A Semantic-Guided Framework for Few-Shot Remote Sensing Object Detection

Chenchen Sun, Yuyu Jia, *Member, IEEE*, Han Han, Qiang Li, *Member, IEEE*, Qi Wang, *Senior Member, IEEE*

***Abstract*—Few-Shot Object Detection (FSOD) aims to recognize novel class targets using limited annotated data. Conventional approaches rely on extensive base class training, followed by fine-tuning where few instances from both base and novel classes are sampled for each category. Although they demonstrate remarkable performance in natural image domains, the specificity of remote sensing scenarios poses two critical challenges for FSOD: 1) The morphological differences between remote sensing images and natural images are significant, leading to a loss of structural priors in the Region Proposal Network (RPN). This makes it difficult for structural priors pretrained on natural images to generalize to remote sensing images, especially for novel class with scarce data; 2) Differences in imaging conditions lead to appearance variations among similar objects, leading to sparse visual features are insufficient to represent the common semantic structure of the entire class. To solve problems above, we introduce an innovative framework named ST-FSOD. Primarily, we introduce the SA-RPN module, which leverages efficient pixel association capability to generate high-quality foreground object proposals. Subsequently, through a text guiding learner module (TGL), we use textual labels of each category to generate image-agnostic text-guided prototypes. The enhanced text prototypes are fused with visual features to complement the sparse visual features. Extensive experiments conducted on the DIOR, NWPU VHR-10 and RSOD benchmarks demonstrate that the proposed method consistently surpasses strong baselines and achieves superior performance compared to previous state-of-the-art (SOTA) approaches. Our project will be open-sourced soon on https://github.com/wdcjhyy/ST-FSOD.**

***Index Terms*—object detection, transfer learning, prototypes.**

I. Introduction

OBJECT detection [1], [2], [3], as a critical challenge in computer vision, aims to quickly and accurately determine the location and category of a target. With the advancement of earth observation technology, large-scale remote sensing datasets [4], [5] have been established, which drives the advancement in deep learning-based remote sensing object detection. However, the fully supervised object detection methods rely on a large amount of labeled data for training. In many practical applications, acquiring vast quantities of labeled data is often time-consuming and expensive. A shortage of data may lead to low detection accuracy or even model overfitting. How to perform remote sensing object detection with only few samples remains a challenge. As an effective solution, many few-shot object detection (FSOD) techniques [6], [7], [8] have emerged. With only few annotated data for novel objects, FSOD can accurately learn to identify novel target categories from a limited number of samples.

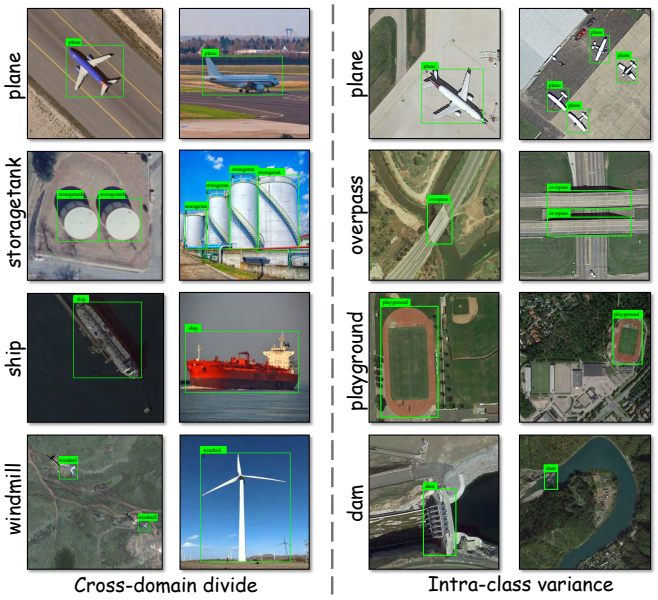

Fig. 1. The comparison highlights key issues in FSOD for remote sensing images: Significant differences in shape, size, and color between natural and remote sensing images show that prior knowledge from natural images cannot be directly applied. Additionally, the diversity of remote sensing images under varying conditions reveals substantial intra-class variations, even for identical object categories.

Specifically, most FSOD studies adopt the Faster R-CNN [1] as their foundational architecture, employing a two-stage training paradigm. The first phase trains a pre-trained detection model on a large base dataset, then freezes all detector weights except the RoI head. In the second phase, the classifier's final layer is fine-tuned using few samples from both base and novel classes for each category. Building on this normal paradigm, numerous improvements in the detector structure have been proposed. Several studies explore the structure of regional proposal networks (RPN). SAE-FSDet [9] significantly improves the quality of the proposal by introducing a two-step regression strategy and a multiscale convolution. [10] integrates an attention mechanism to filter out irrelevant background and category-mismatched proposals. Other research directions involve architectural innovations in classifier head. RepMet [11] addresses category confusion by integrating prototype learning into the classifier head. Efficient-FSOD [12] proposes a knowledge-inheritance classifier initialization strategy to allow rapid adaptation during fine-tuning.

This work was supported in part by the National Natural Science Foundation of China under Grant 62471394 and U21B2041. (Corresponding author: Qi Wang.)

Chenchen Sun, Yuyu Jia, Han Han, Qiang Li, and Qi Wang are with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, P.R. China. (e-mail:scc1234@mail.nwpu.edu.cn; jyy2019@mail.nwpu.edu.cn; hanhan@mail.nwpu.edu.cn; liqmges@gmail.com; crabwq@gmail.com)

Although these methods have shown considerable promise, challenges remain [13]. Considering the significant morphological differences between remote sensing and natural images, this leads to a disruption of the general structural prior knowledge learned by the RPN. In scenarios with scarce samples, directly fine-tuning the RPN can lead to overfitting to the small sample data, resulting in poor generalization performance of the RPN. In addition, remote sensing images are susceptible to variations in imaging conditions, leading to significant intra-class differences in the appearance of similar objects. This poses a challenge in FSOD tasks, as the model struggles to learn representative category features due to the lack of sufficient samples. These sparse features may not adequately represent the semantic characteristics of the category and the potential variations in appearance. Fig.1 directly visualizes these two challenges.

In recent years, universal visual models [14], [15], [16] show impressive zero shot generalization in image processing. Meanwhile, such models create rich visual-semantic connections by integrating visual patterns with linguistic descriptions [17], [18]. These novel componentry is effective for tasks like object description and functional reasoning [19], [20]. Thus, we investigate the generalization capability of the universal vision model and whether the multimodal collaboration mechanism can be effectively applied to FSOD tasks [21], [22].

Building upon the advancements of the above examination, we propose ST-FSOD and suggest incorporating textual modalities into the task to complement sparse visual features. ST-FSOD consists of two components: SA-RPN and the text guiding learner(TGL). SA-RPN aims to obtain high-quality proposals under few-shot fine-tuning conditions. Inspired by SAM's powerful zero-shot generalization ability and flexible prompting mechanism [14], the low-quality proposals are used as prompts to generate additional useful proposals in the feature space, enhancing the generalization performance of the RPN. TGL allows the learning of rich semantic features to complement sparse visual features. It generates textual representations from class names and creates image independent, text-guided prototypes via domain adaptation and relationship decoupling. By aligning these text prototypes with image features in the feature space, TGL can fully exploit the text modality to complement sparse visual features in images.

By combining these two modules, ST-FSOD not only enhances the RPN's generalization ability but also introduces the textual modality to boost the classification quality. Through comprehensive experiments on the FSOD benchmarks i.e., DIOR [4], NWPU VHR-10 [5] and RSOD [23], ST-FSOD achieves precise detection of base classes while significantly improving adaptability to novel classes. Our contributions are summarized as follows:

1) : We propose a new proposal generator (SA-RPN) that uses low quality proposals as prompts to produce high quality regions, thus improving the generalization performance of the RPN structure.

2) : We propose a branch of learning for semantic feature embedding (TGL), which enhances classifier performance by embedding semantic knowledge in visual representations through the establishment of semantic prototypes for specific categories.

3) : Through comprehensive quantitative and qualitative analysis, we demonstrate the effectiveness and superior performance of ST-FSOD on the FSOD benchmark.

II. RELATED WORKS

A. Generic Object Detection

Currently, the advent of deep neural networks leads to extensive research into CNN-based object detection methods. Currently, the methods can be categorized as anchor-based [1], [24], [25] or anchor-free [2], [26], [27], depending on the use of prior boxes. Anchor-based methods rely on prior boxes with varying sizes and aspect ratios. Faster R-CNN uses a Region Proposal Network (RPN) to generate region proposals, perform localization, and classify objects. To enhance multi-scale perception, various feature pyramid structures [28], [29] and linear interpolation algorithms [30] emerge. Addressing issues of prior box generalization, anchor-free methods aim to directly determine target categories and anchor boxes without prior box reliance.

Despite natural image target detection success, applying methods directly to remote sensing images poses difficulties [4] due to their densely distributed and arbitrarily oriented nature. Advanced methods focus on rotation detection to predict target position and angle simultaneously. SCRDet [31] enhances small cluttered object detection by suppressing noise and highlighting target features. FSDet [32] introduces an anchorless frame detector with directional feature refinement and context aggregation to solve feature misalignment issues. LSK [33] adjusts its spatial receptive field dynamically to model varying object contexts in remote sensing scenarios. However, these methods are limited by requiring large-scale data training, hindering adaptation to new scenarios.

B. Few Shot Learning

Few-shot learning attempts to identify new classes with limited examples. Recent research reveals two typical directions: meta-learning methods [34], [35] for learning meta-knowledge from prior tasks, and metric learning methods [36], [37] for computing sample similarity based on distance functions. Moreover, recent studies also emphasize the potential of text modalities in addressing few-shot learning challenges [22], [38]. By utilizing text features as semantic cues, the feature extraction process of the image encoder can be dynamically adjusted, effectively enhancing the discriminative nature of features.

Few-shot learning is also widely applied in the field of remote sensing image processing. For instance, HSL-MINet [39] improves the discriminative ability of model decision boundaries through hard sample learning and multi-view integration methods. MFGNet [40] tackles overfitting issues caused by insufficient samples in the representation space by employing online sample generation. Moreover, MMML [41] introduces a multi-manifold metric learning framework for reducing the impact of intra-class variance and inter-class similarity. While these methods have shown considerable

success in classification tasks, they do not possess the capability to detect and classify remote sensing targets in few-shot scenarios.

C. Few Shot Object Detection

Few-shot object detection (FSOD) accurately detects novel class objects by introducing only a limited number of annotated samples of new categories. As a pioneering work in this field, TFA [42] significantly improves FSOD performance by fine-tuning only the Faster R-CNN classifier and establishes Faster R-CNN as the foundational architecture for FSOD benchmarks. Based on this baseline, remarkable progress emerges. Sun et al. [43] propose the CPE loss function to learn more discriminative feature embeddings, while Qiao et al. [44] design gradient-decoupled layers and prototype calibration modules to mitigate inherent conflicts in traditional classifiers under data-scarce scenarios. In recent studies, state-of-the-art approaches [45] incorporate large language models (LLMs) for context-aware few-shot learning, leveraging LLMs to classify region proposals within contextual understanding.

In the field of FSOD for remote sensing images, remarkable achievements continue to emerge [46], [47], [48], [49], [50], [51]. Widely adopted datasets such as DIOR [5] and NWPU VHR-10 [4] facilitate progress in this domain. Among existing approaches, P-CNN [52] establishes a strong baseline for remote sensing FSOD by constructing a prototype learning network and a prototype-guided region generation network. FSODM [3] enhances the few-shot generalization capability of detectors through a meta-feature extractor and a feature reweighting mechanism. MSOCL [53] significantly improves multi-scale object detection performance by integrating multi-scale contrastive learning. MM-RCNN [46] employs a memory module to reuse historical knowledge and utilizes a cross-category attention mechanism to strengthen inter-class relationships.In contrast, our study focuses on fully exploiting prior information in images and using fine-tuning to achieve semantic knowledge transfer.

III. METHODOLOGY

In this paper, we begin by outlining the FSOD task setup (Section III-A), followed by a detailed overview of the method (Section III-B). Then we improve FSOD framework from two perspectives: RPN and the classifier. For RPN, we introduce a novel proposal generator for high-quality proposals (Section III-C). For the classifier, we embed semantic knowledge into visual representation by creating semantic prototypes for specific categories (Section III-D).

A. Preliminaries

1) Problem Definition: The FSOD task targets to generalize the model's detection capacity from the training set D_{train} to the novel class set D_{novel}, regarding two disjoint sets of categories C_{train} and C_{novel} [7]. As per convention, we follow the two-stage paradigm. Initially, D_{train} is defined as $D_{train} = \{(x_i, y_i)\}_{i=1}^{I}$. Here x_i denotes the i-th image in the image dataset, and y_i denotes the corresponding annotations. Analogously, D_{novel} is defined as $D_{novel} = \{(x_j, y_j)\}_{j=1}^{J}$. Typically, a novel type of dataset contains N categories, with each category having K samples, referred to as N-way K-shot detection. Our goal is to pretrain the model with D_{train} and then fine-tune it with D_{novel} in order to adapt to new categories.

2) Baseline Model: ST-FSOD employs Faster RCNN with the pre-training and fine-tuning paradigm [42] as the basic framework. In the pretrain phase, the model undergoes training on D_{train}, while in the finetune phase, all categories from both D_{train} and D_{novel}, which has limited annotations, are considered. The detection process involves extracting feature maps from an image using the backbone network. These feature maps are then fed into the RPN to generate region proposals, which are refined through classification and regression adjustments. The candidate proposals are then processed by the Region of Interest (RoI) layer. The final output is produced by the RoI head for classification and regression.

In contrast to conventional Faster RCNN architecture, a Gradient Decoupling Layer (GDL) based on DeFRCN [44] is integrated between the model's backbone and the two branches of the RoI Head and RPN. The GDL is specifically defined as follows:

$$\mathcal{G}_{(\mathcal{F},\mu)}(x) = \mathcal{F}(x), \tag{1}$$

$$\frac{\mathrm{d}\mathcal{G}_{(\mathcal{F},\mu)}}{\mathrm{d}x} = \mu\nabla_{\mathcal{F}}, \tag{2}$$

where $\mathcal{F}$ is an affine transformation layer, $\mu \in [0, 1]$ is a decoupling coefficient, and $\nabla_{\mathcal{F}}$ is the Jacobian matrix from the affine layer.

B. Method Overview

The proposed ST-FSOD architecture, as illustrated in Fig. 2, extends the Faster R-CNN by integrating the SA-RPN and TGL modules. The dual-module framework systematically addresses two critical challenges in FSOD: enhancing the generalization of RPN structure and augmenting sparse visual features through multimodal fusion. The training process adheres to the standard FSOD protocol [42]. First of all, input images are processed through the backbone network to generate feature maps, which are subsequently passed to the original RPN to produce initial region proposals. These proposals are then refined by the SA-RPN. The enhanced proposals undergo RoI pooling to extract features, which are subsequently combined with text embeddings within the TGL module. The TGL architecture processes image and text data in parallel, fusing multimodal representations, which are ultimately fed into the detection head for classification and regression tasks.

C. SA-RPN module

The quality of proposals significantly impacts the performance of few-shot detectors. Conventional approaches generally enhance the localization capability of RPNs in complex backgrounds through either region proposal optimization strategies or feature-guided enhancement mechanisms. However, due to limited data, RPN structures are prone to overfitting, especially during fine-tuning stage with few samples. To address this issue, we introduce the SA-RPN module in

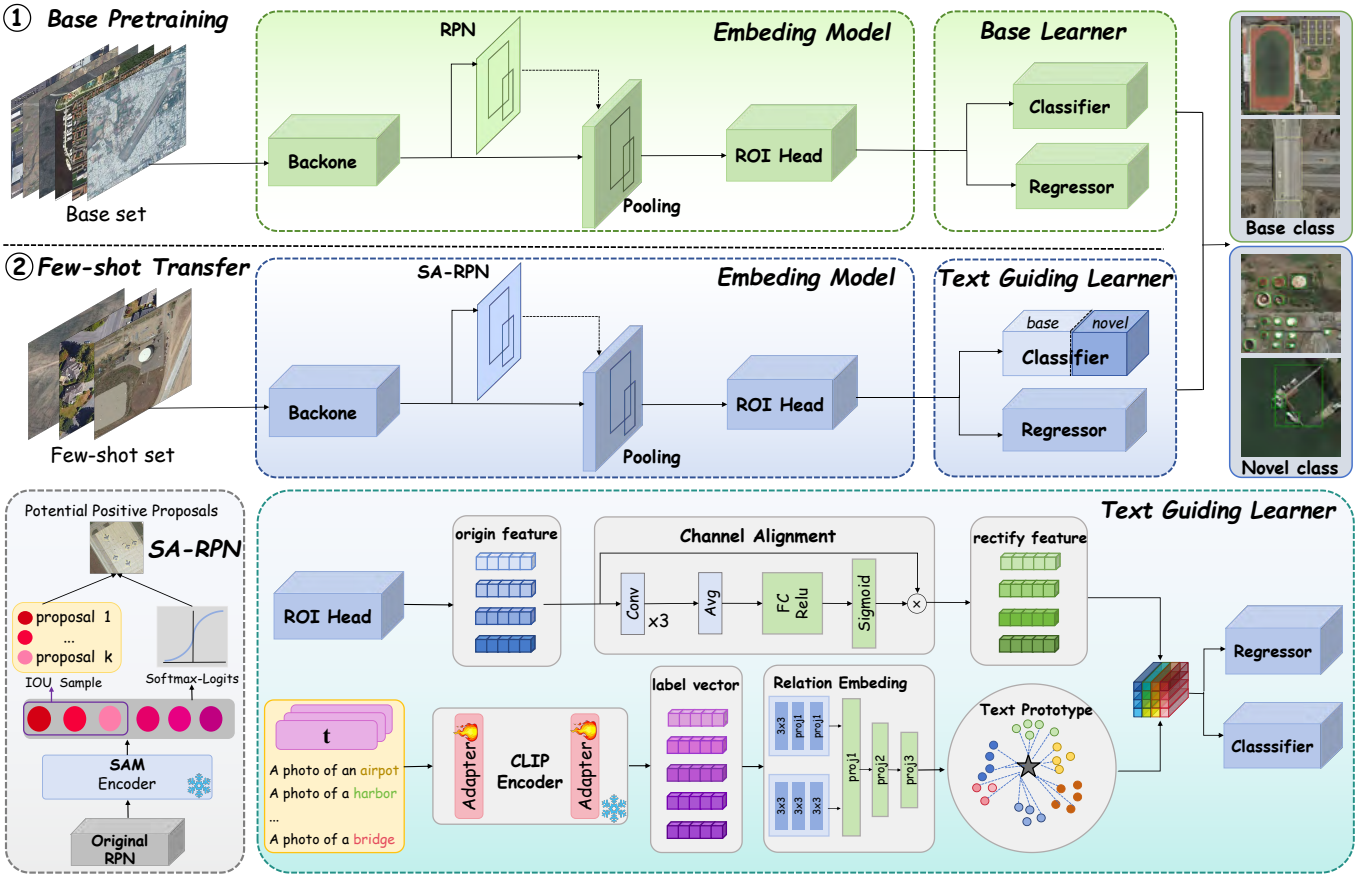

Fig. 2. Our FSOD framework is as follows: We first partition the dataset into D_{base} and D_{novel}. In the base pretraining stage, we train a standard Faster R-CNN, including an embedding model and base learner, solely on D_{base}. After pretraining, we fine-tune the detector using D_{novel} to improve accuracy for both base and novel categories. To generate high-quality proposal boxes and enhance the generalization of the RPN, we build SA-RPN based on the original RPN, generating high-quality proposal boxes with potential foreground for the model. Additionally, to enhance the representational capacity of features, we propose a text guiding learner that integrates image-independent text-guided prototypes with image features, thereby improving the representational ability of class features.

ST-FSOD. This module leverages SAM's powerful reasoning capabilities to generate more additional proposals from the original ones. Our goal is to enhance the generalization ability of RPN by improving the proposals.

Giving the image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, a backbone network extracts high-level features $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$. These features are then delivered to the original RPN to generate a set of initial proposals $\mathcal{P} = \{(x_i, y_i, w, h, s)\}$, where the quaternion (x_i, y_i, w, h) represents the coordinates of the proposals and s denotes the confidence score. We select the top n proposals with the highest confidence scores to form the candidate set $\widehat{\mathcal{P}}$, which is then input into the SAM [14] to generate segmentation masks $\widehat{\mathcal{M}}$. Finally, by calculating the minimum bounding rectangles of the connected components within the mask, a set of proposal $\widehat{\mathcal{P}_f}$ is obtained for the subsequent computations.

$$\widehat{\mathcal{M}}(x_i, y_i, S_\alpha) = SAM(\widehat{\mathcal{P}}), \tag{3}$$

$$\widehat{\mathcal{P}_f} = MinAreaRect(\widehat{\mathcal{M}}(x_i, y_i, S_\alpha)), \tag{4}$$

where $\widehat{\mathcal{M}}(x_i, y_i, S_\alpha)$ denotes the generated semantic mask information and confidence score. $\widehat{\mathcal{P}_f}$ represents the extracted minimum bounding rectangles. However, the generated proposals $\widehat{\mathcal{P}_f}$ face two issues: potential overlaps with the initial proposal $\mathcal{P}$ and the need for a confidence score.

First, we design a proposal selection strategy to filter out redundant proposals, thereby addressing potential overlap issues. Traditional detection methods utilize the Intersection over Union (IoU) to measure the spatial similarity between two proposal. However, this generic metric is highly sensitive to targets of extreme sizes. Inspired by [54] on mining potential proposal settings, we propose a novel approach that combines IoU with the size of the initial proposals to calculate the similarity between the initial proposals and the generated proposals. The similarity formula U_β between $\mathcal{P}$ and $\widehat{\mathcal{P}_f}$ is given as follows:

$$U_\beta = Max\left(\text{IoU}, \ln\frac{\sqrt{w_g \cdot h_g}}{10} + 0.5\cdot\text{IoU}\right) \leq \lambda, \tag{5}$$

where w_g and h_g indicate the width and height of the proposal in $\mathcal{P}$, λ is the threshold factor, and the term 10 actually corresponds to the minimal area definition of DIOR dataset [4], which allows for appropriate threshold settings for objects of different sizes and can be tuned for different datasets. Any proposal in $\widehat{\mathcal{P}_f}$ point with a similarity greater than λ is

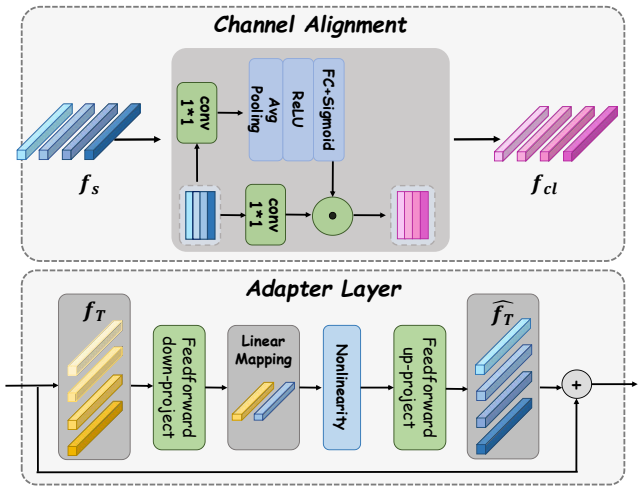

Fig. 3. Structure of the Channel Alignment and the Adapter Layer.

considered too similar to the original proposal and is filtered out.

Next, we employ a strategy based on softmax-logistic normalization to assign a confidence score to each proposal in $\widehat{\mathcal{P}_f}$. Specifically, following the original setting of Faster R-CNN [55], the confidence score S_α generated by SAM is processed through softmax-logistic normalization to produce the confidence score:

$$SCORE = \sigma\left(S_\alpha \cdot \ln(\frac{1-\varepsilon}{1-(1-\varepsilon)}\right), \tag{6}$$

where σ represents the sigmoid function and ε is a constant that approaches 0, which we adopt the Faster R-CNN settings to take as 10^{-6} in our experiment.

Subsequently, all proposals in $\widehat{\mathcal{P}_f}$ and $\mathcal{P}$ are passed as input to the ROI head for subsequent feature processing.

D. TGL module

To enhance detection accuracy, it is crucial to acquire comprehensive visual semantic information from images. Current methods extract sparse visual features from limited images and refine them via prototype learning or contrastive learning, but struggle with prototype bias and semantic ambiguity in sample-scarce few-shot setting. Our goal is to enhance sparse visual features using image-independent category textual representations. Therefore, we propose a dual-branch text-guided learning framework with distinct image and text branches, forming the foundation of our approach.

1) Image Path: As shown in Fig.2, the proposal output by the SA-RPN is mapped to the corresponding region of the image feature map in ROI head. Then, pooling and bilinear interpolation methods convert each region into a fixed-size feature map $f_s \in \mathbb{R}^{C\times H\times W}$.

Inspired by the concept of channel importance [56] and SENet [57], we employ fully connected layers along with 1×1 convolutions to adjust the channel dimension of feature mapsf_s to match the number of target categories. This design captures the channel-category correlations, ultimately generating channel-aligned feature maps f_{cl}. As illustrated in Fig.3:

$$H_{map} = \delta\left(W_\gamma^N(\text{Relu}\left(\text{Avgpool}\left(\text{Conv}\left(f_s\right)\right)\right))\right), \tag{7}$$

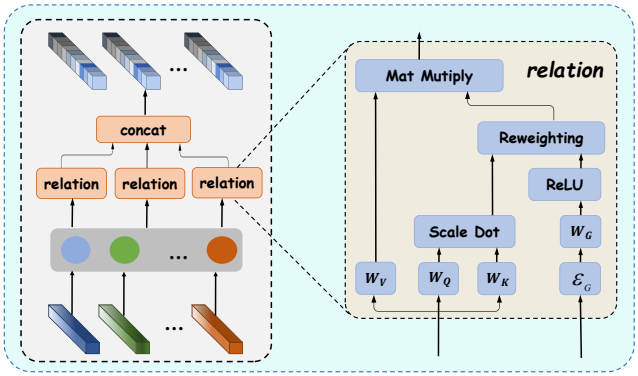

Fig. 4. Structure of the Relation Network.

$$f_{cl} = \text{Conv}(f_s) \otimes H_{\text{map}}, \tag{8}$$

where W_γ^N represents the fully connected layers, Avgpool$(\cdot)$ represents global pooling operation, Conv$(\cdot)$ represents convolution operation and δ represents Sigmoid operation. Through channel reshaping, each channel of the generated feature map f_{cl} encodes information pertaining to a specific category.

2) Text Path: To construct a stable text prototype, we introduce the category labels as additional prior information. For an input phrase T containing n category labels, we utilize the pre-trained CLIP model [17] to embed the feature of each category into an embedding vector z_i. Specifically, we obtain the class name cls corresponding to each category and combine it with the prompt template "a photo of a __". Then, put it into the CLIP encoder to generate a 512-dimensional embedding vector. The detailed generation process is as follows:

$$z_i = \text{“a photo of a \_ ”} + cls, \tag{9}$$

$$T = [z_1, z_2, \ldots, z_n] \quad n = 1.\ldots.C, \tag{10}$$

$$f_T = Encoder(T) \in \mathbb{R}^{512\times1\times1}, \tag{11}$$

where z_i represents the text embedding of each category, T denotes the text vector comprising all categories, and f_T signifies the 512-dimensional feature vector of the input text prompt encoded by the CLIP encoder. These feature vectors are semantically richer, derived from a deep understanding of the text, rather than being simple bag-of-words models or vectors based on statistical methods.

In view of the text encoder are predominantly trained on natural image-text pairs, there may be inherent bias when applied to remote sensing imagery. To address this issue, we propose a lightweight adapter layer [58]. As illustrated in Fig. 3, the adapter consists of a down-projection matrix, a nonlinear activation function, projection operations, and residual connections. The corrected text vector $\widehat{f_T}$, is obtained through the following procedure:

$$\widehat{f_T} = f_T + \sigma\left(W_2 \cdot \phi\left(W_1 \cdot f_T + b_1\right) + b_2\right), \tag{12}$$

where W_1 and W_2 represent the projection matrices in the feedforward neural network, and b_1 and b_2 are the corresponding bias terms. The adapter efficiently mitigates domain bias in feature vectors through the bottleneck layer projection.

We employ a relational network to model the relationships between each class and enhance knowledge transfer among them, as illustrated in Fig. 4. The text vector $\widehat{f_T}$ is transformed into different weight vectors through a weight matrix. Subsequently, calculate the normalized attention weights between different weight vectors. Finally, the attention weights are weighted to the weight vectors to obtain the final text feature $\hat{v}_q$. The relational modeling can be described as:

$$\hat{v}_q = \sigma\left(\frac{\widehat{f_T}\cdot W_\xi^Q \cdot (\widehat{f_T}\cdot W_\xi^K)^\top}{\sqrt{d_K}}\right)\cdot(\widehat{f_T}\cdot W_\xi^V), \tag{13}$$

where W_ξ^Q, W_ξ^K, and W_ξ^V respectively represent learnable linear layers, d_K is a normalization parameter that normalizes the decoupled feature vectors based on the dimension of the linear encoding.

3) Feature Fusion: Finally, we fuse the obtained text prototype with image features. To be more specific, we employ a metric learning approach. The channel alignment feature map $f\mathrm{cl}$ is used as the query feature, and cosine similarity is computed with $\boldsymbol{n}$ text class prototypes $\widehat{v_q}$ to generate class activation maps $B^{(x,y)}$. Subsequently, $B^{(x,y)}$ is concatenated with f_s along the channel dimension and then passed through a convolution to obtain the final output f_{guide}. The detailed calculation process is as follows:

$$B^{(x,y)} = \frac{\hat{v}_q^{\mathrm{T}}\cdot f_{\mathrm{cl}}^{(x,y)}}{\|\hat{v}_q\|\cdot\|f_{\mathrm{cl}}^{(x,y)}\|}, \tag{14}$$

$$f_{\text{guide}} = \mathrm{Conv}(\mathrm{Concat}(f_{\mathrm{s}}, B^{(x,y)}))\oplus f_{\mathrm{s}}\ . \tag{15}$$

The fusion method above effectively integrates semantically aware text prototypes with visual features. By leveraging the correlation between visual and textual attributes, it directs the model's attention to regions pertinent to the designated categories, thereby enhancing the sparse visual representation. Ultimately, the text-guided feature f_{guide} is utilized in the detection head to generate the detection outcomes.

IV. Experiments

In this section, we evaluate the benefits of our proposed method. Firstly, we describe three common benchmarks and provide implementation details. Next, we compare it with the leading current technologies. In addition, we conduct ablation experiments to explore the components that impact the overall performance of the module. Finally, we analyze the experimental results through visualization techniques.

A. Dataset

Our method is evaluated on three widely used RS-FSOD benchmarks: DIOR [4], NWPU VHR-10 [5], and RSOD [23]. We follow the dataset category partitioning approach consistent with [52] and [59], independently dividing the different categories in the dataset into base classes and novel classes.

B. Implementation Details

We use the DeFRCN framework [7] to construct our model and perform experiments using a single GTX 1080ti GPU to accelerate both training and testing processes. Following the original TFA setup, we adopt a ResNet101 [60] model pretrained on ImageNet [61] as the foundational structure of our model. During training, we employ the Adam optimizer with a base learning rate set to 9.5×10^{-4}, a momentum coefficient of 0.88, and a weight decay of 9.5×10^{-6}. In the initial stage of model pre-training, we perform 15,000 iterative training sessions. Subsequently, for the fine-tuning stage, We modify the number of training iterations according to the specific requirements of the task.

C. Experimental Results and Comparisons

To validate the effectiveness of our proposed method, we conducted a thorough comparison of various SOTA FSOD methods, namely FsDetView [8], Meta-RCNN [62], TFA [7], P-CNN [52], DeFRCN [44], VFA [63], ICPE [64], SAE-FSDet [9], MM-RCNN [46], iMTFA [65], Incremental-DETR [66], FRCN-FAHM [67], FRCN-SAAN [59] and G-FSDet [49]. The results are shown on Table I ,Table II and Table III.

1) Experimental results in DIOR: Table I presents the results of the K-shot experiments (where K = 3, 5, 10, 20) carried out on the DIOR dataset. The findings clearly indicate that our approach significantly outperforms other methods. Specifically, in the 5-shot setting, Our method achieves increases in mAP of 9.8%, 9.6%, 9.49%, and 10.84% compared to P-CNN under the K-shot setting. Moreover, our method achieves comparable or superior performance to SAE-FSDet in 8 out of 16 experimental settings, while demonstrating competitive results in the remaining cases. For instance, in Split2, we achieve improvements of 0.62%, 1.29%, 2.52%, and 1.69% under the K-shot setting. Additionally, our model's performance significantly surpasses that of our baseline method, DeFRCN, across all data splits, particularly in split 4, where we achieved improvements of 6.51%, 3.08%, 4.84%, and 4.19% under the K-shot setting. Across most data splits, our model consistently demonstrated the top or second-best performance, showcasing pronounced advantages in situations with very limited sample sizes, such as the 3-shot and 5-shot trials. Furthermore, to comprehensively assess the performance of individual classes, detailed results are also provided. It is evident that our proposed methodology excels across almost all class divisions.

In the experiment, we identify several issues, particularly the uneven improvement of the model across different data set splits. For example, when considering split1 and split3, we observe that varying allocations of novel classes can present certain challenges. Larger targets such as basketball courts and sports fields tend to occupy a significant portion of the images, thus making them relatively easier to detect. On the other hand, smaller targets like bridges and chimneys can pose detection challenges due to their lower spatial resolution. Moreover, the simple shapes and structures of targets like bridges and basketball courts make them relatively easier to detect, whereas the complex structures of targets such as boats and airplanes, combined with factors like different

TABLE I
COMPARISON WITH SOTA ON DIOR DATASET IN MAP UNDER THE FEW SHOT SETTING. RED/BLUE REPRESENTS THE 1ST/2ND BEST PERFORMANCE.

Split	Shot	Meta-RCNN	FsDetView	P-CNN	TFA w/cos	DeFRCN	VFA	ICPE	G-FSDet	MM-RCNN	SAE-FSDet	Ours
Split 1	3-shot	12.02	13.19	18.00	11.35	28.25	21.94	11.68	27.57	19.80	28.80	27.80
	5-shot	13.90	14.29	22.80	11.57	30.30	21.27	12.34	30.52	23.90	32.40	31.23
	10-shot	14.07	18.02	27.60	15.37	32.63	23.32	12.95	37.46	28.80	37.09	37.59
	20-shot	14.45	18.01	29.60	17.96	35.37	24.28	14.33	39.83	30.80	42.46	39.26
Split 2	3-shot	8.44	10.83	14.50	5.77	14.55	12.10	10.92	14.13	15.60	13.99	14.61
	5-shot	10.88	9.63	14.90	8.19	16.45	12.70	10.56	15.84	15.50	15.65	16.94
	10-shot	14.90	13.57	18.90	8.71	14.72	18.40	12.39	20.70	20.10	17.41	19.93
	20-shot	16.71	14.76	22.80	12.18	21.13	15.57	13.18	22.69	23.90	21.34	23.03
Split 3	3-shot	9.10	7.49	16.50	8.36	15.78	11.97	10.56	16.03	16.70	16.74	17.77
	5-shot	12.29	12.61	18.80	10.13	18.73	13.19	11.21	23.25	19.70	19.07	24.27
	10-shot	11.96	11.49	23.30	10.75	20.43	15.45	12.38	26.24	25.00	28.44	27.07
	20-shot	16.14	17.02	28.80	17.99	25.13	17.61	13.08	32.05	30.05	29.88	33.49
Split 4	3-shot	13.94	14.28	15.20	10.42	10.83	15.52	14.45	16.74	16.40	17.27	17.34
	5-shot	15.84	15.95	17.50	14.29	18.62	17.76	14.52	21.03	18.70	20.48	21.70
	10-shot	15.07	15.37	18.90	14.35	21.61	18.62	15.95	25.84	20.30	22.69	26.45
	20-shot	18.17	16.96	25.70	12.01	27.61	20.05	15.61	31.78	27.10	26.75	31.80

TABLE II
COMPARISON WITH SOTA ON NWPU VHR-10 DATASET IN MAP UNDER THE FEW SHOT SETTING. RED/BLUE REPRESENTS THE 1ST/2ND BEST PERFORMANCE.

Split	Shot	Meta-RCNN	FsDetView	P-CNN	TFA w/cos	DeFRCN	iMTFA	Incremental-DETR	G-FSDet	Ours
Split 1	3-shot	20.51	24.56	41.80	8.80	37.90	43.20	50.93	49.05	51.59
	5-shot	21.77	29.55	49.17	9.49	46.08	49.95	53.18	56.10	58.49
	10-shot	26.98	31.77	63.29	9.26	62.95	70.01	69.94	71.82	72.75
	20-shot	28.24	32.73	66.83	10.83	64.61	72.78	73.18	75.41	77.01
Split 2	3-shot	21.41	39.01	39.32	11.14	39.19	46.29	47.35	50.09	54.19
	5-shot	35.34	40.31	46.10	12.46	45.56	54.34	57.27	58.75	60.78
	10-shot	37.14	45.09	55.90	11.35	54.05	62.56	65.63	67.00	66.37
	20-shot	39.47	46.28	58.37	11.56	57.38	66.37	70.58	75.86	76.92

TABLE III
COMPARISON WITH SOTA ON RSOD DATASET IN MAP UNDER THE FEW SHOT SETTING. RED/BLUE REPRESENTS THE 1ST/2ND BEST PERFORMANCE.

Split	Shot	FRCN-FAHM	Meta R-CNN	FRCN-SAAN	Ours
Split 1	3-shot	15.00	18.40	23.20	25.78
	5-shot	20.20	35.03	34.38	39.42
	10-shot	43.50	45.13	45.09	48.67
Split 2	3-shot	51.60	58.45	53.43	61.58
	5-shot	60.92	60.14	67.02	69.45
	10-shot	71.27	81.69	81.42	85.13
Split 3	3-shot	21.01	11.36	10.42	17.61
	5-shot	41.40	32.63	34.09	37.68
	10-shot	59.63	42.66	44.13	50.72
Split 4	3-shot	61.60	62.70	77.72	78.66
	5-shot	77.10	81.33	90.11	91.77
	10-shot	88.98	88.11	96.15	96.83

angles and occlusions, can increase detection difficulty. Upon a thorough examination of such samples, we recognize that our method may not effectively demonstrate its advantages when processing low resolution microscopic spatial elements or in scenarios involving complex nested categories. A detailed discussion of limitations will be presented in Section IV-D.

2) Experimental results in NWPU VHR-10: In the NWPU VHR-10 dataset, the number of images or objects in each category is relatively small. Although this makes the dataset relatively small for the FSOD task, training deep learning models may face more severe overfitting issues compared to the DIOR dataset, posing a challenge for the FSOD task. However, experimental results shown in Table II demonstrate that despite these challenges, our model exhibits excellent generalization performance on the NWPU VHR-10 dataset. Specifically, under the Spilt 1 with K-shot (K = 3, 5, 10, 20) setting, our method has achieved remarkable competitiveness. For example, in the 3-shot scenario, compared to G-FSDet, our method achieved a significant 2.54% improvement in mAP; under the 5-shot setting, our method achieved a 2.39% performance improvement. In the 10-shot and 20-shot settings, our results surpassed P-CNN, with an average mAP increase of 9.71% and 9.68%, respectively. At the same time, compared to our baseline method DeFRCN, we achieve performance improvements of 13.69%, 12.41%, 9.8%, and 12.4% under the K-shot setting. These results clearly demonstrate the superior generalization performance of our model when facing the challenges of relatively small dataset size and limited object quantity.

3) Experimental results in RSOD: As a complement to the DIOR and NWPU VHR-10 datasets, the RSOD dataset presents significant challenges for FSOD tasks, characterized by extreme variations in object scales and complex geometric deformations caused by diverse aerial viewing angles in specific categories. Nevertheless, our method consistently demonstrates outstanding performance under K-shot settings across four novel classes, as shown in Table III. Compared with our main competitor FRCN-SAAN, our approach achieves

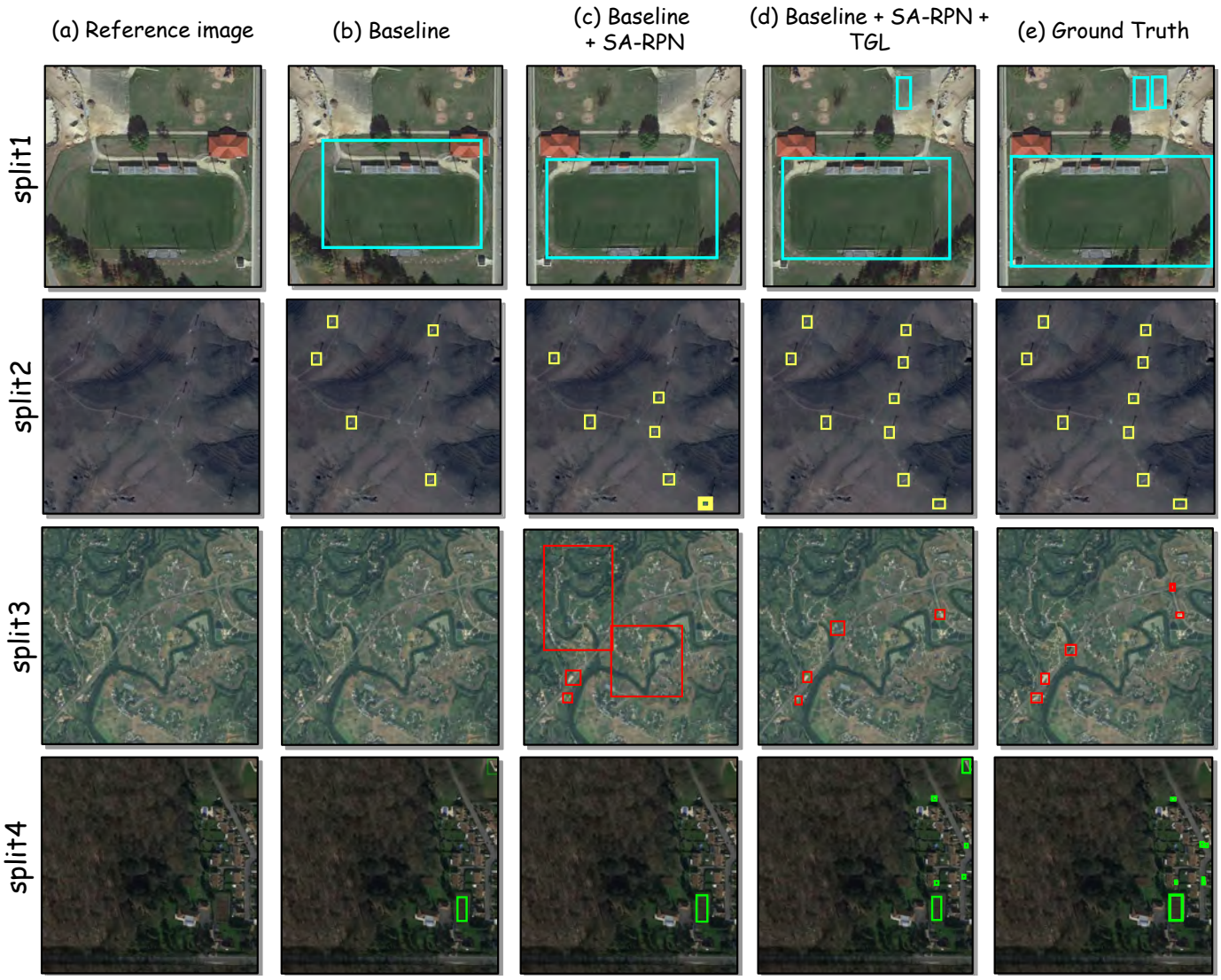

Fig. 5. Formalization of detection results. The first to fourth lines respectively provide examples of different data set partitions. The images in the same column correspond to a certain object category, where there are new categories and base categories. Our method achieved the best detection performance in each dataset split.

accuracy improvements of 8.15%, 2.43%, and 1.71% for the "oil tank" category under 3-shot, 5-shot, and 10-shot configurations, respectively. For the "playground" category, accuracy enhancements of 0.94%, 1.66%, and 0.68% are observed in these three settings. Additionally, our method outperforms Meta R-CNN by 7.38%, 4.4%, and 3.54% in the "playground" category under the corresponding configurations. Notably, even in the challenging "overpass" category, our approach maintains competitive performance, achieving suboptimal yet robust results compared to SOTA methods.

D. Ablation Study and Visualization

In this section, we conduct ablation studies and comprehensive discussions to illustrate the significance of SA-RPN and Text Guiding Learner. Additionally, we determine the optimal settings for our method through these discussions. All experiments in this section are carried out on the DIOR test set.

1) Survey of component design: We conduct ablation experiments to validate the effectiveness of two modules. As shown in Table IV, both our SA-RPN and TGL module consistently improve performance. The incorporation of text-guided branches proves to be advantageous for thoroughly exploring the potential semantic features of images with limited prior knowledge. In conclusion, the combination of the TGL module and SA-RPN produces the most favorable results. The high-quality proposals generated by SA-RPN provide precise guidance for the detection, while the TGL module enhances the representational capability of sparse visual features.

TABLE IV
ABLATION STUDIES OF MAJOR MODEL COMPONENTS USING MAP ON DIOR DATASET.

	SA-RPN	TGL	Baseline	3-shot	5-shot
I			✓	19.43	22.47
II	✓		✓	23.64	27.84
III		✓	✓	23.37	28.71
IV	✓	✓	✓	27.83	31.25

Fig. 6. Visualization of RSOD dataset detection results.

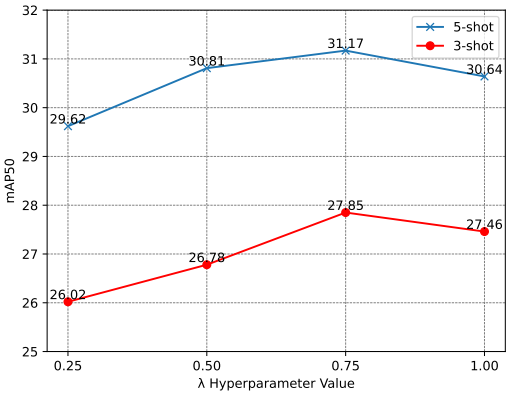

Fig. 7. Impact of the parameter λ for the FSOD mAP.

Fig. 8. Visualization of RPN proposal results.

2) Effect of the of the hyperparameter λ on the performance of SA-RPN: As shown in Fig.7, we conduct an analysis to investigate the impact of the threshold hyperparameter λ. Specifically, the experiments are conducted on the 3-shot and 5-shot settings of the DIOR dataset Split1. We vary λ in the range [0.25, 1] and observe that as λ increases, the network performance improves due to the gradually strengthened supervision. In this data set and setting, SA-RPN achieves its maximum performance when λ is set to 0.75. However, further increasing λ would result in excessive supervision, leading to performance degradation of the SA-RPN structure. This could limit its ability to identify potential proposals, ultimately decreasing overall network performance.

3) Visualization of RPN experiment results: As shown in Fig.8, through comparative visualization experiments of RPN

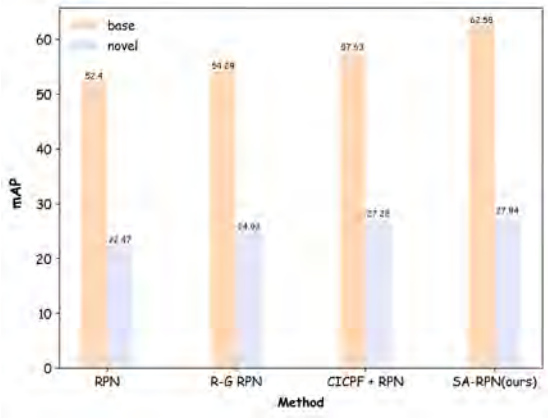

Fig. 9. The effectiveness of the SA-RPN architecture is demonstrated through a bar chart comparing the 5-shot mAP performance with alternative RPN structures.

and SA-RPN in multiple scenarios, it is observed that SA-RPN detects a greater number of potential targets compared to the original RPN, significantly alleviating the issue of RPN's structural neglect of potential targets under few annotation conditions. This validates that our proposed SA-RPN significantly improves the perception of potential targets in complex scenarios through the introduction of SAM's prior knowledge to generate high-quality additional proposals in the feature space, thereby providing higher-quality candidate regions for subsequent detection stages.

4) Comparison with other RPN structures: To demonstrate the superiority of the proposed SA-RPN, we compare its performance with various state-of-the-art RPN architectures. Fig.9 presents the results obtained in the 5-shot setting using RPN [1], R-G RPN [52], CICPF+RPN [48], and SA-RPN. In particular, the SA-RPN consistently achieves superior detection performance by introducing additional high-quality proposals in feature space. It demonstrates exceptional performance on data-rich base classes while simultaneously attaining measurable improvements on data-scarce novel classes, which benefits from the enhanced generalization capability of SA-RPN.

5) Visual validation of TGL module's impact on classification performance: We conduct 5-shot and 10-shot detection experiments on the DIOR dataset and visualize feature distributions using t-SNE [68]. For each novel class, 50 ROI features are selected for analysis. The first and second rows in Fig.10 illustrate the feature distributions of training data without and with the TGL module, respectively. As demonstrated in Fig.10, significant feature overlap among categories is observed when the TGL module is disabled. Although increased support samples alleviate this issue to some extent, intra-class features remain scattered. In contrast, with the TGL module enabled, the intra-class feature distributions become more compact as the number of support samples increases, while inter-class decision boundaries are notably better delineated. These results validate the effectiveness of the TGL module in enhancing feature discriminability.

TABLE V
THE IMPACT OF DIFFERENT IMPLEMENTATIONS OF FEATURE FUSION MECHANISMS IN TGL FOR THE DIOR DATASET.

Method	3-shot	5-shot	10-shot	20-shot
Dot product	21.48	26.45	34.38	37.48
Addition	22.76	27.21	35.13	37.62
Concatenation	22.56	27.95	35.06	37.65
Concatenation + CA	22.94	28.24	35.21	37.98
TGL (Ours)	**23.37**	**28.71**	**35.37**	**38.24**

6) Impact of feature fusion mechanisms on model performance: We systematically investigate the impact of various mechanism of fusion of features on the performance of the model, including element multiplication, element addition, channel concatenation, and channel concatenation combined with a channel attention mechanism. As shown in TableV, both element-wise addition and multiplication operations may induce feature conflicts between modalities and introduce extraneous noise compared to our approach. The integration of the channel attention mechanism effectively mitigates the

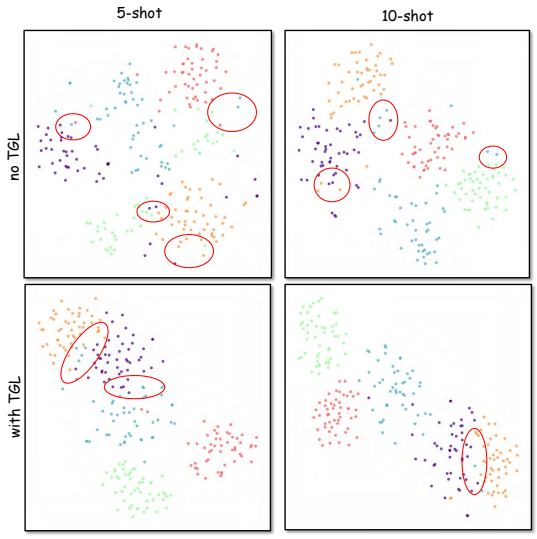

Fig. 10. The t-SNE visualizations of feature embeddings from the DIOR dataset.

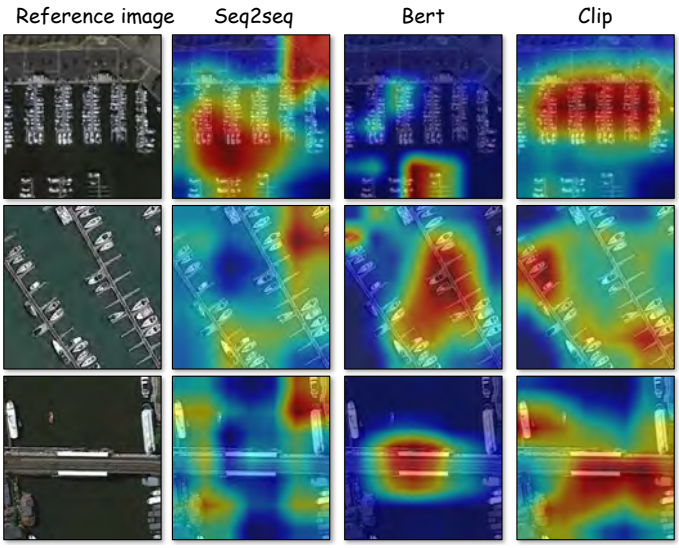

Fig. 11. Under different text encoding settings, select representative activation maps of different object categories, including new and basic categories.

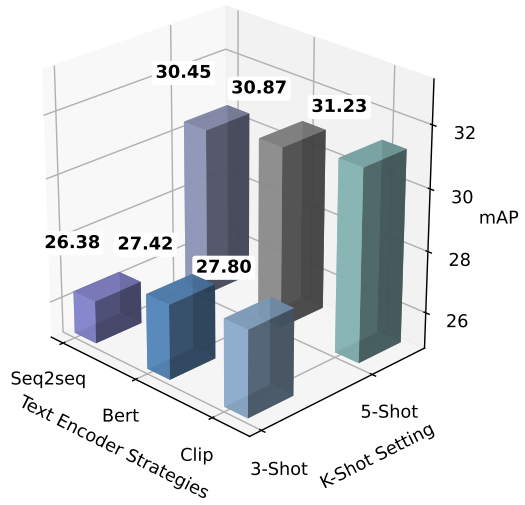

Fig. 12. Ablation study on text encoder selection under 3-shot and 5-shot settings.

heterogeneity between modalities. Our proposed approach also incorporates this structure. Crucially, our design optimally harnesses the synergistic advantages of textual and visual modalities, simultaneously preserving mutually beneficial cross-modal information and suppressing interference-prone information.

7) Selection of language encoder: To systematically investigate the impact of text encoder selection on cross-modal representation learning, we design a comparative experimental framework. As illustrated in Fig. 11, we use Grad-CAM visualization technology to perform reverse mapping of feature activation maps, allowing us to qualitatively analyze the distribution differences of model attention in raw image space under various encoding strategies. The experimental results demonstrate that, compared to seq2seq and BERT, the text-guided visual features generated by CLIP exhibit more focused attention on target objects within images, indicating CLIP's superior capability in capturing semantic category-relevant visual characteristics. Furthermore, we conduct performance comparisons among three text encoders under K-shot scenarios, as illustrated in Fig. 12. Consistent with qualitative findings, CLIP as the text encoder consistently achieves state-of-the-art performance, which may be attributed to its pre-training mechanism that enables better alignment between visual and linguistic concepts.

TABLE VI
COMPARISON OF TIME COSTS WITH OTHER FSOD FRAMEWORKS

Method	Params	FLOPs (InF.)	Time (InF.)
TFA w/cos	60.64M	259.3	0.085
DeFRCN	52.74M	408.9	0.098
Ours	**58.96M**	**464.3**	**0.109**

8) Analysis of the time costs: The time cost directly reflects the overall model efficiency under constrained computational resources. As shown in Table VI, to evaluate the time efficiency of the proposed method, we compare it with two FSOD frameworks in terms of model parameters (Params), computational complexity (FLOPs), and single-image inference time. Compared to the competitive method DeFRCN, our approach achieves significantly superior detection performance while requiring only an increase of 11.8% in parameters, 13.5% in FLOPs, and 11.2% in inference time.

9) Analysis of the reasons for failure: We conduct an analysis to explore the reasons for the experimental failures caused by the limitations of the method. As shown in Fig. 13, our method performs poorly under the following conditions. Specifically, there are two points: firstly, in complex large-scale background interference, our method cannot identify small objects in the images. Secondly, due to the nested relationship of classes, such as in the harbor class in Fig. 13 where there are many objects of the ship class, our network only recognizes the harbor and cannot effectively identify the large number of ships contained within the harbor. This indicates that our method has certain limitations. On one hand, it is because our method does not use a better multi-scale feature fusion method, and on the other hand, it is also due to the limitations of few-shot class information, causing difficulty in recognizing classes, especially those with nested relationships as mentioned earlier. This will be the direction we need to improve in the future.

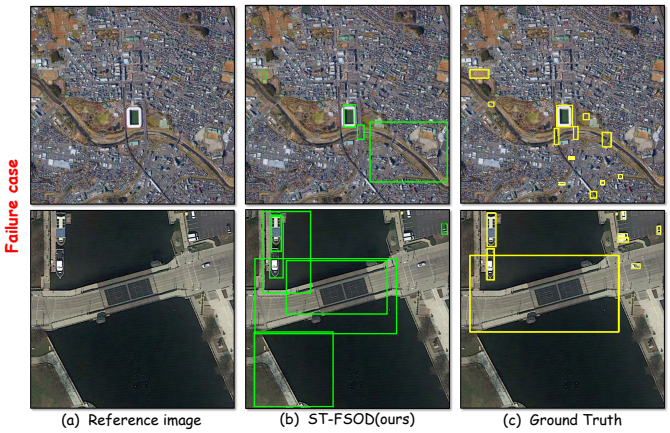

Fig. 13. Example of detection failures by ST-FSOD.

V. CONCLUSION

In this paper, we propose an novel detection framework ST-FSOD for FSOD tasks. Our main contributions include the introduction of the SA-RPN module and the TGL module. Compared to existing transfer learning-based methods, our approach addresses the issues of poor RPN generalization performance and the difficulty of representing sparse features in images, both critical for FSOD tasks. Additionally, we are the first to incorporate the text modality into FSOD tasks to further enhance model performance. Extensive experiments demonstrate that our method performs well on two remote sensing object detection datasets, competing closely with, and sometimes even surpassing and SOTA methods. This method is suitable for scenarios with limited image samples, such as identifying rare buildings, disaster detection and assessment, and military target recognition. In the future, we aim to enhance its interpretability, apply it to datasets beyond the remote sensing domain, and optimize computational efficiency

to expand its applicability and improve performance in various challenging scenarios.

REFERENCES

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[2] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9626–9635.

[3] X. Li, J. Deng, and Y. Fang, "Few-shot object detection on remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[4] 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 Journal of Photogrammetry and Remote Sensing*, vol. 159, pp. 296–307, 2020.

[5] 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 Journal of Photogrammetry and Remote Sensing*, vol. 98, no. dec., pp. 119–132, 2014.

[6] T. Wang, X. Zhang, L. Yuan, and J. Feng, "Few-shot adaptive faster r-cnn," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7166–7175.

[7] X. Wang, T. E. Huang, T. Darrell, J. Gonzalez, and F. Yu, "Frustratingly simple few-shot object detection," *ArXiv*, vol. abs/2003.06957, 2020.

[8] Y. Xiao, V. Lepetit, and R. Marlet, "Few-shot object detection and viewpoint estimation for objects in the wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3090–3106, 2023.

[9] Y. Liu, Z. Pan, J. Yang, B. Zhang, G. Zhou, Y. Hu, and Q. Ye, "Few-shot object detection in remote-sensing images via label-consistent classifier and gradual regression," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2024.

[10] Q. Fan, W. Zhuo, C.-K. Tang, and Y.-W. Tai, "Few-shot object detection with attention-rpn and multi-relation detector," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 4012–4021.

[11] L. Karlinsky, J. Shtok, S. Harary, E. Schwartz, A. Aides, R. Feris, R. Giryes, and A. M. Bronstein, "Repmet: Representative-based metric learning for classification and few-shot object detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5192–5201.

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

[13] G. Huang, I. Laradji, D. Vázquez, S. Lacoste-Julien, and P. Rodríguez, "A survey of self-supervised and few-shot object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4071–4089, 2023.

[14] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 3992–4003.

[15] L. Pang, J. Yao, K. Li, and X. Cao, "Special: Zero-shot hyperspectral image classification with clip," 2025. [Online]. Available: https://arxiv.org/abs/2501.16222

[16] D. Tang, X. Cao, X. Wu, J. Li, J. Yao, X. Bai, D. Jiang, Y. Li, and D. Meng, "Aerogen: Enhancing remote sensing object detection with diffusion-driven data generation," 2025. [Online]. Available: https://arxiv.org/abs/2411.15497

[17] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139, Jul 2021, pp. 8748–8763.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the North American Chapter of the Association for Computational Linguistics*, 2019.

[19] R. Mokady and A. Hertz, "Clipcap: CLIP prefix for image captioning," *ArXiv*, 2021.

[20] C. Yang, Z. Li, and L. Zhang, "Bootstrapping interactive image–text alignment for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–12, 2024.

[21] X. Lu, X. Sun, W. Diao, Y. Mao, J. Li, Y. Zhang, P. Wang, and K. Fu, "Few-shot object detection in aerial imagery guided by text-modal knowledge," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–19, 2023.

[22] Y. Jia, Q. Zhou, W. Huang, J. Gao, and Q. Wang, "Like humans to few-shot learning through knowledge permeation of vision and text," *arXiv preprint arXiv:2405.12543*, 2024.

[23] Y. Long, Y. Gong, Z. Xiao, and Q. Liu, "Accurate object localization in remote sensing images based on convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 5, pp. 2486–2498, 2017.

[24] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6517–6525.

[25] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European Conference on Computer Vision (ECCV)*, 2015.

[26] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6568–6577.

[27] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," *International Journal of Computer Vision*, vol. 128, pp. 642–656, 2018.

[28] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2999–3007.

[29] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *ArXiv*, vol. abs/1804.02767, 2018.

[30] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[31] X. Yang, J. Yang, J. Yan, Y. Zhang, T. Zhang, Z. Guo, X. Sun, and K. Fu, "Scrdet: Towards more robust detection for small, cluttered and rotated objects," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 8231–8240.

[32] Y. Yu, X. Yang, J. Li, and X. Gao, "Object detection for aerial images with feature enhancement and soft label assignment," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[33] Y. Li, Q. Hou, Z. Zheng, M.-M. Cheng, J. Yang, and X. Li, "Large selective kernel network for remote sensing object detection," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 16 748–16 759.

[34] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2017, pp. 1–10.

[35] Z. Li, F. Zhou, F. Chen, and H. Li, "Meta-sgd: Learning to learn quickly for few shot learning," *arXiv:1707.09835*, 2017.

[36] O. Vinyals, C. Blundell, T. P. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Neural Information Processing Systems (NeurIPS)*, 2016.

[37] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems*, 2017.

[38] N. Nashid, M. Sintaha, and A. Mesbah, "Retrieval-based prompt selection for code-related few-shot learning," in *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, 2023, pp. 2450–2462.

[39] Y. Jia, J. Gao, W. Huang, Y. Yuan, and Q. Wang, "Exploring hard samples in multiview for few-shot remote sensing scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[40] X. Zhang, X. Fan, G. Wang, P. Chen, X. Tang, and L. Jiao, "Mfgnet: Multibranch feature generation networks for few-shot remote sensing scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–13, 2023.

[41] X. Chen, G. Zhu, and J. Wei, "Mmml: Multimanifold metric learning for few-shot remote-sensing image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

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

[43] B. Sun, B. Li, S. Cai, Y. Yuan, and C. Zhang, "Fsce: Few-shot object detection via contrastive proposal encoding," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 7348–7358.

[44] L. Qiao, Y. Zhao, Z. Li, X. Qiu, J. Wu, and C. Zhang, "Defrcn: Decoupled faster r-cnn for few-shot object detection," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 8661–8670.
[45] G. Han and S.-N. Lim, "Few-shot object detection with foundation models," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 28 608–28 618.
[46] J. Li, Y. Tian, Y. Xu, X. Hu, Z. Zhang, H. Wang, and Y. Xiao, "Mm-rcnn: Toward few-shot object detection in remote sensing images with meta memory," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.
[47] X. Lu, X. Sun, W. Diao, Y. Mao, J. Li, Y. Zhang, P. Wang, and K. Fu, "Few-shot object detection in aerial imagery guided by text-modal knowledge," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–19, 2023.
[48] L. Li, X. Yao, X. Wang, D. Hong, G. Cheng, and J. Han, "Robust few-shot aerial image object detection via unbiased proposals filtration," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–11, 2023.
[49] T. Zhang, X. Zhang, P. Zhu, X. Jia, X. Tang, and L. Jiao, "Generalized few-shot object detection in remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 195, pp. 353–364, 2023.
[50] B. Yan, C. Lang, G. Cheng, and J. Han, "Understanding negative proposals in generic few-shot object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 7, pp. 5818–5829, 2024.
[51] F. Zhang, Y. Shi, Z. Xiong, and X. X. Zhu, "Few-shot object detection in remote sensing: Lifting the curse of incompletely annotated novel objects," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2024.
[52] G. Cheng, B. Yan, P. Shi, K. Li, X. Yao, L. Guo, and J. Han, "Prototype-cnn for few-shot object detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–10, 2022.
[53] J. Chen, D. Qin, D. Hou, J. Zhang, M. Deng, and G. Sun, "Multiscale object contrastive learning-derived few-shot object detection in vhr imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.
[54] X. Yuan, G. Cheng, K. Yan, Q. Zeng, and J. Han, "Small object detection via coarse-to-fine proposal generation and imitation learning," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 6294–6304.
[55] Z. Zhao, P. Tang, L. Zhao, and Z. Zhang, "Few-shot object detection of remote sensing images via two-stage fine-tuning," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.
[56] X. Luo, J. Xu, and Z. Xu, "Channel importance matters in few-shot image classification," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162, Jul 2022, pp. 14 542–14 559.
[57] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.
[58] R. Zhang, R. Fang, W. Zhang, P. Gao, K. Li, J. Dai, Y. J. Qiao, and H. Li, "Tip-adapter: Training-free clip-adapter for better vision-language modeling," *ArXiv*, vol. abs/2111.03930, 2021.
[59] Z. Xiao, J. Qi, W. Xue, and P. Zhong, "Few-shot object detection with self-adaptive attention network for remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 4854–4865, 2021.
[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
[61] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
[62] 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 *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9576–9585.
[63] J. Han, Y. Ren, J. Ding, K. Yan, and G.-S. Xia, "Few-shot object detection via variational feature aggregation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, Jun 2023, pp. 755–763.
[64] X. Lu, W. Diao, Y. Mao, J. Li, P. Wang, X. Sun, and K. Fu, "Breaking immutable: Information-coupled prototype elaboration for few-shot object detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, pp. 1844–1852, Jun. 2023.
[65] D. A. Ganea, B. Boom, and R. Poppe, "Incremental few-shot instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 1185–1194.
[66] N. Dong, Y. Zhang, M. Ding, and G. H. Lee, "Incremental-detr: Incremental few-shot object detection via self-supervised learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, pp. 543–551, Jun 2023.
[67] Z. Xiao, P. Zhong, Y. Quan, X. Yin, and W. Xue, "Few-shot object detection with feature attention highlight module in remote sensing images," *ArXiv*, vol. abs/2009.01616, 2020.
[68] G. C. Linderman, M. Rachh, J. G. Hoskins, S. Steinerberger, and Y. Kluger, "Efficient algorithms for t-distributed stochastic neighborhood embedding," *ArXiv*, vol. abs/1712.09005, 2017.

Chenchen Sun received the B.E. degree in computer science from Northwestern Polytechnical University, Xi'an, China. He is currently working toward the master's degree at the School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and remote sensing.

Yuyu Jia received the B.E. degree and the M.S. degree in control theory and engineering from the Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China, in 2019 and 2022, respectively. He is currently pursuing the Ph.D. degree at the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include few-shot learning, deep learning, and remote sensing.

Han Han received the B.E. degree in computer science from Northeast Forestry University, Harbin, China in 2023. He is currently working toward the master's degree at the School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and deep learning.

Qiang Li (Member, IEEE) is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University. His research interests include remote sensing image processing, particularly for image quality enhancement, object/change detection.

Qi Wang (Senior Member, IEEE) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, pattern recognition and remote sensing. For more information, visit the link (https://crabwq.github.io/).