Discriminative Prototype Learning for Few-Shot Object Detection in Remote-Sensing Images

Manke Guo, *Graduate Student Member, IEEE*, Yanan You, *Member, IEEE*, and Fang Liu

***Abstract*— Few-shot object detection (FSOD) in remote-sensing images (RSIs), which aims to detect never-seen objects with few training samples, has attracted wide attention. Some recent works leverage meta-learning to tackle this challenging task and achieve promising performance. However, information attenuation during feature extraction and simple prototype representation hamper further improvement in detecting novel classes. In this article, we propose a novel meta-learning-based FSOD approach named DPL-Net. Specifically, within the meta-learning-based framework, performing role-specific feature extraction in query and support branches, DPL-Net adopts a fine-grained information fusion (FIF) module to capture scale-aware information within query regions of interest (RoIs) and a multifrequency information enhancement (MIE) module to retain the spectral information of support samples. Moreover, considering the variability of remote-sensing objects, a discriminative prototype learning (DPL) strategy is developed to rectify the ambiguous distribution of support samples for more representative class-aware prototypes (CPs). Experiments on two benchmark datasets (NWPU VHR-10 and DIOR) demonstrate that our method effectively improves the performance of meta-learning in detecting RSIs with limited training data.**

***Index Terms*— Few-shot object detection (FSOD), meta-learning, prototype learning, remote-sensing images (RSIs).**

I. Introduction

OBJECT detection algorithms in remote-sensing images (RSIs) based on deep convolutional neural networks (CNNs) have made remarkable progress in recent years [1], [2], [3]. However, generic CNN-based object detectors suffer severe overfitting when the training data becomes scarce. In contrast, humans can learn novel concepts from just a few samples. Therefore, object detection with a few samples, that is, few-shot object detection (FSOD), is a significant and challenging task for intelligent agents in the field of RSI processing [4].

Learning base classes with sufficient annotated data and novel classes with only a few annotated samples, a few-shot object detector can simultaneously detect objects from both base and novel classes. Existing studies on FSOD in RSIs mainly consist of transfer learning-based and meta-learning-based methods. Transfer learning [5] achieves the FSOD task by fine-tuning the parameters with few-shot novel classes on the massive data-driven base detector. Aiming to improve the adaptability to novel classes, Zhao et al. [6] aggregated multilevel features from feature pyramids, which shortens the information transmission path by a bottom-up flow to propagate the localization information at low levels. Wang et al. [7] extracted context features of images to identify novel objects from complex backgrounds. Meta-learning [8], [9] subsamples the dataset as episodes, consisting of one support set and one query set, to form a series of few-shot subtasks, expecting the detector to learn how to learn from limited object information. In each subtask, the abstract knowledge learned from the support set is modeled as prototypes or variants, which can help to discriminate the query set. Specifically, on the basis of one-stage Meta YOLO [10], Li et al. [11] produced a set of reweighting vectors to recalibrate query features at multiple scales. Upon the two-stage benchmark Meta R-CNN [12], Zhang et al. [13] expanded the training data through data augmentation to address the arbitrary orientations of objects in RSIs. SAGS-TFS [14] adopts a two-way attention mechanism and fully exploits the knowledge hidden in support images by feeding the similarity map to the detail embeddings of both query and support features.

While promising, compared with transfer learning, experimental performance analysis [4] suggests that meta-learning currently makes finite contributions to FSOD in RSIs. We summarize the existing defects of mainstream Meta R-CNN series [12], [13], [14] into two aspects. For one thing, inevitable information attenuation occurs when extracting object features. Resorting to generic network designs [15], [16], [17], [18], [19] to extract the features of query regions of interest (RoIs) and support samples may not adequately exploit the limited object information. For another, the discriminability of prototypes has not been effectively explored. For each class, the mean of its multiple samples is conventionally adopted as a representative. However, due to the high interclass similarity and intraclass diversity of remote-sensing objects, as shown in Fig. 1, there are large variances among samples in the representation space. In this case, mean-based prototypes (MPs) are not representative enough to stably discriminate query sets, leading to worse detection performance on novel classes.

In this article, inspired by the work of meta-learning over RoIs in Meta R-CNN [12], we propose a novel meta-learning-based method named DPL-Net to alleviate the above-mentioned challenges. First, we design a fine-grained information fusion (FIF) module to maintain multiscale infor-

Manuscript received 4 March 2023; revised 12 July 2023 and 1 October 2023; accepted 8 October 2023. Date of publication 23 October 2023; date of current version 8 November 2023. This work was supported by the National Natural Science Foundation of China under Grant 62101060. *(Corresponding author: Yanan You.)*

The authors are with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: manke.guo@bupt.edu.cn; youyanan@bupt.edu.cn; lindaliu@bupt.edu.cn).

Digital Object Identifier 10.1109/TGRS.2023.3326992

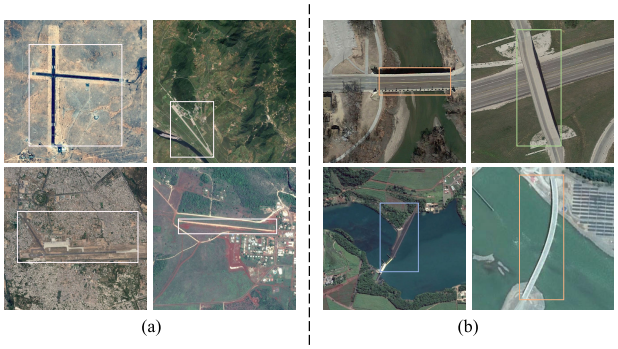

Fig. 1. Characteristics of objects in RSIs. (a) Examples of airports with various appearances (intradiversity). (b) Examples of highly similar categories, such as the overpass, dams, and bridges (intersimilarity).

mation in query RoIs, which aligns corresponding features at different pooling resolutions and provides a more comprehensive focus on the target through dilated convolutions. Second, while extracting global features of support samples, to compensate for the frequency decay caused by the global average pooling (GAP) operation, we propose a multifrequency information enhancement (MIE) module, which utilizes graph propagation to enhance efficient information integration among multiple frequency components. Third, we argue that the key to obtaining representative prototypes is rectifying the ambiguous distribution of samples. This insight motivates us to establish distinguishable sample representations, that is, samples of the same class are close together and samples of different classes are separated. To this end, we propose a discriminative prototype learning (DPL) strategy to mine hard samples in the support set and then utilize cluster-based contrastive objectives [20], [21], [22] to guide the model in learning category-specific knowledge.

The contributions of this article can be summarized as follows.

1) We propose a meta-learning-based approach for FSOD in optical RSIs, which learns the meta-knowledge from a series of episodes to generalize to novel classes.
2) We employ an FIF module and an MIE module to perform role-specific feature extraction within the meta-learning-based framework, focusing on the scale-aware information and spectral information of query RoIs and support samples, respectively.
3) We propose a DPL strategy to cope with the poor discriminability of prototypes caused by the high interclass similarity and intraclass diversity of remote-sensing objects.
4) Experiments on the NWPU-10 and DIOR datasets demonstrate the effectiveness of our DPL-Net for FSOD in RSIs.

II. Related Work

A. Few-Shot Object Detection in Natural Scene Images

To deal with the inherent long-tail distribution of real-world data, research on FSOD, which aims to detect objects belonging to novel classes with only a few annotated samples provided, has made great progress in natural scene images. In general, existing methods of FSOD can be categorized into two branches. First, transfer learning-based methods, which follow standard supervised learning and divide the training into two phases: pretraining on base classes and fine-tuning on both the base and novel classes. For example, TFA [23] only fine-tunes the classifier and the regressor of the detector while freezing the other parameters of the model. MPSR [24] handles the problem of scale variations by generating multiscale positive samples as object pyramids and refining the detectors at different scales. To ease the misclassification issues of novel instances, FSCE [25] introduces supervised contrastive learning to achieve more robust object representations, and CME [26] regularizes the class margins in an adversarial min–max fashion. Second, meta-learning-based methods adopt an episode-based training paradigm to learn how to quickly adapt between subtasks. In particular, one-stage detector Meta YOLO [10] and two-stage detector Meta R-CNN [12] adopt prototypes or variants generated from the support data to take channelwise soft-attention on query features. Subsequently, FsDetView [27] proposes a stronger feature fusion network with multiplication, subtraction, and concatenation subnetworks to replace the channelwise soft-attention layer. Furthermore, DCNet [28] matches query and support features at a pixel-wise level to fully exploit support information. Similarly, Meta faster R-CNN [29] proposes an attention-based feature alignment method to address the spatial misalignment between proposals and prototypes.

B. Few-Shot Object Detection in Remote-Sensing Images

In recent years, the works related to FSOD in RSIs have attracted much attention. For transfer learning-based methods, facing the catastrophic forgetting of base classes, DH-FSDet [30] proposes a double-head predictor in the fine-tuning phase to decouple the predictions of base and novel classes. PAMS-Det [6] aggregates features from all feature levels after FPN. CIR-FSD [7] captures context information from different receptive fields by dense connections. SAM&BFS [31] extracts multiattention maps from base samples as transfer knowledge to improve the localization of novel class objects. TFACSC [32] adopts a metric-based softmax function to calibrate the low-quality classification score.

For meta-learning-based methods, Xiao et al. [33] added more fully connected layers to extract fine-grained support features. Due to the direction arbitrariness of objects in RSIs, OFA [13] augments support set by multiangle clockwise rotation instead of traditional horizontal and vertical flipping. SAGS-TFS [14] calculates the similarity map between the objects in support and query images to strengthen the foreground samples. Based on one-stage multiscale architecture YOLOv3 [19], FSODM [11] recalibrates query features with reweighting vectors at three different scales to enable multiscale object detection. Similarly, SAAN [34] builds a relationship graph between RoIs and support images to integrate the support information into the query feature in a self-adaptive way. P-CNN [35] proposes a P-G region proposal network (RPN) to effectively identify foreground objects from the complex background under the guidance of prototypes.

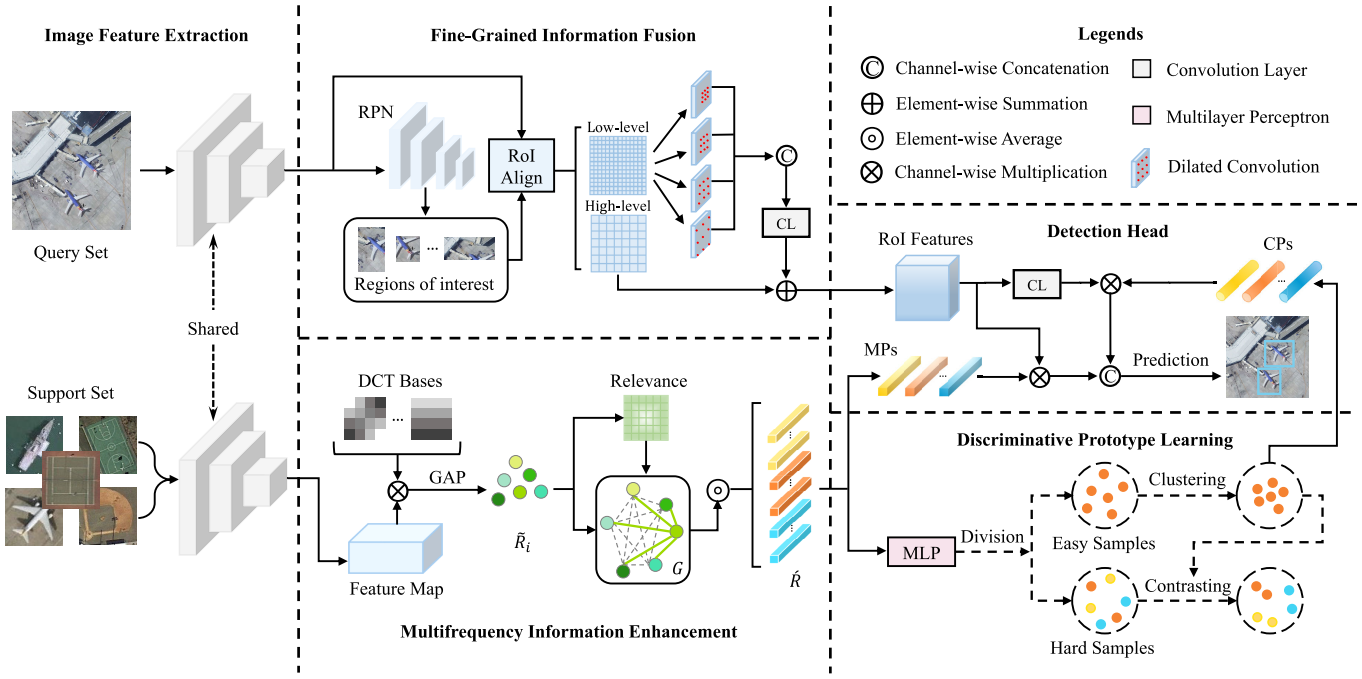

Fig. 2. Overview of our method DPL-Net, which consists of a query branch and a support branch. An image feature extraction module primarily extracts image features. Then, an FIF module and an MIE module perform feature extraction at the object level, respectively. Under the training of a DPL strategy, a DH module utilizes CPs and MPs to refine RoI features, which are fed into a classifier and a box regressor to predict objects.

Although meta-learning-based detectors are numerically superior, they usually perform inferior to transfer learning-based detectors. One reason is that existing network designs lack adjustments for meta-learning, which misses the information contained in limited objects. The other is that there is no effective training mechanism to explore the discriminability of prototypes, which inhibits the contribution of the support set. In this work, we aim to solve these problems regarding network architecture and learning strategy and thus improve the detection accuracy of novel classes.

III. Proposed Method

A. Problem Formulation

We follow common few-shot settings [12], [23] in our work. The training dataset is divided into a base dataset D_{base} with abundant annotated images for base classes and a novel dataset D_{novel} with only a few samples for novel classes. Instead of transfer learning which iteratively trains on mini-batches of training samples, meta-learning reconstructs the input as a series of episodes. Each episode consists of a query set Q and a support set S

$$\begin{aligned} Q &= \left\{\left(x^q, y^q\right), x_q \in I, y_q \in L\right\} \\ S &= \left\{\left(x^s, y^s\right), x_s \in I, y_s \in T\right\} \end{aligned} \tag{1}$$

where (x^q, y^q) represents query images and their corresponding annotations, serving as supervisory information to optimize model parameters, assisting the detector in learning to localize and classify objects. (x^s, y^s) denotes object instances cropped from support images and their class labels, providing category-specific information to the detector for discriminating objects in query images. I represents the RGB color space, and L and T denotes image- and instance-level annotation space, respectively.

B. Overall Architecture

The proposed DPL-Net is built on the popular detector Meta R-CNN [12]. As shown in Fig. 2, DPL-Net consists of a query branch (top) and a support branch (bottom), each performing role-specific feature extraction behind a shared backbone network [16].

On the query branch, taking image-level query features as input, the proposed FIF module first uses the RPN [15] to predict RoIs that may contain objects. Then, RoI features enriched with scale-aware information are extracted through a multiscale alignment design, which provides a more comprehensive focus on various objects.

On the support branch, the proposed MIE module employs frequency analysis to reassess the compression process of instance-level support features and integrates multiple frequency information into global features through graph propagation. Then, to obtain more representative class-aware prototypes (CPs), imitating the human learning process from simplicity to complexity, we introduce a DPL strategy to perform sample division and amend large variances among sample representations.

Ultimately, the detection head (DH) module exploits abstract knowledge from prototypes to detect query objects. Specifically, category-related information encoded as MPs are incorporated into RoI features, and CPs focus on discriminative characteristics of the target category. These refined RoIs are then fed into a classifier and a box regressor for prediction.

C. Fine-Grained Information Fusion

In this section, we introduce how query RoI features are extracted. As input, query feature maps produced by the shared feature extractor are fed into the RPN to generate various RoIs.

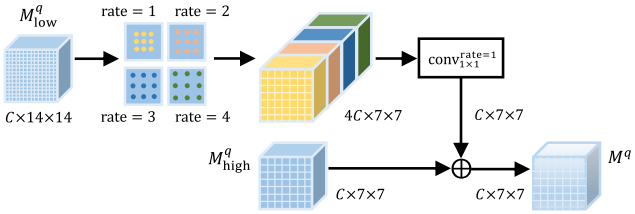

Fig. 3. Illustration of multilevel information fusion. Multiple convolutional layers are adopted to obtain different receptive fields.

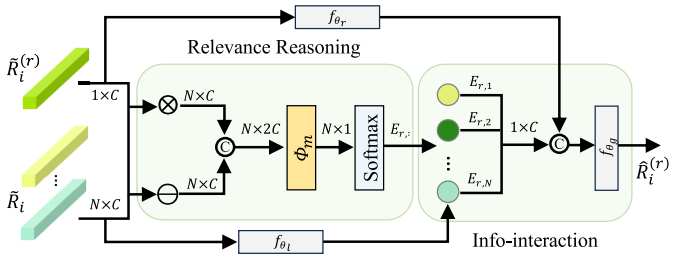

Fig. 4. Detailed flowchart of information integration. We use the measuring function ϕ_m to calculate the correlation between node r and its neighborhoods, followed by the softmax function for normalization.

Generally, to obtain RoI features with consistent size, the RoI Align module [36] subdivides each RoI into spatial bins and then uses bilinear interpolation [37] to compute the quantized values of each RoI bin. However, for RSIs with a wide range of object size variations, only using a fixed pooling resolution 7×7 during RoI subdivision will likely lose details of RoIs. Limited by data scarcity, we designed an optimized alignment to overcome this detail reduction.

Specifically, two pooling resolutions of 14×14 and 7×7 are chosen to execute RoI Align simultaneously, extracting $M^q_{\text{low}} \in \mathbb{R}^{C\times 14\times 14}$ and $M^q_{\text{high}} \in \mathbb{R}^{C\times 7\times 7}$ within each RoI. 14×14 pooling resolution is more suitable for capturing rich low-level details of objects, such as textures and shapes, and 7×7 pooling resolution mainly focuses on the overall features on a large scale, representing high-level semantics such as attributes and categories.

Next, we fuse the above multilevel information, as shown in Fig. 3. At first, M^q_{low} are fed into multibranch 3×3 convolutional layers with different dilated rates [38] of 1, 2, 3, and 4 in parallel to extract more comprehensive features from various receptive fields. Then, these features are aggregated by channelwise concatenation and a 1×1 convolutional layer to adjust the channel dimension as follows:

$$\tilde{M}^q_{\text{low}} = \text{conv}^{\text{rate}=1}_{1\times 1}\left[M^q_{\text{low}_1}, M^q_{\text{low}_2}, M^q_{\text{low}_3}, M^q_{\text{low}_4}\right] \tag{2}$$

$$M^q_{\text{low}_\lambda} = \text{conv}^{\text{rate}=\lambda}_{3\times 3}\left(M^q_{\text{low}}\right). \tag{3}$$

For small or confusable objects, the above design effectively bridges the gap between details and semantics. Finally, $\tilde{M}^q_{\text{low}}$ is added by M^q_{high} through elementwise summation to obtain RoI feature $M^q \in \mathbb{R}^{C\times 7\times 7}$, which maintains the identical size of the original RoI Align results.

D. Multifrequency Information Enhancement

In parallel with the FIF module, the MIE module serves to extract global features of support samples. Based on frequency analysis, our notion is to optimize the GAP operation, which downsamples the feature map $M_i \in \mathbb{R}^{C\times H\times W}$ of support sample i into a vector $R_i \in \mathbb{R}^C$

$$R_i = \text{gap}(M_i). \tag{4}$$

From the data compression perspective, GAP is a simplified case of 2-D discrete cosine transform (DCT), and its result is proportional to the lowest frequency component of 2-D DCT [39], [40]. In addition, several former studies [41], [42] have shown that low-frequency components tend to be more informative. To this end, we propose to integrate information of the Top-N (default $N = 6$) low-frequency components of M_i.

Concretely, we first calculate the DCT bases $B_{\text{DCT}} \in \mathbb{R}^{N\times 1\times H\times W}$

$$B_{\text{DCT}} = \left\{B^{h_0,w_0}_{i,j}, B^{h_1,w_1}_{i,j}, \ldots, B^{h_N,w_N}_{i,j}\right\} \tag{5}$$

$$B^{h_n,w_n}_{i,j} = \cos\left(\frac{\pi h_n}{H}\left(i+\frac{1}{2}\right)\right)\cos\left(\frac{\pi w_n}{W}\left(j+\frac{1}{2}\right)\right)$$
$$\text{s.t. } i \in \{0,1,\ldots,H-1\}, \quad j \in \{0,1,\ldots,W-1\} \tag{6}$$

where $B^{h_n,w_n}_{i,j} \in \mathbb{R}^{1\times H\times W}$ denotes the basis function of 2-D DCT, and (h_n, w_n) denotes 2-D indices of the Top-N low frequencies (0, 0), (1, 0), (0, 1), (1, 1), (2, 0), (0, 2), respectively. Then, multiple frequency components $\tilde{R}_i \in \mathbb{R}^{N\times C}$ can be obtained by

$$\tilde{R}_i = \text{gap}(B_{\text{DCT}} \otimes M_i) \tag{7}$$

in which $\otimes$ represents elementwise multiplication. Compared with (4), B_{DCT} is used to measure the informativeness of M_i on the frequency spectrum.

Second, a complete graph $G = (V, E)$ is built to enhance information integration. Here, nodes $V = \{V_r\}^N_{r=1}$ represents $\tilde{R}_i = \{\tilde{R}^{(r)}_i \in \mathbb{R}^C\}^N_{r=1}$, and edges $E \in \mathbb{R}^{N\times N}$ are calculated by the correlation between nodes

$$E_{rl} = \frac{\mathbb{I}_{r\neq l}\cdot e^{v_{rl}}}{\sum^N_{k=1}\mathbb{I}_{k\neq r}\cdot e^{v_{rk}}} \tag{8}$$

$$v_{rl} = \phi_m\left[\tilde{R}^{(r)}_i \otimes \tilde{R}^{(l)}_i, \tilde{R}^{(r)}_i \ominus \tilde{R}^{(l)}_i\right] \tag{9}$$

where ϕ_m, $\ominus$, and $[,]$ denote the measuring function, channelwise subtraction, and concatenation, respectively [27]. Next, we implement graph convolutions [43] to update each node

$$\hat{R}^{(r)}_i = f_{\theta_g}\left[f_{\theta_r}\left(\tilde{R}^{(r)}_i\right), \sum_l E_{rl} f_{\theta_l}\left(\tilde{R}^{(l)}_i\right)\right] \tag{10}$$

where f_{θ_g}, f_{θ_r}, and f_{θ_l} denote different fully connected layers. As shown in Fig. 4, this process is considered as relevant reasoning and information interaction of multiple frequency components. Finally, we perform elementwise average on $\hat{R}_i = \{\hat{R}^{(r)}_i \in \mathbb{R}^C\}^N_{r=1}$ in the channel dimension and obtain the enhanced vector $\hat{R}_i \in \mathbb{R}^C$.

E. Discriminative Prototype Learning

Upon the results of the MIE module, we develop a DPL strategy, which aims to estimate more discriminative CPs.

Formally, after we achieve enhanced vectors $\acute{R}$, MP representations [44] are calculated for each class as follows:

$$p_j = \frac{1}{|S^j|} \sum_{(x_i^s, y_i^s) \in S^j} \acute{R}_i \tag{11}$$

where S^j denotes the set of support samples from class j.

In practice, MPs may be far from ground-truth centers because of large variances in the distribution of support samples [45]. Although Meta R-CNN proposes a meta-classifier [12] to diversify support samples, results show that learning sophisticated decision boundaries only partially adapts to remote-sensing objects with high intraclass diversity and interclass similarity. Instead, our DPL strategy tends to enable the model to establish distinguishable support sample representations. Concretely, we first stack a multilayer perceptron $g_{\theta_d}(\cdot)$ to map enhanced vectors to a discriminative space

$$z_i = g_{\theta_d}(\acute{R}_i) \tag{12}$$

where $z_i \in \mathbb{R}^{D_d}$, by default $D_d = 512$. Subsequently, three steps are sequentially adopted in the discriminative space to cluster samples from the same class and expand the margins between different classes. The full process of the DPL strategy is summarized in Algorithm 1.

1) Division Step: Previous works [12], [13], [44] typically allocate uniform attention to the representation learning of each support sample. Nevertheless, for remote-sensing objects, variability implies varying degrees of confusion. A more reasonable pattern is to imitate human learning by commencing with easy tasks and advancing to challenging ones.

For each episode, we dynamically divide all samples into easy and hard samples. Suppose support classes are initially characterized by the centroids of their corresponding N_s samples in the discriminative space, denoted as $\mu = \{\mu_j\}_{j=1}^{N_{\text{cls}}}$, where N_{cls} denotes the number of support classes. We use Student's t-distribution [46] as a kernel to measure the similarity a_{Ij} between support sample representation z_i and class-centroid μ_j

$$a_{ij} = \frac{\left(1 + z_i \cdot \mu_j / \alpha\right)^{\frac{\alpha+1}{2}}}{\sum_{j'} \left(1 + z_i \cdot \mu_{j'} / \alpha\right)^{\frac{\alpha+1}{2}}} \tag{13}$$

$$\mu_j = \frac{1}{N_s} \sum_{i=1}^{N_s} z_i^{(j)}. \tag{14}$$

Here, $\alpha = 1$ denotes the degree of freedom of the Student's t-distribution. $z_i^{(j)}$ represents support samples belonging to class j. a_{ij} can be considered as the normalized probability of assigning sample i to class j. Then, we assign pseudo-labels [20] $\tilde{y}^s$ to all support samples according to $A = [a_{ij}]$

$$\tilde{y}_i^s = \arg\max_j a_{ij} \tag{15}$$

which reflects the distribution of support samples in the discriminative space. Based on the congruence between $\tilde{y}^s$ and the ground-truth distribution y^s, all support samples can be divided into easy and hard samples as follows:

$$\varphi(z_i) = \begin{cases} 0 \text{ (easy)}, & \text{if } \tilde{y}_i^s = y_i^s \\ -1 \text{ (hard)}, & \text{if } \tilde{y}_i^s \neq y_i^s. \end{cases} \tag{16}$$

Algorithm 1 DPL Strategy

Input:
Data: Enhanced vectors $\acute{R}$;
Number of support classes: N_{cls};
Output:
Clustering loss $\mathcal{L}_{\text{Cluster}}$ and contrasting loss $\mathcal{L}_{\text{Contrast}}$;
Construct a discriminative space $g_{\theta_d}(\cdot)$ by Eq. (12);
Initialize the centroids $\mu = \{\mu_1, \mu_2, \ldots, \mu_{N_{\text{cls}}}\}$ of support classes by Eq. (14);
for each support sample $(x_i^s, y_i^s) \in S$ **do**
  Generate pseudo-label $\tilde{y}_i^s$ based on the similarity between z_i and μ by Eq. (15);
  Compare $\tilde{y}_i^s$ and y_i^s to divide easy and hard samples by Eq. (16);
end for
Calculate target distribution P by Eq. (17);
Calculate clustering loss $\mathcal{L}_{\text{Cluster}}$ by Eq. (18)-(19);
Update μ with easy samples by Eq. (20);
Calculate contrasting loss $\mathcal{L}_{\text{Contrast}}$ by Eq. (21)-(23);
return $\mathcal{L}_{\text{Cluster}}$, $\mathcal{L}_{\text{Contrast}}$;

Obviously, easy samples are closer to the corresponding class centroids of their ground truth, while hard samples are the opposite.

2) Clustering Step: In this step, easy sample representations are strengthened by self-supervised deep clustering to form tighter clusters. First, an auxiliary target distribution [47] P for easy samples is calculated as follows:

$$p_{ij} = \frac{a_{ij}^2 / f_j}{\sum_{j'} a_{ij'}^2 / f_{j'}} \tag{17}$$

where $f_j = \sum_i a_{ij}$ denote soft cluster frequencies. Compared with A, after square and normalization, P has higher confidence assignments. Then, A is pushed toward P by optimizing the clustering loss

$$\mathcal{L}_{\text{Cluster}} = \frac{1}{N_{\text{easy}}} \sum_{i=1}^{N_{\text{cls}} N_s} \varphi(z_i) \ell_i \tag{18}$$

$$\ell_i = \text{KL}\left[p_i \| a_i\right] = \sum_{j=1}^{N_{\text{cls}}} p_{ij} \log \frac{p_{ij}}{a_{ij}}. \tag{19}$$

Here, N_{easy} denotes the number of easy samples.

In addition, due to the hard sample bias introduced in the initialization, class centroids μ are updated using only easy samples

$$\tilde{\mu}_j = \frac{1}{N_{\text{easy}}^{(j)}} \sum_{i=1}^{N_s} \mathbb{I}\left\{\varphi\left(z_i^{(j)}\right) = 0\right\} \cdot z_i^{(j)} \tag{20}$$

where $N_{\text{easy}}^{(j)}$ denotes the number of easy samples belonging to class j.

3) Contrasting Step: Next, for hard samples, we introduce a contrasting loss $\mathcal{L}_{\text{Contrast}}$ to rectify its ambiguous distribution in the discriminative space. Inspired by supervised contrastive learning [25], [48], $\mathcal{L}_{\text{Contrast}}$ defined between updated class

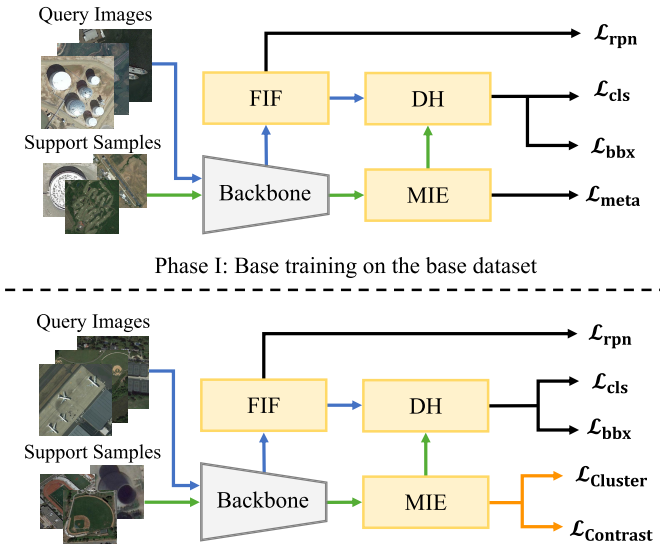

Fig. 5. Two-phase training scheme of DPL-Net. The query branch, support branch, and the proposed DPL strategy are marked with blue, green, and orange lines, respectively.

centroids $\tilde{\mu}$ and hard samples is

$$\mathcal{L}_{\text{Contrast}} = \frac{1}{N_{\text{cls}}} \sum_{j=1}^{N_{\text{cls}}} \ell_{\tilde{\mu}_j} \tag{21}$$

$$\ell_{\tilde{\mu}_j} = \frac{1}{N_{\text{hard}}^{(j)}} \sum_{i=1}^{N_s} \varphi\left(z_i^{(j)}\right) \log \frac{\exp\left(\tilde{\mu}_j \cdot z_i^{(j)}/\tau\right)}{\sum_{c=1}^{N_{\text{cls}}} \mathbb{I}_{c\neq j} \cdot \exp\left(\tilde{\mu}_j \cdot \tilde{\mu}_c/\tau\right)} \tag{22}$$

where $N_{\text{hard}}^{(j)}$ denotes the number of hard samples belonging to class j, and $\tau = 0.1$ is responsible for regulating the sensitivity of $\mathcal{L}_{\text{Contrast}}$ to hard samples. In the process of optimizing $\mathcal{L}_{\text{Contrast}}$, hard samples gradually approach the correct class centroids. Especially, when $N_{\text{hard}}^{(j)} = 0$, we adjust $\ell_{\tilde{\mu}_j}$ as

$$\ell_{\tilde{\mu}_j} = -\log \frac{\exp(1/\tau)}{\sum_{c=1}^{N_{\text{cls}}} \mathbb{I}_{c\neq j} \cdot \exp\left(\tilde{\mu}_j \cdot \tilde{\mu}_c/\tau\right)} \tag{23}$$

to maintain the discrimination between class centroids.

Through our DPL strategy, the few-shot detector learns essential class characteristics from limited samples, reflected in the discriminative space as reduced intraclass variance and declining interclass overlap. As a result, class centroids generated by (20) are representative enough to discriminate the class properties of query RoIs, so we rename them CPs.

F. Detection Head

The final prediction of RoIs generated from the RPN is completed in two steps, as shown in Fig. 2. First, to improve the generalization of our detector, MPs p_j and CPs $\tilde{\mu}_j$ are jointly used to refine query RoI features M^q as follows:

$$\acute{M}_j^q = \left[M^q \otimes p_j, f_{\theta_d}\left(M^q\right) \otimes \tilde{\mu}_j\right] \tag{24}$$

where $[,]$ and $f_{\theta_d}(\cdot)$ denote channelwise concatenation and a 1×1 convolutional layer. In this manner, MPs project RoI features into the category embedding space, and CPs further highlight the discriminative characteristics of the target

TABLE I

FEW-SHOT DETECTION PERFORMANCE ON NOVEL CLASSES OF THE NWPU VHR-10 DATASET. THE mAP (IoU = 0.5) IS CONSIDERED AN EVALUATION METRIC. THE BEST PERFORMANCE IS MARKED IN BOLD

Shots	Method	Novel Class			
		APL	BD	TC	mAP
3	TFA [23]	0.12	0.61	0.13	0.29
	PAMS-Det [6]	0.21	0.76	0.16	0.37
	CIR-FSD [7]	0.52	0.79	0.31	0.54
	Meta YOLO [10]	0.13	0.12	0.11	0.12
	Meta R-CNN [12]	0.20	0.58	0.22	0.33
	FSODM [11]	0.15	0.57	0.25	0.32
	OFA [13]	0.18	**0.87**	0.24	0.43
	SAGS-TFS [14]	0.35	0.76	**0.43**	0.51
	DPL-Net	**0.56**	0.71	0.37	**0.55**
5	TFA [23]	0.51	0.78	0.19	0.49
	PAMS-Det [6]	0.55	0.88	0.20	0.55
	CIR-FSD [7]	0.67	0.88	0.37	0.64
	Meta YOLO [10]	0.24	0.39	0.11	0.24
	Meta R-CNN [12]	0.51	0.80	0.23	0.51
	FSODM [11]	0.58	0.84	0.16	0.53
	OFA [13]	0.28	**0.89**	**0.65**	0.60
	SAGS-TFS [14]	0.64	0.82	0.52	0.66
	DPL-Net	**0.69**	0.85	0.49	**0.68**
10	TFA [23]	0.60	0.85	0.49	0.65
	PAMS-Det [6]	0.61	0.88	0.50	0.66
	CIR-FSD [7]	0.71	0.88	0.53	0.70
	Meta YOLO [10]	0.20	0.74	0.26	0.40
	Meta R-CNN [12]	0.54	0.82	0.46	0.61
	FSODM [11]	0.60	0.88	0.48	0.65
	OFA [13]	0.43	0.89	**0.68**	0.67
	SAGS-TFS [14]	0.66	0.87	0.64	0.72
	DPL-Net	**0.72**	**0.89**	0.62	**0.74**

TABLE II

FEW-SHOT DETECTION PERFORMANCE ON BASE CLASSES OF THE NWPU VHR-10 DATASET. THE mAP (IoU = 0.5) IS CONSIDERED AN EVALUATION METRIC. THE BEST PERFORMANCE IS MARKED IN BOLD

Base Class	Method					
	PAMS-Det [6]	CIR-FSD [7]	Meta YOLO [10]	FSODM [11]	Meta R-CNN [12]	DPL-Net
SP	0.88	0.91	0.77	0.72	0.84	0.90
ST	0.89	0.88	0.80	0.71	0.85	0.90
BC	0.90	0.91	0.51	0.72	0.81	0.91
GTF	0.99	0.99	0.94	0.91	0.94	0.97
HB	0.84	0.80	0.86	0.87	0.79	0.87
BR	0.80	0.87	0.77	0.76	0.82	0.89
VC	0.89	0.89	0.68	0.76	0.77	0.88
mAP	0.88	0.89	0.76	0.78	0.83	**0.90**

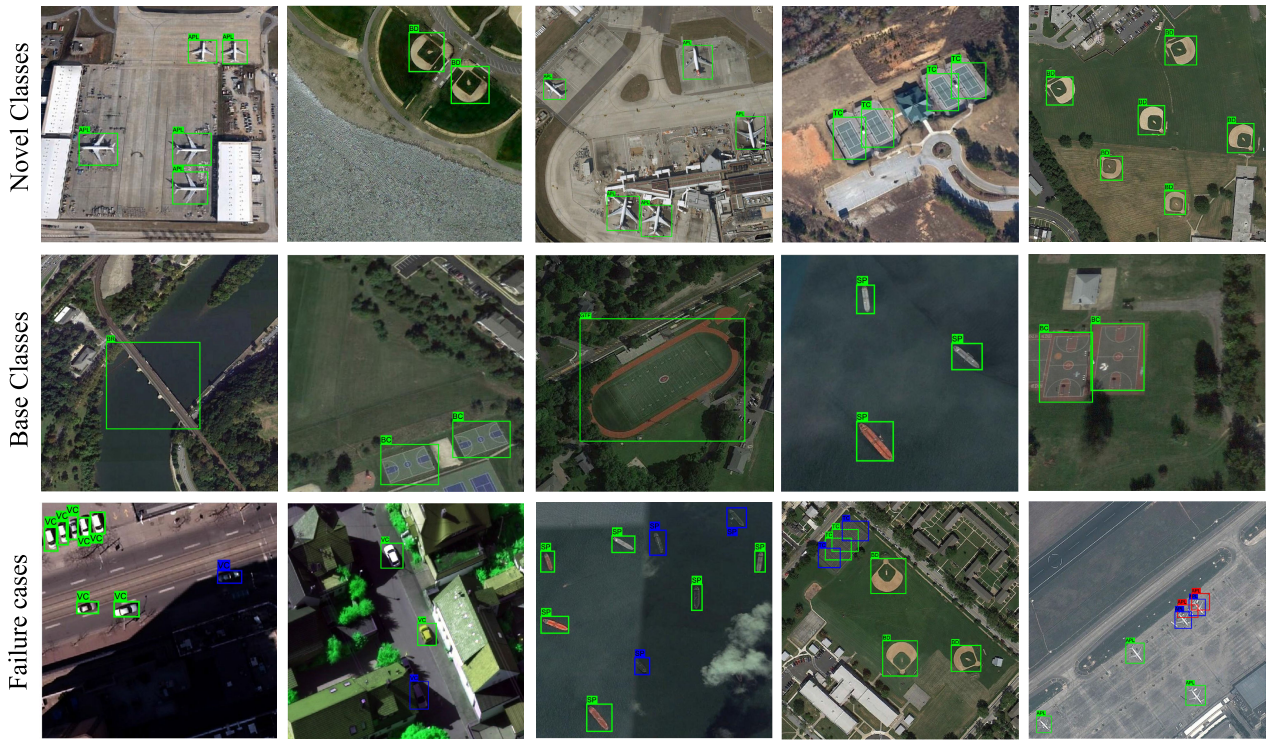

Fig. 6. Visualization of the detection results in our DLP-Net on the NWPU-10 dataset under a ten-shot setting. The first row shows the results for novel classes, the second row shows the results for base classes, and the last row shows some typical failure cases. Green, blue, and red boxes denote true positive, false negative, and false positive, respectively.

category. Then, the refined query RoI features are fed into the classifier and box regressor for prediction.

G. Training Objective

As shown in Fig. 5, the training process is divided into two phases. In the first phase (base training), our DPL-Net is trained on D_{base} with abundant annotations to obtain a base detector. Following Meta R-CNN, the loss of base training $\mathcal{L}_{\text{bt}}$ is defined as

$$\mathcal{L}_{\text{bt}} = \mathcal{L}_{\text{rpn}} + \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{bbx}} + \mathcal{L}_{\text{meta}} \tag{25}$$

where $\mathcal{L}_{\text{rpn}}$ is proposed in Faster R-CNN [15] to train the RPN for better foreground recognition. $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{bbx}}$ are proposed in Fast R-CNN [49] to train the classifier and box regressor. $\mathcal{L}_{\text{meta}}$ is proposed in Meta R-CNN [12] to learn the decision boundary of the base classes.

In the second phase (few-shot training), to improve the generalization ability of the base detector to novel classes and prevent forgetting base classes, each episode consists of one query image and N_s support samples for each class. The proposed DPL strategy is adopted to train our few-shot detector that can detect both base and novel objects. Formally, the loss of few-shot training $\mathcal{L}_{\text{ft}}$ is defined as

$$\mathcal{L}_{\text{ft}} = \mathcal{L}_{\text{rpn}} + \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{bbx}} + \lambda_1 \mathcal{L}_{\text{Cluster}} + \lambda_2 \mathcal{L}_{\text{Contrast}} \tag{26}$$

where $\mathcal{L}_{\text{Cluster}}$ is defined in (18) and (19), and $\mathcal{L}_{\text{Contrast}}$ is defined in (21)–(23). For stable training, we choose relatively smaller weights as $\lambda_1 = 0.2$ and $\lambda_2 = 0.1$.

IV. Experiments and Results

A. Datasets

NWPU VHR-10 [50] is a very-high-resolution RSI dataset collected from Google Earth and the ISPRS Vaihingen dataset. This dataset has 800 RSIs ranging in size from 500×500 to 1200×1200 pixels, where 715 RGB images with spatial resolutions from 0.5 to 2 m and 85 pansharpened color infrared images with spatial resolutions of 0.08 m. There are ten geospatial object classes: airplane (APL), baseball diamond (BD), basketball court (BC), bridge (BR), ground track field (GTF), harbor (HB), ship (SP), storage tank (ST), tennis court (TC), and vehicle (VC), with a total of 3775 object instances manually annotated in a horizontal bounding box format. NWPU VHR-10 is divided into a negative image set containing 150 images and a positive image set containing 650 images. All images from the negative image set do not contain any objects of the given object categories, while each image in the positive image set contains at least one object to be detected.

DIOR [2] is a large-scale RSI benchmark dataset collected from Google Earth. The dataset includes 23 463 optical RSIs with a size of 800×800 pixels and spatial resolutions ranging from 0.5 to 30 m. DIOR contains 20 geospatial object classes: airplane (APL), airport (APO), baseball field (BF), basketball court (BC), bridge (BR), chimney (CM), dam (DM), expressway service area (ESA), expressway toll station (ETS), harbor (HB), golf course (GC), ground track field (GTF), overpass (OP), ship (SP), stadium (SD), storage tank (ST), tennis court (TC), train station (TS), vehicle (VC), and windmill (WM), with a total of 1 92 472 object instances manually annotated

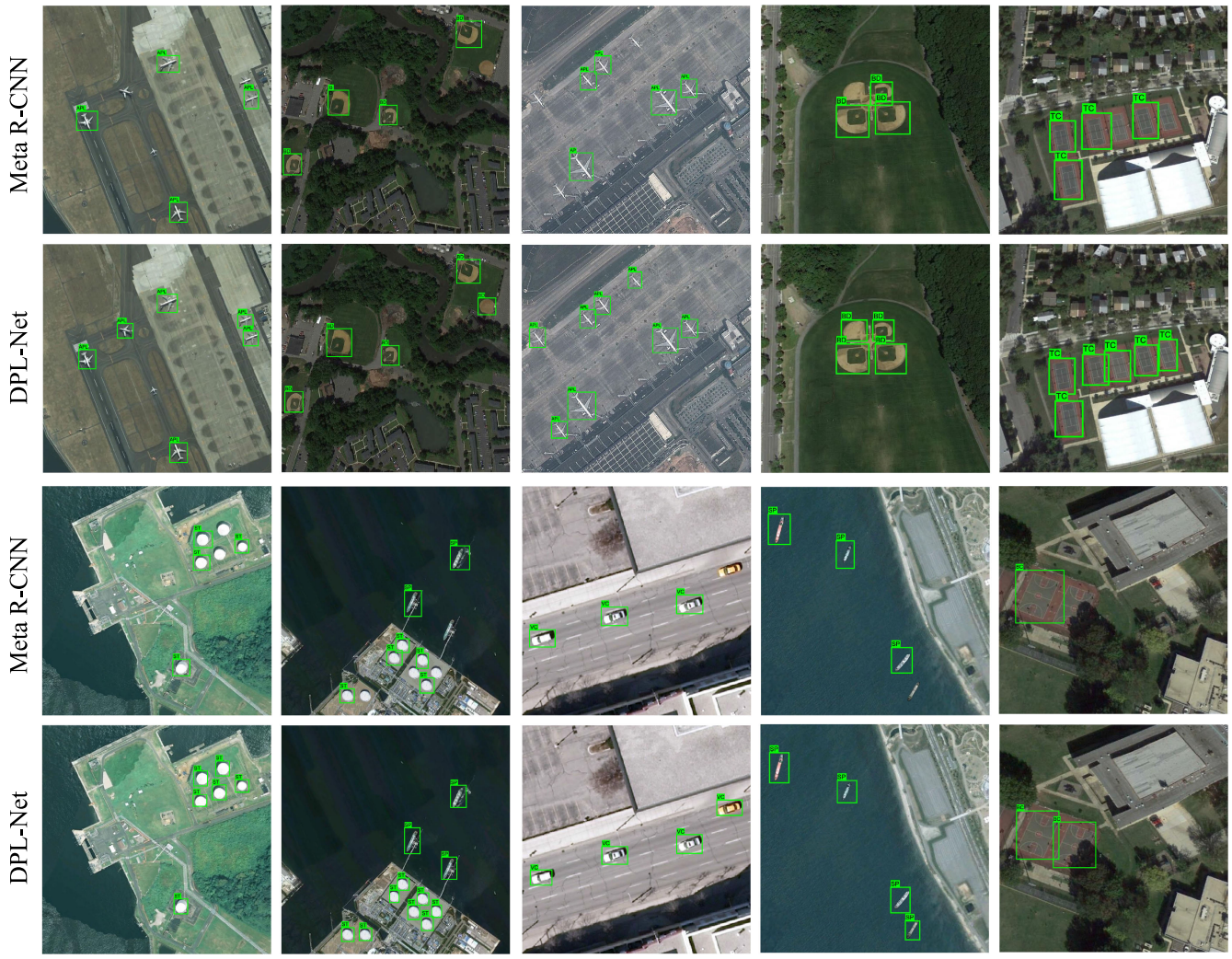

Fig. 7. Visualization of the comparison results between Meta R-CNN and our DLP-Net on the NWPU-10 dataset under a ten-shot setting. The first two rows show the results for novel classes, and the third and fourth rows show the results for base classes.

with horizontal bounding boxes. The dataset is further divided into two subsets: train-val set and test set, which contain 11 725 and 11 738 images, respectively.

B. Implementation Details

To evaluate our DPL-Net in few-shot scenarios, we split the novel and base classes in the same way as the existing work [11], that is, three novel classes (airplane, baseball diamond, and tennis court) in the NWPU VHR-10 dataset and five novel classes (airplane, baseball field, tennis court, train station, and windmill) in the DIOR dataset. The size of images in NWPU VHR-10 is scaled proportionally, where the long side is 1024 pixels. In the DIOR dataset, all images remain 800×800 pixels. Following the meta-learning paradigm, the training data is divided into a query set and a support set during each episode. In the few-shot training phase, each episode consists of one query image and N_s support samples cropped from K support images for each class, where $K = 3$, 5, and 10 for NWPU VHR-10 and $K = 5$, 10, and 20 for DIOR.

The proposed DPL-Net is implemented on the PyTorch framework. Nvidia Tesla T4 GPUs are used for acceleration in training and testing. ResNet-101 [16] pretrained on ImageNet [51] is adopted as a backbone network. During training, the SGD optimizer updates the network parameters with a momentum of 0.9 and a weight decay of 0.001. The initial learning rate is 0.001, reduced by 0.1 every five epochs. In the base training phase, we train our DPL-Net 15 epochs for the NWPU VHR-10 dataset and 20 epochs for the DIOR dataset. Ten epochs of training in the few-shot training phase are enough for the base detector to generalize from base classes to novel classes. Notably, inspired by Wang et al. [7] and Huang et al. [31], we make some adjustments to the hyperparameters of the RPN to obtain enough RoIs of novel classes. Specifically, the intersection over union (IoU) threshold between positive anchors and ground-truth boxes is reduced to 0.4, and the maximum number of candidate boxes is doubled to 5000. The predictions for base and novel classes are decoupled [30] in the few-shot training phase.

C. Comparing Methods

We comprehensively compare the proposed DPL-Net with current FSOD methods such as transfer learning-based

TABLE III

FEW-SHOT DETECTION PERFORMANCE ON NOVEL CLASSES OF THE DIOR DATASET. THE mAP (IoU = 0.5) IS CONSIDERED AN EVALUATION METRIC. THE BEST PERFORMANCE IS MARKED IN BOLD

Shots	Method	Novel Class					
		APL	BF	TC	TS	WM	mAP
5	TFA [23]	0.13	0.51	0.24	0.13	0.25	0.25
	PAMS-Det [6]	0.14	0.54	0.24	0.17	**0.31**	0.28
	CIR-FSD [7]	0.20	0.50	0.50	0.24	0.20	0.33
	Meta YOLO [10]	0.09	0.33	0.47	0.09	0.13	0.22
	Meta R-CNN [12]	0.14	0.41	0.36	0.15	0.15	0.24
	FSODM [11]	0.09	0.27	0.57	0.11	0.19	0.25
	SAGS-TFS [14]	0.17	0.50	0.62	0.17	0.23	0.34
	OFA [13]	0.26	0.60	**0.69**	0.09	0.25	0.38
	DPL-Net	**0.29**	**0.66**	0.60	**0.25**	0.18	**0.40**
10	TFA [23]	0.17	0.53	0.41	0.15	0.30	0.31
	PAMS-Det [6]	0.17	0.55	0.41	0.17	0.34	0.33
	CIR-FSD [7]	0.20	0.55	0.50	0.23	**0.36**	0.38
	Meta YOLO [10]	0.15	0.45	0.54	0.07	0.18	0.28
	Meta R-CNN [12]	0.21	0.48	0.56	0.15	0.25	0.33
	FSODM [11]	0.16	0.46	0.60	0.14	0.24	0.32
	SAGS-TFS [14]	0.24	0.51	0.63	0.22	0.24	0.37
	OFA [13]	0.30	0.63	**0.70**	0.11	0.31	0.41
	DPL-Net	**0.34**	**0.67**	0.63	**0.27**	0.26	**0.43**
20	TFA [23]	0.24	0.56	0.50	0.21	0.33	0.37
	PAMS-Det [6]	0.25	0.58	0.50	0.23	0.36	0.38
	CIR-FSD [7]	0.27	0.62	0.55	0.28	**0.37**	0.43
	Meta YOLO [10]	0.19	0.52	0.55	0.18	0.26	0.34
	Meta R-CNN [12]	0.25	0.51	0.58	0.20	0.28	0.36
	FSODM [11]	0.22	0.50	0.66	0.16	0.29	0.36
	SAGS-TFS [14]	0.33	0.53	0.63	0.28	0.35	0.42
	DPL-Net	**0.35**	**0.68**	**0.66**	**0.30**	0.32	**0.45**

methods (TFA [23], PAMS-Det [6], and CIR-FSD [7]) and meta-learning-based methods (Meta YOLO [10], Meta R-CNN [12], FSODM [11], OFA [13], and SAGS-TFS [14]).

Reliably, we adopt the mean average precision (mAP) as a metric to evaluate the performance of our DPL-Net through the PASCAL VOC2007 [52] development kit.

D. Results on NWPU-10

Table I lists the FSOD performance on novel classes of the proposed DPL-Net and the comparison methods on the NWPU VHR-10 dataset. It can be seen that DPL-Net outperforms baseline Meta R-CNN by a large margin. Compared with transfer learning-based CIR-FSD, DPL-Net achieves an overall improvement, with 1%, 4%, and 4% higher mAP in the three-shot, five-shot, and ten-shot settings, respectively. Compared with the meta-learning-based method SAGS-TFS, DPL-Net demonstrates remarkable performance, especially at lower-shot scenarios, where the mAP increases by 4%, 2%, and 2% in three-shot, five-shot, and ten-shot, respectively.

TABLE IV

FEW-SHOT DETECTION PERFORMANCE ON BASE CLASSES OF THE DIOR DATASET. THE mAP (IoU = 0.5) IS CONSIDERED AN EVALUATION METRIC. THE BEST PERFORMANCE IS MARKED IN BOLD

Base Class	Method					
	PAMS-Det [6]	CIR-FSD [7]	Meta YOLO [10]	FSODM [11]	Meta R-CNN [12]	DPL-Net
APO	0.78	0.87	0.59	0.63	0.70	0.81
BC	0.79	0.88	0.74	0.80	0.80	0.84
BR	0.52	0.55	0.29	0.32	0.49	0.55
CM	0.69	0.79	0.70	0.72	0.67	0.72
DM	0.55	0.72	0.52	0.45	0.53	0.61
ESA	0.67	0.86	0.63	0.63	0.66	0.75
ETS	0.62	0.78	0.48	0.60	0.62	0.72
GC	0.81	0.84	0.61	0.61	0.68	0.82
GTF	0.78	0.83	0.54	0.61	0.74	0.79
HB	0.50	0.57	0.52	0.43	0.48	0.54
OP	0.51	0.64	0.49	0.46	0.50	0.58
SP	0.67	0.72	0.33	0.50	0.62	0.69
SD	0.76	0.77	0.52	0.45	0.69	0.76
ST	0.57	0.70	0.26	0.43	0.58	0.64
VC	0.54	0.56	0.29	0.39	0.47	0.58
mAP	0.65	**0.74**	0.50	0.54	0.62	0.69

Furthermore, our DPL-Net achieves a more balanced performance among novel classes. In a three-shot setting, OFA gets 0.87 AP on the baseball diamond, but 0.18 AP on the airplane. In comparison, DPL-Net achieves 0.71 AP and 0.56 AP in those two classes, which demonstrates excellent generalization to novel classes. As the number of samples increased, the performance of DPL-Net improved significantly. In a ten-shot setting, DPL-Net only underperforms OFA and SAGS-TFS on the baseball diamond.

In the end, we test the detection performance on the base classes, as shown in Table II. The proposed DPL-Net effectively avoids catastrophic forgetting after few-shot training. Some detection results on base and novel classes are shown in Fig. 6. The first two rows of detection results demonstrate the effectiveness of our method. In the last row, we visualize some failure cases. It can be seen that insufficient lighting conditions cause some vehicles and ships to be missed. In the future, low-light image enhancement is a promising solution. Besides, our DPL-Net sometimes misses some densely distributed tennis courts and airplanes. However, the proposed method makes significant progress compared with Meta R-CNN, and the comparative results are visualized in Fig. 7.

E. Results on DIOR

Compared with the NWPU-10 dataset, the DIOR dataset has richer image variations. Therefore, we increase the number of annotated samples for novel classes.

The FSOD performance of the proposed DPL-Net and the comparison methods on the DIOR dataset is shown in Tables III and IV. It can be observed that the complex dataset is a great challenge for current few-shot detectors. However,

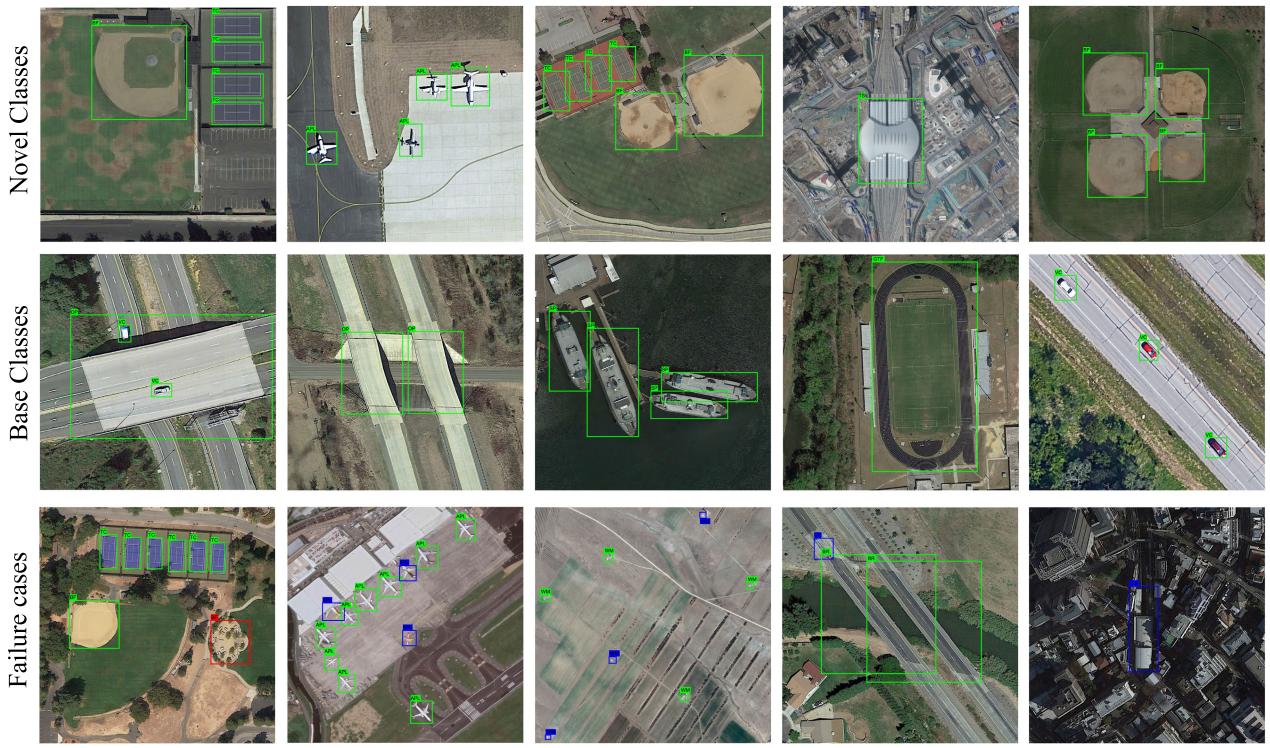

Fig. 8. Visualization of the detection results in our DLP-Net on the DIOR dataset under a 20-shot setting. The first row shows the results for novel classes, the second row shows the results for base classes, and the last row shows some typical failure cases. Green, blue, and red boxes denote true positive, false negative, and false positive, respectively.

TABLE V

ABLATION EXPERIMENT RESULTS OF OPTIMIZED FEATURE EXTRACTION DESIGNS. THE MAP (IOU = 0.5) IS CONSIDERED AN EVALUATION METRIC. THE BEST PERFORMANCE IS MARKED IN BOLD

FIF	MIE	NWPU VHR-10			DIOR		
		3-shot	5-shot	10-shot	5-shot	10-shot	20-shot
✓		0.523	0.652	0.690	0.385	0.406	0.433
	✓	0.511	0.645	0.688	0.379	0.414	0.421
✓	✓	**0.545**	**0.681**	**0.737**	**0.402**	**0.426**	**0.447**

TABLE VI

ABLATION EXPERIMENT RESULTS OF USING DIFFERENT PROTOTYPES TO DISCRIMINATE THE QUERY SET. "MPS" DENOTES MEAN-BASED PROTOTYPES, AND "CPS" DENOTES CLASS-AWARE PROTOTYPES. THE MAP (IOU = 0.5) IS CONSIDERED AN EVALUATION METRIC. THE BEST PERFORMANCE IS MARKED IN BOLD

MPs	CPs	NWPU VHR-10			DIOR		
		3-shot	5-shot	10-shot	5-shot	10-shot	20-shot
✓		0.527	0.640	0.701	0.387	0.408	0.430
	✓	0.522	0.663	0.715	0.384	0.416	0.429
✓	✓	**0.545**	**0.681**	**0.737**	**0.402**	**0.426**	**0.447**

DPL-Net still maintains superiority in any shot setting. Compared with meta-learning-based SAG-TFS, DPL-Net improves mAP by 6%, 6%, and 3% in five-shot, ten-shot, and 20-shot settings, respectively. Compared with strong competitor OFA, DPL-Net performs better in mAP in five-shot and ten-shot settings, with 2% and 2% improvements, respectively.

Specific to each novel class, DPL-Net consistently performs best on the airplane, the baseball field, and the train station. As for the tennis court, DPL-Net achieves significant progress with the help of more samples and surpasses all reported comparison methods in a ten-shot setting. On the windmill, transfer learning-based methods usually perform better. Meanwhile, DPL-Net outperforms most comparison methods on base classes.

In addition, we visualize some detection results in Fig. 8. The first row shows the results for novel classes. The second row shows the results for base classes. It can be seen that our DPL-Net demonstrates excellent performance. However, there are still some failure cases, as shown in the last row. First, one fake baseball field successfully deceives the detector. Second, a partially occluded airplane is lost by its surroundings. Third, some airplanes and windmills are missed because of their small size. Finally, due to the inconspicuous appearance, one vehicle and one train station are mistaken for the background. To be fair, we also visualize the comparative results between Meta R-CNN and our DLP-Net in Fig. 9.

F. Ablation Studies

To verify the effectiveness of our DPL-Net, we perform ablation experiments with different shot settings, and the results on two datasets are shown in Tables V and VI.

1) Impact of the FIF Module: The FIF module fuses the details and semantics of RoIs, which improves the adaptability of our detector to size variations of remote-sensing objects. After using the FIF module, the performance improves by 3.4%–4.9% on the NWPU VHR-10 dataset and by 1.2%–2.6% on the DIOR dataset.

Fig. 9. Visualization of the comparison results between Meta R-CNN and our DLP-Net on the DIOR dataset under a 20-shot setting. The first two rows show the results for novel classes, and the third and fourth rows show the results for base classes.

2) Impact of the MIE Module: The MIE module optimizes extracting global features of support samples from the frequency domain perspective, where helpful information in multifrequency components is preserved. With the MIE module, the performance improves by 2.2%–4.7% on the NWPU VHR-10 dataset and by 1.4%–2.0% on the DIOR dataset.

3) Impact of the DPL Strategy: Our DPL strategy enables the model to establish distinguishable support sample representations and obtain more representative CPs. It can be seen that our DPL strategy effectively improves the performance of the network. As the number of training samples increases, the impact of our DPL strategy follows the law of diminishing marginal utility, which indicates its significance in lower-shot scenarios. Compared with using MPs to refine the query set, the performance of our DPL-Net improves by 4.1% in the five-shot setting on the NWPU VHR-10 dataset and by 1.8% in the ten-shot setting on the DIOR dataset. Fig. 10 shows the cosine similarity difference matrices between MPs and CPs on the DIOR dataset. It can be found that the DPL strategy effectively reduces the similarity between prototype representations, which facilitates the discrimination of query RoIs.

4) Ablation for Hyperparameters of Loss $\mathcal{L}_{ft}$: In the second phase (few-shot training), the proposed clustering loss $\mathcal{L}_{\text{Cluster}}$ and contrasting loss $\mathcal{L}_{\text{Contrast}}$ are introduced into the training objective $\mathcal{L}_{\text{ft}}$ to optimize the model parameters. Considering that classification and regression are the main tasks of our few-shot detector, the proportion of auxiliary loss should be controlled within a small value range. The impact of the weight of $\mathcal{L}_{\text{Cluster}}$ and $\mathcal{L}_{\text{Contrast}}$ on detection performance is studied, and the results are shown in Table VII. In a five-shot setting, the detection performance is better when both λ_1 and λ_2 are 0.2. In a ten-shot setting, reducing λ_2 can further improve the detection performance. Specifically, we take $\lambda_1 = 0.2$ and $\lambda_2 = 0.1$ in our implementation.

5) Analysis of Parameters and Computational Complexity: As shown in Table VIII, we evaluate the model parameters (Params) and floating-point operations (FLOPs) of our DPL-Net relative to Meta R-CNN on the NWPU-10 and DIOR datasets, respectively. It can be found that the proposed method inevitably leads to a small increase in the parameters and computational complexity of our model, but effectively improves the detection performance, which indicates the superiority of our method.

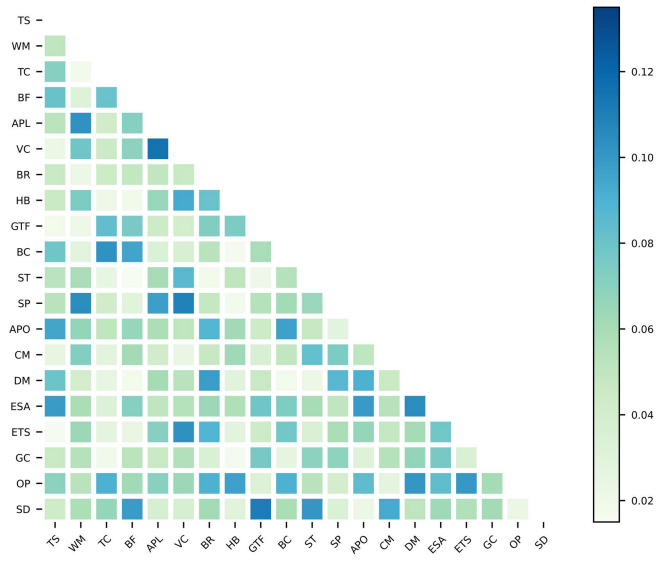

Fig. 10. Similarity difference matrices between MPs and CPs. The darker the color, the better the improvement in discrimination.

TABLE VII

ABLATION FOR DIFFERENT HYPERPARAMETERS OF THE TRAINING OBJECTIVE $\mathcal{L}_{\text{ft}}$ ON THE DIOR DATASET IN FIVE-SHOT AND TEN-SHOT SETTINGS, RESPECTIVELY. λ_1 AND λ_2 REPRESENT THE WEIGHTS OF CLUSTERING LOSS $\mathcal{L}_{\text{Cluster}}$ AND CONTRASTING LOSS $\mathcal{L}_{\text{Contrast}}$ IN (26), RESPECTIVELY. THE mAP (IoU = 0.5) IS CONSIDERED AN EVALUATION METRIC. THE BEST PERFORMANCE IS MARKED IN BOLD

λ_1	λ_2	Shots		λ_1	λ_2	Shots	
		5	10			5	10
0.05	0.1	0.379	0.401	0.1	0.2	0.397	0.410
0.1	0.1	0.385	0.418	0.2	0.2	**0.405**	0.420
0.1	0.05	0.394	0.425	0.2	0.1	0.402	**0.426**

TABLE VIII

MODEL PARAMETERS AND COMPUTATIONAL COMPLEXITY BETWEEN META R-CNN AND DPL-NET

Method	NWPU VHR-10		DIOR	
	Params (M)	FLOPs (G)	Params (M)	FLOPs (G)
Meta R-CNN	75.612	185.851	75.623	185.875
DPL-Net	77.356	186.235	77.369	186.254

V. CONCLUSION

This article proposes a novel meta-learning-based method named DPL-Net to realize FSOD in RSIs. At first, to better capture object information during feature extraction, we propose an FIF module and an MIE module for query RoIs and support samples, respectively, where the FIF module can mine scale-aware information within query RoIs, and the MIE module enhances the informativeness of support global features in the frequency domain. Furthermore, we propose a DPL strategy to improve the discriminability of prototypes by establishing distinguishable support sample representations, to stably discriminate the query set. As a result, DPL-Net effectively mitigates the information attenuation of object features and can identify novel classes from a few annotated samples. Experiments on two public benchmark datasets demonstrate that our method performs well for the FSOD task in remote-sensing scenes.

Nevertheless, modeling prototypes as channel attention vectors do not effectively explore the spatial correlation between support samples and query RoIs. In the future, to cope with complicated object patterns in RSIs, we will continue to focus on modeling prototypes in a collaborative manner with spatial and channel attention.

REFERENCES

[1] Z. Li et al., "Deep learning-based object detection techniques for remote sensing images: A survey," *Remote Sens.*, vol. 14, no. 10, p. 2385, May 2022.

[2] 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.

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

[4] S. Liu, Y. You, H. Su, G. Meng, W. Yang, and F. Liu, "Few-shot object detection in remote sensing image interpretation: Opportunities and challenges," *Remote Sens.*, vol. 14, no. 18, p. 4435, Sep. 2022.

[5] F. Zhuang et al., "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021.

[6] 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.

[7] 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.

[8] T. Liu et al., "Recent few-shot object detection algorithms: A survey with performance comparison," 2022, *arXiv:2203.14205*.

[9] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5149–5169, Sep. 2022.

[10] 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. 8420–8429.

[11] 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. 5601614.

[12] X. Yan, Z. Chen, A. Xu, X. Wang, X. Liang, and L. Lin, "Meta R-CNN: Towards general solver for instance-level low-shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9577–9586.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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. 2117–2125.

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

[19] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[20] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 132–149.

[21] Y. Li, P. Hu, Z. Liu, D. Peng, J. T. Zhou, and X. Peng, "Contrastive clustering," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 8547–8555.

[22] D. Zhang et al., "Supporting clustering with contrastive learning," 2021, *arXiv:2103.12953*.
[23] X. Wang, T. E. Huang, T. Darrell, J. E. Gonzalez, and F. Yu, "Frustratingly simple few-shot object detection," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 9919–9928.
[24] J. Wu, S. Liu, D. Huang, and Y. Wang, "Multi-scale positive sample refinement for few-shot object detection," in *Proc. Eur. Conf. Comput. Vis.* Glasgow, U.K.: Springer, Aug. 2020, pp. 456–472.
[25] B. Sun, B. Li, S. Cai, Y. Yuan, and C. Zhang, "FSCE: Few-shot object detection via contrastive proposal encoding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7352–7362.
[26] B. Li, B. Yang, C. Liu, F. Liu, R. Ji, and Q. Ye, "Beyond max-margin: Class margin equilibrium for few-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7363–7372.
[27] 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.
[28] H. Hu, S. Bai, A. Li, J. Cui, and L. Wang, "Dense relation distillation with context-aware aggregation for few-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10185–10194.
[29] G. Han, S. Huang, J. Ma, Y. He, and S.-F. Chang, "Meta Faster R-CNN: Towards accurate few-shot object detection with attentive feature alignment," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 780–789.
[30] 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.
[31] 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.
[32] 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.
[33] Z. Xiao, P. Zhong, Y. Quan, X. Yin, and W. Xue, "Few-shot object detection with feature attention highlight module in remote sensing images," in *Proc. Int. Conf. Image, Video Process. Artif. Intell.*, Nov. 2020, pp. 217–223.
[34] Z. Xiao, J. Qi, W. Xue, and P. Zhong, "Few-shot object detection with self-adaptive attention network for remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4854–4865, 2021.
[35] G. Cheng et al., "Prototype-CNN for few-shot object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–10, 2022.
[36] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.
[37] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 2017–2025.
[38] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016, pp. 1–13.
[39] Z. Qin, P. Zhang, F. Wu, and X. Li, "FcaNet: Frequency channel attention networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 783–792.
[40] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Trans. Comput.*, vol. COM-100, no. 1, pp. 90–93, Jan. 1974.
[41] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[42] K. Xu, M. Qin, F. Sun, Y. Wang, Y.-K. Chen, and F. Ren, "Learning in the frequency domain," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1740–1749.
[43] G. Han, Y. He, S. Huang, J. Ma, and S.-F. Chang, "Query adaptive few-shot object detection with heterogeneous graph convolutional networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3263–3272.
[44] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 4077–4087.
[45] B. Zhang, X. Li, Y. Ye, Z. Huang, and L. Zhang, "Prototype completion with primitive knowledge for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3754–3762.
[46] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 1–27, 2008.
[47] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 478–487.
[48] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.
[49] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
[50] 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.
[51] 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.
[52] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

Manke Guo (Graduate Student Member, IEEE) received the B.S. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2021, where he is currently pursuing the master's degree.

His main research interests include computer vision and remote-sensing image understanding.

Yanan You (Member, IEEE) received the Ph.D. degree from the School of Electronic and Information Engineering, Beihang University, Beijing, China, in 2015.

He held a post-doctoral position at Beihang University from 2015 to 2017. He is currently an Associate Professor with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing. His research interests include remote-sensing image processing, deep learning, imaging detection, and intelligent perception.

Fang Liu received the Ph.D. degree from Nankai University, Tianjin, China, in 1997.

She is currently an Associate Professor with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China. Her research interests include broadband IP networks, network traffic monitoring, machine learning, and data mining.