Few-Shot Object Detection in Remote Sensing: Lifting the Curse of Incompletely Annotated Novel Objects

Fahong Zhang, Yilei Shi, *Member, IEEE*, Zhitong Xiong, *Member, IEEE*, and Xiao Xiang Zhu, *Fellow, IEEE*

***Abstract*— Object detection (OD) is an essential and fundamental task in computer vision (CV) and satellite image processing. Existing deep learning methods have achieved impressive performance thanks to the availability of large-scale annotated datasets. Yet, in real-world applications, the availability of labels is limited. In this article, few-shot OD (FSOD) has emerged as a promising direction, which aims at enabling the model to detect novel objects with only few of them annotated. However, many existing FSOD algorithms overlook a critical issue: when an input image contains multiple novel objects and only a subset of them are annotated, the unlabeled objects will be considered as background during training. This can cause confusions and severely impact the model's ability to recall novel objects. To address this issue, we propose a self-training-based FSOD (ST-FSOD) approach, which incorporates the self-training mechanism into the few-shot fine-tuning process. ST-FSOD aims to enable the discovery of novel objects that are not annotated and take them into account during training. On the one hand, we devise a two-branch region proposal networks (RPNs) to separate the proposal extraction of base and novel objects. On the another hand, we incorporate the student-teacher mechanism into RPN and the region-of-interest (RoI) head to include those highly confident yet unlabeled targets as pseudolabels. Experimental results demonstrate that our proposed method outperforms the state of the art in various FSOD settings by a large margin. The codes will be publicly available at: https://github.com/zhu-xlab/ST-FSOD.**

***Index Terms*— Few-shot learning, object detection (OD), remote sensing image processing, self-training.**

Manuscript received 20 June 2023; revised 28 September 2023 and 9 November 2023; accepted 21 November 2023. Date of publication 8 January 2024; date of current version 10 January 2024. This work was supported in part by the German Research Foundation (DFG GZ: ZH 498/18-1) under Project 519016653; in part by the German Federal Ministry of Education and Research (BMBF) within the framework of the International Future AI Laboratory "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" under Grant 01DD20001; in part by the German Federal Ministry for Economic Affairs and Climate Action within the framework of the "National Center of Excellence ML4Earth" under Grant 50EE2201C; in part by the German Federal Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection (BMUV) Based on a Resolution of the German Bundestag (EKAPEx) under Grant 67KI32002B; and in part by the Munich Center for Machine Learning. *(Corresponding author: Xiao Xiang Zhu.)*

Fahong Zhang and Zhitong Xiong are with the Chair of Data Science in Earth Observation, Technical University of Munich, 80333 Munich, Germany (e-mail: fahong.zhang@tum.de; zhitong.xiong@tum.de).

Yilei Shi is with the School of Engineering and Design, Technical University of Munich (TUM), 80333 Munich, Germany (e-mail: yilei.shi@tum.de).

Xiao Xiang Zhu is with the Chair of Data Science in Earth Observation, Technical University of Munich, 80333 Munich, Germany, and also with the Munich Center for Machine Learning, 80333 Munich, Germany (e-mail: xiaoxiang.zhu@tum.de).

Digital Object Identifier 10.1109/TGRS.2023.3347329

I. Introduction

OBJECT detection (OD) is a critical task in computer vision (CV) as well as remote sensing image processing, which enables the automatic identification and localization of objects of interest within an image. With the rapid development of deep learning techniques [2], [3], [4] and the emergence of large-scale human-annotated data [5], [6], the performance of the state-of-the-art OD approaches has been pushed to a new stage. These approaches have achieved remarkable success in detecting objects in various domains, including remote sensing [7], [8], [9]. However, traditional OD methods rely on a large amount of labeled data for training, which can be challenging and time consuming to obtain in remote sensing image processing, particularly in scenarios where novel or rare objects are involved.

This results in the need for few-shot learning [10], [11], [12], a paradigm that aims to overcome the data scarcity issue by enabling models to generalize and detect objects with only a limited number of labeled examples. Few-shot learning achieves this by leveraging knowledge acquired from previously seen categories to adapt and recognize novel objects efficiently.

Concurrently with the achievements in few-shot classification [13] and few-shot semantic segmentation [14], [15], few-shot object detection (FSOD) [16], [17], [18] has emerged as a compelling research area in recent years. In the conventional FSOD framework, the model undergoes a two-stage training process: first, it is trained on a large-scale labeled dataset consisting of base objects, and subsequently, it is fine tuned on a fine-tuning set with only a few labeled novel object instances.

However, when there are multiple novel objects in a single image, it is possible that only a part of them are provided with labels during the fine-tuning stage. As a result, these incomplete annotations can negatively impact the training toward novel classes and hinder the discovery of novel objects. This issue, illustrated in Fig. 1, can be referred as the incompletely annotated novel objects (IANOs) issue.

While the challenge of IANO has been investigated in the field of FSOD for natural images [19], [20], it still remains unexplored in remote sensing. However, objects in the remote sensing images are usually smaller, and scenes often exhibit higher levels of congestion, particularly in contexts with vehicles, planes, and ships. This phenomenon can be

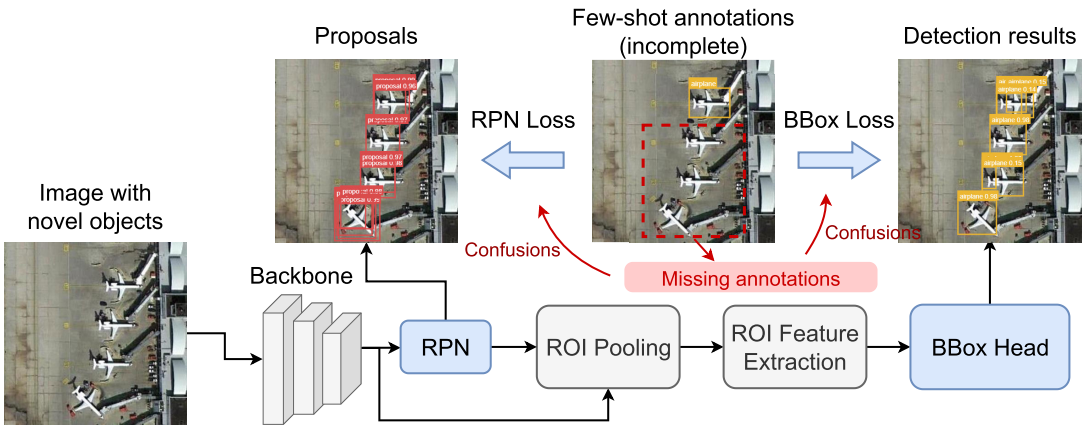


Fig. 1. Illustration of the IANOs issue. In the standard FSOD protocol, only a few bounding boxes for the novel object are provided for the few-shot fine-tuning stage [1]. Let us consider a scenario where our goal is detecting the presence of a "plane" as the novel object, and we are provided with just one bounding box annotation for a plane. As depicted in this figure, the challenge arises when multiple instances of planes are present within a single image. In such cases, some of the planes within the image are left unannotated. This can mislead the detector since RPN loss and bounding box classification loss are calculated based on these incomplete annotations.

readily observed from different OD datasets tailored for remote sensing applications [21]. As a result, the IANO problem takes on an even more formidable and pressing nature when considered in the context of remote sensing.

To tackle this issue, it is necessary to establish a mechanism capable of identifying and subsequently excluding potential unannotated novel objects during the background sampling process. A promising solution that emerges for mitigating this concern is self-training, a well-established technique in the field of domain adaptation [22], [23], [24] and semi-supervised learning [9], [25]. This type of methods first generate pseudolabels on unlabeled data using a pretrained model and then fine tune the model using these pseudolabels. The underlying philosophy of generating pseudolabels and leveraging the unlabeled data aligns with our objective of identifying potential unannotated novel objects, as they are hidden in the background and remain unannotated. Therefore, we propose using self-training as a feasible approach to tackle this issue in remote sensing imagery.

We build our self-training-based FSOD (ST-FSOD) method based on a popular OD framework: faster region-based convolutional neural networks (R-CNN) [2], which consists of two stages. In the first stage, a region proposal network (RPN) is used to generate a set of candidate object proposals, each with a corresponding objectness score and a bounding box regression score. In the second stage, the candidate object proposals are refined and classified using a fully connected layer network, also known as a region-of-interest (RoI) pooling layer [26].

We incorporate the self-training mechanism into the RPN and the bounding box head (BBH) of the RoI layer, and devise a self-training RPN (ST-RPN) and a self-training BBH (ST-BBH) modules accordingly. In these two modules, a momentum-based teacher-student modeling strategy [27] is used to filter out highly confident potential novel objects and refine the loss calculation. More specifically, the ST-RPN module uses the teacher–student modeling strategy to filter out highly confident proposals for potential novel classes. Meanwhile, the ST-BBH module filters out highly confident novel bounding boxes. Both modules refine the loss calculation to improve the accuracy of the detection model. Our contributions can be summarized as follows.

1) Our study highlights and examines the challenge of the IANO issue in FSOD for remote sensing imagery. To the best of the authors' knowledge, this is the first work that discusses and tackles the issue in the field of remote sensing, which is neglected in many existing FSOD methods for remote sensing imagery and needs to be addressed to advance the field.
2) To handle the IANO issue in OD, we propose to apply self-training technique. To this end, ST-RPN and ST-BBH modules are devised to identify proposals or bounding boxes that are likely to include a novel object, even in the absence of novel annotations.
3) We conduct extensive experiments on three publicly available datasets and evaluate our proposed method under various FSOD settings. Experimental results demonstrate that our approach outperforms the state-of-the-art methods by a significant margin.

II. Related Works

A. OD

OD refers to identifying and localizing objects within an image, which has been one of the main research tasks in CV. Traditional OD techniques are based on, e.g., feature extraction [28], object recognition, and template matching [29]. State-of-the-art OD techniques are mostly based on deep learning, due to its overwhelming performance on large-scale OD benchmarks. Deep learning-based OD methods utilize convolutional neural networks (CNNs) to perform OD directly from raw image pixels, without the need for hand-crafted feature engineering. They can be categorized into two-stage and single-stage detectors. While two-stage detectors aim to first generate object proposals and

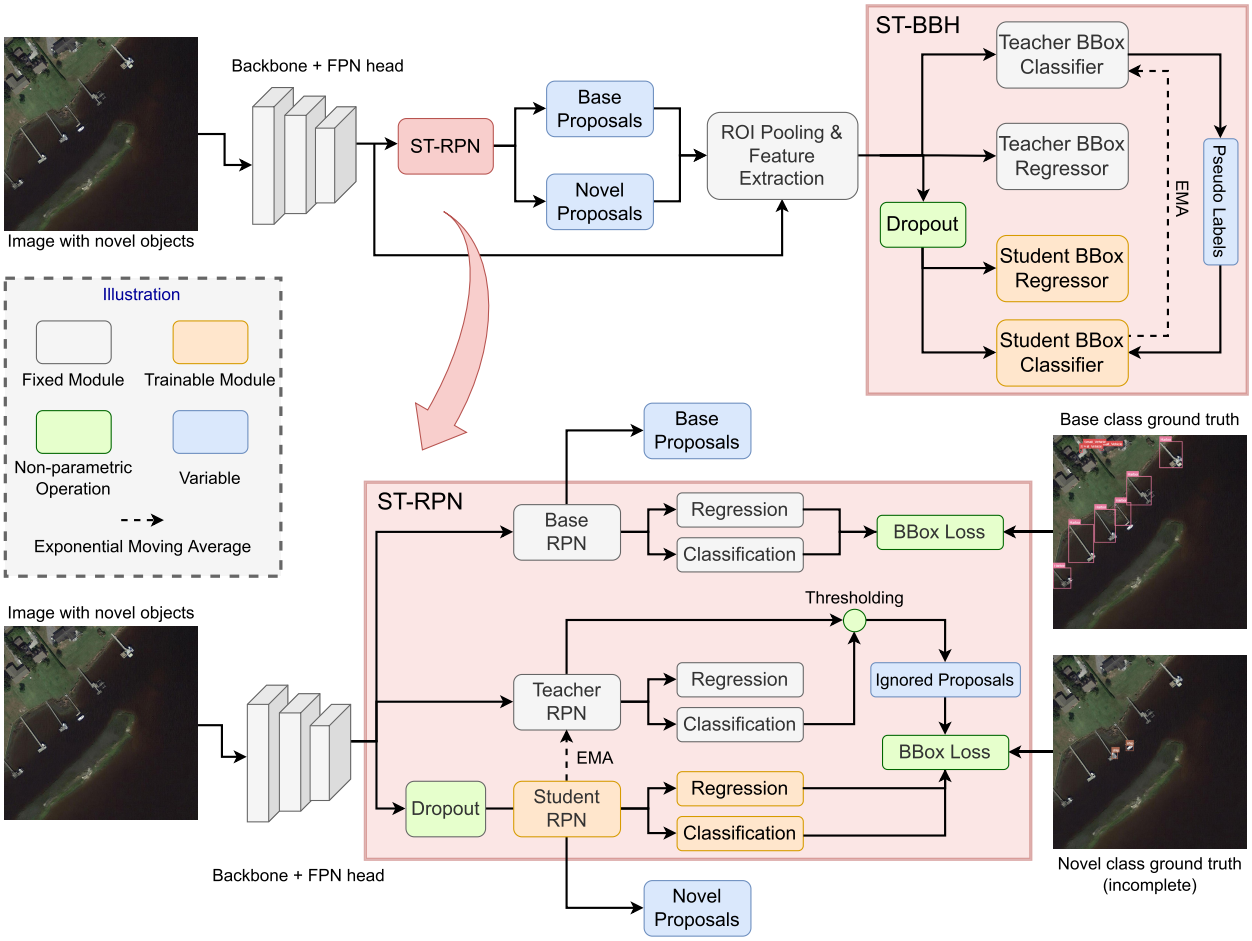


Fig. 2. Overall architecture of the proposed method. The FPN following the backbone is not illustrated in the figures for the sake of simplicity.

then classify them, single-stage detectors perform both tasks simultaneously. A typical example of two-stage detectors is faster R-CNN [2], which generates a set of region proposals and then extracts features from each proposal using an RoI pooling layer before classifying the object within the proposal. One well-known single-stage detector is you only look once (YOLO) [3], which applies a CNN to the entire image to simultaneously predict bounding boxes and class probabilities for each object without any proposal generation step.

B. FSOD in Computer Vision

FSOD aims at recognizing novel or unseen object classes based only on a few examples of them, by fine tuning on a model trained on many labeled examples of base classes. FSOD methods can be roughly categorized into fine-tuning-based methods, meta-learning-based methods, and metric-learning-based methods.

1) Fine-Tuning-Based Methods: Fine-tuning-based methods are popular for FSOD, which first train on a large number of base class examples and then perform few-shot fine tuning on a smaller support set that includes both base and novel classes.

One such method, low-shot transfer detector (LSTD) [16], uses a flexible deep architecture that integrates the advantages of both single shot multibox detector (SSD) and faster R-CNN in a unified deep framework. It also introduces transfer knowledge (TK) and background depression (BD) regularizations to leverage object knowledge from source and target domains during the fine-tuning stage.

Another fine-tuning-based method, two-stage fine-tuning approach (TFA) [17], indicates that even a simple fine tuning of only the last layer of a faster R-CNN detector to novel classes can achieve better performance compared with the previous meta-learning-based methods. In addition, TFA replaces the fully connected classification heads of faster R-CNN with cosine similarities. The aim is to reduce intraclass variances and preserve performance in base classes through this feature normalization technique.

Yang et al. [30] introduce a pretrain-transfer framework (PTF) that utilizes a knowledge inheritance approach to initialize the weights for the box classifier. In addition, they develop an adaptive length rescaling strategy to ensure consistency of the dimensions of the pretrained weights for both the base and novel classes. This helps to improve the efficiency and effectiveness of the fine-tuning process.

2) Metric-Learning-Based Methods: Metric-learning-based methods aim to reduce the dimensionality of each sample and learn a feature representation such that similar samples are closer to each other, while dissimilar ones are easier to discriminate.

RepMet [31] is a metric-learning approach suitable for both few-shot classification and OD. It uses a collection of Gaussian

mixture models, each with multiple modes, to describe the base and novel classes. During base training, an embedding loss is employed to ensure a margin between the distance of each query feature and its respective class representative, as well as the distance to the nearest representative of an incorrect class.

Fan et al. [32] introduce an attention-based RPN (attention-RPN) that uses support features to enhance the proposal generation process and eliminate nonmatching or background proposals. In addition, the authors develop a multirelation detector for feature representation learning that measures the similarity between the RoI features of query and support objects.

3) Meta-Learning-Based Methods: Meta-learning-based methods are designed to learn quickly from a few examples, using a meta-learner that has been trained on a diverse set of tasks. During the base training stage, the meta-learner is trained on a meta-dataset composed of various tasks. Once trained, the meta-learner can quickly adapt to new tasks or generate a learner that is customized to the target task.

One example of meta-learning-based methods is Meta-YOLO [33], which improves the query feature of a model by using a set of weighting coefficients generated during the meta-learning phase. These coefficients are based on the support samples and allow the model to effectively learn the intrinsic importance of features for detecting objects. This enables the reweighting coefficients of novel classes to be learned with only a few support examples.

Another approach is FsDetView [34], where a novel technique for aggregating query and support features is introduced. Instead of feature reweighting, this technique involves performing element-wise multiplication, subtraction, and concatenation between the two sets of features. This approach has shown promising results in FSOD tasks.

C. FSOD in Remote Sensing

Remote sensing images have unique characteristics that distinguish them from natural scene images, such as complex backgrounds, objects with multiple orientations, and dense and small objects. Thus, designing an FSOD algorithm that accounts for these distinct features is crucial.

In shared attention module (SAM) and Balanced Fine-tuning Strategy (BFS), Huang et al. [35] propose a shared attention module that leverages class-agnostic prior knowledge gained during the base training stage to aid in detecting novel objects with significant size variations. In DH-FSDet, Wolf et al. [36] suggest using a balanced sampling strategy (BSS) between base and novel samples, taking all the base samples into consideration. In addition, they propose separating the classification and regression heads in the RoI layer according to base and novel classes for better balance in detection. Zhang et al. [37] introduce the generalized FSOD task to remote sensing. On one hand, they propose a metric-based discriminative loss to reduce the intraclass diversity and increase the interclass separability of the base and novel objects. On the other hand, they replace the RoI feature extractor with a representation compensation module to prevent the model from catastrophic forgetting.

More recently, researchers have explored integrating text data into the visual learning pipeline to improve FSOD performance. Models such as text-modal knowledge (TEMO) [38] and text semantic fusion relation graph reasoning (TSF-RGR) [39] leverage TEMO extractors to provide prior knowledge on the relationship between base and novel classes, resulting in improved FSOD performance.

D. Self-Training

Self-training [40], [41], [42], [43] is a widely used approach for domain adaptation [44] and semi-supervised learning [25]. This technique involves generating pseudolabels for the target domain or unlabeled data and then using them to fine-tune the network. Self-training has shown promising results in various tasks, including semi-supervised classification [45] and semantic segmentation [46].

In semi-supervised classification, FixMatch [47] is a popular approach for self-training. It generates pseudolabels by using a model's predictions on weakly augmented and unlabeled images. These pseudolabels are then used to supervise the model's predictions on a strongly augmented view of the same image. This approach has been shown to achieve state-of-the-art performance on several benchmark datasets.

For semantic segmentation tasks, self-supervised augmentation consistency (SAC) [27] is a self-training approach that employs a momentum network as the teacher network. The teacher network generates pseudolabels on the weakly augmented images, which are used to supervise the student network that receives the strongly augmented images. The momentum network is updated based on the exponential moving average of the student network. This approach has also shown promising results on several benchmark datasets.

E. IANOs Issue

Li et al. [48] are the first to identify the IANO issue in their work, pointing out that the base set images might contain unlabeled novel objects, leading to false positives. They address this challenge by introducing a distractor utilization loss. More specifically, for each annotated bounding box, a cropped image centered at the object will be fed into a few-shot correction network. This network generate corresponding pseudolabels, which are subsequently used to identify potential novel objects and adjust the loss calculation with respect to the RoI head. On a similar note, Qiao et al. [19] emphasize that the IANO issue also exists when multiple novel objects were present within a single image. To resolve this concern, they propose a label calibration method. This method recalibrated the predicted targets of background objects based on their predicted confidence. As a result, unannotated novel objects are assigned with lower weights during the loss calculation, mitigating their negative impact.

Despite these valuable contributions, it is worth noting that the aforementioned approaches primarily addressed the impact of unannotated objects on bounding box classification, neglecting their influence on the training of RPNs for a two-stage object detector. Specifically, when calculating the RPN loss, the incompleteness of novel object annotations could cause the RPN to mistakenly predict unannotated novel objects

as background, thereby reducing the overall performance. In our method, we propose a comprehensive approach to mitigate the IANO issue. We apply advanced self-training techniques not only to the bounding box classification head but also to the RPN. By doing so, we extend our efforts to tackle this challenge on a broader scale, addressing the impact of unannotated objects throughout the entire OD pipeline.

III. Methods

In this section, we introduce the proposed ST-FSOD method, which consists of two major components: the ST-RPN and the self-training bounding box head (ST-BBH). The overall architecture is illustrated in Fig. 2. The ST-RPN module takes the multilevel features extracted from the backbone and feature pyramid networks (FPNs) head [49] as input and generates two sets of object proposals, namely, the base and novel proposals, corresponding to the base and novel categories, respectively. These proposals are merged and fed into the ROI pooling and feature extraction layers to obtain the ROI features. The ST-BBH module takes the ROI features as input and produces the final detection results. Specifically, it detects potential unannotated novel class objects and uses them as pseudolabels to recall more novel class objects, thereby improving the model's performance.

A. Problem Formulation

In this section, we present the formulation of the standard FSOD setting. We assume that we have access to a base set containing base class annotations denoted by $\mathcal{D}_{\text{base}} = \{(\mathbf{I}_i, \mathcal{B}_i^{\text{base}})\}$ and a novel set containing novel class annotations denoted by $\mathcal{D}_{\text{nov}} = \{(\mathbf{I}_i, \mathcal{B}_i^{\text{nov}})\}$, where $\mathbf{I}_i$ represents an image and $\mathcal{B}_i = \{(x, y, h, w, c)\}$ represents a set of object bounding boxes within the image. Here, x, y, w, and h denote the locations of the bounding box and c denotes the class label.

Furthermore, let $\mathcal{C}_{\text{base}}$ and $\mathcal{C}_{\text{nov}}$ be the label set of $\mathcal{B}_{\text{base}}$ and $\mathcal{B}_{\text{nov}}$, respectively, and they satisfy the condition $\mathcal{C}_{\text{base}} \cap \mathcal{C}_{\text{nov}} = \varnothing$, indicating that there are no common classes between the base and novel classes.

Our proposed ST-FSOD method is established on the classic TFA [17]. In the first stage of TFA, a base detector is trained using the base set $\mathcal{D}_{\text{base}}$, following the same procedure as in a regular object detector. In the second stage, a few-shot object detector is initialized using the weights obtained in the first stage and fine-tuned on a K-shot fine-tuning set denoted by $\mathcal{D}_{\text{ft}} = (\mathbf{I}_i, \mathcal{B}_i^{\text{base}}, \mathcal{B}_i^{\text{nov}})$, where the number of novel object bounding boxes for each novel class is K. More formally, for each novel class $c_{\text{nov}} \in \mathcal{C}_{\text{nov}}$, we have

$$\forall c_{\text{nov}} \in \mathcal{C}_{\text{nov}}, \quad \sum_{\mathbf{I}_i, \mathcal{B}_i^{\text{base}}, \mathcal{B}_i^{\text{nov}} \in \mathcal{D}_{\text{ft}}} |\{(x, y, h, w, c) \in \mathcal{B}_i^{\text{nov}} | \ c = c_{\text{nov}}\}| = K. \quad (1)$$

Here, $|\cdot|$ denotes the number of elements of the set.

B. Balanced Sampling Strategy

One of the key questions in constructing the few-shot fine-tuning set $\mathcal{D}_{\text{ft}}$ is how many base class objects to sample. In the original TFA, [17] proposed sampling exactly K shots of base objects for each base class to maintain a better balance between base and novel classes. However, recent studies have shown that this strategy may not be the optimal for remote sensing images [35], [36]. For example, Wolf et al. [36] found that using more base objects and oversampling the few-shot novel objects can improve the overall performance.

Inspired by these findings, we propose a BSS as follows. First, we randomly sample the K-shot novel objects as usual. Next, we include all base class images and annotations in $\mathcal{D}_{\text{base}}$ into the fine-tuning set $\mathcal{D}_{\text{ft}}$. Finally, when sampling images for fine tuning, we increase the probability of sampling images $\mathbf{I}_i$ with a corresponding nonempty $\mathcal{B}_i^{\text{nov}}$ to ensure they are sampled with the same probability as images that do not contain any novel annotations. By adopting the BSS, we achieve a more balanced fine-tuning set between base and novel classes, while also making full use of all the available base annotations.

C. Self-Training-Based RPNs

In the faster R-CNN network architecture, the RPN is utilized to generate a set of proposals $\mathcal{P} = \{p = (t_x, t_y, t_w, t_h, o)\}$ for each input image I using multiscale features extracted from I. The parameters t_x, t_y, t_w, and t_h represent the coordinates of each proposal, and o is an objectness score that indicates the probability of the proposal containing an object. Fig. 2 illustrates the architecture of ST-RPN, which consists of three submodules: base RPN, teacher RPN, and student RPN. All of these submodules follow the original RPN architecture. ST-RPN generates two sets of proposals, $\mathcal{P}^{\text{base}}$ and $\mathcal{P}^{\text{nov}}$, which are obtained as follows.

1) The base RPN is responsible for extracting base proposals $\mathcal{P}^{\text{base}}$ from the input image $\mathbf{I}$. For fine-tuning the base RPN module, only base annotations $\mathcal{B}^{\text{base}}$ are used to calculate the RPN loss
$$\mathcal{L}_{\text{rpn}}^{\text{base}}(\mathcal{P}^{\text{base}}, \mathcal{B}^{\text{base}}) = \sum_{p_i \in \mathcal{P}^{\text{base}}} L(p_i, p_i^*) \quad (2)$$
where p_i^* is the regression and classification targets achieved by matching $\mathcal{P}^{\text{base}}$ with $\mathcal{B}^{\text{base}}$ according to the intersection over union (IoU) between each pair of the proposal and the ground-truth bounding box [2].
2) The teacher RPN generates a set of ignored proposals $\mathcal{P}^{\text{ign}}$, which includes output proposals from this module that have an objectness score o greater than a given threshold τ_{rpn}. Please refer to Section IV-B for the selection of τ_{rpn}.
3) The student RPN receives the extracted features from the backbone and the FPN head [49] with dropout [50] applied and outputs a set of novel proposals $\mathcal{P}^{\text{nov}}$. This module is trainable and is supervised only by the few-shot novel annotations $\mathcal{B}^{\text{nov}}$. During the calculation of the regression and classification target of each output proposal, those with a high overlap with the ignored proposals in $\mathcal{P}^{\text{ign}}$ are excluded from the loss calculation. The loss function for the student RPN can be formulated as follows:
$$\mathcal{L}_{\text{rpn}}^{\text{nov}}(\mathcal{P}^{\text{nov}}, \mathcal{P}^{\text{ign}}, \mathcal{B}^{\text{nov}}) = \sum_{p_i \in \mathcal{P}^{\text{nov}}} w_i L(p_i, p_i^*). \quad (3)$$

Here, p_i^* is the box targets achieved by assigning $\mathcal{P}^{\text{nov}}$ to $\mathcal{B}^{\text{nov}}$. w_i is a weighting coefficient that is used to ignore the highly confident proposals that might contain novel objects

$$w_i = \begin{cases} 0, \text{if } \exists\ p_j \in \mathcal{P}^{ign}, \text{IoU}(p_j, p_i) > 0.7, \\ 1, \text{otherwise.} \end{cases} \tag{4}$$

Here, $\text{IoU}(\cdot, \cdot)$ denotes the IoU value of the two proposals.

The separation of base and novel proposals has two benefits. First, during fine tuning, the pretrained weights obtained from the base training stage could be negatively affected, which can impact the quality of the extracted base proposals. Separating proposal extraction prevents the fine tuning toward novel objects from biasing the extraction of base proposals. Second, previous FSOD works in remote sensing [35] have found that the RPN module achieved from the base training stage often fails to recall novel objects successfully. Thus, training an additional RPN module from scratch can lead to a better network state in terms of novel OD.

During the extraction of novel proposals $\mathcal{P}^{\text{nov}}$, using a student–teacher-based self-training mechanism and introducing ignored proposals into the loss calculation can prevent potential unannotated novel objects from being misclassified as background. This helps the RPN module to recall more novel objects and improves detection performance.

It is worth noting that there could be cases where a positive novel proposal, which is associated with a ground-truth bounding box, ends up being excluded according to (4) during the few-shot fine-tuning stage. However, in practice, we have observed that this does not significantly impact the generation of novel proposals. This is because the ignored proposals in $\mathcal{P}^{ign}$ are already of very high confidence, indicating that the learning process toward that specific ground-truth bounding box has already reached saturation.

D. Self-Training-Based Bounding Box Head

In the proposed framework, the extracted base and novel proposals are merged and passed through an RoI pooling layer and an RoI feature extraction layer to obtain the corresponding RoI features. These features are then forwarded to the ST-BBH for achieving the final detection results. The ST-BBH consists of a teacher and a student BBHs. Each BBH contains a bounding box classifier and a regressor, following the same architecture as the one proposed in faster R-CNN [26]. While processing each RoI, the feature is input to both the teacher and student classifier heads. In order to improve the robustness, the student's head receives the feature with dropout [50] applied as the input. Let $\mathbf{u}^{stu}$ and $\mathbf{u}^{tch}$ be the output probability of the student and teacher BBH and v be the corresponding ground-truth label. The student classifier's classification loss is calculated using the following equation:

$$\mathcal{L}_{bbh}(\mathbf{u}^{\text{stu}}, \mathbf{u}^{\text{tch}}, v) = \begin{cases} L_{\text{cls}}(\mathbf{u}^{\text{stu}}, \hat{u}^{\text{tch}}), & \text{if } v = 0 \text{ and } \max(\mathbf{u}^{\text{tch}}) > \tau_{\text{bbox}} \\ L_{\text{cls}}(\mathbf{u}^{\text{stu}}, v), & \text{otherwise.} \end{cases} \tag{5}$$

Here, $\hat{u}^{\text{tch}}$ represents the class index with the highest value in $\mathbf{u}^{\text{tch}}$. The assigned ground-truth class label of the RoI is denoted by v, with $v = 0$ indicating that the RoI is considered as background. The threshold τ_{bbox} is used to determine when to use the prediction from the teacher BBH as the pseudolabel. Please refer to Section IV-B for the selection of τ_{bbox}.

Overall, ST-RPN and ST-BBH share the same self-training philosophy with a momentum network, represented in our context as the teacher module. This teacher module maintains a slowly updated copy of the original module, ensuring stable yet recent targets (or pseudolabels) for model updates, as discussed in [27]. However, it is important to note that ST-RPN and ST-BBH are technically distinct: ST-RPN is employed to extract class-agnostic proposals, while ST-BBH is specifically designed for the classification and regression of class-specific bounding boxes. Furthermore, we emphasize that in ST-RPN, we explicitly separate the extraction of base and novel proposals, whereas, in ST-BBH, the classification and regression heads for both base and novel classes share the same ROI features.

E. Weights Initialization and Update

As depicted in Fig. 2, different trainable and fixed modules have been highlighted in different colors. Among these modules, the backbone, FPN head, base RPN, and RoI feature extraction layer will be initialized by the pretrained weights obtained from the base training stage. For the classifier and regressor of both student and teacher BBH, the entries of their weight matrices corresponding to the base classes will be initialized based on the pretrained weights, while entries for the novel classes will be randomly initialized. The student and teacher RPN module of the ST-RPN will be randomly initialized.

It is worth noting that the weights for the teacher RPN and teacher BBH are not trainable. Instead, they will be updated by the corresponding student module's weights using exponential moving average [27]

$$\theta_{\mathcal{T}}^{(t)} = \alpha\theta_{\mathcal{T}}^{(t-1)} + (1-\alpha)\theta_{\mathcal{S}}^{(t)}. \tag{6}$$

Here, $\theta_{\mathcal{T}}^{(t)}$ and $\theta_{\mathcal{S}}^{(t)}$ denote the weights of the teacher networks and their corresponding student networks at time step t during the training stage, respectively. α is a decay weight that is set to 0.999 following [27].

IV. Experiments

In this section, we will present the experimental results of our proposed method on various benchmarks for FSOD in remote sensing.

A. Experimental Settings

We evaluate the proposed method on three large-scale public OD datasets in remote sensing, including NWPU-VHR10 v2 [51], DIOR [52], and instance segmentation in serial images dataset (iSAID) [21]. NWPU-VHR10 v2 dataset comprises a total of 1172 images, each with dimension 400×400 and are divided into ten categories. Following the previous works [53],

three categories "airplane," "baseball diamond," and "tennis court" are adopted as novel classes, while the others are as the base classes. In line with previous researches, we employ the training and validation sets to fine tune the model and report the performance on the test set, which contains 293 images.

The DIOR dataset contains 23 463 images and over 190 000 instances, with an image size of 800×800. All objects are categorized into 20 classes. In the previous literature, two commonly used settings are adopted. The first setting, proposed by Li et al. [20], uses five categories (i.e., "plane," "baseball field," "tennis court," "train station," and "wind mill") as the novel categories and the remaining as the base categories. In this setting, the training set is used for base training and few-shot fine tuning, while the validation set is used for evaluation. The second setting, proposed in [54], includes a total of four base-novel class splits, each containing five novel categories and 15 base categories. In this setting, both the training and validation sets are used for base training and fine-tuning, and the test set is used for evaluation.

iSAID is a large-scale instance segmentation dataset for remote sensing. It is built on the same image set as DOTA [55], but provides the instance-level mask annotations, and also finer bounding box annotations. iSAID contains 2806 images, whose sizes range from 800×800 to $20\,000 \times 20\,000$. In total, there are 655 451 annotated objects, which are classified into 15 categories. We follow the official data preprocessing pipeline to crop the images into 800×800 patches, with an overlap of 25%. We follow the FSOD setting of [36], which uses three different base-novel class splits, and sets the number of shots for each split to 10, 50, and 100.

To make a fair comparison with the previous works, we adopt the mean average precision (mAP) with an IoU threshold at 0.5 as the evaluation metric following the common practices.

B. Implementation Details

The proposed method uses the faster R-CNN architecture [2] with an ResNet101 [56] backbone that is pretrained on the ImageNet dataset. An FPNs [49] is used to generate multiscale features. The AdamW optimizer [57] with a weight decay of 0.01 and a learning rate of $1e-4$ is used to train the model on all settings. The base training stage of NWPU-VHR10 v2, DIOR and iSAID datasets lasts for 10 000, 40 000, and 80 000 iterations, respectively. Learning rate decay with a factor of 0.1 is applied at 5000 and 8000 for the NWPU-VHR10 v2 dataset, 24 000 and 32 000 iterations for the DIOR dataset, and 40 000 and 60 000 for iSAID dataset. The few-shot fine-tuning lasts for 2000 iterations for the NWPU-VHR10 v2 dataset and 10 000 iterations for the other settings.

For data preprocessing and augmentation, image patches are randomly cropped to sizes of 400×400 for NWPU-VHR10 v2 dataset and 608×608 for the others. Multiscale training with a range from 0.5 to 2.0, random flipping, and random rotation with degrees of 90, 180, and 270 are applied. The batch size is set to 16 for base training and 8 for fine tuning.

The momentum parameter is set to $\alpha = 0.999$ when updating the networks' weights by exponential moving average [27]. The thresholds τ_{rpn} and τ_{bbh} used in ST-ROI and ST-BBH are set to 0.8. In Section IV-F, sensitivity analyses to these hyperparameters are provided. Our codes are based on PyTorch, EarthNets [58], and MMDetection [59] platform. For more information, please refer to our published codes.

TABLE I

AVERAGE PRECISION (AP) (IN %) AT AN IOU THRESHOLD OF 0.5 OF DIFFERENT METHODS ON NWPU-VHR V2 DATASET, WHERE THE BASE-NOVEL CLASS SPLIT FOLLOWS [53]. AVERAGED RESULTS AND STANDARD DEVIATIONS OF THREE DIFFERENT RUNS ARE REPORTED FOR THE PROPOSED METHODS

Method / Shots	Novel Classes			
	3	5	10	20
OFA [53]	43.2	60.4	66.7	-
FSODM [20]	32	53	65	-
SAM&BFS [35]	47.0	61.6	74.9	-
PAMS-Det [60]	37	55	66	-
CIR-FSD [61]	54	64	70	-
TFACSC [62]	47	**67**	72	-
SAGS&TFS [63]	51	66	72	-
TSF-RGR [39]	57	66	**77**	-
G-FSOD [37]	50.1	58.8	67.0	75.9
Ours	**60.7** $\pm$ **2.1**	**67.2** $\pm$ **1.2**	**77.2** $\pm$ **2.5**	**83.3** $\pm$ **1.5**

TABLE II

AP (IN %) AT AN IOU THRESHOLD OF 0.5 OF DIFFERENT METHODS ON DIOR DATASET, WHERE THE BASE-NOVEL CLASS SPLIT FOLLOWS [20]. AVERAGED RESULTS AND THE STANDARD DEVIATIONS OF THREE DIFFERENT RUNS ARE REPORTED FOR THE PROPOSED METHODS

Method / Shots	Novel Classes			
	3	5	10	20
OFA [53]	32.8	37.9	40.7	-
FSODM [20]	-	25	32	36
SAM&BFS [35]	-	38.3	47.3	50.9
PAMS-Det [60]	28	33	38	-
CIR-FSD [61]	-	33	38	43
TFACSC [62]	38	42	47	-
SAGS&TFS [63]	-	34	37	42
TSF-RGR [39]	-	42	49	54
Ours	**43.5** $\pm$ **5.5**	**48.3** $\pm$ **1.3**	**55.8** $\pm$ **1.5**	**61.3** $\pm$ **2.0**

C. Quantitative Results

The quantitative results for the NWPU-VHR10 v2, DIOR, and iSAID datasets are presented in Tables I–IV, respectively. Overall, our proposed method achieves superior or comparable performance across all datasets and various settings related to novel objects.

On the NWPU-VHR10 v2 dataset, our method attains state-of-the-art performance in both 3- and 20-shot settings, surpassing the second-best results by a notable margin ranging from 3% to 7%. In addition, it performs comparably to state-of-the-art models in the five- and ten-shot settings.

Regarding the DIOR dataset, as shown in Table II, our method consistently outperforms the second-best approach by a substantial margin of 5%–7%. In Table III, our method

TABLE III

AP (IN %) AT AN IOU THRESHOLD OF 0.5 OF DIFFERENT METHODS ON DIOR DATASET, WHERE THE FOUR BASE-NOVEL CLASS SPLITS FOLLOW [54]. AVERAGED RESULTS AND THE STANDARD DEVIATIONS OF THREE DIFFERENT RUNS ARE REPORTED FOR THE PROPOSED METHODS. THE RESULTS ON THE BASE CLASSES ARE ALSO REPORTED

Shots	Methods	Novel Classes				Base Classes			
		split1	split2	split3	split4	split1	split2	split3	split4
3	P-CNN [54]	18.0	14.5	16.5	15.2	47.0	48.9	49.5	49.8
	SAGS&TFS [63]	29.3	12.6	**20.9**	17.5	-	-	-	-
	G-FSOD [37]	27.6	14.1	16.0	16.7	68.9	69.2	71.1	69.0
	Ours	**41.9 ± 0.6**	**17.7 ± 2.0**	**20.9 ± 0.4**	**20.4 ± 3.6**	**73.5 ± 0.5**	**72.5 ± 0.5**	**75.2 ± 0.4**	**73.3 ± 0.6**
5	P-CNN [54]	22.8	14.9	18.8	17.5	48.4	49.1	49.9	49.9
	SAGS&TFS [63]	31.6	15.5	24.8	19.7	-	-	-	-
	G-FSOD [37]	30.5	15.8	23.3	21.0	69.5	69.3	70.2	68.0
	Ours	**45.7 ± 1.6**	**20.7 ± 2.8**	**26.0 ± 2.5**	**25.2 ± 4.5**	**73.3 ± 0.4**	**72.7 ± 0.4**	**75.6 ± 0.6**	**73.5 ± 0.4**
10	P-CNN [54]	27.6	18.9	23.3	18.9	50.9	52.5	52.1	51.7
	SAGS&TFS [63]	31.6	15.5	24.8	19.7	-	-	-	-
	G-FSOD [37]	37.5	20.7	26.2	25.8	69.0	68.7	71.1	68.6
	Ours	**50.0 ± 1.5**	**27.3 ± 1.1**	**31.3 ± 0.3**	**33.4 ± 1.1**	**72.6 ± 0.3**	**72.3 ± 0.5**	**75.7 ± 0.4**	**73.9 ± 0.2**
20	P-CNN [54]	29.6	22.8	28.8	25.7	52.2	51.6	53.1	52.3
	SAGS&TFS [63]	40.2	23.8	**36.1**	27.7	-	-	-	-
	G-FSOD [37]	39.8	22.7	32.1	31.8	69.8	68.2	71.3	67.7
	Ours	**53.7 ± 1.1**	**33.4 ± 0.4**	34.6 ± 1.9	**38.2 ± 2.0**	**73.3 ± 0.5**	**73.3 ± 0.5**	**75.5 ± 0.2**	**73.8 ± 0.2**

TABLE IV

AP (IN %) AT AN IOU THRESHOLD OF 0.5 OF DIFFERENT METHODS ON ISAID DATASET, WHERE THE THREE BASE-NOVEL CLASS SPLITS FOLLOWS [36]. AVERAGED RESULTS AND THE STANDARD DEVIATIONS OF THREE DIFFERENT RUNS ARE REPORTED FOR THE PROPOSED METHODS. RESULTS OF FSDETVIEW AND TFA ARE CITED FROM [36]. THE RESULTS ON THE BASE CLASSES ARE ALSO REPORTED

Shots	Methods	Novel Classes			Base Classes		
		split1	split2	split3	split1	split2	split3
10	FSDetView [34]	1.3 ± 0.3	8.7 ± 2.1	4.6 ± 1.2	33.8 ± 0.5	29.8 ± 1.6	32.9 ± 3.4
	TFA [17]	3.3 ± 0.8	9.0 ± 2.6	3.8 ± 1.1	58.6 ± 0.3	56.5 ± 0.8	59.0 ± 1.5
	DH-FSDet [36]	5.2 ± 0.8	14.5 ± 1.7	9.7 ± 2.2	**65.0 ± 0.2**	**64.5 ± 0.1**	**67.8 ± 0.1**
	Ours	**10.2 ± 3.3**	**17.7 ± 3.8**	**14.0 ± 2.1**	63.7 ± 0.4	62.4 ± 0.4	66.1 ± 0.7
50	FSDetView [34]	7.2 ± 2.3	26.8 ± 2.8	17.1 ± 1.1	35.3 ± 0.5	30.0 ± 1.1	34.6 ± 1.1
	TFA [17]	4.7 ± 0.0	12.1 ± 1.9	5.6 ± 1.4	60.7 ± 0.5	58.5 ± 0.8	60.9 ± 0.3
	DH-FSDet [36]	12.8 ± 0.8	28.9 ± 3.4	19.6 ± 2.4	**65.1 ± 0.1**	**64.7 ± 0.1**	**68.0 ± 0.1**
	Ours	**24.8 ± 2.1**	**39.3 ± 2.1**	**31.1 ± 1.2**	62.8 ± 0.4	62.6 ± 0.3	65.8 ± 0.1
100	FSDetView [34]	10.2 ± 1.2	32.8 ± 2.0	24.1 ± 1.1	36.4 ± 0.6	30.4 ± 0.4	34.5 ± 1.3
	TFA [17]	5.0 ± 0.3	14.4 ± 1.5	5.4 ± 1.1	61.4 ± 0.3	59.2 ± 0.2	61.6 ± 0.4
	DH-FSDet [36]	16.7 ± 1.7	36.0 ± 1.7	23.1 ± 0.9	**65.2 ± 0.1**	**64.8 ± 0.1**	**68.1 ± 0.1**
	Ours	**34.3 ± 1.9**	**45.0 ± 1.0**	**33.0 ± 1.3**	63.3 ± 0.3	62.9 ± 0.2	65.6 ± 0.1

excels in all settings except for the 20-shot setting of split 3. Remarkably, our results in split 1 demonstrate improvements exceeding 10%.

On the iSAID dataset, our proposed method delivers significant enhancements, ranging from 5% to over 10%. These improvements remain consistent across various splits and different numbers of shots.

These results not only demonstrate the effectiveness of our proposed method but also underscore the significance of addressing the issue of IANOs. It is worth noting that performance variance tends to be higher in cases with fewer shots, suggesting that FSOD performance is somewhat sensitive to the sampling of annotated novel objects, especially when the available shots are limited.

D. Qualitative Results

We present the visualization results on NWPU-VHR10 v2, DIOR, and iSAID datasets in Figs. 3–5. Some observations and conclusions can be made from the results as follows.

1) The proposed method plays a crucial role in alleviating the challenge posed by IANOs. This issue becomes particularly pronounced in scenarios featuring multiple small objects within a single image. For instance, in the first column of Fig. 4, ships can be easily mistaken as part of the background by a few-shot object detector. This occurs because, during the few-shot fine-tuning stage, a significant number of ship annotations will be missing, as depicted in Fig. 1. In contrast to this, our detection results show demonstrate that our method

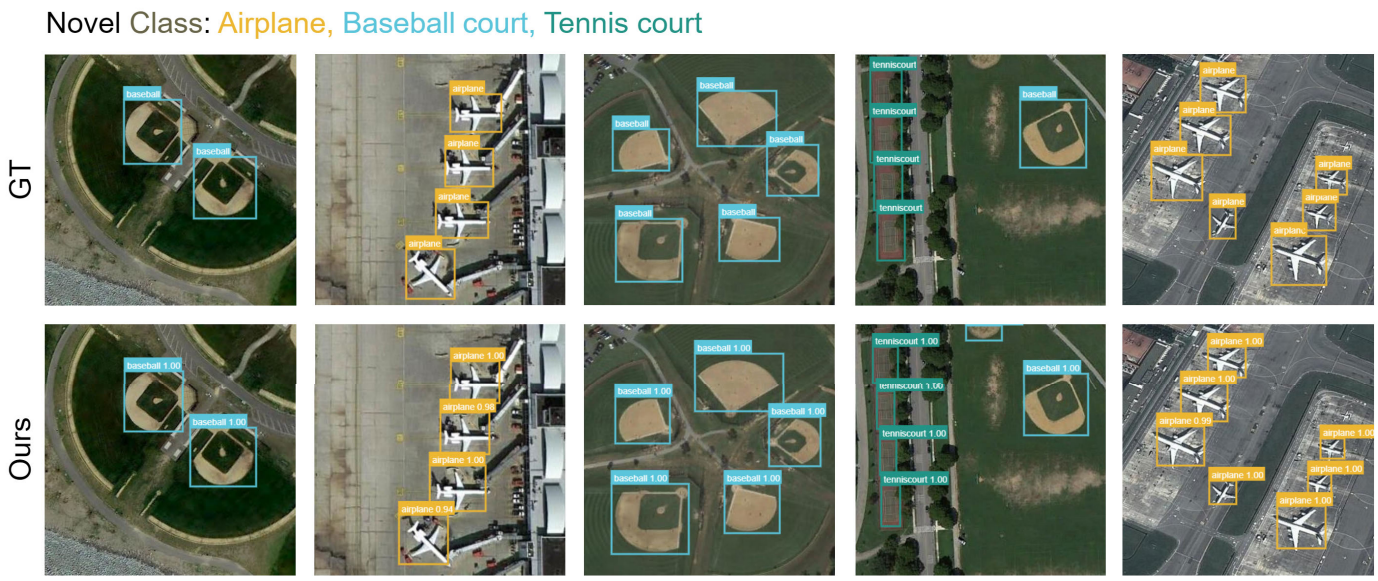


Fig. 3. Visualized FSOD results of the proposed methods under $K = 20$ shots setting on NWPU-VHR10 v2 dataset. The base-novel class split follows [53].

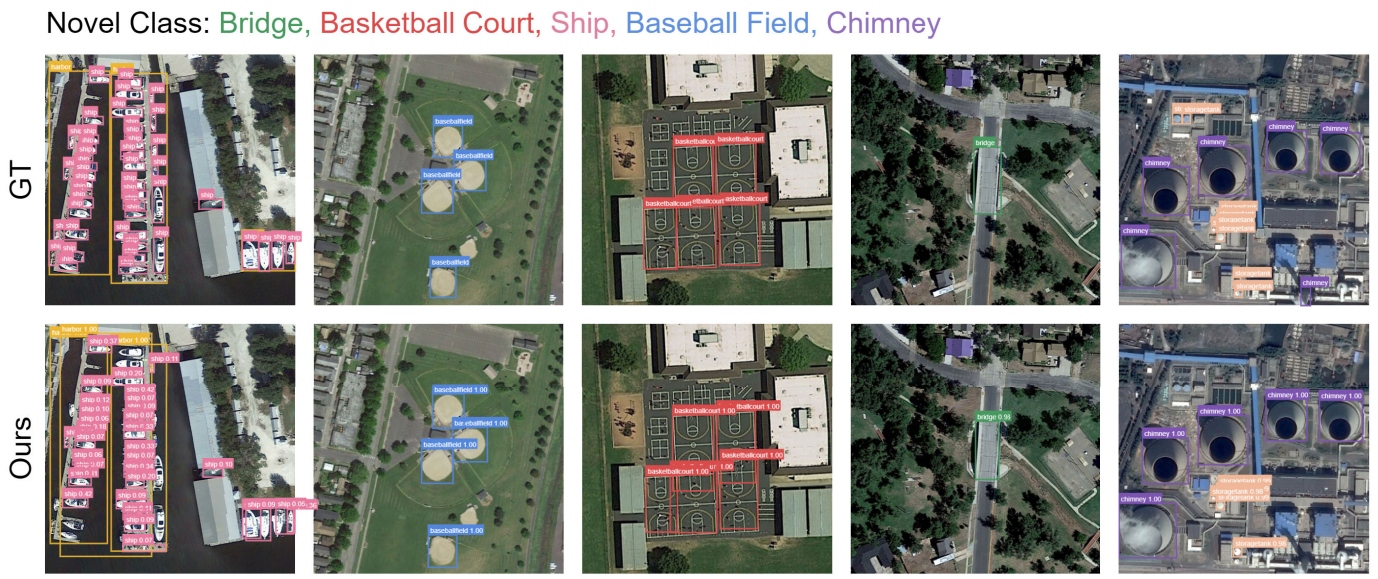


Fig. 4. Visualized FSOD results of the proposed methods under $K = 20$ shots setting on DIOR dataset. The base-novel class split follows the split 2 in [54].

excels in recalling many ship objects. This verifies the effectiveness of employing self-training techniques to solve the IANO issue.

2) By separating the proposals for base- and novel-class objects, our approach effectively preserves the performance of the few-shot detector on the base class. As elaborated in Section III-C, this separation of region proposals serves to prevent the few-shot fine-tuning process from negatively influencing the detection of base class objects. As evidenced by the results depicted in the fifth column of Fig. 4 and the third column of Fig. 5, our method is able to detect challenging, small base class objects such as storage tanks and vehicles.

E. Ablation Studies

We conduct ablation studies on the first split of the iSAID dataset to better understand the effect of all the components used in the proposed method. The results are presented in Table V and Fig. 6 for quantitative and qualitative analyses, respectively. The following observations can be made from the results.

1) The naive fine-tuning strategy introduced in [17] has a detrimental effect on the detection of base class objects. A noticeable performance decline becomes evident when comparing the results in the first row to those in the second row, according to Table V. This decline in performance stems from the fact that this particular strategy employs only a limited subset of base class annotations for model fine tuning, leading to incomplete annotations for the base class.
2) BSS proves to be highly effective in preserving the base class performance after the few-shot fine-tuning process, at the cost of a slight trade-off in novel class performance. This highlights the critical importance of

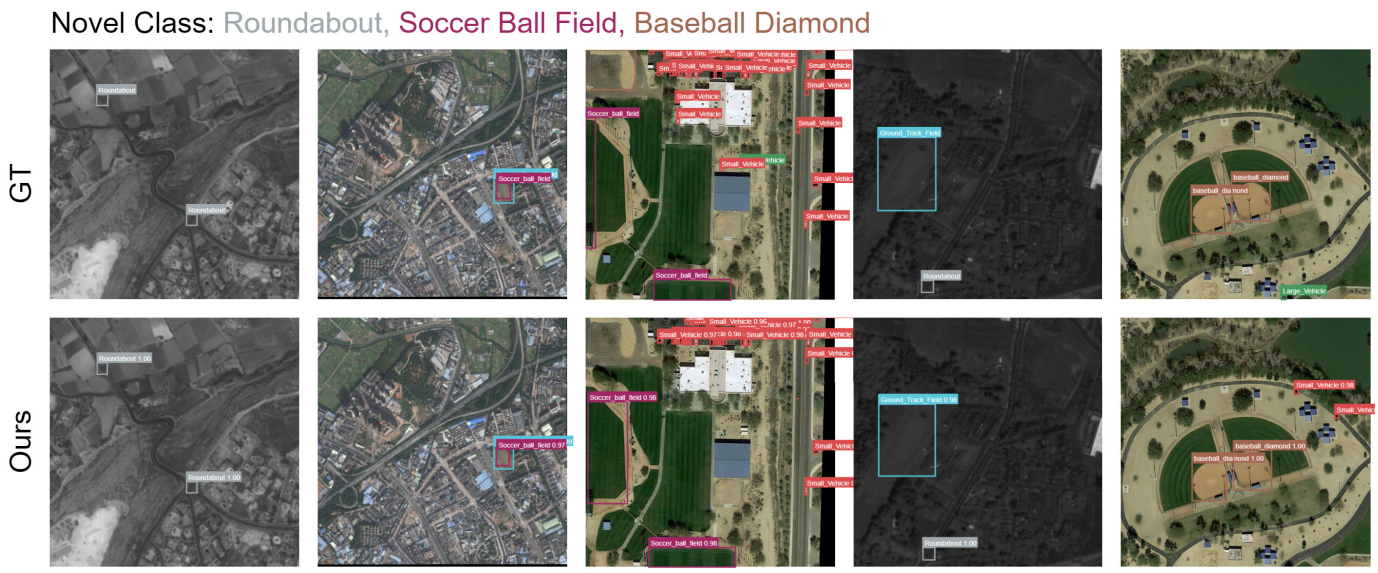


Fig. 5. Visualized FSOD results of the proposed methods under $K = 100$ shots setting on iSAID dataset. The base-novel class split follows the split 2 in [36].

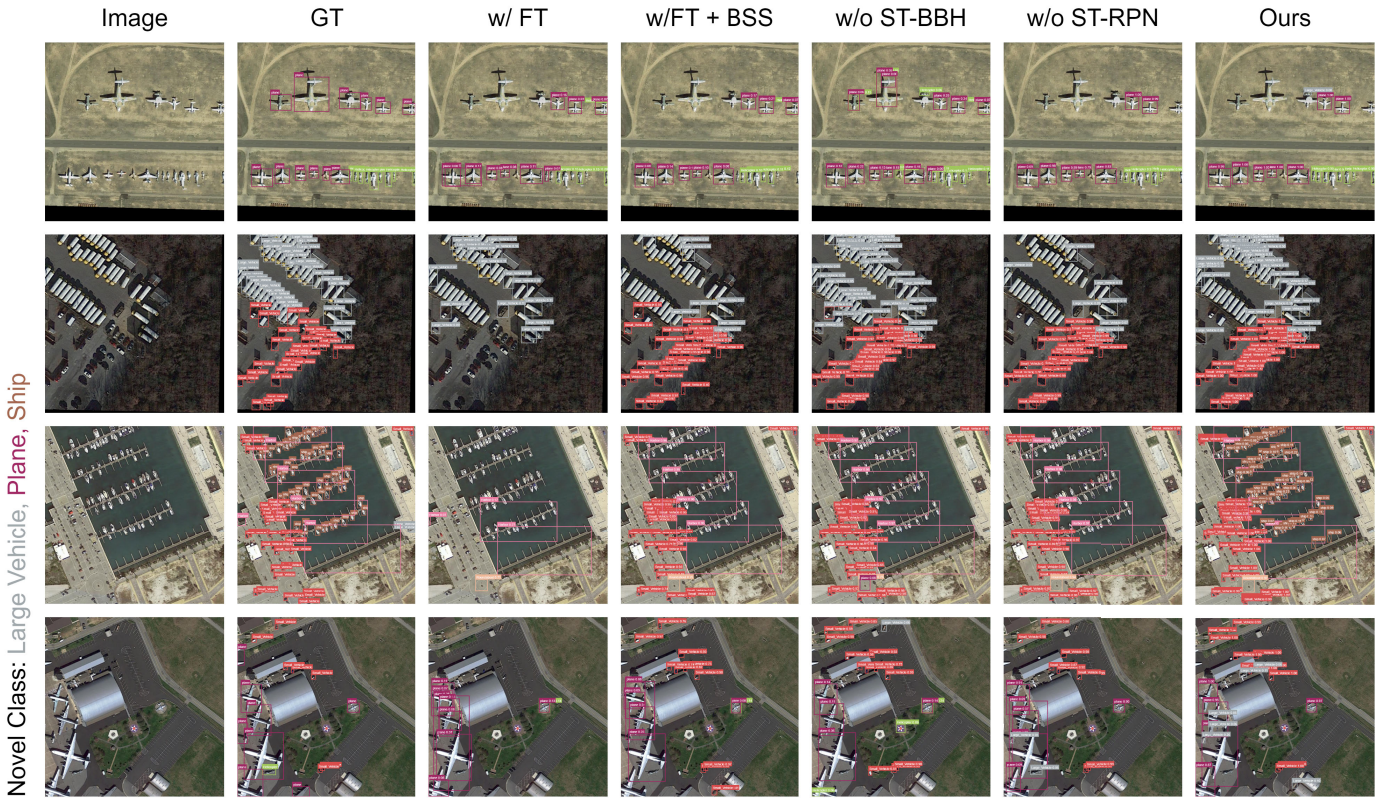


Fig. 6. Visualized FSOD results of the proposed methods under different ablations on the iSAID dataset. The number of novel shots is set to $K = 100$. The base-novel class split is the split 1 in [36].

fine tuning with a complete base annotation set for the maintenance of base class performance in satellite image-based FSOD.

3) Applying the proposed ST-RPN module does not influence the base class performance, due to the separation of the extraction of base and novel proposals. In addition, applying the proposed ST-RPN module is beneficial in improving the performance on novel classes.

4) After applying the proposed ST-BBH module, there is a slight performance decay on the base classes. One possible reason is that since the base and novel classes share the same BBH, the fine tuning on novel objects affects the general feature extraction within BBH and further affects the detection of base classes. However, there is a consistent improvement in the novel classes after applying ST-BBH, ranging from 3% to even 10%, at different numbers of shots.

TABLE V

ABLATION STUDY RESULTS ON SPLIT 1 OF THE iSAID BENCHMARK. THE SAME SEED FOR IMAGE SAMPLING AND OTHER RANDOMIZED PROCESS IS USED TO MAKE A FAIR COMPARISON."w/ FT" DENOTES WHETHER THE FEW-SHOT FINE-TUNING APPROACH PROPOSED IN [17] IS APPLIED. "BSS" DENOTES THE BALANCED SAMPLING STRATEGY AS DESCRIBED IN SECTION III-B

w/ FT	BSS	ST-RPN	ST-BBH	Novel Classes			Base Classes		
				10	50	100	10	50	100
				-	-	-	65.8	65.8	65.8
✓				3.7	13.1	17.0	39.6	54.2	57.8
✓	✓			4.0	12.3	16.2	**65.9**	65.9	66.0
✓	✓	✓		5.2	16.3	22.3	65.5	**66.0**	**66.4**
✓	✓		✓	6.5	17.8	27.0	65.1	64.9	65.0
✓	✓	✓	✓	**10.3**	**23.4**	**35.5**	63.0	62.3	63.0

TABLE VI

SENSITIVITY ANALYSES OF THE HYPERPARAMETERS ON SPLIT 1 OF THE iSAID BENCHMARK

τ_{rpn}	τ_{bbh}	α	Novel Classes			Base Classes		
			10	50	100	10	50	100
0.8	0.8	0.999	10.3	23.4	35.5	63.0	62.3	63.0
0.8	0.6	0.999	10.1	**25.1**	34.4	63.8	**63.2**	62.9
0.8	0.9	0.999	10.0	22.8	**36.5**	**64.0**	62.4	62.9
0.6	0.8	0.999	**10.5**	22.4	35.0	63.4	62.4	62.5
0.9	0.8	0.999	10.5	23.7	36.4	63.9	61.9	**63.3**
0.8	0.8	0.99	10.4	23.9	35.3	63.5	62.7	**63.3**

5) By combining all the modules, the highest novel class accuracy is achieved, with a margin of 6% to nearly 20% compared with the Naïve fine-tuning strategy. This verifies the effectiveness of the proposed ST-RPN and ST-BBH modules in solving the IANO issue.

Fig. 6 shows the visualized results for different ablated models. The second row shows that using the ST-RPN module helps to recall more large vehicle objects. However, using ST-RPN alone without ST-BBH may lead to a higher false positive rate, as can be seen from the first row. By combining the ST-RPN and ST-BBH, the best visualization quality is achieved. The third row demonstrates that this combination can help to detect small ship objects with high accuracy. These results further demonstrate the significance of incorporating the self-training mechanism to solve the unlabeled novel object issue.

F. Sensitivity Analyses of Hyperparameters

We conducted a sensitivity analysis to evaluate the impact of hyperparameter selection on the proposed method. The hyperparameters we tested include the two self-training thresholds τ_{rpn} and τ_{bbh} used in ST-RPN and ST-BBH, and the momentum α used when updating the teacher networks via EMA [27]. The results are presented in Table VI. We observed slight performance fluctuations (generally less than 2%) with different hyperparameter values. However, compared to the variances caused by different sampling seeds as shown in Tables II–IV, such fluctuations are not significant. Therefore, we can conclude that the proposed method is not highly sensitive to the values of the aforementioned hyperparameters.

TABLE VII

TRAINING TIMES (SECONDS PER ITERATION) AND INFERENCE FPS OF THE PROPOSED METHOD EVALUATED ON iSAID DATASET. MAP (%) IS ALSO REPORTED FOR COMPARISON

ST-RPN	ST-BBH	mAP	Training Times	Inference FPS
		16.2	0.258	18.8
✓		22.3	0.458	16.0
	✓	27.0	0.260	18.7
✓	✓	35.5	0.479	15.9

G. Computational Efficiency

To assess the incremental computational demands introduced by our two proposed modules, ST-RPN and ST-BBH, we conducted an evaluation based on training times (measured in seconds per iteration) during the fine-tuning stage and inference frames/s (FPS) when these modules were integrated. The outcomes, as illustrated in Table VII, indicate that the inclusion of ST-BBH results in a negligible computational overhead, both during the training and inference phases. This outcome aligns with our expectations, as ST-BBH simply introduces a pair of additional bounding box regressor and classifier layers on top of the RoI features.

In contrast, the integration of ST-RPN introduces a more substantial computational load during the training phase. However, it is crucial to note that the fine-tuning phase in FSOD algorithms typically involves significantly fewer iterations compared with the base training phase (see our settings in Section IV-B). As such, the additional training costs remain within acceptable bounds. Furthermore, while ST-RPN does marginally reduce inference speed, it is essential to consider this cost in light of the performance gains it offers.

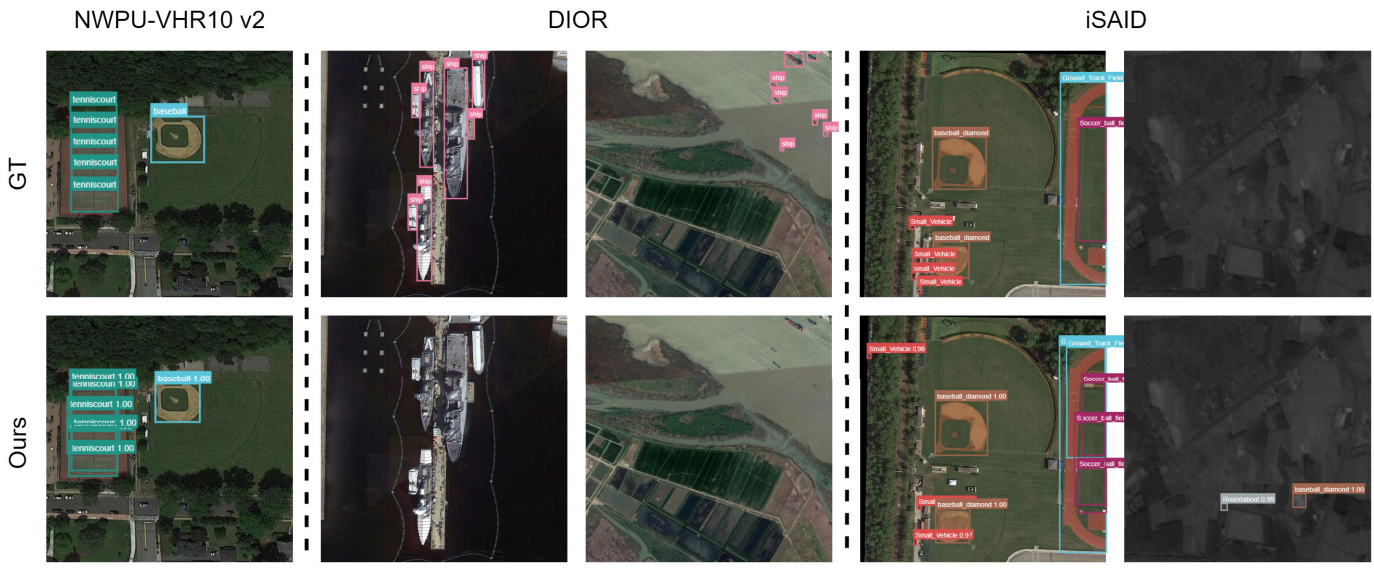


Fig. 7. Visualized failure cases of the proposed method on NWPU-VHR10 v2, DIOR, and iSAID datasets. Base-novel class splits are the same as the splits in Figs. 3–5.

This expense is manageable given the corresponding boost in detection performance. In addition, upon closer investigation, we found that a substantial portion of the computational cost introduced by ST-RPN is attributed to the additional bounding box operations, such as nonmaximum suppression (NMS). These options can be optimized and accelerated through the utilization of compute unified device architecture (CUDA) implementations.

H. Failure Cases

To gain a better understanding of the limitations of the proposed FSOD method, we visualize some failure cases in different FSOD settings, as shown in Fig. 7. Based on the figure, we can observe that the majority of the missed detections are primarily due to the small size of the objects ("ships" in the third column) or objects with large size variance compared with the training data ("ships" in the second column). In addition, there may be duplicated detected boxes for objects that lack clear boundaries, such as the "tennis court" in the first column or "soccerball field" in the fourth column. While these issues are prevalent in general OD [64], improved techniques related to addressing them can also be utilized to enhance the performance of FSOD.

V. Conclusion

In this article, we analyze the current FSOD setting for remote sensing and identified the issue of IANOs that can negatively impact the performance of FSOD methods. To address this issue, we propose to incorporate the self-training mechanism into the classical two-stage fine-tuning-based FSOD pipeline. Our approach includes an ST-RPN module, which generates a set of novel proposals by excluding some proposals from the loss calculation that are likely to be novel objects but cannot be assigned to an existing few-shot annotation. In addition, we designed an ST-BBH module that leverages the pseudolabels generated from a teacher BBH to filter out potential novel bounding boxes that are unlabeled and use them to supervise the student BBH to recall more novel objects.

While our proposed method significantly improved the novel class FSOD performance in remote sensing, the base class performance may slightly decrease compared with the base model. Future works could focus on designing a generalized FSOD method that prevents the model from forgetting the previously learned base knowledge while improving the performance in detecting novel classes.

References

[1] G. Huang, I. Laradji, D. Vázquez, S. Lacoste-Julien, and P. Rodríguez, "A survey of self-supervised and few-shot object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4071–4089, Apr. 2023.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 91–99.

[3] 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.

[4] N. Carion et al., "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K., Jul. 2020, pp. 213–229.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[6] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, Zurich, Switzerland. Cham, Switzerland: Springer, Sep. 2014, pp. 740–755.

[7] J. Zhang, J. Lei, W. Xie, Z. Fang, Y. Li, and Q. Du, "SuperYOLO: Super resolution assisted object detection in multimodal remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5605415.

[8] L. Dai, H. Liu, H. Tang, Z. Wu, and P. Song, "AO2-DETR: Arbitrary-oriented object detection transformer," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 5, pp. 2342–2356, May 2023.

[9] X. Zhang, Y. Feng, S. Zhang, N. Wang, S. Mei, and M. He, "Semi-supervised person detection in aerial images with instance segmentation and maximum mean discrepancy distance," *Remote Sens.*, vol. 15, no. 11, p. 2928, Jun. 2023.

[10] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Comput. Surv.*, vol. 53, no. 3, pp. 1–34, May 2021.

[11] C. Lang, G. Cheng, B. Tu, and J. Han, "Learning what not to segment: A new perspective on few-shot segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8047–8057.

[12] Z. Xiong, H. Li, and X. X. Zhu, "Doubly deformable aggregation of covariance matrices for few-shot segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 133–150.

[13] O. Vinyals, C. Blundell, T. Lillicrap, and D. Wierstra, "Matching networks for one shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 3630–3638.

[14] G. Cheng, C. Lang, and J. Han, "Holistic prototype activation for few-shot segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4650–4666, Apr. 2023.

[15] C. Lang, G. Cheng, B. Tu, C. Li, and J. Han, "Base and meta: A new perspective on few-shot segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10669–10686, Aug. 2023.

[16] H. Chen, Y. Wang, G. Wang, and Y. Qiao, "LSTD: A low-shot transfer detector for object detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018, pp. 2836–2843.

[17] X. Wang, T. E. Huang, T. Darrell, J. E. Gonzalez, and F. Yu, "Frustratingly simple few-shot object detection," 2020, *arXiv:2003.06957*.

[18] L. Wang, S. Zhang, Z. Han, Y. Feng, J. Wei, and S. Mei, "Diversity measurement-based meta-learning for few-shot object detection of remote sensing images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2022, pp. 3087–3090.

[19] B. Qiao, H. Zhou, L. Yang, and X. Xie, "Few shot object detection with incompletely annotated samples," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jun. 2023, pp. 1–7.

[20] X. Li, J. Deng, and Y. Fang, "Few-shot object detection on remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3145483.

[21] S. W. Zamir et al., "iSAID: A large-scale dataset for instance segmentation in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2019, pp. 28–37.

[22] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, "A brief review of domain adaptation," in *Advances in Data Science and Information Engineering*. Cham, Switzerland: Springer, 2021, pp. 877–894.

[23] M. Wurm, T. Stark, X. X. Zhu, M. Weigand, and H. Taubenböck, "Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 150, pp. 59–69, Apr. 2019.

[24] Q. Xu, Y. Shi, X. Yuan, and X. X. Zhu, "Universal domain adaptation for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 3235988.

[25] C. Wei, K. Sohn, C. Mellina, A. Yuille, and F. Yang, "CReST: A class-rebalancing self-training framework for imbalanced semi-supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10852–10861.

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

[27] N. Araslanov and S. Roth, "Self-supervised augmentation consistency for adapting semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15379–15389.

[28] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Dec. 2001, p. 1.

[29] N. D. Thanh, W. Li, and P. Ogunbona, "An improved template matching method for object detection," in *Proc. Asian Conf. Comput. Vis.*, Xi'an, China. Cham, Switzerland: Springer, 2010, pp. 193–202.

[30] Z. Yang, C. Zhang, R. Li, Y. Xu, and G. Lin, "Efficient few-shot object detection via knowledge inheritance," *IEEE Trans. Image Process.*, vol. 32, pp. 321–334, 2023.

[31] L. Karlinsky et al., "RepMet: Representative-based metric learning for classification and few-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5192–5201.

[32] Q. Fan, W. Zhuo, C.-K. Tang, and Y.-W. Tai, "Few-shot object detection with attention-RPN and multi-relation detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4012–4021.

[33] B. Kang, Z. Liu, X. Wang, F. Yu, J. Feng, and T. Darrell, "Few-shot object detection via feature reweighting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8419–8428.

[34] Y. Xiao, V. Lepetit, and R. Marlet, "Few-shot object detection and viewpoint estimation for objects in the wild," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3090–3106, Mar. 2023.

[35] X. Huang, B. He, M. Tong, D. Wang, and C. He, "Few-shot object detection on remote sensing images via shared attention module and balanced fine-tuning strategy," *Remote Sens.*, vol. 13, no. 19, p. 3816, Sep. 2021.

[36] S. Wolf, J. Meier, L. Sommer, and J. Beyerer, "Double head predictor based few-shot object detection for aerial imagery," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 721–731.

[37] T. Zhang, X. Zhang, P. Zhu, X. Jia, X. Tang, and L. Jiao, "Generalized few-shot object detection in remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 195, pp. 353–364, Jan. 2023.

[38] X. Lu et al., "Few-shot object detection in aerial imagery guided by text-modal knowledge," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 3250448.

[39] S. Zhang et al., "Text semantic fusion relation graph reasoning for few-shot object detection on remote sensing images," *Remote Sens.*, vol. 15, no. 5, p. 1187, Feb. 2023.

[40] Y. Yang and S. Soatto, "FDA: Fourier domain adaptation for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4084–4094.

[41] M. N. Subhani and M. Ali, "Learning from scale-invariant examples for domain adaptation in semantic segmentation," in *Proc. ECCV*, Glasgow, U.K., Aug. 2020, pp. 290–306.

[42] K. Mei, C. Zhu, J. Zou, and S. Zhang, "Instance adaptive self-training for unsupervised domain adaptation," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K., Cham, Switzerland: Springer, Aug. 2020, pp. 415–430.

[43] I. Shin, S. Woo, F. Pan, and I. S. Kweon, "Two-phase pseudo label densification for self-training based domain adaptation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 532–548.

[44] F. Zhang, Y. Shi, Z. Xiong, W. Huang, and X. X. Zhu, "Pseudo features-guided self-training for domain adaptive semantic segmentation of satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5612414.

[45] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "MixMatch: A holistic approach to semi-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 5050–5060.

[46] L. Hoyer, D. Dai, and L. Van Gool, "DAFormer: Improving network architectures and training strategies for domain-adaptive semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9914–9925.

[47] K. Sohn et al., "FixMatch: Simplifying semi-supervised learning with consistency and confidence," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 596–608.

[48] Y. Li et al., "Few-shot object detection via classification refinement and distractor retreatment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15390–15398.

[49] T.-Y. Lin, P. Dollár, 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. 936–944.

[50] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014.

[51] 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.

[52] 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.

[53] Z. Zhang, J. Hao, C. Pan, and G. Ji, "Oriented feature augmentation for few-shot object detection in remote sensing images," in *Proc. IEEE Int. Conf. Comput. Sci., Electron. Inf. Eng. Intell. Control Technol. (CEI)*, Sep. 2021, pp. 359–366.

[54] G. Cheng et al., "Prototype-CNN for few-shot object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3105575.

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

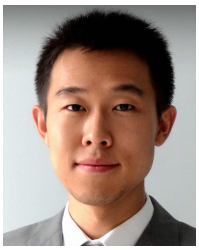
[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[57] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

[58] Z. Xiong, F. Zhang, Y. Wang, Y. Shi, and X. X. Zhu, "EarthNets: Empowering AI in Earth observation," 2022, *arXiv:2210.04936*.
[59] K. Chen et al., "MMDetection: Open MMLab detection toolbox and benchmark," 2019, *arXiv:1906.07155*.
[60] Z. Zhao, P. Tang, L. Zhao, and Z. Zhang, "Few-shot object detection of remote sensing images via two-stage fine-tuning," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[61] Y. Wang, C. Xu, C. Liu, and Z. Li, "Context information refinement for few-shot object detection in remote sensing images," *Remote Sens.*, vol. 14, no. 14, p. 3255, Jul. 2022.
[62] R. Li, Y. Zeng, J. Wu, Y. Wang, and X. Zhang, "Few-shot object detection of remote sensing image via calibration," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[63] Y. Zhang, B. Zhang, and B. Wang, "Few-shot object detection with self-adaptive global similarity and two-way foreground stimulator in remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 7263–7276, 2022.
[64] J. Rabbi, N. Ray, M. Schubert, S. Chowdhury, and D. Chao, "Small-object detection in remote sensing images with end-to-end edge-enhanced GAN and object detector network," *Remote Sens.*, vol. 12, no. 9, p. 1432, May 2020.

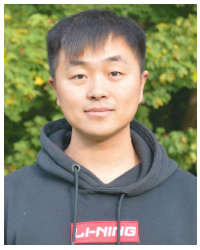
Fahong Zhang received the B.E. degree in software engineering from Northwestern Polytechnical University, Xi'an, China, in 2017, and the M.S. degree in computer science from the Center for OPTical IMagery Analysis and Learning, Northwestern Polytechnical University in 2020. He is currently pursuing the Ph.D. degree with the Department of Aerospace and Geodesy, Chair of Data Science in Earth Observation, Technical University of Munich, Munich, Germany.

His research interests include computer vision and satellite image processing.



Yilei Shi (Member, IEEE) received the Dipl.-Ing. degree in mechanical engineering and the Dr.-Ing. degree in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2010 and 2019, respectively.

He is currently a Senior Scientist with the Chair of Remote Sensing Technology, TUM. His research interests include fast solver and parallel computing for large-scale problems, high-performance computing, and computational intelligence, advanced methods on synthetic aperture radar (SAR) and interferometric synthetic aperture radar (InSAR) processing, machine learning, and deep learning for a variety of data sources, such as SAR, optical images, and medical images, and partial differential equation (PDE)-related numerical modeling and computing.



Zhitong Xiong (Member, IEEE) received the Ph.D. degree in computer science and technology from Northwestern Polytechnical University, Xi'an, China, in 2021.

He is currently a Research Scientist and leads the ML4Earth working group with the Chair of Data Science in Earth Observation, Technical University of Munich (TUM), Munich, Germany. His research interests include computer vision, machine learning, Earth observation, and Earth system modeling.

Xiao Xiang Zhu (Fellow, IEEE) received the M.Sc., Dr.-Ing., and Habilitation degrees in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011, and 2013, respectively.

She is the Chair Professor for data science in earth observation with the Technical University of Munich (TUM) and was the Founding Head of the Department "EO Data Science" at the Remote Sensing Technology Institute, German Aerospace Center (DLR). Since May 2020, she has been a PI and the Director of the international future AI lab "AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond", Munich, Germany. Since October 2020, she has been also serves as the Director of the Munich Data Science Institute (MDSI), TUM. She was a Guest Scientist or Visiting Professor at the Italian National Research Council (CNR-IREA), Naples, Italy, Fudan University, Shanghai, China, the University of Tokyo, Tokyo, Japan, and University of California, Los Angeles, USA, in 2009, 2014, 2015, and 2016, respectively. She is currently a Visiting AI Professor at ESA's Phi-lab, Frascati, Italy. From 2019 to 2022, she has a co-coordinator of the Munich Data Science Research School (www.mu-ds.de) and the Head of the Helmholtz Artificial Intelligence – Research Field "Aeronautics, Space and Transport". Her main research interests are remote sensing and earth observation, signal processing, machine learning and data science, with their applications in tackling societal grand challenges, e.g. global urbanization, united nations sustainable development goals (UN's SDGs), and climate change.

Dr. Zhu has been a member of Young Academy (Junge Akademie/Junges Kolleg) at the Berlin-Brandenburg Academy of Sciences and Humanities, Berlin, Germany; and the German National Academy of Sciences Leopoldina, Schweinfurt, Germany; and the Bavarian Academy of Sciences and Humanities, Munich. She serves at the Scientific Advisory Board in several research organizations, among others the German Research Center for Geosciences (GFZ, 2020-2023) and Potsdam Institute for Climate Impact Research (PIK). She is an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, Pattern Recognition, and serves as an Area Editor responsible for special issues of IEEE *Signal Processing Magazine*.