This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3115796, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

Scale-aware Anchor-free Object Detection via Curriculum Learning for Remote Sensing Images

Wandi Cai, Bo Zhang, and Bin Wang, *Senior Member, IEEE*

***Abstract*—Accurate detection of multi-class instance objects in remote sensing images (RSIs) is a fundamental but challenging task in the field of aviation and satellite image processing, which plays a crucial role in a wide range of practical applications. Compared with the natural image-based object detection task, RSIs-based object detection still faces two main challenges: 1) The instance objects often present large variations in object size, and they are densely arranged in the given input images; 2) Complex background distributions around instance objects tend to cause boundary blurring, making it difficult to distinguish instance objects from the background, resulting in undesired feature learning interference. In this paper, to address the above challenges, we propose a novel RSI anchor-free object detection framework that consists of two key components: a cross-channel feature pyramid network (CFPN) and multiple foreground-attentive detection heads (FDHs). First, an anchor-free baseline detector with the CFPN structure is developed to extract features from different convolutional layers and incorporates these multi-scale features through parameterized cross-channel learning processes, learning the semantic relations across different scales and levels. Next, each FDH is designed to predict an attention map to enhance the features of the foreground region in RSIs. Furthermore, under this scale-aware anchor-free baseline detector structure, we design a curriculum-style optimization objective that dynamically reweights training instances during the current training epoch, enabling the detector to receive relatively easy instances that match with its current ability. Experimental results on three publicly available object detection datasets demonstrate that the proposed method outperforms existing object detection methods.**

***Index Terms*—Remote sensing images, anchor-free object detection, feature pyramid structure, foreground attention, curriculum learning.**

I. Introduction

OBJECT detection in remote sensing images (RSIs) aims to recognize and localize multi-class remote sensing objects from given satellite or aerial images, which plays an important role in a wide scope of applications, such as intelligent monitoring, urban planning, precision agriculture, and geographic information system (GIS) [1]. Benefitting from the rapid development of deep neural networks (DNNs) [2]-[4] in the computer vision community, the ability to learn a robust detector which can predict the region of interests (RoIs) from the input image has been pushed forward a lot.

Recently, researchers have explored a large number of DNNs-based detectors to address generic object detection task. These detectors can be roughly divided into two categories: anchor-based detectors [5]-[14] and anchor-free detectors [15]-[20].

Typical anchor-based detectors such as Faster RCNN [7], FPN [9], and SSD [10], *etc*., first predefine a set of region proposals sampled from the input image, and then learn to predict the category and position information of each region proposal, *via* a sparse prediction way such as the two-stage detection framework [5]-[9] or a dense prediction way such as the one-stage detection framework [10]-[14]. On the one hand, the two-stage detection methods aim to generate pre-default proposals for potential foreground objects, and then classify and regress these proposals by a following proposal refinement process that can be achieved *via* a fully-connected network. For example, Faster RCNN [7] develops a region proposal network (RPN) that learns to generate region proposals using a DNNs-based network, and FPN [9] tries to recognize objects with multiple scales through a feature pyramid structure where small objects are often recognized by the shallow layer, and large objects are usually detected by the high layer. On the other hand, the one-stage detection methods regard the detection task as a one-shot problem (dense prediction process) without relying on the region proposal generation process. Overall, the two-stage detectors usually have relatively higher detection accuracy than the one-stage detectors owing to the pre-generated proposals and the subsequent refinement process for these proposals. However, both the two-stage and the one-stage detectors need to predefine a set of region proposals. Such a large number of region proposals introduce many hyperparameters and instability for model learning, including the size, number, and aspect ratio of region proposals, especially for RSIs whose instance objects are usually arranged in a spatially dense and multi-scale way. This could result in extra computational cost and design choices, and further increase the risk of overfitting.

In contrast, anchor-free detectors [15]-[20] decouple the predefined region proposals from the typical object detection frameworks, and thus reduce redundant computation related to proposals of some negative instance regions, by predicting a set of keypoints to represent the ROIs of a given image [16], [17]. For example, based on an effective keypoint estimation network, CornerNet [16] aims to find a pair of keypoints (the top-left and bottom-right corners) to detect an instance object

This manuscript was first submitted on Jul. 23, 2021 for review. This work was supported by the National Natural Science Foundation of China under Grant 61971141 and Grant 61731021. *(Corresponding author: Bin Wang)*

The authors are with the Key Laboratory for Information Science of Electromagnetic Waves (MoE), Fudan University, Shanghai 200433, China, and also with the Research Center of Smart Networks and Systems, School of Information Science and Technology, Fudan University, Shanghai 200433, China (e-mail: wangbin@fudan.edu.cn).

without using any predefined hyperparameters of region proposals, achieving a real-time and accurate detection framework. Different from CornerNet that produces some corner points from the given image, CenterNet [17] directly predicts the center point of each instance object and further regresses its object size and localization offset according to the size and offset prediction branch. In CornerNet and CenterNet, only the corner or center points of instances are positive, resulting in the imbalance of positive and negative samples. To alleviate this problem, FoveaBox [19] introduces a positive area where the label of each point is positive. However, the above keypoint-based approaches mainly focus on learning the generic representations to recognize and localize instance objects in natural images, which do not have sufficient generalization and representation ability for the instance objects in RSIs. This is mainly because the constantly variable object size and densely arranged instance objects in RSIs usually bring disturbances and interferences for feature learning and localization of instance objects.

Driven by the success of DNNs-based object detection methods [5]-[20] for natural images, some works [21]-[27] try to extend the study of object detection from the natural images to the RSIs, by leveraging the contextual information surrounding the instance objects to alleviate the semantic ambiguity caused by the variable scene layout. For example, to address various sizes of instance objects in RSIs, feature-merged single-shot detection (FMSSD) [25] is designed on the basis of SSD [10], utilizing an atrous spatial feature reconstruction module to fuse context information and a reweighted loss function to focus on the feature learning of small objects. Besides, contextual bidirectional enhancement (CBD-E) [26] and decoupled classification localization network (DCL-Net) [27] adopt a spatial-level attention mechanism to enhance the features of objects and boost the spatial consistency of prediction results. However, these works for RSIs still suffer from three main limitations as follows.

Firstly, all the above detectors still rely on predefined default boxes (region proposals), where the final prediction results are often sensitive to their hyperparameter settings. Especially for instance objects in RSIs, their scales and layouts change frequently. This may increase the difficulty of presetting such default boxes, and further result in the scale mismatch between the predefined default boxes and instance objects.

Secondly, it is inevitable that for RSIs, some unexpected background disturbances surrounding the instance objects result in decreased detection accuracy. For example, the background distribution surrounding objects belonging to the same class often varies greatly in different images. This may make it difficult to accurately predict the size of foreground objects from the complex background information.

Thirdly, due to that RSIs often present more complicated scene layout and background distribution, some hard-to-learn instance objects or hard-to-distinguish complex backgrounds may cause ambiguity in the feature learning process during the early training stage of the object detection framework. However, previous methods ignore this problem and sample the training instances in a random learning order.

To address the above problems, we propose a novel anchor-free object detection framework for RSIs, whose network structure mainly consists of a designed cross-channel feature pyramid network (CFPN) and multiple developed foreground-attentive detection heads (FDHs). Specifically, an input image is fed into a selected anchor-free baseline detector with the CFPN that can be aware of the scale variations of potential instance objects during the whole detection process. In the CFPN, each prediction layer that combines rich inter-layer semantic relationships can detect objects with a certain size, which effectively deals with instance objects in multiple scales and encodes their semantic relations across different layers in the anchor-free baseline detector. Next, each FDH is designed to predict the center, size, and offset of an instance object by using the output of CFPN. With the aid of pixel-level annotations that is converted from the bounding-box level annotations, the FDH can learn to predict a spatial-wise attention map to emphasize the foreground object of a given bounding-box prediction, by a soft attention enhancement achieved through an element-wise multiplication operation.

Furthermore, in order to provide the ability of learning instance features from easy to hard for the anchor-free detection framework, we design a dynamic curriculum-style optimization objective that can be directly connected after the common object detection loss such as focal loss [13], and reweight each instance object according to a dynamically changing easy-hard sample threshold, further improving the initial training stability and final detection accuracy of the anchor-free detection framework. Compared with the previous methods, the proposed curriculum-style optimization objective can be easily integrated with the off-the-shelf object detection loss to achieve the curriculum learning for object detection task, which does not require any extra learnable hyperparameters nor any change of the training procedure.

We conduct extensive ablation studies and experiments on three public benchmark datasets, including DIOR [28], NWPU VHR-10 [1], and RSOD [29]. The experimental results on these datasets and insightful analyses demonstrate that the proposed framework can lead to consistent detection accuracy improvements.

The main contributions of this work can be concisely summarized as follows:

1) Aiming at the densely-arranged objects with variable scales and complex background disturbances in RSIs, we propose a scale-aware anchor-free detection framework without the need to predefine extra hyperparameters of the region proposals such as the number and aspect ratio of the proposals, which mainly consists of a CFPN and multiple FDHs, where the CFPN can learn cross-layer/channel semantic relationships to predict multiple densely-arranged and scale-variable instance objects while each FDH is designed to learn a soft foreground mask to emphasize foreground region and decrease background interferences.
2) To generate the object instances matched with the current model in different epochs of model training, a dynamic optimization objective is designed to reweight each sample

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3115796, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

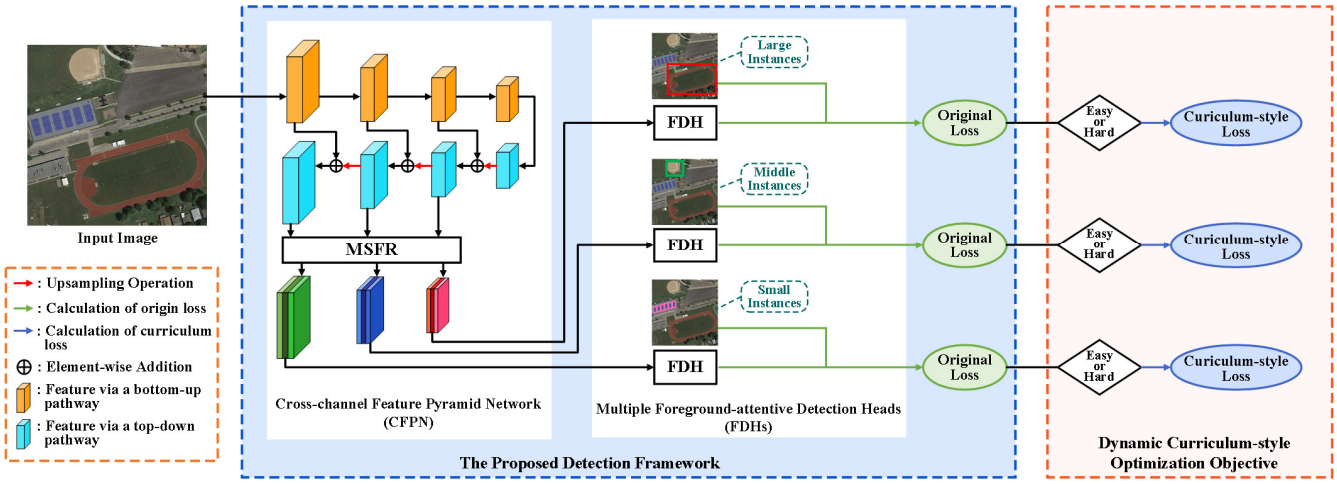

Fig. 1. The overview of the proposed detection framework and dynamic curriculum learning strategy. The detection framework is mainly composed of a cross-channel feature pyramid network and multiple foreground-attentive detection heads.

according to the initially calculated detection loss, further boosting the training stability and detection accuracy of the designed detection framework.

3) Extensive experiments have been conducted on the three publicly available object detection datasets, demonstrating that the superiority of the proposed method. Besides, thorough ablation studies further verify the effectiveness of each designed module.

The remainder of this paper is organized as follows: In Section II, the related works are briefly reviewed. In Section III, the proposed framework including CFPN and FDHs is described in detail, and then, the curriculum-style loss function is described. In Section IV, the experimental results, insightful analyses, and detailed discussions are given. Concluding remarks are drawn in Section V.

II. Related Works

A. Anchor-free Detectors

Anchor-free detectors generally consist of two parts: detection backbone and detection head. The detection backbone encodes the input images into high-level semantic features, while the detection head converts the above features into category prediction and position offset prediction according to the given object annotations. Unlike the anchor-based detections employing a predefined set of region proposals, the anchor-free detections decouple the predefined region proposals from the DNNs-based detection structure.

As a representative anchor-free detector, CenterNet [17] firstly extracts features from the input image $I \in \mathbb{R}^{W\times H\times 3}$ of width W and height H using the detection backbone. Then the detection head produces three prediction maps, which can be regarded as a center point heatmap $\hat{Y}\in[0,1]^{\frac{W}{R}\times\frac{H}{R}\times C}$, an object size map $\hat{S}\in\mathbb{R}^{\frac{W}{R}\times\frac{H}{R}\times 2}$, and an offset map $\hat{O}\in\mathbb{R}^{\frac{W}{R}\times\frac{H}{R}\times 2}$, where C is the number of classes and $R=4$ represents the downsampling factor of the extracted features. Based on the three predicted maps, CenterNet can produce a bounding box as follows:

$$\begin{aligned}&\left(\hat{x}_i+\delta\hat{x}_i-\hat{w}_i/2,\ \hat{y}_i+\delta\hat{y}_i-\hat{h}_i/2,\right.\\&\left.\hat{x}_i+\delta\hat{x}_i+\hat{w}_i/2,\ \hat{y}_i+\delta\hat{y}_i+\hat{h}_i/2\right)\end{aligned}\tag{1}$$

where $(\hat{x}_i,\hat{y}_i)$ denotes the prediction result of the center point, $(\hat{w}_i,\hat{h}_i)$ represents the prediction result of the object size, and $(\delta\hat{x}_i,\delta\hat{y}_i)$ denotes the prediction result of the offset. Specifically, the goal of the center point heatmap $\hat{Y}$ is to predict the keypoints of a certain class, and the object size map $\hat{S}$ aims to produce the corresponding scale predictions of an object. Besides, the offset is designed to recover the position offset of the center point location caused by the downsampling factor R. For example, a center point (x,y) in the input image is mapped to $(\lfloor x/R\rfloor,\lfloor y/R\rfloor)$ in the $\hat{Y}$, where $\lfloor\ \rfloor$ refers to round down. A real position offset $(x/R-\lfloor x/R\rfloor, y/R-\lfloor y/R\rfloor)$ needs to be predicted by the offset map $\hat{O}$.

B. Feature Pyramid Network

Due to that object detection task is often required to recognize and localize multiple instances with large scale variations, feature learning for scale-invariant patterns is crucial during the whole model training process [9], [30], [31]. Feature pyramid network (FPN) [9], as a representative backbone structure of multi-scale feature learning, aims to build multi-scale features from different high levels having rich semantics. Specifically, the FPN adopts a top-down pathway and lateral connections to construct a feature pyramid structure by combining two adjacent layers. In this architecture, a feature hierarchy consisting of multi-scale feature maps is firstly constructed by the backbone ConvNet (such as ResNets). Then the low-resolution, semantically strong features are upsampled in the top-down pathway and enhanced with features of the same spatial size from the backbone *via* the above lateral connections, in the form of element-wise addition [9]. Finally, different pyramidal layers are used to predict objects with different scales. Specifically, the layer with a relatively higher

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3115796, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

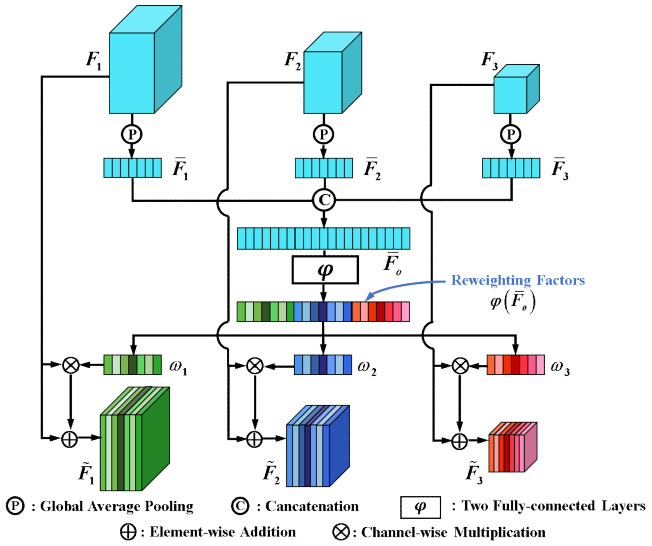

Fig. 2. The illustration of the multi-scale feature reweighting (MSFR).

resolution predicts smaller objects while larger objects are recognized in a lower resolution layer. Moreover, stacked discriminative sparse autoencoder (SDSAE) [32] improves the performance of land-use classification by introducing a weakly supervised feature transferring annotation framework to learn the relations of features from different scales, and discriminative CNNs (D-CNNs) [33] exploits a new object function to learn more scale-invariant and discriminative representations that have small within-class scatter and large between-class separation.

C. Curriculum Learning

There is an obvious phenomenon in the natural world that the learning process of humans or animals generally starts from relatively easier concepts and gradually progresses to complex ones. Motivated by this fact, curriculum learning [34]-[36] aims to obtain an order of sample learning, and thus boost the convergence speed of the sample training process. The main challenge in curriculum learning is how to accurately evaluate the *learning difficulty* of training samples. In earlier works [34], [36], the learning difficulty can be given using a predefined heuristic threshold that is determined according to the prior knowledge obtained by the input dataset. Moreover, self-paced learning [37] adopts another concept to evaluate the learning difficulty for each sample, where the difficulty is determined by the current status of the trained model. More recently, some natural image-based object detection methods using self-paced learning concept are designed [38]-[43]. For example, by assigning a learnable parameter (governing the importance of each instance) to each class and instance, a dynamic curriculum with learnable parameters is constructed so that the baseline detector can be optimized using the learned curriculum-related parameters [41]. Moreover, to select the most reliable samples for training in current status, the prediction scores of the previous iterations are comprehensively considered and further represented as the confidence of predicted boxes [42]. It should be noted that, in the field of RSIs-based weakly-supervised object detection, a few attempts also have been made to design an entropy-based easy-hard instance learning network by means of the self-paced learning concept [44].

III. The Proposed Method

To better deal with large variations of object sizes and complex background information in RSIs, a novel anchor-free detection framework is proposed with two main parts: a cross-channel feature pyramid network (CFPN) and multiple foreground-attentive detection heads (FDHs). Meanwhile, to avoid poor local optima in the training of models [34], a dynamic curriculum-style optimization is designed. The overall framework is shown in Fig. 1. Details of the proposed framework are elaborated as follows.

A. Cross-channel Feature Pyramid Network

Considering that the scales and layouts of objects in RSIs vary greatly, it is hard for anchor-based methods to find a suitable match between the default boxes to be predefined and the instance objects to be predicted. In order to eliminate these predefined sets of default boxes, we resort to the anchor-free detectors and select CenterNet [17] as our baseline model.

However, such an anchor-free detection backbone, CenterNet, only utilizes the single-scale features from the highest semantic level to predict potential instance objects in RSIs. As a result, the above single-scale feature representations will lose many spatial details of objects due to the downsampling operation in the backbone network, which may give rise to large recognition and localization errors for small or clustered objects, especially when instance objects are densely arranged. On the other hand, the semantic relations between features of different resolutions are neglected. To address the above issues, a CFPN is developed, which employs a pyramid feature structure (FPN) to extract features from different scales and adopts multi-scale feature reweighting (MSFR) to learn semantic relations across these multi-scale features.

As illustrated in Fig. 1, CFPN is a designed multi-layer anchor-free network. Firstly, to fully consider low-level spatial-detailed information and high-level semantically-rich information, FPN is used to fuse features from different semantic levels and generate pyramidal feature representations. With the employment of FPN, the multi-scale features from all scales are semantically strong, and thus objects of different scales could be predicted using the corresponding pyramidal layers which combine both spatial details and rich semantics. Meanwhile, in the multi-layer prediction of FPN, spatially adjacent objects could be assigned to different prediction layers, and thus the interferences from densely arranged instances could be greatly reduced.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3115796, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

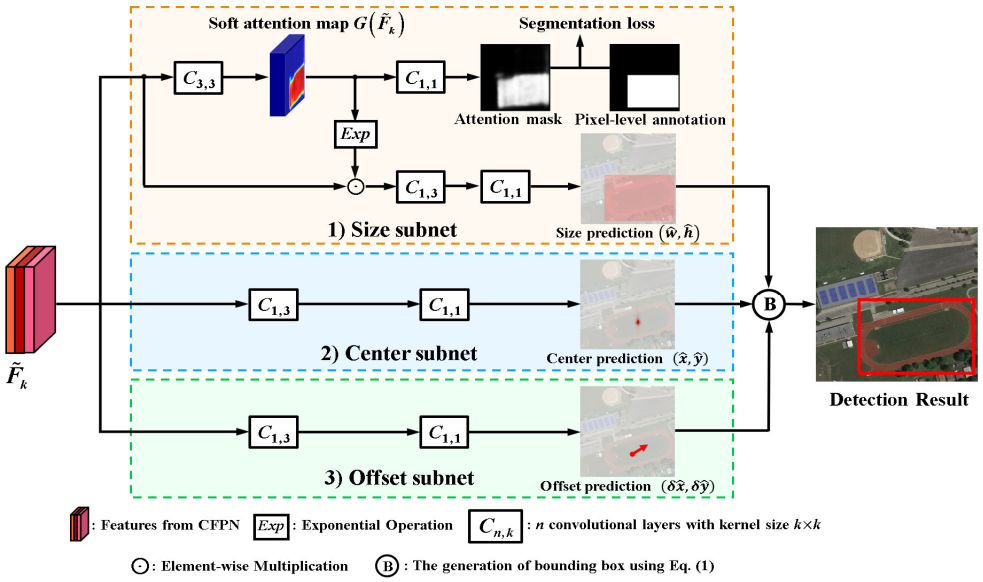

Fig. 3. The illustration of each FDH, which is composed of a size subnet with foreground enhancement, a center subnet, and an offset subnet.

Although FPN generates multi-scale feature representations, it considers each layer independently and ignores the relations across different scales. For example, "ship" and "harbor" categories often appear in the same RSI but are assigned to different network prediction layers, due to the large object size differences. However, under the FPN architecture, the above two categories with different object sizes are likely to be predicted independently, without the semantic-level information interaction across sizes/layers, which implies that the object detection network should have a good ability to encode the inter-scale/layer semantic relations. To obtain more discriminative feature maps across different scales/layers, MSFR is designed to learn inter-layer semantic relationships and reweight the input features at each scale.

Specifically, let $F_k \in \mathbb{R}^{H_k \times W_k \times C}\ (k=1,2,...,K)$ denote the k-th layer features extracted by the FPN, where K denotes the total layer number of pyramidal feature representations, and H_k, W_k and C represent the height, width, and channel number of the k-th layer features F_k, respectively. As shown in Fig. 2, a global average pooling operation is firstly performed on features F_k in each scale, generating channel-wise intra-layer statistics $\overline{F}_k \in \mathbb{R}^{1\times 1\times C}$. And then, cross-layer channel-wise representations $\overline{F}_o \in \mathbb{R}^{1\times 1\times KC}$ can be obtained by concatenating them as follows:

$$\overline{F}_o = [\overline{F}_1, \overline{F}_2, ..., \overline{F}_K] \tag{2}$$

where $[...,...]$ denotes the concatenation operation which concatenates K vectors with C dimensions along the channel direction to generate a vector with KC dimensions. Based on the cross-layer channel-wise representations $\overline{F}_o$, two fully-connected layers with learnable parameters φ are cascaded to learn inter-layer semantic dependencies across different scales, which can be represented by initial reweighting factors $\varphi(\overline{F}_o)$. Note that, in order to fully capture the relations between all channels, both of the two fully-connected layers preserve the channel dimension. Then, the final reweighting factors $\omega_k \in \mathbb{R}^{1\times 1\times C}$ which are split from the initial reweighting factors $\varphi(\overline{F}_o)$ are multiplied with the corresponding input features F_k in channel dimension to produce reweighted features. Finally, the reweighted features are added to the input features F_k to calculate the output features $\tilde{F}_k \in \mathbb{R}^{H_k \times W_k \times C}$ of the CFPN in each scale as follows:

$$Split(\varphi(\overline{F}_o)) = \omega_1, \omega_2, ..., \omega_K \tag{3}$$

$$\tilde{F}_k = (\omega_k \otimes F_k) \oplus F_k \tag{4}$$

where the $Split(\bullet)$ denotes the split operation which splits a vector with KC dimensions to K vectors with C dimensions along the channel direction, ω_k represents the final reweighted factors in the k-th layer, and $\otimes$ and $\oplus$ are the channel-wise multiplication and element-wise addition operations, respectively.

In MSFR block, with the guidance of such a channel-wise attention that can enforce the detection network to focus on the inter-scale semantic relationships, the features of strong correlations under different scales will be emphasized, and therefore the refined features contain more accurate information of all scales.

B. Foreground-attentive Detection Head

In RSIs, it is common that the background distribution is very complex and diversified. Even for objects belonging to the same class, the background representations in their neighborhood can vary greatly. Therefore, a robust detector should have a better ability to focus on the foreground region

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3115796, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

learning so that it can successfully detect foreground objects from the complex background.

To constrain the model to focus on learning foreground information in RSIs, the FDHs are proposed, as shown in Fig. 3. The motivation of the FDHs is to highlight the features from foreground objects and decrease the influence of cluttered background information. In order to realize it, an attention map is learned to identify the foreground region.

As shown in Fig. 3, each FDH consists of three subnets: size subnet, center subnet, and offset subnet. For the size subnet, given the output features $\tilde{F}_k \in \mathbb{R}^{H_k \times W_k \times C}$ that are calculated by the above CFPN, a soft attention map $G(\tilde{F}_k) \in \mathbb{R}^{H_k \times W_k \times C}$ can be obtained through a three-layer convolution operation where each layer has a kernel size of 3×3. Then the soft attention map $G(\tilde{F}_k)$ is split into two branches: 1) one branch produces a *single-channel attention mask* through a convolution layer with a kernel size of 1×1, which is supervised by the segmentation loss with a pixel-level annotation; 2) Based on the calculated soft attention map $G(\tilde{F}_k)$, the other branch can predict the size $(\hat{w},\hat{h})$ of an object by a convolution layer with a kernel size of 3×3 followed by a convolution layer with a kernel size of 1×1 as follows:

$$\hat{F}_k = \tilde{F}_k \odot Exp(G(\tilde{F}_k)) \tag{5}$$

where $\odot$ and Exp denote the element-wise multiplication operation and exponential operation, respectively, and the purpose of the exponential operation is to enhance the foreground regions in a soft attention way, avoiding the loss of useful information in feature maps. Besides, for the center subnet and size subset, they utilize a convolution layer with a kernel size of 3×3 followed by a convolution layer with a kernel size of 1×1 to generate the center prediction $(\hat{x},\hat{y})$ and offset prediction $(\delta\hat{x},\delta\hat{y})$, respectively.

Foreground enhancement is not applied to the center subnet and offset subnet, since the two subnets should focus on recognizing and localizing the center point of an object, but actually, the soft attention map $G(\tilde{F}_k)$ is used to highlight the object size in a region.

In particular, inspired by a pixel-level classification way in semantic segmentation, the pixel-level annotation for the foreground region can be produced from the bounding-box level annotation in RSIs. The pixel-level annotation is generated by labelling all pixels inside the object bounding box as 1 and others as 0, emphasizing all pixels of foreground objects without requiring any additional manual labelling information. Finally, with the aid of the predicted attention map, the features of foreground objects are enhanced and the prediction for instance objects becomes more accurate.

C. Initial Object Detection Loss Function

We employ a loss function that combines object detection loss and semantic segmentation loss to jointly optimize the parameters of the proposed framework. The overall loss function, which includes center loss L_{ch}, size loss L_{size}, offset loss L_{off}, and segmentation loss L_{seg}, can be formulated as follows:

$$L = \sum_k L_{ch}^{(k)} + \lambda_{size}\sum_k L_{size}^{(k)} + \lambda_{off}\sum_k L_{off}^{(k)} + \lambda_{seg}\sum_k L_{seg}^{(k)} \tag{6}$$

where k is the index of the feature pyramid level.

For the center subnet, focal loss [13] is used to assign larger weights to some hard samples during training as follows:

$$L_{ch} = -\frac{1}{N}\sum_{xyc}\begin{cases}\left(1-\hat{Y}_{xyc}\right)^{\alpha}\log\left(\hat{Y}_{xyc}\right) & , \quad \text{if } Y_{xyc}=1\\ \left(1-Y_{xyc}\right)^{\beta}\left(\hat{Y}_{xyc}\right)^{\alpha}\log\left(1-\hat{Y}_{xyc}\right) & , \quad \text{otherwise}\end{cases} \tag{7}$$

where Y_{xyc} and $\hat{Y}_{xyc}$ are the ground truth and prediction results of object centers, respectively. α and β are hyperparameters of the focal loss [13], and N denotes the number of keypoints in an input image. Besides, size and offset predictions are trained *via* a ℓ_1 regression loss as follows:

$$L_{size} = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{S}_i - S_i\right| \tag{8}$$

$$L_{off} = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{O}_i - O_i\right| \tag{9}$$

where $S_i = \left(x_2^{(i)}-x_1^{(i)}, y_2^{(i)}-y_1^{(i)}\right)$ is the ground truth of object size, $O_i = \left(x_c^{(i)}/R - \lfloor x_c^{(i)}/R \rfloor, y_c^{(i)}/R - \lfloor y_c^{(i)}/R \rfloor\right)$ is the ground truth of offset, and $\lfloor\ \rfloor$ refers to round down operation. Besides, $\hat{S}_i$ and $\hat{O}_i$ represent the prediction results of object size and offset, respectively. Further, the segmentation branch is trained using the binary cross-entropy loss as follows:

$$L_{seg} = -\frac{1}{N}\sum_{i=1}^{N}\hat{y}_i\log y_i + \left(1-\hat{y}_i\right)\log\left(1-y_i\right) \tag{10}$$

where y_i denotes the value of *i*-th pixel in pixel-level annotation and $\hat{y}_i$ represents the corresponding prediction result.

D. Final Curriculum-style Loss Function

In the early training stage of deep neural networks, some noise samples or outliers may lead the model to a bad local minimum, which is more serious in RSIs since the complex background and scene layout of RSIs will result in more noise samples and outliers [34]. If the learning order for training samples can be performed in an easy-to-hard process, the generalization capability of the trained model should be further boosted.

In order to provide the detection framework with samples that can match with current model ability, a dynamic curriculum-style optimization objective is designed. Inspired by the fact that detection loss intuitively reflects the learning difficulty of the sample instance in the current state, we use the loss in the current training iteration as the evaluation criterion of learning difficulty of a sample, and then transform it so as to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3115796, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

(a) (b) (c)

Fig. 4. Representative examples of three datasets. (a) DIOR. (b) NWPU VHR-10 (The first column is an RGB image and the second and the third columns are two pan-sharpened color infrared images). (c) RSOD.

TABLE I
EVALUATION RESULTS (MAP) ON DIOR, NWPU VHR-10 AND RSOD DATASETS. COO DENOTES CURRICULUM-STYLE OPTIMIZATION OBJECTIVE.

Method		DIOR	NWPU VHR-10	RSOD
Anchor-based	Faster RCNN [7]	54.1	87.3	92.0
	SSD [10]	58.6	83.1	87.9
	YOLOv3 [12]	57.1	87.3	\
	RetinaNet [13]	66.1	89.6	91.2
	FMSSD [25]	\	90.4	\
	CBD-E [26]	67.8	95.0	94.2
	DCL-Net [27]	\	94.6	94.6
Anchor-free	CenterNet [17]	57.7	83.7	78.3
	CornerNet [16]	64.9	\	\
	FCOS [18]	\	92.1	93.7
	FoveaBox [19]	69.0	91.4	94.0
	CFPN+FDHs+COO	**73.5**	**96.8**	**95.6**

TABLE II
RESULTS OF ABLATION STUDIES ON DIOR AND NWPU VHR-10 DATASETS. COO DENOTES CURRICULUM-STYLE OPTIMIZATION OBJECTIVE.

Method					DIOR		NWPU VHR-10		Params	Times (s)
Baseline	FPN	CFPN	FDHs	COO	mAP	mF1	mAP	mF1		
√					57.7	53.2	83.7	77.8	204.34M	0.065
√	√				68.3	63.6	93.7	87.4	209.35M	0.071
√		√			69.6	64.8	94.8	88.7	213.85M	0.071
√			√		60.2	56.5	86.8	79.9	209.44M	0.066
√		√	√		71.7	66.3	96.2	91.2	261.88M	0.078
√		√	√	√	**73.5**	**67.6**	**96.8**	**92.5**	261.88M	0.078

emphasize easy samples in the early stage of training. The designed curriculum-style loss is formulated as follows:

$$L_{cl} = \mu^{-\mathcal{M}(L-\tau)} \bullet L \quad (\mu > 1)$$
$$\mathcal{M}(x) = \begin{cases} 1, & x > 0 \\ -1, & x \le 0 \end{cases} \tag{11}$$

where L is the initial object detection loss calculated by Eq. (6), L_{cl} is the curriculum-style loss inserted after the object detection loss L, and τ is the threshold to distinguish easy samples from hard ones. The designed curriculum-style loss function can dynamically cherry-pick easy samples, since we preset a threshold τ that can be automatically calculated as the average of all losses from the training beginning up to now. The dynamic threshold τ is consistent with the global status of all stages so far and can accurately identify easy and hard samples.

For samples with losses smaller than the threshold τ, *i.e.*, $L - \tau \le 0$, we regard them as easy samples and the loss of them would be amplified since the weighting factor $\mu^{-\mathcal{M}(L-\tau)}$ would equal to μ *via* the designed function $\mathcal{M}(x)$. In contrast, hard samples with the initial object detection losses bigger than τ would be weakened, as the weighting factor equals to $1/\mu$.

With the aid of the designed curriculum-style loss function, easy samples can contribute more optimization loss for the update of model parameters during the earlier stage of the training process. As training continues, the predictive ability of the detection model will increase as the number of simple samples increases, so that the model can learn the curriculum-related knowledge of all samples globally.

Through such a reweighting of loss, the dynamic curriculum-style optimization objective successfully realizes a curriculum that pays more attention to easy samples in the early stage, without adding any extra learnable hyperparameters or computational cost. Meanwhile, the designed curriculum-style optimization objective always can fit the current status of the training model since the change of loss is synchronous with that of the model.

IV. EXPERIMENTAL RESULTS AND ANALYSES

A. Dataset Description

Our method is evaluated on the three public remote sensing object detection datasets, including DIOR [28], NWPU VHR-10 [1], and RSOD [29]. DIOR is a large-scale dataset with a wide range of object size variations. RSOD involves both urban and suburban backgrounds, which enriches the diversity of dataset. Unlike the above two datasets which contain only RGB images, NWPU VHR-10 consists of 715 RGB images and 85 pan-sharpened color infrared images. Representative instances of these three datasets are displayed in Fig. 4.

1) DIOR: The dataset is a very large-scale open-source data-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3115796, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

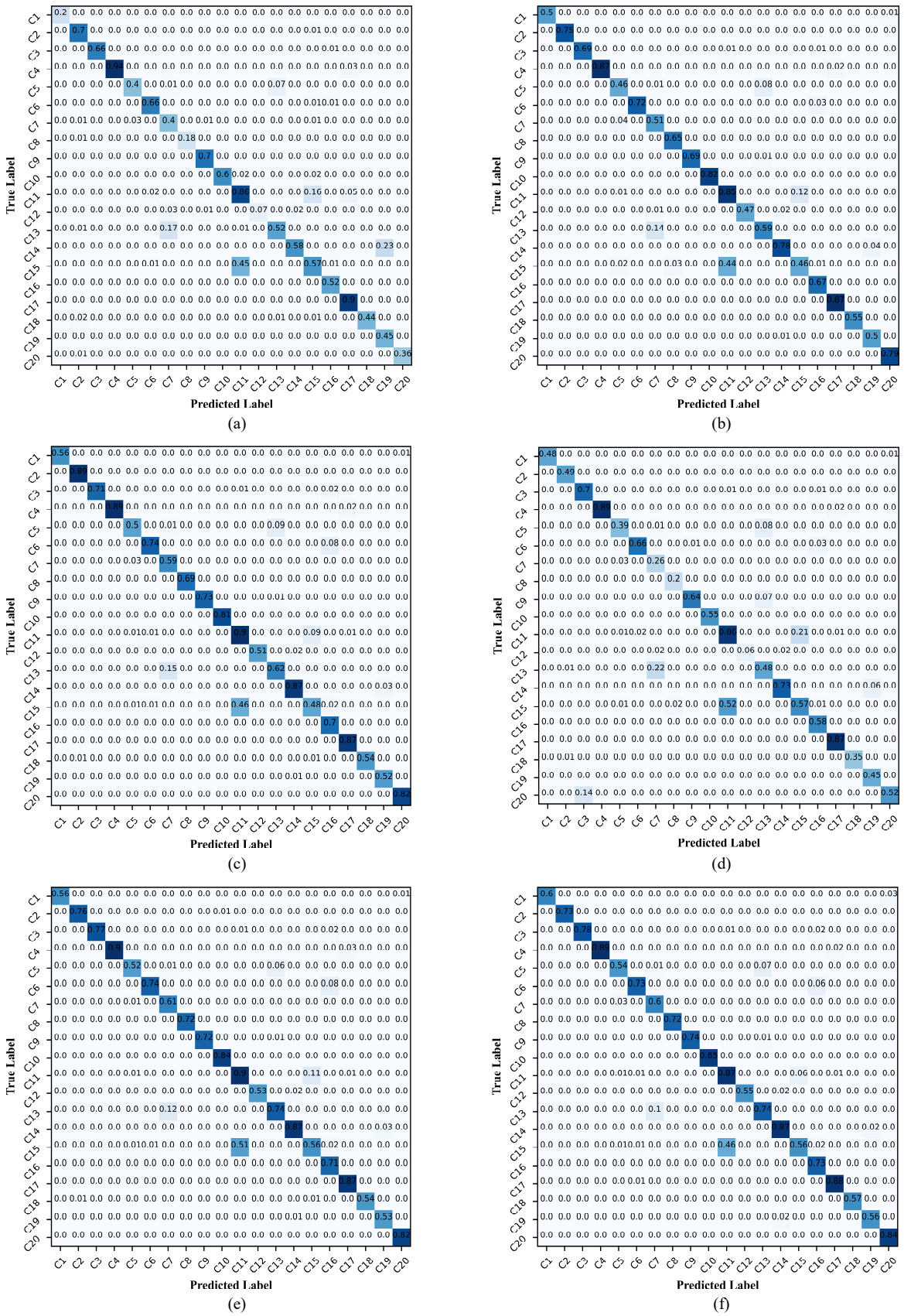

Fig. 5. The confusion matrices of ablation experiments on the DIOR dataset. (a) Baseline. (b) Baseline + FPN. (c) Baseline + CFPN. (d) Baseline + FDHs. (e) Baseline + CFPN + FHDs. (f) Baseline + CFPN + FHDs + COO, COO denotes curriculum-style optimization objective. The categories are listed in Table III.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3115796, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

TABLE III
20 OBJECT CATEGORIES IN THE DIOR DATASET.

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20
airplane	airport	baseball field	basket-ball court	bridge	chimney	dam	express-way service area	express-way toll station	golf field	ground track field	harbor	overpass	ship	stadium	storage tank	tennis court	train station	vehicle	wind mill

TABLE IV
EVALUATION RESULTS (MAP %) OF DIFFERENT BACKBONES ON DIOR AND NWPU VHR-10 DATASETS. OURS DENOTES CFPN+FDHs+COO.

Dataset	ResNet-50		ResNet-101		DLA-34	
	Baseline	Ours	Baseline	Ours	Baseline	Ours
DIOR	57.0	**70.7**	57.7	**73.5**	57.1	**72.9**
NWPU	82.3	**95.0**	83.7	**96.8**	83.3	**95.6**

set containing 20 kinds of objects categories and 23464 images with a unified size of 800×800 pixels, in which 192472 instances are annotated. Besides, we use the officially divided train set and validation set for model training, and test set for testing use.

2) NWPU VHR-10: It is a challenging ten-class geospatial object detection dataset. NWPU VHR-10 consists of 800 high-resolution RSIs, including 650 positive images and 150 negative images having no targets of the given object classes. In this paper, we randomly select 80% samples for model training and 20% samples for testing use from its positive image set, which is consistent with the splitting way in [25], [26].

3) RSOD: It is an open dataset for object detection in RSIs, which includes 4 kinds of objects: aircraft, playground, overpass, and oiltank. The dataset contains a total of 976 RSIs and 6950 instances. It is randomly split into train set and test set according to the proportion of 8 : 2, which is consistent with the splitting way in [26], [27].

B. Experimental Setup

We select CenterNet [17] as our anchor-free baseline model and adopt ResNet-101 [2] as the backbone network for feature extraction. For experiments on all benchmark datasets, following the hyperparameter setting in CenterNet [17], we set λ_{size}, λ_{off}, and λ_{seg} to 0.1, 1, and 1, respectively, and a Focal-loss parameters of $\alpha = 2$ and $\beta = 4$ for the center heatmap learning. During the training process, learning rate is set to 5e-5 and decays by a factor of 0.1 at epoch 20 and 30, respectively. For all datasets, the overall training process ends when 40 epochs are reached.

For the image size, the input images of three datasets are all resized to 800×800 pixels in the training phase and keep the original resolution during the inference stage. For all datasets, we use cropping, random flip and random scaling (between 0.6 and 1.3), as data augmentation in the training stage, which is same with the settings in CenterNet [17]. As for inference, flip test is utilized for augmentation.

For all the following experiments, mean average precision (mAP) and mean F1-socre (mF1) under the threshold of 0.5 is used to evaluate the model's average detection accuracy across all classes, which is consistent with the evaluation setup in [25]-[27].

C. Comparison with Other Methods

In this part, we compare the proposed methods with seven representative anchor-based object detection methods: Faster RCNN [7], SSD [10], YOLOv3 [12], RetinaNet [13], FMSSD [25], CBD-E [26], and DCL-Net [27]; and four state-of-the-art anchor-free detectors: CenterNet [17], CornerNet [16], FCOS [18] and FoveaBox [19]. Table I reports the comparison results of detection accuracy with the ten state-of-the-art methods on DIOR, NWPU VHR-10, and RSOD datasets, respectively. The experimental results demonstrate that the proposed framework (CFPN+FDHs+COO) achieves the highest mAP on all the benchmark datasets.

D. Ablation Studies

We perform ablation studies to verify the effectiveness of each designed part and the corresponding experimental results are reported in Table II. It can be seen from Table II that the combination of CFPN and FDHs is beneficial to improve detection accuracy. Meanwhile, the dynamic curriculum-style optimization objective further enhances the detection ability of the model. Compared with the baseline, the proposed framework significantly improves the mAP on DIOR and NWPU VHR-10 by 15.8% and 13.1%, respectively. Meanwhile, the results of mF1 on these two datasets also gain substantial improvements by 14.4% and 14.7%, respetively.

In order to analyze the effect of each part in detail, we draw the heatmap confusion matrix of each ablation experiment on DIOR dataset in Fig. 5. Comparing Fig. 5(a) and Fig. 5(c), it can be observed that the detection accuracy of almost all classes has been improved after introducing CFPN. The detection accuracy of some classes having large variations in object sizes, such as airplane, expressway service area, harbor, wind mill, vastly increases by even more than 30%, indicating that CFPN can cope well with the variable object size in RSIs. Moreover, for most classes, the detection results of CFPN are better than that of FPN, which can be concluded by comparing Fig. 5(b) and Fig. 5(c), especially for airport, dam, *etc*, indicating that the semantic relations are often helpful for identifying objects, and verifying the effectiveness of the proposed CFPN. As for FDHs, it can be observed from Fig. 5(d) and Fig. 5(e) that the detection accuracy can be further improved, especially for classes with a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3115796, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

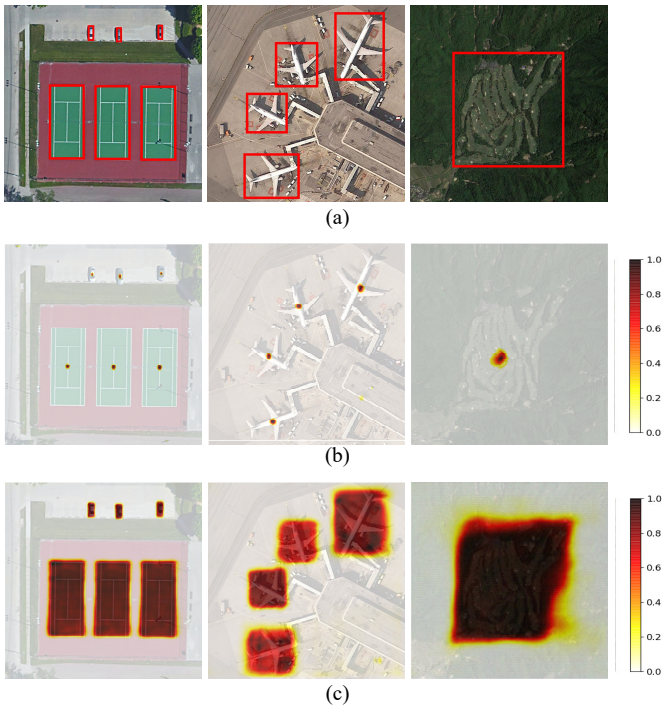

Fig. 6. The visualization of center points heatmaps and attention masks. (a) Ground truth. (b) Center points heatmap. (c) Attention mask.

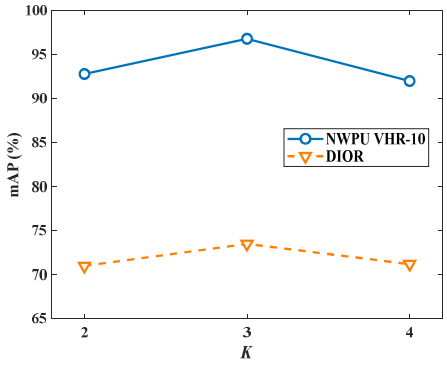

Fig. 7. Detection accuracy (mAP %) of employing different settings of the total layer number of pyramidal feature representations K .

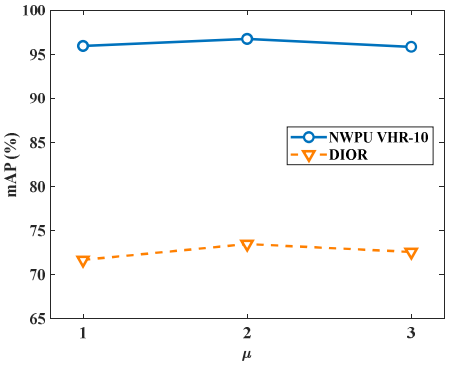

Fig. 8. Detection accuracy (mAP %) of employing different values of the weighting factor μ in Eq. (11).

relatively complex background, such as airplane, overpass, ship, and stadium. Besides, with the aid of the dynamic curriculum-style optimization objective, the overwhelming majority of classes gain higher detection accuracy, which can be observed through the differences between Fig. 5(e) and Fig. 5(f), demonstrating that a model with better generalization capability is obtained.

Moreover, due to that objects with the similar appearances widely exist in DIOR dataset, such as ground track field and stadium, dam and overpass, chimney and storage tank, some mis-predicted classes are often pairwise-related, which means that a certain hard-to-recognize class may be misclassified as another semantically-similar class and vice versa.

In the proposed framework, the backbone can extract features from the input images and may influence the detection results. To evaluate the effect of different backbones, we perform experiments on three representative backbones: ResNet-50 [2], ResNet-101 [2] and DLA-34 [17], and the corresponding experimental results are reported in Table IV, where the best results are marked in bold. It can be observed that all three backbones achieve state-of-the-art results in both datasets, which demonstrates the superiority and generalization of the proposed method. Moreover, ResNet-101 yields the highest mAP among the three backbones and is utilized in the following experiments.

E. Insightful Analyses and Visual Examples

In each FDH, the center subnet focuses on the center points of objects while the size subnet pays attention to the whole object by enhancing the foreground representations. To visualize the mechanism of them, we show the center points heatmaps and attention masks in Fig. 6. It can be seen that, for objects with various scales, the center points heatmaps highlight the centers of the target objects and the attention masks precisely activate the entire foreground regions.

Although the anchor-free detection backbone with a pyramid structure can bring considerable performance gains for object detection in RSIs, fusing the features from different semantic levels is still crucial and needs to be further studied. We investigate the impact of employing different numbers of prediction layers K on the detection accuracy. It can be observed from Fig. 7 that the optimal mAP can be obtained when the number of prediction layers equals to 3.

For the dynamic curriculum-style loss, the weighting factor μ affects the importance of different samples, as the weighting factor μ becomes larger, the attention on easy samples will be relatively added. We investigate the influence of the weighting factor μ in Fig. 8 and it can be observed that the best detection results are achieved when $\mu = 2$ for both datasets.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3115796, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

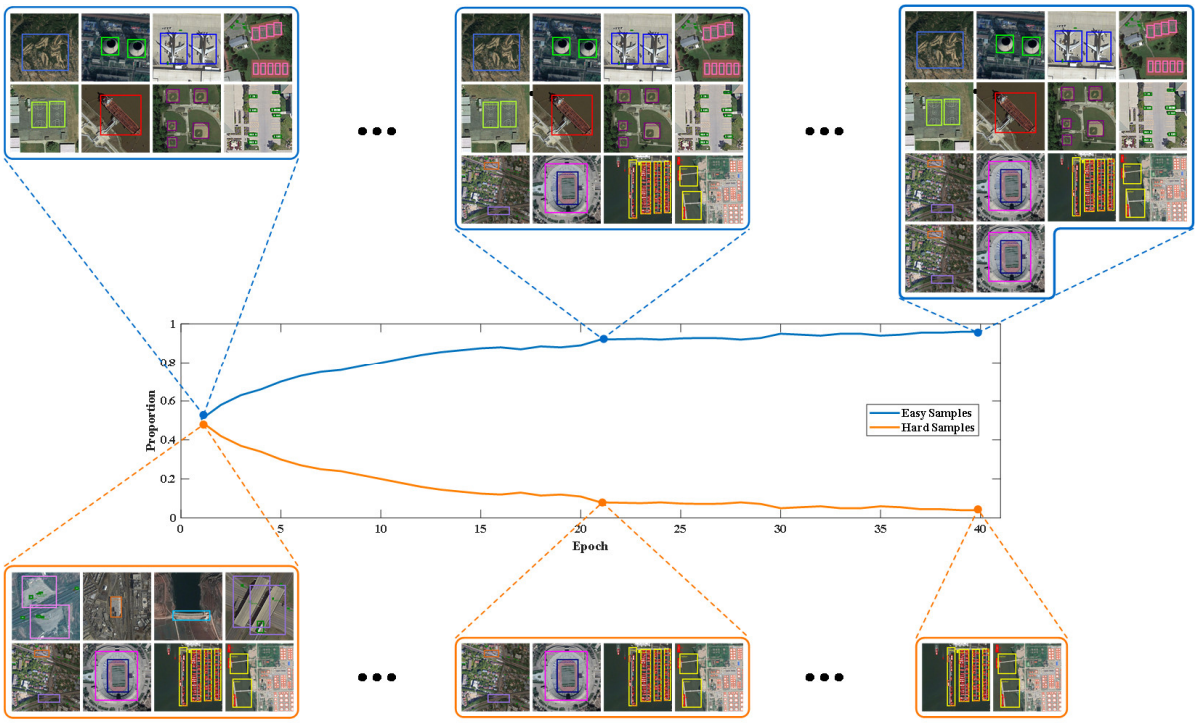

Fig. 9. The changing trend of the proportion of easy and hard samples.

Fig. 10. Visual detection results on DIOR dataset. (a) The ground-truth annotations. (b) Detection results predicted by CenterNet [17]. (c) Detection results predicted by the proposed framework.

Furthermore, in order to show the effect of dynamic curriculum-style optimization objective in the training process intuitively, we plot the variation of the proportion of easy and hard samples in Fig. 9. At the beginning of the training process, the easy and hard samples are about evenly divided because we use the average loss as the threshold. With the progress of training, the proportion of easy samples continues to increase while that of hard samples decreases,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3115796, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

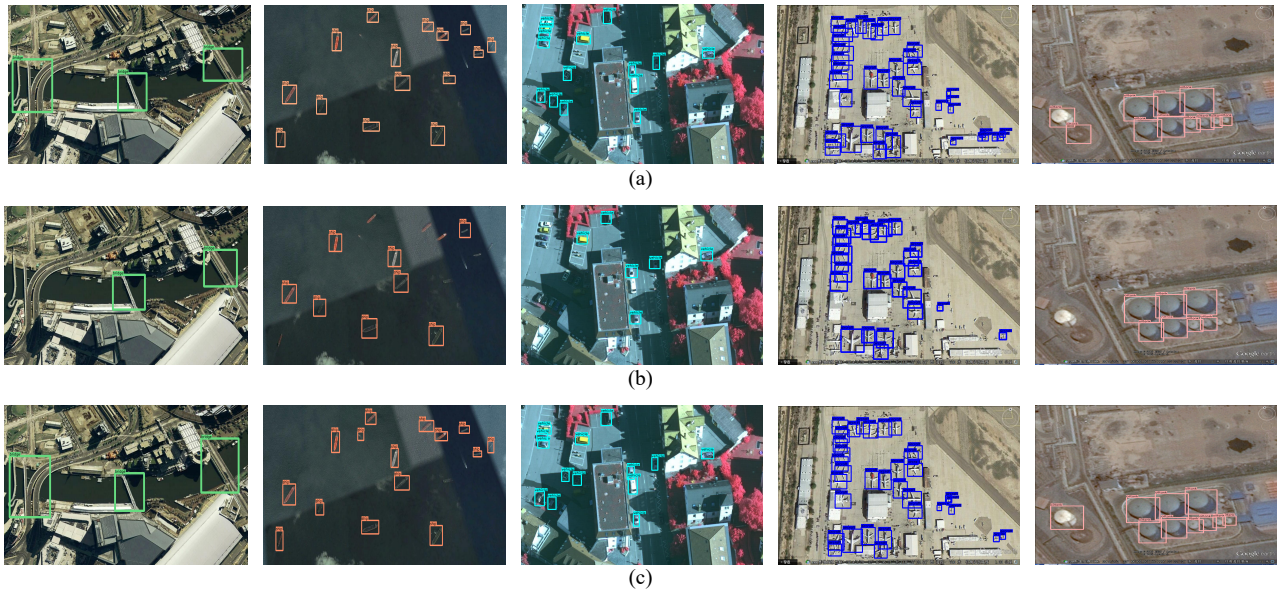

Fig. 11. Visual detection results on NWPU VHR-10 (3 columns on the left) and RSOD (2 columns on the right) datasets. (a) The ground-truth annotations. (b) Detection results predicted by CenterNet [17]. (c) Detection results predicted by the proposed framework.

and finally, most samples become easy. It is consistent with the idea in curriculum learning that easy samples should be emphasized first and then gradually transfer to overall samples.

Besides, Fig. 10 shows the visual detection results of different methods on DIOR. It can be observed that our method shows better visualization results for the object detection task in RSIs. Compared with the baseline [17], the proposed framework has a better ability to capture the very small objects (shown in columns 1 and 2), instance objects with the complicated background (shown in column 3) and the dense arrangement (shown in columns 4). Besides, for closely related species, like ship and harbor (column 4), vehicle and express service area (column 5), our framework which utilizes the relationship across different scales can boost the detection accuracy of these classes. Meanwhile, multi-layer prediction and foreground size attention ensure that objects can be predicted accurately.

Moreover, Fig. 11 shows the visual detection results of different methods on NWPU VHR-10 and RSOD. The three columns on the left are the detection results on NWPU VHR-10 and the two columns on the right are the detection results on RSOD. It can be observed that our method shows better visualization results than the baseline CenterNet [17] on both datasets.

V. Conclusion

In this work, we have proposed a novel region-free detection framework consisting of a cross-channel feature pyramid network (CFPN) and multiple foreground-attentive detection heads (FDHs), and have designed a dynamic curriculum-style optimization objective, towards detecting multi-class objects in optical RSIs. Considering that the object size and scene layout in RSIs often change and objects across different have semantic relations, the CFPN is proposed to predict multi-class objects with different scales and layouts without the requirement of predefining a set of region proposals. Due to the severe background interferences in RSIs, each FDH is developed to predict an attention map according to the semantic features produced by the CFPN, further enhancing the foreground representations by means of the calculated attention map. Meanwhile, a dynamic curriculum-style optimization objective is designed to learn features from easy to hard by reweighting samples and further boost the generalization capability of the detection model. The experimental results on the three public benchmark datasets demonstrate the effectiveness and superiority of the proposed method over several state-of-the-art object detection methods.

When the orientation of an object is arbitrary, it is difficult for the traditional horizontal bounding box to provide the accurate location of the object, since the horizontal bounding box covers a larger region of background. On the contrary, oriented bounding boxes could overcome this disadvantage by tightly surrounding the boundary of the target object. In future work, we would try to supplement an angle prediction branch into our framework and extend it to detect objects in various orientations.

References

[1] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 11–28, Jul. 2016.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, Apr. 2015, pp. 1-10.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classificationwith deep convolutional neural networks," *in Anvances in neural information processing systems*, vol 25, pp. 1097-1105, 2012.

[5] 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. (CVPR)*, Jun. 2014, pp. 580-587.

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

[7] 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.

[8] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask-RCNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Jun. 2017, pp. 2961-2969.

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

[10] W. Liu *et al*., "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 21–37.

[11] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779-788.

[12] J. Redmon, and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv*: 1804.02767. [Online]. Available: http://arxiv.org/abs/1804.02767

[13] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[14] B. Zhang, T. Chen, B. Wang, X. Wu, L. Zhang and J. Fan, "Densely semantic enhancement for domain adaptive region-free detectors," *IEEE Trans. Circuits Syst. Video Technol.*, doi: 10.1109/TCSVT.2021.3069034

[15] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[16] H. Law and J. Deng. "Cornernet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 734–750.

[17] X. Zhou, D. Wang, and Krähenbühl, "Object as points," 2019, *arXiv:*1904.07850, [Online]. Available: http://arxiv.org/abs/1904.07850

[18] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9627-9636.

[19] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, and J. Shi, "FoveaBox: Beyond anchor-based object detector," 2019, *arXiv*:1904.03797. [Online]. Available: http://arxiv.org/abs/1904.03793

[20] C. Zhu, Y. He, and M. Savvides, "Feature selective anchor-free module for single-shot object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 840-849.

[21] 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. 2016.

[22] 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.

[23] P. Ding, Y. Zhang, W. Deng, P. Jia, and A. Kuijper, "A light and faster regional convolutional neural network for object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens*., vol. 141, pp. 208–218, Jul. 2018.

[24] 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, Dec. 2019.

[25] 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.

[26] J. Zhang, C. Xie, X. Xu, and Z Shi, "A contextual bidirectional enhancement method for remote sensing image object detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens*., vol. 13, pp. 4518–4531, Aug. 2020.

[27] E. Liu, Y. Zheng, B. Pan, X. Xu, and Z. Shi, "DCL-Net: Augmenting the capability of classification and localization for remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, doi: 10.1109/TGRS.2020.3048384.

[28] 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.

[29] Y. Long, Y. Gong, Z. Xiao, and Q. Liu, "Accurate object localization in remote sensing images based on convolutional neural networks," *IEEE Trans. Geosci. and Remote Sens.*, vol. 55, no. 5, pp. 2486-2498, May 2017.

[30] 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. (CVPR)*, Jun. 2018, pp. 8759-8768.

[31] G. Ghiasi, T. Lin, and Q. V. Le, "NAS-FPN: Learning Scalable Feature Pyramid Architecture for Object Detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7029-7038.

[32] X. Yao, J. Han, G. Cheng, X. Qian and L. Guo, "Semantic annotation of high-resolution satellite images via weakly supervised learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3660-3671, Jun. 2016.

[33] G. Cheng, C. Yang, X. Yao, L. Guo and J. Han, "When deep learning meets metric learning: remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811-2821, May 2018.

[34] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. 26th Annu. Int. Conf. Mach. Learn. (ICML)*, Jun. 2009, pp. 41–48.

[35] G. Hacohen and. Weinshall, "On the power of curriculum learning in training deep networks," in *Proc. 36th Annu. Int. Conf. Mach. Learn. (ICML)*, May 2019, pp: 2535-2544.

[36] F. Khan, B. Mutlu, and J. Zhu, "How do humans teach: On curriculum learning and teaching dimension," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Dec, 2011 pp: 1449-1457.

[37] M. P. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Dec, 2010, pp.1189-1197.

[38] L. Jiang, Z. Zhou, T. Leung, L. LI, and F. Li, "Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels," in *Proc. 35th Annu. Int. Conf. Mach. Learn. (ICML)*, Jul. 2018, pp: 2304-2313.

[39] J. Han, P. Luo, and X. Wang, "Deep self-learning from noisy labels," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2019, pp. 1440-1448.

[40] Y. Wang, W. Gan, J. Yang, W. Wu, and J. Yan, "Dynamic curriculum learning for imbalanced data classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2019, pp. 5017-5026.

[41] S. Saxena, O. Tuzel, and D. DeCoste, "Data parameters: A new family of parameters for learning a differentiable curriculum," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Dec, 2019, pp: 11095-11105.

[42] E. Sangineto, M. Nabi, D. Culibrk, and N. Sebe, "Self paced deep learning for weakly supervised object detection," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 41, no. 3, pp. 712–725, Mar. 2019.

[43] P. Soviany, R. Ionescu, P. Rota, ans N. Sebe, "Curriculum self-paced learning for cross-domain object detection," 2019, *arXiv:1911.06849*. [Online]. Available: http://arxiv.org/abs/1911.06849

[44] X. Yao, X. Feng, J. Han, G. Cheng, and L. Guo, "Automatic weakly supervised object detection from high spatial resolution remote sensing images via dynamic curriculum learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 675-685, Jan. 2021.