, Sep. 2010.

[38] 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 Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3325–3337, Jun. 2014.

[39] Y. Miao, Z. Lin, X. Ma, G. Ding, and J. Han, "Learning transformation-invariant local descriptors with low-coupling binary codes," *IEEE Trans. Image Process.*, vol. 30, pp. 7554–7566, 2021.

[40] Y. Liu, D. Zhang, Q. Zhang, and J. Han, "Part-object relational visual saliency," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3688–3704, Jan. 2022.

[41] J. Han et al., "Efficient, simultaneous detection of multi-class geospatial targets based on visual saliency modeling and discriminative learning of sparse coding," *ISPRS J. Photogramm. Remote Sens.*, vol. 89, pp. 37–48, Mar. 2014.

[42] Q. Li, G. Wang, J. Liu, and S. Chen, "Robust scale-invariant feature matching for remote sensing image registration," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 2, pp. 287–291, Apr. 2009.

[43] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, no. 1, Jun. 2005, pp. 886–893.

[44] X. Ye, F. Xiong, J. Lu, J. Zhou, and Y. Qian, "F^3-Net: Feature fusion and filtration network for object detection in optical remote sensing images," *Remote Sens.*, vol. 12, no. 24, p. 4027, Dec. 2020.

[45] Y. Wu, K. Zhang, J. Wang, Y. Wang, Q. Wang, and X. Li, "GCWNet: A global context-weaving network for object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Mar. 2022, Art. no. 5619912.

[46] Y. Zhou, S. Chen, J. Zhao, R. Yao, Y. Xue, and A. E. Saddik, "CLT-Det: Correlation learning based on transformer for detecting dense objects in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 2022, Art. no. 4708915.

[47] K. Li, G. Cheng, S. Bu, and X. You, "Rotation-insensitive and context-augmented object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2337–2348, Apr. 2017.

[48] J. Liu, S. Li, C. Zhou, X. Cao, Y. Gao, and B. Wang, "SRAF-Net: A scene-relevant anchor-free object detection network in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Nov. 2022, Art. no. 5405914.

[49] W. Ma et al., "Feature split–merge–enhancement network for remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 5616217.

[50] X. Yang, J. Yan, Z. Feng, and T. He, "R3Det: Refined single-stage detector with feature refinement for rotating object," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 4, May 2021, pp. 3163–3171.

[51] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 7263–7271.

[52] S. Chen, B. Wang, X. Tan, and X. Hu, "Embedding attention and residual network for accurate salient object detection," *IEEE Trans. Cybern.*, vol. 50, no. 5, pp. 2050–2062, May 2018.

[53] Z. Huang, W. Li, X.-G. Xia, X. Wu, Z. Cai, and R. Tao, "A novel nonlocal-aware pyramid and multiscale multitask refinement detector for object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Feb. 2021, Art. no. 56019205601920.

[54] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 4438–4446.

[55] X. Zhu, D. Cheng, Z. Zhang, S. Lin, and J. Dai, "An empirical study of spatial attention mechanisms in deep networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 6688–6697.

[56] J. Hu, L. Shen, and G. Sun, "Squeeze- and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[57] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13–19.

[58] Z. Qin, P. Zhang, F. Wu, and X. Li, "FcaNet: Frequency channel attention networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 783–792.

[59] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[60] X. Yang et al., "SCRDet: Towards more robust detection for small, cluttered and rotated objects," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 8232–8241.

[61] S. Liu, L. Zhang, H. Lu, and Y. He, "Center-boundary dual attention for oriented object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 2022, Art. no. 5603914.

[62] G. Sumbul, R. G. Cinbis, and S. Aksoy, "Multisource region attention network for fine-grained object recognition in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4929–4937, Jul. 2019.

[63] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7794–7803.

[64] L. Zhu et al., "Unifying nonlocal blocks for neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2021, pp. 12292–12301.

[65] K. Yue, M. Sun, Y. Yuan, F. Zhou, E. Ding, and F. Xu, "Compact generalized non-local network," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 6511–6520.

[66] Y. Tao, Q. Sun, Q. Du, and W. Liu, "Nonlocal neural networks, nonlocal diffusion and nonlocal modeling," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

[67] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning RoI transformer for oriented object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 2849–2858.

[68] J. Yan, H. Wang, M. Yan, D. Wenhui, 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 Sens.*, vol. 11, no. 3, p. 286, Feb. 2019.

[69] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
[70] J. Wang, K. Chen, S. Yang, C. C. Loy, and D. Lin, "Region proposal by guided anchoring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 2965–2974.
[71] X. Zhang, F. Wan, C. Liu, R. Ji, and Q. Ye, "Freeanchor: Learning to match anchors for visual object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–9.
[72] Q. Ming, Z. Zhou, L. Miao, H. Zhang, and L. Li, "Dynamic anchor learning for arbitrary-oriented object detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 3, 2021, pp. 2355–2363.
[73] Z. Xiao, K. Wang, Q. Wan, X. Tan, C. Xu, and F. Xia, "A2S-Det: Efficiency anchor matching in aerial image oriented object detection," *Remote Sens.*, vol. 13, no. 1, p. 73, Dec. 2021.
[74] Q. Ming, L. Miao, Z. Zhou, J. Song, and X. Yang, "Sparse label assignment for oriented object detection in aerial images," *Remote Sens.*, vol. 13, no. 14, p. 2664, Jul. 2021.
[75] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9759–9768.
[76] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9627–9636.
[77] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, and J. Shi, "FoveaBox: Beyound anchor-based object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 7389–7398, 2020.
[78] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.
[79] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[80] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Trans. Comput.*, vol. C-100, no. 1, pp. 90–93, Jan. 1974.
[81] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 3844–3852.
[82] Z. Huang, Y. Wei, X. Wang, W. Liu, T. S. Huang, and H. Shi, "AlignSeg: Feature-aligned segmentation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 550–557, Mar. 2021.
[83] J. Dai et al., "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2017, pp. 764–773.
[84] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets v2: More deformable, better results," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 9308–9316.
[85] G. Wang et al., "FSoD-Net: Full-scale object detection from optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Mar. 2021, Art. no. 5602918.
[86] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1097–1105.
[87] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 734–750.
[88] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.
[89] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768.
[90] G. Cheng, Y. Si, H. Hong, X. Yao, and L. Guo, "Cross-scale feature fusion for object detection in optical remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 3, pp. 431–435, Mar. 2020.
[91] Z. Teng, Y. Duan, Y. Liu, B. Zhang, and J. Fan, "Global to local: Clip-LSTM-based object detection from remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Mar. 2022, Art. no. 5603113.
[92] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[93] T. Xu, X. Sun, W. Diao, L. Zhao, K. Fu, and H. Wang, "ASSD: Feature aligned single-shot detection for multiscale objects in aerial imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jun. 2022, Art. no. 5607117.
[94] P. Wang, X. Sun, W. Diao, and K. Fu, "FMSSD: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3377–3390, May 2020.
[95] Y. Zhu, J. Du, and X. Wu, "Adaptive period embedding for representing oriented objects in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7247–7257, Oct. 2020.
[96] H. Wei, Y. Zhang, Z. Chang, H. Li, H. Wang, and X. Sun, "Oriented objects as pairs of middle lines," *ISPRS J. Photogramm. Remote Sens.*, vol. 169, pp. 268–279, Nov. 2020.

Cong Zhang (Graduate Student Member, IEEE) received the B.E. degree from the School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China, in 2018, and the M.E. degree from the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, in 2021. He is currently pursuing the Ph.D. degree with the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong.

His research interests include remote sensing, computer vision, and machine learning.

Kin-Man Lam (Senior Member, IEEE) received the Associateship (Hons.) in electronic engineering from The Hong Kong Polytechnic University (formerly called Hong Kong Polytechnic), Hong Kong, in 1986, the M.Sc. degree in communication engineering from the Department of Electrical Engineering, Imperial College of Science, Technology and Medicine, London, U.K., in 1987, and the Ph.D. degree from the Department of Electrical Engineering, University of Sydney, Sydney, NSW, Australia, in 1996.

From 1990 to 1993, he was a Lecturer at the Department of Electronic Engineering, The Hong Kong Polytechnic University. He joined the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, again as an Assistant Professor in 1996, where he became an Associate Professor in 1999 and has been a Professor since 2010. He is currently the Associate Dean of the Faculty of Engineering, The Hong Kong Polytechnic University. He was actively involved in professional activities. He has been a member of the organizing committee or program committee of many international conferences. His research interests include image processing, computer vision, and human face analysis and recognition.

Dr. Lam was the Chairperson of the IEEE Hong Kong Chapter of Signal Processing from 2006 to 2008. He was the General Co-Chair of the 2012 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC 2012), the APSIPA Annual and Summit 2015, and the 2017 IEEE International Conference on Multimedia and Expo (ICME 2017), which were held in Hong Kong, and the Technical Chair of the 2020 IEEE International Conference on Visual Communications and Image Processing. He was the Director-Student Services and the Director-Membership Services of the IEEE Signal Processing Society from 2012 to 2014 and from 2015 to 2017, respectively. He was also the VP-Member Relations and Development and VP-Publications of the Asia-Pacific Signal and Information Processing Association (APSIPA) from 2014 to 2017 and from 2017 to 2021, respectively. He was an Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING from 2009 to 2014 and *Digital Signal Processing* from 2014 to 2018. He was an Editor of *HKIE Transactions* from 2013 to 2018 and an Area Editor of the *IEEE Signal Processing Magazine* from 2015 to 2017. He is also the VP-Membership of IEEE SPS and a Member-at-Large of APSIPA. He serves as a Senior Editorial Board Member for *APSIPA Transactions on Signal and Information Processing* and an Associate Editor for *EURASIP International Journal on Image and Video Processing*.

Qi Wang (Senior Member, IEEE) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively.

He is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. He is also with the Key Laboratory of Intelligent Interaction and Applications, Ministry of Industry and Information Technology, Northwestern Polytechnical University. His research interests include computer vision and pattern recognition.