ISPRS Journal of Photogrammetry and Remote Sensing 195 (2023) 353–364

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Generalized few-shot object detection in remote sensing images ☆

Tianyang Zhang [a], Xiangrong Zhang [a,*], Peng Zhu [a], Xiuping Jia [b], Xu Tang [a], Licheng Jiao [a]

[a] Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, Xidian University, Xi'an, Shaanxi Province 710071, China
[b] School of Engineering and Information Technology, University of New South Wales, Canberra, ACT 2612, Australia

ARTICLE INFO

Keywords:
Generalized few-shot object detection
Remote sensing images
Transfer-learning
Metric learning

ABSTRACT

Recently few-shot object detection (FSOD) in remote sensing images (RSIs) has drawn increasing attention. However, the current FSOD methods in RSIs merely focus on the detection performance of few-shot novel classes while ignoring the severe degradation of the base class performance. Generalized few-shot object detection (G-FSOD) aims to solve the FSOD problem without forgetting previous knowledge. In this paper, we focus on the G-FSOD in RSIs and propose a Generalized Few-Shot Detector (G-FSDet) that can learn novel knowledge without forgetting. Through the comprehensive analysis of each component in the detector, a novel efficient transfer-learning framework is presented as the foundation of our G-FSDet, which is more suitable for FSOD in remote sensing scenes. Considering the greater intra-class diversity and lower inter-class separability of geospatial objects, we design a metric-based discriminative loss to learn a more discriminative classifier in the few-shot fine-tuning stage. Furthermore, a representation compensation module is proposed to alleviate the catastrophic forgetting problem by decoupling the representation learning of previous and novel knowledge. Extensive experiments on DIOR and NWPU VHR-10.v2 datasets demonstrate that our proposed G-FSDet achieves competitive novel class performance with minor degradation in the base class, reaching state-of-the-art overall performance among all few-shot settings. The source code is available at (https://github.com/RSer-XDU/G-FSDet).

1. Introduction

Benefiting from the rapid advances in remote sensing technology, remote sensing images (RSIs) have entered the era of big data, and the automatic analysis and understanding of these abundant RSIs have become an active field in the remote sensing community (Audebert et al., 2018; Ding et al., 2022; Sun et al., 2022; Xia et al., 2017). As a fundamental but essential task in RSIs interpretation, object detection can provide both instance-level category and location information, which serves wide applications (Ma et al., 2019), such as traffic monitoring, urban planning, and precision agriculture.

With the development of deep learning, a series of advanced deep learning based detectors (Lin et al., 2017a,b; Redmon and Farhadi, 2017; Ren et al., 2015) have been developed for common objects in natural scenes, making a qualitative leap in object detection performance. Drawing on the robust feature representation capabilities of deep learning methods, remote sensing community researchers have also presented various detectors to address the huge scale variations, arbitrary orientations, and high background complexity of geospatial objects and achieved impressive results (Cheng et al., 2016; Pang et al., 2019; Zhong et al., 2018). However, the current state-of-the-art geospatial object detectors typically need sufficient samples to achieve good performance and are prone to overfitting with limited samples. In practical applications, it is unrealistic to collect sufficient training data due to extensive human labor for annotation, especially for geospatial object detection (e.g., huge scale variation of objects, a large number of objects per image, and the experts' domain knowledge for annotation.), and the real-world data distribution is long-tailed where some categories may only have a few instances. Hence, exploring efficient object detection in the few-shot scenario is critical.

Inspired by the human ability to learn new concepts rapidly from very few instructions, few-shot learning has been proposed and successfully applied in the various few-shot tasks (Snell et al., 2017; Tian et al., 2022; Wang et al., 2020). Although few-shot object detection (FSOD) has achieved excellent results in natural scene images, it is unsuitable for directly applying off-the-shelf methods to RSIs. Unlike generic objects, geospatial objects have greater intra-class diversity and lower inter-class separability, which exacerbates the challenges of few-shot geospatial object detection and requires the detector to learn a

☆ This work was supported in part by the National Natural Science Foundation of China under Grants 62276197, 61871306, and 62171332; the Key Research and Development Program in the Shaanxi Province of China under Grant 2019ZDLGY03-08.

* Corresponding author.
E-mail address: xrzhang@mail.xidian.edu.cn (X. Zhang).

https://doi.org/10.1016/j.isprsjprs.2022.12.004
Received 20 July 2022; Received in revised form 4 December 2022; Accepted 4 December 2022
Available online 17 December 2022

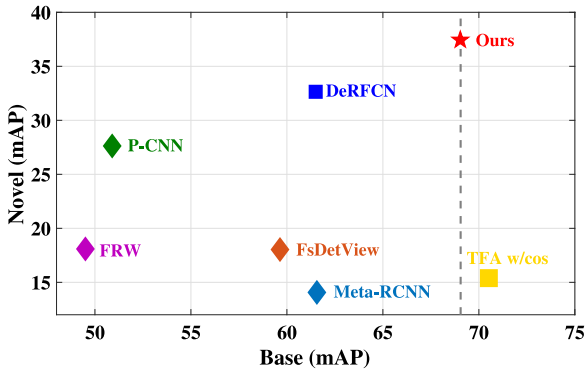

Fig. 1. Performance comparison of our G-FSDet with previous FSOD methods on the DIOR test set in the first novel/base split under the 10-shot setting. The gray dash line indicates the base mAP (69.05) of pre-trained Faster-RCNN.

more discriminative classifier under few-shot settings. In addition, most of the existing FSOD methods (Cheng et al., 2022; Huang et al., 2021; Li et al., 2022b; Xiao et al., 2021; Zhao et al., 2022) merely focus on the performance of novel classes while ignoring the degradation of the base classes, resulting in the catastrophic forgetting problem. In contrast, generalized few-shot object detection (G-FSOD) targets to tackle the FSOD without forgetting previously seen classes, which is more suitable for practical application (Li et al., 2021). For example, in search and rescue missions, the search systems should be able to detect novel objects without forgetting the old ones.

Considering the above problems, we focus on exploring G-FSOD in RSIs and propose a Generalized Few-Shot Detector (G-FSDet) to improve the detection performance of few-shot novel classes while avoiding the catastrophic forgetting of base classes. We first analyze the properties of each component in the pre-trained detector and propose a simple yet effective transfer-learning framework for FSOD in RSIs. In order to learn a more discriminative classifier with limited training samples, we design a discriminative loss based on metric learning, where the samples are encouraged to be more similar to their corresponding class prototypes than to other class prototypes. To solve the catastrophic forgetting of base classes in FSOD, we design a representation compensation module, which decouples the representation learning of previous and novel knowledge and further regularizes the novel knowledge learning in the fine-tuning stage.

The main contributions of this paper can be summarized as follows:

1. A highly efficient transfer-learning framework is proposed by the comprehensive analysis of different components in the pre-trained detector, more suitable for few-shot geospatial object detection.

2. A metric-based discriminative loss is designed to enforce the intra-class compactness and inter-class separability during the fine-tuning stage, leading to a more discriminative classifier in the few-shot scenario.

3. To the best of our knowledge, this paper is the first work in the remote sensing community to focus on the G-FSOD task and devises a representation compensation module to tackle the catastrophic forgetting problem of base classes.

4. Our proposed G-FSDet achieves the state-of-the-art overall performance of the two remote sensing FSOD benchmarks, where it obtains competitive performance in the few-shot novel classes with minor degradation in base class performance (as shown in Fig. 1).

The remainder of this paper is organized as follows. Section 2 gives a brief introduction to the related work. Section 3 provides preliminary knowledge about few-shot object detection. Section 4 describes our proposed method in detail. Section 5 reports and analyzes the comparative results on two remote sensing few-shot object detection datasets. Finally, the conclusion is detailed in Section 6.

2. Related works

2.1. Object detection

Benefiting from the powerful feature representation capabilities of deep learning techniques, deep learning based object detectors have made impressive improvements in the past decade. In the deep learning era, object detection is mainly divided into two streams: two-stage detectors (Dai et al., 2016; Lin et al., 2017a; Ren et al., 2015) and one-stage detectors (Lin et al., 2017b; Liu et al., 2016; Redmon and Farhadi, 2017). The two-stage detectors follow a “coarse-to-fine” detection process. Specifically, the detectors first generate several region proposals to capture the potential foreground regions and then refine these region proposals to obtain the final detected bounding boxes. This “coarse-to-fine” paradigm guarantees superior accuracy. In contrast, one-stage detectors discard the region proposals and directly make dense predictions on the deep features of the image. Thanks to the “complete in one step” framework, the one-stage detectors are faster in speed.

Along with the rapid development of detectors in the nature scene, object detection in the remote sensing community has also achieved tremendous progress (Li et al., 2020a). Unlike natural images, geospatial objects have their characteristics, making object detection in RSIs more challenging. Due to the different ground sampling distances (GSDs) of remote sensing satellites, there are huge scale variations for geospatial objects, which is a challenging but active research topic in geospatial object detection (Deng et al., 2018; Li et al., 2018; Zhang et al., 2020, 2021; Zheng et al., 2020). For example, Deng et al. (2018) proposed a multi-scale object proposal network (MS-OPN) consisting of three proposal branches to predict multi-scale proposals. Zheng et al. (2020) designed the hyper-scale blocks to learn the robust scale-invariant feature representation and achieve better performance than previous works. The unique bird views of RSIs lead to arbitrary orientations of geospatial objects. Cheng et al. proposed the RICNN (Cheng et al., 2016), which learns rotation-invariant features with a new rotation-invariant layer. Han et al. (2021) designed the ReDet to encode rotation equivariance and rotation invariance explicitly. Considering the misalignment problem between the horizontal bounding box and the rotated object, many researchers (Ding et al., 2019; Fu et al., 2020; Xu et al., 2021; Yu et al., 2020) have introduced oriented bounding boxes to eliminate this problem. For instance, Ding et al. (2019) designed a Rotated RoI learner to model the geometry transformation and relieve the misalignment efficiently. Yu et al. (2020) introduced an orientation-guided anchoring scheme to generate high-quality, rotated anchors. In addition, object detection in RSIs also suffers from the problem of high background complexity, and several studies (Feng et al., 2021; Yang et al., 2019; Zhang et al., 2022b) focus on strengthening the feature representations of the foreground regions.

Detectors, as mentioned above, require a sufficient number of bounding box annotations to achieve satisfactory performance and are prone to overfitting when the annotated data becomes scarce. However, collecting volumes of well-labeled data is expensive, especially for specific tasks or categories.

2.2. Few shot object detection

Since the annotation task is time-consuming and labor-intensive, there is an ongoing research effort to design few-shot object detectors that can achieve accurate detection. These recent FSOD works can be grouped into meta-learning methods (Fan et al., 2020; Kang et al., 2019; Karlinsky et al., 2019; Xiao and Marlet, 2020; Yan et al., 2019) and transfer-learning methods (Chen et al., 2018; Qiao et al., 2021; Sun et al., 2021; Wang et al., 2020; Yang et al., 2022).

The core of the meta-learning method is “learning to learn”. It acquires task-level knowledge by simulating and solving various few-shot

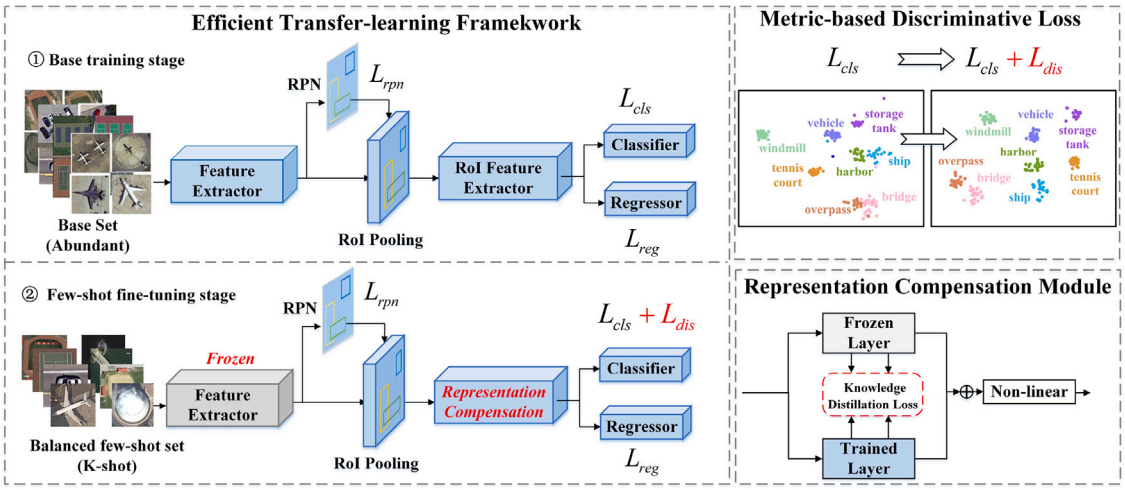

Fig. 2. The overall structure of our proposed G-FSDet. Based on the proposed efficient transfer-learning framework, our G-FSDet adopts a two-stage training process. The entire detector is first trained on the abundant base set and then fine-tuned on the balanced few-shot set with a frozen feature extractor. Besides, in the second fine-tuning stage, a metric-based discriminative loss is introduced to learn a more discriminative classifier in the few-shot scenario, and a representation compensation module is applied to alleviate the catastrophic forgetting of base classes. Details can be found in Section 5.

learning tasks and generalizes this knowledge to tackle few-shot learning of novel categories. Yan et al. (2019) introduce a Predictor-head Remodeling Network (PRN) to perform meta-learning over Region-of-Interest (RoI) features. The PRN infers the class attentive vectors of low-shot categories and remodels the R-CNN head with a channel-wise soft-attention. Fan et al. (2020) introduced metric learning under the meta-learning framework, which exploits the relationship between object pairs with the Attention RPN and Multi-Relation Detector.

Transfer-learning methods aim at fine-tuning the common knowledge learned from the abundant annotated data to the novel categories with limited data. Wang et al. (2020) proposed a concise transfer-learning method, which only fine-tunes the box classifier with instance-level feature normalization and achieves remarkable performance. Considering the inconsistency among the optimization goals of RPN and RCNN, Qiao et al. (2021) designed the Gradient Decoupled Layer to scale or stop the gradient of these two modules, leading to a more suitable fine-tuning strategy for FSOD.

Recently, a few researchers (Cheng et al., 2022; Huang et al., 2021; Li et al., 2022b; Xiao et al., 2021; Zhao et al., 2022) have begun to focus on the FSOD task in RSIs. Li et al. (2022b) introduced the meta-learner (Kang et al., 2019) into the multi-scale object detection architecture for the remote sensing FSOD task. Cheng et al. (2022) proposed a Prototype-CNN, which mainly consists of a prototype learning network (PLN), a prototype-guided RPN (P-G RPN), and a novel detection head. Concretely, the PLN learns the class-specific prototypes, which are fed into the later modules to generate better foreground proposals and class-aware RoI features. Huang et al. (2021) designed a novel balanced fine-tuning strategy to mitigate the severe data imbalance between the novel class and base class and introduced a shared attention module to better utilize the rich ground information in RSIs. Additionally, Li et al. (2022a) attempted to study the more challenging one-shot object detection in RSIs.

Unlike the above methods, our method proposes a powerful yet efficient transfer-learning framework for FSOD in RSIs and devises a metric-based discriminative loss to tackle the classification confusion caused by the large intra-class variances of geospatial objects. Furthermore, to the best of our knowledge, we are the first to explore the G-FSOD task in the remote sensing community.

3. Preliminary knowledge

3.1. Problem setting

As in previous literature (Cheng et al., 2022; Wang et al., 2020; Yang et al., 2022), we follow the standard problem settings of FSOD in our paper. Concretely, we split the training set into a base class sub-datasets $\mathcal{D}_{\text{base}} = \{(x_i, y_i), y_i \in C_{base}\}_{i=1}^{I}$ and a novel class sub-datasets $\mathcal{D}_{\text{novel}} = \{(x_j, y_j), y_j \in C_{novel}\}_{j=1}^{J}$, where x_* is the input image and y_* is the corresponding annotations. Note that the base classes C_{base} and the novel classes C_{novel} are non-overlapping, namely $C_{\text{base}} \cap C_{\text{novel}} = \emptyset$. $\mathcal{D}_{\text{base}}$ generally contains sufficient annotated data, while $\mathcal{D}_{\text{novel}}$ has only K-shot instances annotations for each class. For the inference phase in FSOD, the test set contains both base and novel classes ($C_{\text{test}} = C_{\text{base}} + C_{\text{novel}}$), and the detector is required to detect all classes of objects. Therefore, a balanced few-shot dataset $\mathcal{D}_{\text{few}} = \{(x_n, y_n), y_n \in C_{\text{base}} + C_{\text{novel}}\}_{n=1}^{K}$ containing K-shot instances annotations of each base and novel class is constructed for few-shot transfer.

3.2. Revisiting TFA

TFA (Wang et al., 2020) is a widely adopted baseline for transfer-learning few-shot object detectors and achieves advanced performance in natural scenes. TFA follows a simple two-stage training pipeline to leverage the knowledge of base classes. In the base training stage, the model is trained on the base class sub-datasets to establish prior knowledge. In the novel fine-tuning stage, randomly initialized weights are first assigned to the box predictor for the novel classes, which enables the pre-trained detectors to predict the novel class. Then, only the box predictor of this initialized detector is fine-tuned with the balanced few-shot set. Besides, the TFA replaces the FC-based classifier with the cosine similarity-based classifier, which can be given as:

$$p_{i,y_i} = \frac{e^{s_{i,y_i}}}{\sum_{j=0}^{C} e^{s_{i,j}}}, \quad s_{i,j} = \frac{\alpha \mathcal{F}(x)_i^{\top} w_j}{\left\|\mathcal{F}(x)_i\right\| \left\|w_j\right\|} \tag{1}$$

where p_{i,y_i} represents the prediction probability that the ith RoI feature belongs to class j. w_j stands for the weight vector of class j. $\alpha = 20$ is the fixed scaling factor. $C = C_{base} + C_{novel}$ denotes the total number of classes in the dataset. Thanks to the instance-level feature normalization in the cosine similarity-based classifier, the TFA reduces the massive intra-class variance and brings considerable gains compared with previous works.

Considering the simplicity of TFA, our proposed G-FSDet also adopts the two-stage fine-tuning paradigm, and we design an efficient transfer-learning framework for FSOD in remote sensing scenes.

4. Methodology

Fig. 2 shows the overall structure of our proposed G-FSDet. Based on the proposed efficient transfer-learning framework, the G-FSDet involves a simple two-stage training. Specifically, the Faster R-CNN (Ren et al., 2015) with Feature Pyramid Network (FPN) (Lin et al., 2017a) is first trained on the abundant base class sub-datasets and then fine-tuned on the balanced few-shot set with our well-designed fine-tuning strategy. Moreover, in the few-shot fine-tuning stage, we also introduced a metric-based discriminative loss and a representation compensation module to facilitate learning the novel knowledge and preserving the previous knowledge.

4.1. Efficient transfer-learning framework

Although the existing meta-learning methods have promoted the FSOD performance in the remote sensing scene, episodic learning in meta-learning suffers from the inefficient memory problem as the number of classes increases in the support set (Wang et al., 2020). Besides, the TFA, which achieves superior performance in the nature scene, undergoes severe degeneration in the remote sensing scene. With the aforementioned consideration, we propose an efficient transfer learning framework for FSOD in RSIs. Our proposed transfer-learning framework consists of a base model training stage and a few-shot fine-tuning stage, where the base model training stage aims to learn the general knowledge from sufficient annotated base class sub-datasets D_{base} while the few-shot fine-tuning stage transfers the prior knowledge to assist the learning of the novel class. We employ the Faster R-CNN with FPN, a widely used object detection method in RSIs, as our base detector.

Base model training stage. In this stage, we train all the parameters of the base detector on the base class sub-datasets D_{base} to better establish the general knowledge, following the standard joint loss function adopted in (Ren et al., 2015).

$$\begin{aligned}\theta^*, W^* = &\arg\min_{\theta,W} L_{rpn}\left(f\left(D_{\text{base}};\theta\right);w_{rpn}\right)+\\ &L_{rcnn}\left(f\left(D_{\text{base}};\theta\right);w_{rcnn}\right)\end{aligned} \tag{2}$$

where θ represents the parameters of the feature extractor (i.e., backbone and FPN) and $W = \{w_{rpn}, w_{rcnn}\}$ represents the parameters of Region Proposal Network (RPN) and RCNN.

Few-shot fine-tuning stage. In TFA, only the box predictor is fine-tuned with novel data, and other components are frozen. We find such a fine-tuning strategy framework severely restricts the knowledge transfer, leading to the poor few-shot novel class performance in the remote sensing scene. Since different components play specific roles in the detector, we comprehensively analyze the properties of each component in the fine-tuning stage and propose an efficient fine-tuning strategy for FSOD in RSIs.

Specifically, we follow the same initialization and fine-tuning strategy for the box predictor as TFA (Wang et al., 2020). Since the RoI feature extractor encodes the high-level semantic features (Sun et al., 2021), we unfreeze it to capture high-level features of the novel classes in the fine-tuning stage. The RPN is responsible for discovering the foreground regions in the massive anchors. During the base training stage, the novel class objects are discarded, and the RPN only treats the base class objects as foreground, making the RPN biased towards base classes, which fails to generalize well to novel classes. Therefore, we propose to fine-tune the RPN so that it can discover novel class objects. The experiments in Section 5 demonstrate the essential of fine-tuning the RPN in RSIs. As for the feature extractor, we freeze it during the fine-tuning stage to relieve the forgetting of base class knowledge while increasing the efficiency of the fine-tuning stage. In summary, we **only freeze the feature extractor and fine-tune the other components with the balanced few-shot set.**

$$\begin{aligned}W_{ft}^* = &\arg\min_{W_{ft}} L_{rpn}\left(f\left(D_{\text{few}};\widetilde{\theta}^*\right);w_{rpn}\right)+\\ &L_{rcnn}\left(f\left(D_{\text{few}};\widetilde{\theta}^*\right);w_{rcnn}\right)\end{aligned} \tag{3}$$

where $\widetilde{\theta}^*$ represents the frozen pre-trained parameters of the feature extractor, and $W_{ft} = \{w_{rpn}, w_{rcnn}\}$ defines the parameters of RPN and RCNN in the fine-tuning stage. D_{few} denotes the balanced few-shot set.

4.2. Metric-based discriminative loss

The geospatial objects in remote sensing scenes have the characteristics of greater intra-class diversity and low inter-class separability (Xia et al., 2017), which becomes more challenging with limited training samples. Therefore, we devise a discriminative loss based on the cosine similarity score metric to learn a more discriminative classifier in the few-shot scenario. The intuition behind our proposed loss function is that the similarity between a sample and its corresponding class prototype should be greater than the similarity with other class prototypes.

Following the typical conventions in cosine similarity-based classifiers (Chen et al., 2019, 2021; Wang et al., 2018), we define the learned weight vectors of classifiers as the class prototypes and adopt the prediction probability that sample x belongs to class c as the cosine similarity score between sample x and the prototypes of class c. Formally, the cosine similarity score between the sample x_i and the prototype of its corresponding class y_i can be written as follows:

$$score_{x_i,y_i} = \frac{\exp\left(\alpha \cdot \left\langle x_i, w_{y_i}\right\rangle\right)}{\sum_{j=0}^{C}\exp\left(\alpha \cdot \left\langle x_i, w_j\right\rangle\right)} \tag{4}$$

where $\langle *, * \rangle$ denotes the cosine similarity, and w_j represents the prototypes of class j.

Drawing on the above cosine similarity score, we encourage samples to have a much higher cosine similarity score with their corresponding class prototypes than the cosine similarity score with other class prototypes, which can be described as:

$$\mathcal{L}_{dis_i} = \sum_{j=0, j\neq y_i}^{C} -\log\left(max\left\{score_{x_i,y_i} - score_{x_i,j}, 0\right\} + \epsilon\right) \tag{5}$$

where $\epsilon = 1e^{-7}$ to keep numerical stability. This metric-based discriminative loss enforces the intra-class compactness and inter-class separability and builds up a more discriminative classifier in the few-shot scenario.

Besides, in the fine-tuning stage, the few-shot object detector has an inherent learning bias that the base class samples tend to have higher predicted scores for their ground truth class (Cao et al., 2021). These easy samples in the base class may lead to an inefficient optimization of novel class samples in our proposed loss. Inspired by the focal loss (Lin et al., 2017b), we design a focusing factor for each sample to adjust their loss contribution dynamically, which is defined as follows:

$$factor_i = (1 - p_i)^{\gamma} \tag{6}$$

where p_i represents the predicted probability of its corresponding ground truth class for the ith sample, and $\gamma = 4$ is the focusing parameter.

With the focusing factor, the few-shot object detector puts more focus on the novel hard samples, and the final formula of our proposed metric-based discriminative loss can be denoted as:

$$\mathcal{L}_{dis} = \frac{1}{N}\sum_{i=1}^{N}(\mathcal{L}_{dis_i} \cdot factor_i) \tag{7}$$

where N represents the number of RoI samples in the training batch size.

4.3. Representation compensation module

Recent few-shot object detectors have achieved acceptable performance in learning few-shot novel data (Cheng et al., 2022; Li et al., 2022b). However, the catastrophic forgetting of base classes has been

Table 1
Four different novel/base split settings on the DIOR dataset.

Split	Novel					Base
1	Baseball field	Basketball court	Bridge	Chimney	Ship	Rest
2	Airplane	Airport	Expressway toll station	Harbor	Ground track field	Rest
3	Dam	Golf course	Storage tank	Tennis court	Vehicle	Rest
4	Express service area	Overpass	Stadium	Train station	Windmill	Rest

Table 2
Two different novel/base split settings on the NWPU VHR.v2 dataset.

Split	Novel			Base
1	Airplane	Baseball diamond	Tennis court	Rest
2	Basketball	Ground track field	Vehicle	Rest

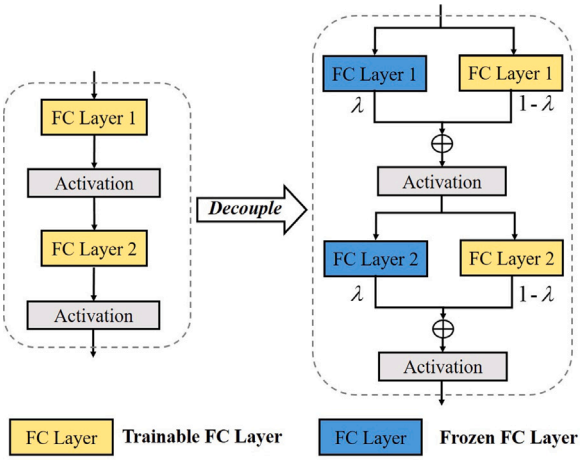

Fig. 3. Illustration of proposed representation decoupling mechanism. The representation decoupling mechanism modifies the single branch FC layer into parallel branches and aggregates the features from the two branches before the activation layer. In the fine-tuning stage, the frozen and trainable layers are responsible for protecting previous knowledge and learning novel knowledge, respectively.

neglected. Inspired by incremental learning methods (Zhang et al., 2022a; Zhu et al., 2021), we devise a representation compensation module to learn novel knowledge while remembering previous knowledge.

(1) Representation Decoupling Mechanism. As shown in Fig. 3, we introduce the representation decoupling mechanism into the RoI feature extractor, which modifies the single branch with the parallel branches to decouple its representation of previous knowledge and novel knowledge. Concretely, for the two consecutive fully connected (FC) layers in the RoI feature extractor, we add a parallel FC layer with pre-trained parameters for each component to represent the previous knowledge learned from the base training stage.

In the fine-tuning stage, the newly added branch is frozen to protect the previous knowledge, and the other branch is trainable to learn the novel knowledge. Besides, by fusing the output of two parallel FC layers, the RoI feature extractor will consider both the previous knowledge from the frozen layers and the novel knowledge from the trainable layers. Formally, the calculation of each FC layer equipped with the representation decoupling mechanism can be described as follows:

$$F_o = act(\lambda \cdot \left(\widetilde{W}_{fc} F_i + \widetilde{b}_{fc}\right) + (1-\lambda) \cdot \left(W_{fc} F_i + b_{fc}\right)) \tag{8}$$

where $\widetilde{W}_{fc}$, $\widetilde{b}_{fc}$ indicate the parameters in the frozen branch, and W_{fc}, b_{fc} mean the trainable parameters to learn the novel knowledge. $\lambda = 0.5$ is a weight vector to balance the impact of previous knowledge and novel knowledge. F_i and F_o are the input and output features in each FC layer. $act(*)$ stands for the non-linear activation function.

In the inference phase, we adopt the structural re-parameterization (Ding et al., 2021) to merge the parameters of the two parallel branches, which can be denoted as:

$$\begin{aligned} F_o &= act(\lambda \cdot \left(\widetilde{W}_{fc} F_i + \widetilde{b}_{fc}\right) + (1-\lambda) \cdot \left(W_{fc} F_i + b_{fc}\right)) \\ &= act(\left(\lambda \cdot \widetilde{W}_{fc} + (1-\lambda) \cdot W_{fc}\right) F_i + \lambda \cdot \widetilde{b}_{fc} + (1-\lambda) \cdot b_{fc}) \\ &= act(\hat{W}_{fc} F_i + \hat{b}_{fc}) \end{aligned} \tag{9}$$

where $\hat{W}_{fc}$, $\hat{b}_{fc}$ are the merged parameters. With the above re-parameterization technique, our proposed method does not increase any extra computational overhead and parameters during inference.

(2) Knowledge Distillation. To further alleviate the forgetting of previous knowledge, we employ knowledge distillation (Hinton et al., 2015) between intermediate layers in the proposed representation decoupling mechanism. Specifically, we regularize the trainable branches by matching the fused features with the output features of the frozen branches, and the knowledge distillation loss is denoted as:

$$\mathcal{L}_{kd} = \frac{1}{L}\frac{1}{N}\sum_{i=1}^{L}\sum_{j=1}^{N}\left\| F_o^{i,j} - \widetilde{F}_o^{i,j} \right\| \tag{10}$$

where $L = 2$ represents two FC layers. $F_o^{i,j}$ denotes the fused features, and $\widetilde{F}_o^{i,j}$ indicates the output features of the frozen branches.

Unlike the previous work (Chen et al., 2018; Fan et al., 2021) that performs knowledge distillation on the probability distribution over base classes, we employ the knowledge distillation between the intermediate features to alleviate the catastrophic forgetting of base classes.

4.4. Loss function of G-FSDet

With the aforementioned metric-based instance-level discriminative loss and knowledge distillation loss, the overall loss function of our proposed G-FSDet in the fine-tuning stage can be written as:

$$\mathcal{L}_{ft} = \mathcal{L}_{rpn} + \mathcal{L}_{cls} + \beta \cdot \mathcal{L}_{reg} + \mathcal{L}_{dis} + \eta \cdot \mathcal{L}_{kd} \tag{11}$$

where $\mathcal{L}_{rpn}$ indicates the standard RPN loss adopted in Ren et al. (2015). The term $\mathcal{L}_{cls}$ means the cross-entropy loss for RCNN classification, and the term $\mathcal{L}_{reg}$ denotes the smooth-L1 loss for RCNN regression. $\mathcal{L}_{dis}$ and $\mathcal{L}_{kd}$ indicate the proposed metric-based discriminative loss and knowledge distillation loss, respectively. Since we adopt the $\mathcal{L}_{dis}$ as an auxiliary loss in the RCNN classification task, we increase the regression loss weight β to 2 to balance the two tasks. As for the knowledge distillation loss weight η, we set it as 0.025.

5. Experiments

5.1. Datasets

DIOR (Li et al., 2020a). The DIOR dataset, a large-scale publicly available object detection dataset in RSIs, contains 23 463 images collected from Google Earth. The entire dataset is divided into three parts: the training set, validation set, and testing set, including 5862, 5863, and 11 738 images, respectively. The spatial resolutions of this dataset range from 0.5 to 30 m, and the size of all images is 800 × 800 pixels. Following the few-shot object detection setup in Cheng et al. (2022), we adopt the four different random novel/base splits, each of which includes 15 base classes and 5 novel classes, detailed in Table 1.

Table 3
Few-shot object detection performance on the DIOR test set under 3, 5, 10, 20-shot settings. The red denotes the best results, and the blue represents the second-best results.

Split	Method	3-shot			5-shot			10-shot			20-shot		
		Base	Novel	All	Base	Novel	All	Base	Novel	All	Base	Novel	All
1	FRW	49.4	7.5	38.9	49.7	12.1	40.3	49.5	18.1	41.7	50.0	22.0	43.0
	Meta-RCNN	60.62	12.02	48.47	62.01	13.09	49.78	61.55	14.07	49.68	63.21	14.45	51.02
	FsDetView	59.54	13.19	47.95	58.58	14.29	47.51	59.64	18.02	49.24	62.69	18.01	51.52
	P-CNN	47.0	18.0	39.8	48.4	22.8	42.0	50.9	27.6	45.1	52.2	29.6	46.8
	TFA w/cos	**70.32**	11.35	**55.58**	**70.51**	11.57	**55.78**	**70.52**	15.37	**56.73**	**71.07**	17.96	**57.79**
	DeRFCN	61.90	**28.25**	53.48	62.04	**30.30**	54.11	61.51	**32.64**	54.29	62.44	**35.37**	55.67
	Ours	**68.94**	**27.57**	**58.61**	**69.52**	**30.52**	**59.72**	**69.03**	**37.46**	**61.16**	**69.80**	**39.83**	**62.31**
2	FRW	48.5	4.8	37.6	46.8	7.0	36.9	46.4	9.0	37.1	43.5	14.1	36.2
	Meta-RCNN	62.55	8.84	49.12	63.14	10.88	50.07	63.28	14.90	51.18	63.86	16.71	52.07
	FsDetView	58.88	10.83	46.87	60.31	9.63	47.64	61.16	13.57	49.26	61.16	14.76	49.56
	P-CNN	48.9	**14.5**	40.3	49.1	14.9	40.6	52.5	**18.9**	44.1	51.6	**22.8**	44.4
	TFA w/cos	**70.75**	5.77	**54.51**	**70.79**	8.19	**55.14**	**69.93**	8.71	**54.63**	**70.02**	12.18	**55.56**
	DeRFCN	61.41	**14.55**	49.70	61.79	**16.45**	50.45	61.42	18.4	50.69	62.79	21.13	52.38
	Ours	**69.20**	14.13	**55.43**	**69.25**	**15.84**	**55.87**	**68.71**	**20.70**	**56.70**	**68.18**	**22.69**	**56.86**
3	FRW	45.5	7.8	36.1	47.9	13.7	39.3	44.5	13.8	36.8	43.5	18.5	37.3
	Meta-RCNN	61.93	9.10	48.72	63.44	12.29	50.66	62.57	11.96	49.92	65.53	16.14	53.18
	FsDetView	61.00	7.49	47.63	61.33	12.61	49.15	61.94	11.49	49.32	65.17	17.02	53.14
	P-CNN	49.5	**16.5**	41.3	49.9	**18.8**	42.1	52.1	**23.3**	44.9	53.1	**28.8**	47.0
	TFA w/cos	**71.95**	8.36	**56.05**	**71.64**	10.13	**56.26**	**72.56**	10.75	**57.11**	**73.13**	17.99	**59.35**
	DeRFCN	63.56	15.78	51.47	63.25	18.73	52.11	64.55	20.43	53.52	64.56	25.13	54.71
	Ours	**71.10**	**16.03**	**57.34**	**70.18**	**23.25**	**58.43**	**71.08**	**26.24**	**59.87**	**71.26**	**32.05**	**61.46**
4	FRW	48.2	3.7	37.1	48.5	6.8	38.1	45.7	7.2	36.1	44.4	12.2	36.4
	Meta-RCNN	61.73	13.94	49.78	62.60	15.84	50.91	62.23	15.07	50.44	63.24	18.17	51.98
	FsDetView	58.90	14.28	47.75	58.97	15.95	48.22	60.37	15.37	49.12	60.89	16.96	49.91
	P-CNN	49.8	**15.2**	41.2	49.9	17.5	41.8	51.7	18.9	43.5	52.3	25.7	45.7
	TFA w/cos	**68.57**	10.42	**54.03**	**68.85**	14.29	**55.21**	**68.58**	14.35	**55.03**	**68.86**	12.01	**54.65**
	DeRFCN	59.81	10.83	47.50	59.71	**18.62**	49.44	59.54	**21.61**	49.99	60.09	**27.61**	51.97
	Ours	**69.01**	**16.74**	**55.95**	**67.96**	**21.03**	**56.30**	**68.55**	**25.84**	**57.87**	**67.73**	**31.78**	**58.75**

Fig. 4. Visualization of the detection results on DIOR test set in the first novel/base split under 10-shot setting.

We randomly select $K = 3$, 5, 10, and 20 instances annotations for each class from the training and validation set for the K-shot detection task. For the inference phase, we evaluate our method on the testing set with both base and novel classes.

NWPU VHR-10.v2 (Cheng et al., 2014). This dataset contains 10 categories of geospatial objects with a total of 1172 images, and the size of images in this dataset is 400 × 400 pixels. The object categories include airplane, baseball diamond, basketball, bridge, ground track field, harbor, ship, storage tank, tennis court, and vehicle. As depicted in Table 2, we follow the random novel/base splits (i.e., 7 base classes and 3 novel classes) in Li et al. (2022a) to create the few-shot learning setup. We set K to be 3, 5, 10, and 20 for few-shot learning and evaluate the performance of all classes in the test set.

5.2. Evaluation metric

Following previous work (Cheng et al., 2022), we employ the mean Average Precision (mAP) with an IoU threshold of 0.5 to evaluate the performance of the few-shot object detector. The mAP denotes the mean of Average Precision (AP) in all categories. The higher the mAP value, the better the performance.

The AP metric is calculated by the average value of Precision (P) in the interval of Recall (R) from 0 to 1 under a certain IoU threshold in a single category, which can be formulated as follows:

$$AP = \sum_{R\in(0,0.1,\ldots,1)} \frac{1}{11} P(R) \tag{12}$$

Different from previous works (Cheng et al., 2022; Li et al., 2022b) that only evaluate the performance of novel classes, this work not only

focuses on the performance of the few-shot novel classes but also on the performance of base classes. Therefore, we choose the mAP of the base classes, the novel classes, and all classes as evaluation metrics to evaluate the generalized few-shot object detection performance of different methods.

5.3. Implementation details

Our proposed G-FSDet is based on Faster-RCNN (Ren et al., 2015) and adopts the pre-trained ResNet101 (He et al., 2016) with an FPN (Lin et al., 2017a) as the backbone. For the base training stage, the model is trained by 36k iterations in total, where the learning rate starts from 0.01 and decreases to 0.001 and 0.0001 at the 24k iterations and 32k iterations. During the fine-tuning stage, we randomly initialize the weight of the box predictor for the novel classes (Wang et al., 2020) and fine-tune the pre-trained model by 6k iterations with a learning rate of 0.001. In both the base training and fine-tuning stage, we train the model using the SGD optimizer with a weight decay of 0.0001 and a momentum of 0.9. The training batch size for all experiments is set to 4 on two NVIDIA GeForce GTX 1080Ti GPUs. In the inference phase, we set the confidence threshold and IoU threshold in non-maximum suppression (NMS) to 0.05 and 0.5. Besides, the data augmentation strategies are not utilized in both the training and inference phase. We run each experiment of our proposed G-FSDet five times and report the average results computed over five repeated experiments.

5.4. Comparison with the state-of-the-arts

To demonstrate the effectiveness of our proposed G-FSDet, we report the comparative performance of our method with several state-of-the-art methods on DIOR and NWPU VHR-10.v2 datasets. The compared few-shot object detection methods include four meta-learning methods (FRW (Li et al., 2022b), Meta-RCNN (Yan et al., 2019), FsDetView (Xiao and Marlet, 2020), P-CNN (Cheng et al., 2022)) and two transfer-learning methods (TFA (Wang et al., 2020), DeRFCN (Qiao et al., 2021)). For a fair comparison, we discard the multi-scale training strategy in TFA and DeRFCN.

5.4.1. Results on DIOR dataset

Table 3 presents the performance comparison on the DIOR test set. As the baseline framework of our proposed G-FSDet, TFA achieves the best base class performance while suffering from the low performance of the novel class. We attribute the above problem to the fine-tuning strategy in TFA, where only the box predictor is trained in the fine-tuning stage, which well preserves the base class knowledge but severely limits the detection performance of the novel class. By contrast, our proposed method adopts a more efficient transfer-learning framework and employs the proposed metric-based discriminative loss function, leading to a significant performance improvement of novel class in all few-shot scenarios. Furthermore, equipped with the proposed representation compensation module, our method can maintain excellent novel class performance with only slight base class performance degradation, which confirms that the proposed G-FSDet can effectively learn novel knowledge while protecting the previous knowledge. Compared with meta-learning methods (FRW, Meta-RCNN, and FsDetView), our approach shows superior performance under different few-shot settings. Besides, based on the transfer-learning framework, our G-FSDet discards the complicated episodic training in the meta-learning methods and has a simpler and more efficient training process. In comparison with the state-of-the-art few-shot detectors (DeRFCN and P-CNN), our proposed G-FSDet also achieves better detection performance in most few-shot scenarios. Notably, our method promotes novel class performance and effectively alleviates the catastrophic forgetting problem of base classes neglected in DeRFCN and P-CNN. Therefore, the proposed G-FSDet is more suitable for real-world generalized few-shot object detection.

Apart from the above qualitative analysis, we also visualize the detection results under the 10-shot setting in the first novel/base split in Fig. 4. Our proposed G-FSDet has the ability of generalized few-shot object detection, which can simultaneously detect both novel and base objects.

5.4.2. Results on NWPU VHR-10.v2 dataset

Table 4 reports the detection performance between our G-FSDet with comparison methods on the NWPU VHR-10.v2 test set. Specifically, TFA still shows the best performance in the base class but has trailed the detection performance of our proposed G-FSDet in the novel class and overall class. Compared with the state-of-the-art remote sensing few-shot detector P-CNN, our G-FSDet achieves more than 5% improvement in novel class performance under all few-shot settings and effectively overcomes the base class performance degradation, which confirms the superior performance of our G-FSDet in remote sensing FSOD task once more. Moreover, due to the low variations of objects in the NWPU VHR-10.v2 dataset, the comparison methods also maintain good performance in the base class, which covers up the effectiveness of our proposed G-FSDet in overcoming the catastrophic forgetting problem. The detection results under the 10-shot setting in the first novel/base split are sketched in Fig. 5.

5.4.3. Complexity comparison

Apart from the detection performance analysis, we also consider the model complexity comparison between the proposed G-FSDet and other FSOD methods. As shown in Table 5, our proposed G-FSDet method achieves a significant reduction in the model complexity after adopting the re-parameterization technique while maintaining the same detection performance. Thanks to the re-parameterization technique, the G-FSDet has the same model complexity as the baseline TFA method but exhibits better performance in few-shot object detection. Besides, compared with other few-shot detectors (DeRFCN and P-CNN), our G-FSDet achieves the best detection performance with only a slight increase in the number of model parameters.

5.5. Ablation analysis

We also perform an extensive ablation analysis to better understand the effectiveness of each proposed module in our G-FSDet. All experiments are conducted on the DIOR test set in the first novel/base split.

5.5.1. Ablation study on G-FSDet

To analyze the importance of each proposed module, we perform an ablation study on the G-FSDet and show the performance in Table 6. Firstly, we set TFA (Wang et al., 2020), an advanced few-shot detector for nature images, as the baseline model. However, TFA suffers severe performance degradation in the remote sensing scene. Then, we substitute the TFA transfer-learning approach (Wang et al., 2020) with the proposed efficient transfer-learning framework (ETF), and the performance on the few-shot novel classes shows a significant improvement of 15.85%/19.07% under the 5-shot/10-shot setting, which confirms that our proposed ETF is more suitable for FSOD in RSIs. Next, we employ our designed metric-based discriminative loss (DL) as an auxiliary loss to learn a more discriminative classifier in the few-shot scenario. As illustrated in the third row of Table 6, it further boosts the performance by 1.74%/2.15% for the few-shot novel classes. It is worth noting that the proposed DL also leads to performance degradation of the base class. We attribute it to the inter-class difference in focusing factors where the base class samples tend to have smaller focusing factors while the novel class samples prefer the larger ones, which makes the few-shot detector focus more on the novel class and suffer the performance degradation of the base class. Finally, with the help of the proposed representation compensation module (RC), the catastrophic forgetting problem in the few-shot object detection is

Table 4
Few-shot object detection performance on the NWPU VHR-10.v2 test set under 3, 5, 10, 20-shot settings. The red denotes the best results, and the blue represents the second-best results.

Split	Method	3-shot			5-shot			10-shot			20-shot		
		Base	Novel	All	Base	Novel	All	Base	Novel	All	Base	Novel	All
1	FRW	83.13	15.35	62.80	82.78	16.24	62.82	83.89	24.00	65.92	82.80	27.16	66.11
	Meta-RCNN	87.00	20.51	67.05	85.74	21.77	66.55	87.01	26.98	69.00	87.29	28.24	69.57
	FsDetView	87.68	24.56	68.75	87.77	29.55	70.31	87.75	31.77	70.96	87.83	32.73	71.30
	P-CNN	82.84	**41.80**	70.53	82.89	**49.17**	72.79	83.05	**63.29**	78.11	83.59	**66.83**	78.55
	TFA w/cos	**89.35**	8.80	65.19	**89.60**	9.49	64.65	**89.95**	9.26	65.74	**89.62**	10.83	65.98
	DeRFCN	87.22	37.90	**73.33**	86.87	46.08	**74.63**	87.29	62.95	**79.99**	87.26	64.61	**80.46**
	Ours	**89.11**	**49.05**	**77.01**	**88.37**	**56.10**	**78.64**	**88.40**	**71.82**	**83.43**	**89.73**	**75.41**	**85.44**
2	FRW	85.34	28.61	68.32	85.75	31.45	69.45	85.49	32.53	69.60	86.58	33.93	70.79
	Meta-RCNN	86.86	21.41	67.23	87.38	35.34	71.77	87.56	37.14	72.43	87.26	39.47	72.92
	FsDetView	88.11	39.01	73.38	89.34	40.31	74.63	89.34	45.09	76.07	89.31	46.28	76.40
	P-CNN	81.03	39.32	68.52	81.18	**46.10**	70.70	80.93	**55.90**	73.41	81.21	**58.37**	75.50
	TFA w/cos	**90.14**	11.14	66.44	**91.19**	12.46	67.57	**90.79**	11.35	66.96	**90.37**	11.56	66.73
	DeRFCN	88.47	**39.19**	**73.69**	88.17	45.56	**75.29**	88.25	54.05	**77.99**	88.05	57.38	**78.83**
	Ours	**89.99**	**50.09**	**78.02**	**90.52**	**58.75**	**80.99**	**89.23**	**67.00**	**82.56**	**90.61**	**75.86**	**86.13**

Table 5
Model complexity comparison between our G-FSDet and other FSOD approaches.

Model	FRW	Meta-RCNN	FsDetView	P-CNN	TFA w/cos	DeFRCN	G-FSDet	G-FSDet[a]
#Params	66.39M	74.95M	79.36M	56.53M	60.19M	52.37M	74.08M	60.19M

[a]Represents the re-parameterization technique.

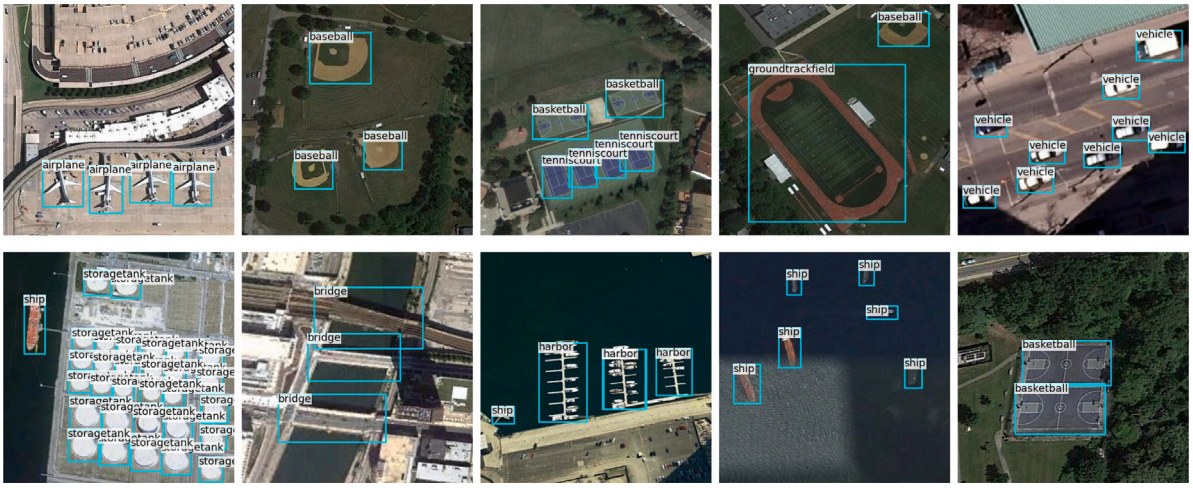

Fig. 5. Visualization of the detection results on NWPU VHR-10.v2 test set in the first novel/base split under 10-shot setting.

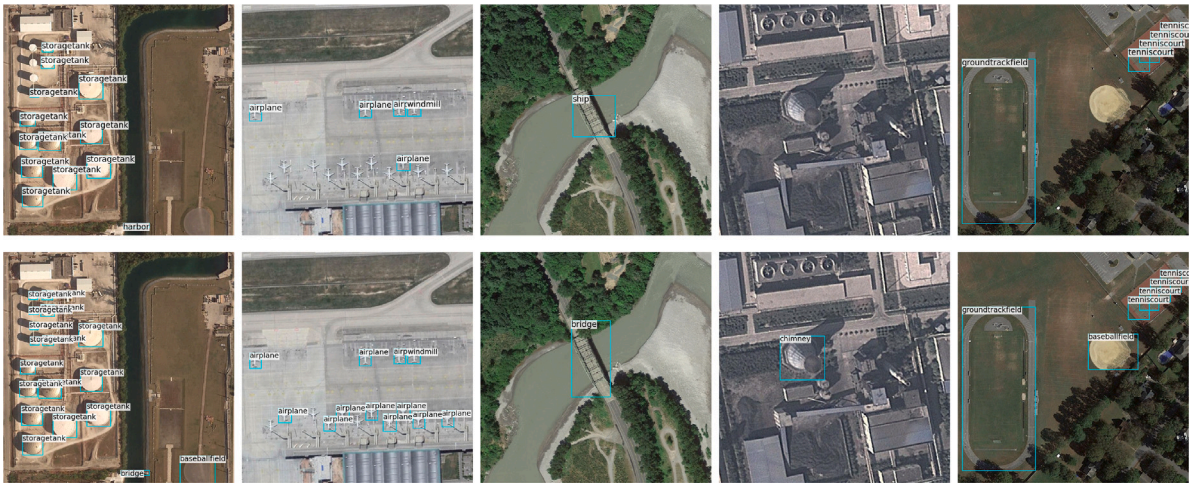

Fig. 6. Visualization of the comparison results between the ETF (first row) and G-FSDet (second row) on the DIOR test set in the first novel/base split under 10-shot setting.

Table 6
Ablation study of each proposed component in our G-FSDet on DIOR test set in the first novel/base split.

Method	ETF	DL	RC	5-shot			10-shot		
				Base	Novel	All	Base	Novel	All
TFA w/cos				**70.51**	11.57	55.78	**70.52**	15.37	56.73
G-FSDet	✓			66.24	27.42	56.54	65.15	34.44	57.48
	✓	✓		64.33	29.16	55.54	63.89	36.59	57.06
	✓	✓	✓	69.52	**30.52**	**59.72**	69.03	**37.46**	**61.16**

Table 7
Ablation study on the proposed ETF on DIOR test set in the first novel/base split.

Method	Fine-tuned			5-shot			10-shot		
	RoI-FE	RPN	FE	Base	Novel	All	Base	Novel	All
TFA w/cos				**70.51**	11.57	55.78	**70.52**	15.37	56.73
ETF	✓			65.81	17.08	53.63	64.99	21.61	54.15
	✓	✓		66.24	**27.42**	**56.54**	65.15	**34.44**	**57.48**
	✓	✓	✓	58.80	20.95	49.34	58.79	32.43	52.20

Table 8
Average recall comparisons between frozen and un-frozen RPN of the novel classes on the DIOR test set in the first novel/base split under the 10-shot setting.

RPN/Metric	AR@100	AR@300	AR@1000
Frozen	13.7	20.3	31.5
Un-frozen	**21.0**	**27.3**	**37.7**

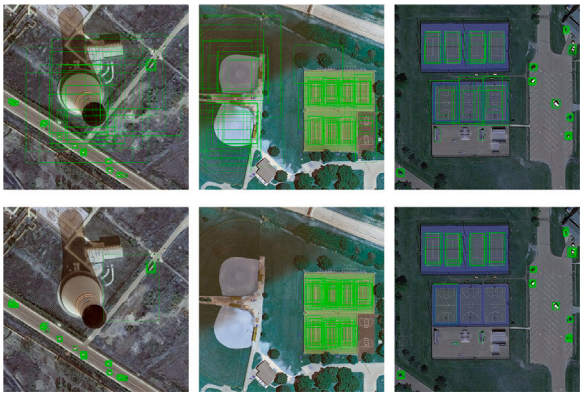

Fig. 7. Anchor prediction on the DIOR test set in the first novel/base split. Up: Un-frozen RPN. Bottom: Frozen RPN.

alleviated, leading to a remarkable performance improvement for the base classes. Besides the above quantitative analysis, we also visualize the comparative results in Fig. 6. As shown in the first two columns, our proposed G-FSDet maintains good detection performance of the base class and avoids the miss detection caused by the catastrophic forgetting problem. Meanwhile, the comparative results in the last three columns demonstrate the superiority of our method in the detection performance of novel classes.

5.5.2. Ablation study on ETF

As shown in Table 7, we conduct the ablation study of the proposed ETF to analyze the properties of different components in the detector during the fine-tuning stage. For the RoI feature extractor, we unfreeze it during the fine-tuning stage, and the novel class performance is improved by 5.51%/6.24%. We attribute this improvement to the fact that the high-level semantic features encoded in the RoI feature extractor are conducive to the classification task. Previous works (Wang et al., 2020; Xiao and Marlet, 2020; Yan et al., 2019) treat the RPN as a class-agnostic component and freeze it during the fine-tuning stage. However, in the base training phase, the RPN is only interested in the base class objects, and this bias leads to RPN ignoring the novel class objects in the fine-tuning stage, resulting in miss detection for the novel class. As shown in Table 8, we compare the average recall (AR) of the frozen RPN with the fine-tuned RPN on novel classes, and the fine-tuned one achieves better recall performance. In addition, we visualize the output proposal of the frozen and fine-tuned RPN in Fig. 7. It is clear that the fine-tuning strategy endows the RPN to capture more proposals for novel classes (e.g., chimneys, baseball fields, and basketball courts.) compared with the frozen one. Based on the above qualitative and quantitative results, we propose to unfreeze the RPN in the fine-tuning stage. This modification boosts the detection performance of novel classes by 10.34%/12.83% under the 5-shot/10-shot setting. Furthermore, we also adopt the fine-tuning RPN strategy to the meta-learning FSOD methods and achieve notable improvements for the few-shot novel classes, as shown in Fig. 8, which confirms that fine-tuning the RPN is indispensable for remote sensing FSOD. As for the feature extractor, we keep it frozen in the fine-tuning stage to alleviate the forgetting problem of base classes and reduce the overfitting phenomenon with fewer trainable parameters.

5.5.3. Hyper-parameter analysis on metric-based discriminative loss

The metric-based discriminative loss is designed as an auxiliary loss for the category classification to learn a more discriminative classifier in the few-shot scenario. As shown in Table 9, we analyze the sensitivity of focusing parameter γ. We can see that the proposed metric-based discriminative loss is relatively robust to different focusing parameters and achieves a satisfactory performance improvement for the novel class. Considering the overall detection performance of all classes, we adopt γ to 4 in our proposed metric-based discriminative loss.

5.5.4. Ablation study on RC module

The proposed RC module aims to alleviate the catastrophic forgetting problem in the fine-tuning stage. As shown in Fig. 9, we study the effectiveness of the representation decoupling mechanism and knowledge distillation loss in the RC module. It can be observed that applying the representation decoupling mechanism prevents the performance deterioration of the base classes to a certain extent, and further adding the knowledge distillation loss effectively alleviates the forgetting of previous knowledge. Besides, the proposed RC module does not hurt the performance of the novel classes, which demonstrates the proposed RC module not only remembers the previous knowledge but also effectively learns novel knowledge.

In the RC module, we design a knowledge distillation loss to further regularize the learning of novel knowledge in the fine-tuning stage. As shown in Table 10, we analyze the few-shot detection performance under different degrees of knowledge distillation. We can see that a larger knowledge distillation loss weight η limits the learning of novel knowledge and harms the performance of the novel classes. In our experiments, we set η as 0.025 to achieve the best performance over all classes.

6. Conclusion

In this paper, we focus on the G-FSOD in the remote sensing scenes and propose a novel generalized few-shot detector, namely G-FSDet, based on the transfer-learning framework. Drawing on the comprehensive analysis of each component in the detector, we propose a highly efficient transfer-learning framework for the FSOD in RSIs. Taking into account the classification confusion problem exacerbated by limited available training samples, we devise a metric-based discriminative loss to learn a more discriminative classifier in the few-shot scenario. Furthermore, we propose a representation compensation module to alleviate the catastrophic forgetting problem of base classes. Extensive experiments on two remote sensing FSOD datasets demonstrate the effectiveness of our proposed G-FSDet.

The number of available samples is the key that limits the performance in few-shot learning. Therefore, in future work, we will exploit the advanced generative models, such as AutoEncoder (Schwartz

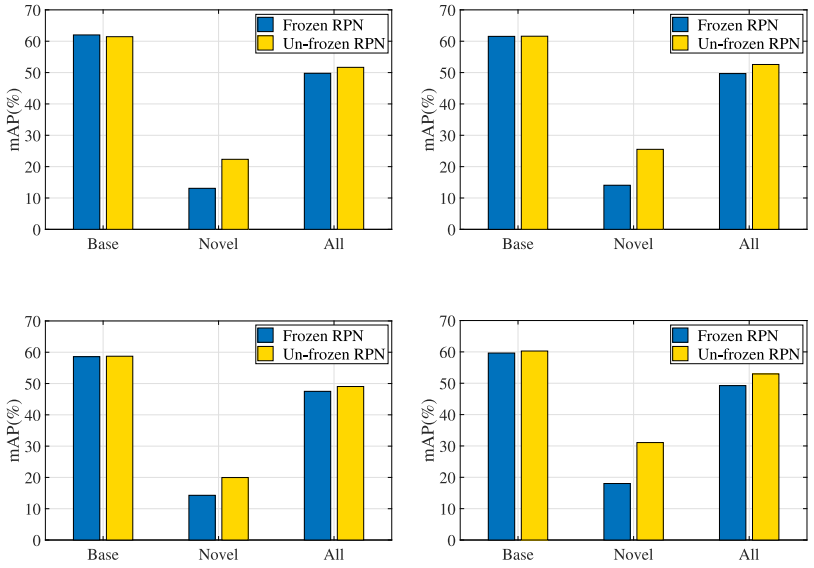

Fig. 8. The comparison performance with or without fine-tuning RPN strategy for the meta-learning methods Meta-RCNN (Yan et al., 2019) and FsDetView (Xiao and Marlet, 2020), under 5-shot (left) and 10-shot (right) settings on the DIOR test set in the first novel/base split.

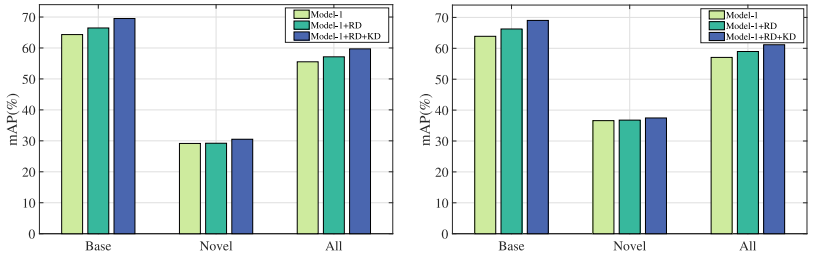

Fig. 9. The ablation study on the proposed RC module under 5-shot (left) and 10-shot (right) settings on the DIOR test set in the first novel/base split. Model-1 represents our proposed ETF+DL model (the third row in Table 6). RD and KD denote the representation decoupling mechanism and knowledge distillation loss, respectively.

Table 9
Sensitivity analysis of focusing parameter γ on the DIOR test set in the first novel/base split. For each metric, we report the average and 95% confidence interval computed over five repeated experiments.

γ	5-shot			10-shot		
	Base	Novel	All	Base	Novel	All
1	63.24 ± 0.051	29.24 ± 0.070	54.74 ± 0.044	62.48 ± 0.068	36.56 ± 0.093	56.01 ± 0.050
2	63.87 ± 0.056	**29.20 ± 0.104**	55.21 ± 0.045	63.62 ± 0.027	36.24 ± 0.087	56.78 ± 0.030
3	64.08 ± 0.042	29.18 ± 0.113	55.39 ± 0.036	63.76 ± 0.033	36.51 ± 0.061	56.94 ± 0.019
4	**64.33 ± 0.027**	29.16 ± 0.079	**55.54 ± 0.035**	**63.89 ± 0.040**	**36.59 ± 0.063**	**57.06 ± 0.030**
5	64.21 ± 0.041	28.97 ± 0.112	55.39 ± 0.058	63.84 ± 0.025	36.24 ± 0.072	56.95 ± 0.028

Table 10
Sensitivity analysis of knowledge distillation loss weight η on the DIOR test set in the first novel/base split. For each metric, we report the average and 95% confidence interval computed over five repeated experiments.

η	5-shot			10-shot		
	Base	Novel	All	Base	Novel	All
0.01	69.48 ± 0.034	30.23 ± 0.111	59.63 ± 0.057	68.56 ± 0.034	37.40 ± 0.065	60.78 ± 0.040
0.025	**69.52 ± 0.030**	**30.52 ± 0.076**	**59.79 ± 0.025**	69.03 ± 0.042	**37.46 ± 0.056**	**61.16 ± 0.023**
0.05	69.41 ± 0.032	30.44 ± 0.103	59.67 ± 0.021	69.18 ± 0.056	36.95 ± 0.097	61.13 ± 0.021
0.075	69.31 ± 0.040	30.21 ± 0.091	59.54 ± 0.035	**69.27 ± 0.029**	36.57 ± 0.091	61.13 ± 0.017
0.1	69.26 ± 0.032	30.17 ± 0.082	59.49 ± 0.025	69.25 ± 0.029	36.16 ± 0.083	60.98 ± 0.028

et al., 2018), Generative Adversarial Network (Li et al., 2020b), and Flow (Shen et al., 2020), to generate diverse samples or features to further promote the detection performance in G-FSOD.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Audebert, N., Le Saux, B., Lefèvre, S., 2018. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. ISPRS J. Photogramm. Remote Sens. 140, 20–32.

Cao, Y., Wang, J., Jin, Y., Wu, T., Chen, K., Liu, Z., Lin, D., 2021. Few-shot object detection via association and discrimination. Adv. Neural Inf. Process. Syst. 34, 16570–16581.

Chen, W., Liu, Y., Kira, Z., Wang, Y.F., Huang, J., 2019. A closer look at few-shot classification. In: 7th International Conference on Learning Representations, ICLR 2019.

Chen, Y., Liu, Z., Xu, H., Darrell, T., Wang, X., 2021. Meta-baseline: Exploring simple meta-learning for few-shot learning. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9042–9051.

Chen, H., Wang, Y., Wang, G., Qiao, Y., 2018. LSTD: A low-shot transfer detector for object detection. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. pp. 2836–2843.

Cheng, G., Han, J., Zhou, P., Guo, L., 2014. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. ISPRS J. Photogramm. Remote Sens. 98, 119–132.

Cheng, G., Yan, B., Shi, P., Li, K., Yao, X., Guo, L., Han, J., 2022. Prototype-CNN for few-shot object detection in remote sensing images. IEEE Trans. Geosci. Remote Sens. 60, 1–10. http://dx.doi.org/10.1109/TGRS.2021.3078507.

Cheng, G., Zhou, P., Han, J., 2016. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. IEEE Trans. Geosci. Remote Sens. 54 (12), 7405–7415.

Dai, J., Li, Y., He, K., Sun, J., 2016. R-FCN: Object detection via region-based fully convolutional networks. In: Advances in Neural Information Processing Systems. pp. 379–387.

Deng, Z., Sun, H., Zhou, S., Zhao, J., Lei, L., Zou, H., 2018. Multi-scale object detection in remote sensing imagery with convolutional neural networks. ISPRS J. Photogram. Remote Sens. 145, 3–22.

Ding, J., Xue, N., Long, Y., Xia, G.-S., Lu, Q., 2019. Learning RoI transformer for detecting oriented objects in aerial images. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 2849–2858.

Ding, J., Xue, N., Xia, G., Bai, X., Yang, W., Yang, M.Y., Belongie, S.J., Luo, J., Datcu, M., Pelillo, M., Zhang, L., 2022. Object detection in aerial images: A large-scale benchmark and challenges. IEEE Trans. Pattern Anal. Mach. Intell. 44 (11), 7778–7796.

Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., Sun, J., 2021. RepVGG: Making VGG-style ConvNets great again. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 13733–13742.

Fan, Z., Ma, Y., Li, Z., Sun, J., 2021. Generalized few-shot object detection without forgetting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4527–4536.

Fan, Q., Zhuo, W., Tang, C., Tai, Y., 2020. Few-shot object detection with attention-RPN and multi-relation detector. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4012–4021.

Feng, X., Han, J., Yao, X., Cheng, G., 2021. TCANet: Triple context-aware network for weakly supervised object detection in remote sensing images. IEEE Trans. Geosci. Remote Sens. 59 (8), 6946–6955.

Fu, K., Chang, Z., Zhang, Y., Xu, G., Zhang, K., Sun, X., 2020. Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images. ISPRS J. Photogramm. Remote Sens. 161, 294–308.

Han, J., Ding, J., Xue, N., Xia, G., 2021. ReDet: A rotation-equivariant detector for aerial object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2786–2795.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Hinton, G.E., Vinyals, O., Dean, J., 2015. Distilling the knowledge in a neural network. URL: http://arxiv.org/abs/1503.02531.

Huang, X., He, B., Tong, M., Wang, D., He, C., 2021. Few-shot object detection on remote sensing images via shared attention module and balanced fine-tuning strategy. Remote Sens. 13 (19), 3816.

Kang, B., Liu, Z., Wang, X., Yu, F., Feng, J., Darrell, T., 2019. Few-shot object detection via feature reweighting. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8419–8428.

Karlinsky, L., Shtok, J., Harary, S., Schwartz, E., Aides, A., Feris, R.S., Giryes, R., Bronstein, A.M., 2019. RepMet: Representative-based metric learning for classification and few-shot object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5197–5206.

Li, X., Deng, J., Fang, Y., 2022b. Few-shot object detection on remote sensing images. IEEE Trans. Geosci. Remote Sens. 60, 1–14. http://dx.doi.org/10.1109/TGRS.2021.3051383.

Li, Y., Kong, D., Zhang, Y., Tan, Y., Chen, L., 2021. Robust deep alignment network with remote sensing knowledge graph for zero-shot and generalized zero-shot remote sensing image scene classification. ISPRS J. Photogramm. Remote Sens. 179, 145–158.

Li, Q., Mou, L., Liu, Q., Wang, Y., Zhu, X.X., 2018. HSF-Net: Multiscale deep feature embedding for ship detection in optical remote sensing imagery. IEEE Trans. Geosci. Remote Sens. 56 (12), 7147–7161.

Li, K., Wan, G., Cheng, G., Meng, L., Han, J., 2020a. Object detection in optical remote sensing images: A survey and a new benchmark. ISPRS J. Photogramm. Remote Sens. 159, 296–307.

Li, L., Yao, X., Cheng, G., Xu, M., Han, J., Han, J., 2022a. Solo-to-collaborative dual-attention network for one-shot object detection in remote sensing images. IEEE Trans. Geosci. Remote Sens. 60, 1–11. http://dx.doi.org/10.1109/TGRS.2021.3091003.

Li, K., Zhang, Y., Li, K., Fu, Y., 2020b. Adversarial feature hallucination networks for few-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 13467–13476.

Lin, T., Dollár, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J., 2017a. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 936–944.

Lin, T., Goyal, P., Girshick, R.B., He, K., Dollár, P., 2017b. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2999–3007.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S.E., Fu, C., Berg, A.C., 2016. SSD: Single shot MultiBox detector. In: European Conference Computer Vision. pp. 21–37.

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., Johnson, B.A., 2019. Deep learning in remote sensing applications: A meta-analysis and review. ISPRS J. Photogramm. Remote Sens. 152, 166–177.

Pang, J., Li, C., Shi, J., Xu, Z., Feng, H., 2019. $\mathcal{R}^2$-CNN: Fast tiny object detection in large-scale remote sensing images. IEEE Trans. Geosci. Remote Sens. 57 (8), 5512–5524.

Qiao, L., Zhao, Y., Li, Z., Qiu, X., Wu, J., Zhang, C., 2021. DeFRCN: Decoupled faster R-CNN for few-shot object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8661–8670.

Redmon, J., Farhadi, A., 2017. YOLO9000: Better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6517–6525.

Ren, S., He, K., Girshick, R.B., Sun, J., 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems. pp. 91–99.

Schwartz, E., Karlinsky, L., Shtok, J., Harary, S., Marder, M., Kumar, A., Feris, R.S., Giryes, R., Bronstein, A.M., 2018. Delta-encoder: an effective sample synthesis method for few-shot object recognition. Adv. Neural Inf. Process. Syst. 2850–2860.

Shen, Y., Qin, J., Huang, L., Liu, L., Zhu, F., Shao, L., 2020. Invertible zero-shot recognition flows. In: European Conference Computer Vision, Vol. 12361. pp. 614–631.

Snell, J., Swersky, K., Zemel, R.S., 2017. Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems. pp. 4077–4087.

Sun, B., Li, B., Cai, S., Yuan, Y., Zhang, C., 2021. FSCE: Few-shot object detection via contrastive proposal encoding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7352–7362.

Sun, X., Wang, P., Yan, Z., Xu, F., Wang, R., Diao, W., Chen, J., Li, J., Feng, Y., Xu, T., Weinmann, M., Hinz, S., Wang, C., Fu, K., 2022. FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. ISPRS J. Photogramm. Remote Sens. 184, 116–130.

Tian, Z., Zhao, H., Shu, M., Yang, Z., Li, R., Jia, J., 2022. Prior guided feature enrichment network for few-shot segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 44 (2), 1050–1065.

Wang, X., Huang, T.E., Gonzalez, J., Darrell, T., Yu, F., 2020. Frustratingly simple few-shot object detection. In: Proceedings of the International Conference on Machine Learning, Vol. 119. pp. 9919–9928.

Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W., 2018. CosFace: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5265–5274.

Xia, G., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., Lu, X., 2017. AID: A benchmark data set for performance evaluation of aerial scene classification. IEEE Trans. Geosci. Remote Sens. 55 (7), 3965–3981.

Xiao, Y., Marlet, R., 2020. Few-shot object detection and viewpoint estimation for objects in the wild. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (Eds.), European Conference Computer Vision, Vol. 12362. pp. 192–210.

Xiao, Z., Qi, J., Xue, W., Zhong, P., 2021. Few-shot object detection with self-adaptive attention network for remote sensing images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 4854–4865.

Xu, Y., Fu, M., Wang, Q., Wang, Y., Chen, K., Xia, G., Bai, X., 2021. Gliding vertex on the horizontal bounding box for multi-oriented object detection. IEEE Trans. Pattern Anal. Mach. Intell. 43 (4), 1452–1459.

Yan, X., Chen, Z., Xu, A., Wang, X., Liang, X., Lin, L., 2019. Meta R-CNN: towards general solver for instance-level low-shot learning. In: Proceedings of the IEEE International Conference on Computer Vision,. pp. 9576–9585.

Yang, X., Yang, J., Yan, J., Zhang, Y., Zhang, T., Guo, Z., Sun, X., Fu, K., 2019. SCRDet: Towards more robust detection for small, cluttered and rotated objects. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8231–8240.

Yang, Z., Zhang, C., Li, R., Lin, G., 2022. Efficient few-shot object detection via knowledge inheritance. URL: https://doi.org/10.48550/arXiv.2203.12224.

Yu, Y., Guan, H., Li, D., Gu, T., Tang, E., Li, A., 2020. Orientation guided anchoring for geospatial object detection from remote sensing imagery. ISPRS J. Photogramm. Remote Sens. 160, 67–82.

Zhang, X., Wang, G., Zhu, P., Zhang, T., Li, C., Jiao, L., 2021. GRS-Det: An anchor-free rotation ship detector based on Gaussian-mask in remote sensing images. IEEE Trans. Geosci. Remote Sens. 59 (4), 3518–3531.

Zhang, C., Xiao, J., Liu, X., Chen, Y., Cheng, M., 2022a. Representation compensation networks for continual semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7053–7064.

Zhang, T., Zhang, X., Shi, J., Wei, S., 2020. HyperLi-Net: A hyper-light deep learning network for high-accurate and high-speed ship detection from synthetic aperture radar imagery. ISPRS J. Photogramm. Remote Sens. 167, 123–153.

Zhang, T., Zhang, X., Zhu, P., Chen, P., Tang, X., Li, C., Jiao, L., 2022b. Foreground refinement network for rotated object detection in remote sensing images. IEEE Trans. Geosci. Remote Sens. 60, 1–13. http://dx.doi.org/10.1109/TGRS.2021.3109145.

Zhao, Z., Tang, P., Zhao, L., Zhang, Z., 2022. Few-shot object detection of remote sensing images via two-stage fine-tuning. IEEE Geosci. Remote Sens. Lett. 19, 1–5. http://dx.doi.org/10.1109/LGRS.2021.3116858.

Zheng, Z., Zhong, Y., Ma, A., Han, X., Zhao, J., Liu, Y., Zhang, L., 2020. HyNet: Hyper-scale object detection network framework for multiple spatial resolution remote sensing imagery. ISPRS J. Photogramm. Remote Sens. 166, 1–14.

Zhong, Y., Han, X., Zhang, L., 2018. Multi-class geospatial object detection based on a position-sensitive balancing framework for high spatial resolution remote sensing imagery. ISPRS J. Photogramm. Remote Sens. 138, 281–294.

Zhu, F., Zhang, X., Wang, C., Yin, F., Liu, C., 2021. Prototype augmentation and self-supervision for incremental learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5871–5880.