remote sensing

MDPI

Article

Multi-Modal Prototypes for Few-Shot Object Detection in Remote Sensing Images

Yanxing Liu [1,2,3,4], Zongxu Pan [1,2,3,4,*], Jianwei Yang [1,2,3,4], Peiling Zhou [1,2,3,4] and Bingchen Zhang [1,2,3,4]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China; liuyanxing21@mails.ucas.ac.cn (Y.L.); yangjianwei20@mails.ucas.ac.cn (J.Y.); zhoupeiling21@mails.ucas.ac.cn (P.Z.); zhangbc@aircas.ac.cn (B.Z.)
[2] Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, Chinese Academy of Sciences, Beijing 100190, China
[3] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China
[4] Key Laboratory of Target Cognition and Application Technology (TCAT), Chinese Academy of Sciences, Beijing 100190, China
* Correspondence: zxpan@mail.ie.ac.cn

Abstract: Few-shot object detection has attracted extensive attention due to the abomination of time-consuming or even impractical large-scale data labeling. Current studies attempted to employ prototype-matching approaches for object detection, constructing class prototypes from textual or visual features. However, single visual prototypes exhibit limited generalization in few-shot scenarios, while single textual prototypes lack the spatial details of remote sensing targets. Therefore, to achieve the best of both worlds, we propose a prototype aggregating module to integrate textual and visual prototypes, leveraging both semantics from textual prototypes and spatial details from visual prototypes. In addition, the transferability of multi-modal few-shot detectors from natural scenarios to remote sensing scenarios remains unexplored, and previous training strategies for FSOD do not adequately consider the characteristics of text encoders. To address the issue, we have conducted extensive ablation studies on different feature extractors of the detector and propose an efficient two-stage training strategy, which takes the characteristics of the text feature extractor into account. Experiments on two common few-shot detection benchmarks demonstrate the effectiveness of our proposed method. In four widely used data splits of DIOR, our method significantly outperforms previous state-of-the-art methods by at most 8.7%.

Keywords: few-shot learning; object detection; remote sensing images

Citation: Liu, Y.; Pan, Z.; Yang, J.; Zhou, P.; Zhang, B. Multi-Modal Prototypes for Few-Shot Object Detection in Remote Sensing Images. *Remote Sens.* **2024**, *16*, 4693. https://doi.org/10.3390/rs16244693

Academic Editors: Qian Du, Yanni Dong and Xiaochen Yang

Received: 4 November 2024
Revised: 10 December 2024
Accepted: 10 December 2024
Published: 16 December 2024

1. Introduction

Benefiting from the rapid development of deep learning, significant progress has been achieved in the realm of remote-sensing object detection. Nonetheless, nearly all deep learning detectors rely heavily on large-scale annotated datasets to achieve satisfactory performance. Unfortunately, collecting sufficient labeled data is time-consuming or even unfeasible, particularly for remote sensing images (RSIs) with large image sizes and complex backgrounds. In this case, existing detectors will suffer from severe overfitting. To mitigate this limitation, few-shot object detection (FSOD) has been proposed and has garnered increasing attention from researchers.

FSOD aims to leverage knowledge learned from data-rich base datasets to enhance the performance of novel classes with limited data. Concretely, most FSOD approaches can be divided into fine-tuning methods and prototype-based methods. Fine-tuning methods [1–4] transfer the knowledge from base classes to novel classes through fine-tuning. Methods based on fine-tuning are more straightforward, yet their detection performance remains unsatisfactory. Prototype-based methods aim to construct a class prototype for

each class and detect few-shot classes by matching candidate regions with corresponding class prototypes. Pioneering works [5–8] construct class prototypes from few-shot support images to detect objects in query images, as illustrated in Figure 1a. Inspired by the rapid development of natural language processing, current researchers are exploring the construction of class prototypes from textual class names [9,10] or class descriptions [11], as illustrated in Figure 1b. Owing to the robustness of class prototypes, prototype-based approaches have demonstrated superior detection performance compared to fine-tuning methods in numerous applications.

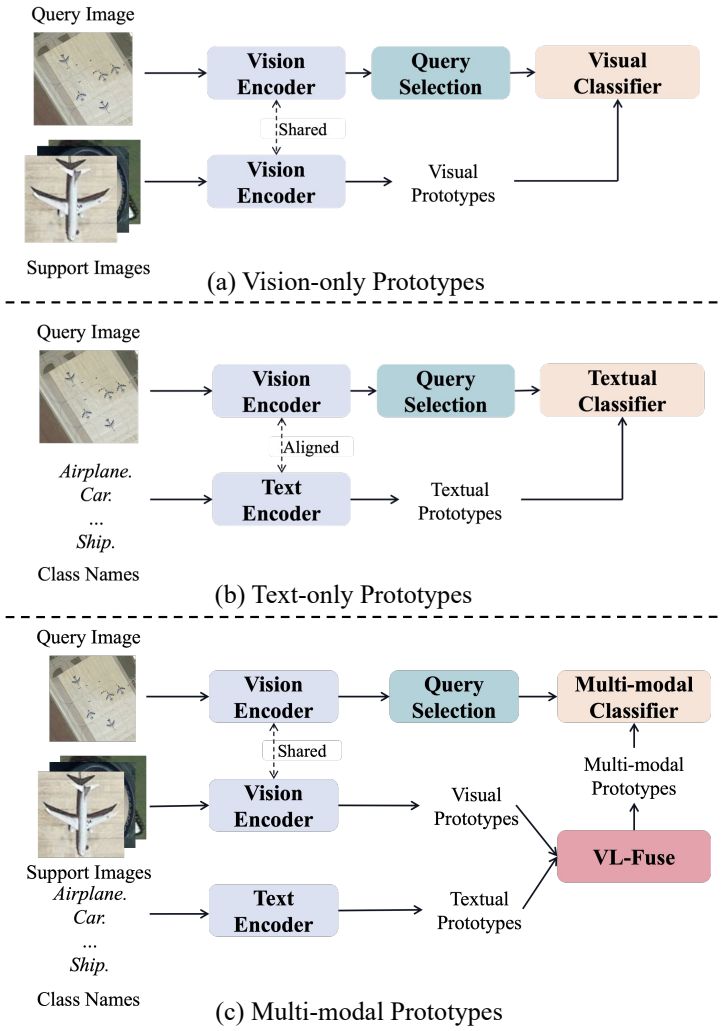

Figure 1. (**a**) Detectors based on vision-only prototypes construct class prototypes from the visual features of few-shot support images. (**b**) Detectors based on text-only prototypes build their class prototypes from class names. (**c**) Our approach constructs multi-modal prototypes to leverage the prior knowledge of textual prototypes and the spatial details of visual prototypes.

However, most prototype-based object detectors [5–11] construct class prototypes from either a single-modal image feature or a class name, as depicted in Figure 1a,b. These single-modal features constrain the feature extraction capabilities of prototypes. Actually, both visual and textual prototypes possess distinct strengths and drawbacks. On the one hand, textual prototypes generated from the text encoder contain rich prior knowledge about target classes and exhibit strong generalization capabilities, as the text encoder has been pre-trained on a large-scale natural language corpus. Nonetheless, since the textual prototypes are domain-agnostic, they cannot provide specific spatial details about target classes in RSIs, thereby limiting their applications in the field of remote-sensing

object detection. On the other hand, a picture paints a thousand words. Compared to textual prototypes, visual prototypes generated from RSIs can provide richer details about target classes. Nevertheless, the visual prototypes generated from limited support images lack generalization capabilities in few-shot scenarios, which will significantly degrade detection performance. In summary, single-modal visual and textual prototypes have distinct strengths and limitations.

Therefore, to further boost the performance of FSOD in RSIs, we propose a framework that leverages both visual and textual prototypes and we introduce a prototype aggregation module (PAM) that integrates the generalized prior knowledge of textual prototypes with the spatial details of visual prototypes. To be specific, as illustrated in Figure 1c, few-shot support images and textual class names are processed through their respective feature extractors to derive individual visual and textual prototypes. The highly generalizable textual prototypes and detail-rich visual prototypes are then aggregated by a prototype aggregator to obtain multi-modal prototypes. The produced multi-modal prototype retains both the generalized prior knowledge of textual prototypes and the detailed spatial information of visual prototypes. This multi-modal prototype is then integrated into the encoder and decoder of GroundingDINO [10] to detect targets.

The introduction of pre-trained text encoders makes previous few-shot object detection training strategies no longer suitable for our approach, as these strategies have not taken the text encoder into consideration and the text encoder is inherently different from the vision encoder. Therefore, we have conducted extensive ablation studies on the generalization capability of the pre-trained text encoder for FSOD in RSIs. We surprisingly find that the pre-trained text feature extractor exhibits remarkable robustness in FSOD tasks on RSIs. Even without any fine-tuning of the text encoder, a detector based on the pre-trained text encoder can achieve comparable performance to previous state-of-the-art FSOD methods. We argue that the reason for this is that the text encoder is domain-agnostic. Unlike images that exhibit different characteristics across domains, textual semantic information typically maintains similarity across various domains. Based on ablation studies and previous findings, we propose a novel efficient two-stage training strategy (ETS) for FSOD on RSIs, which is the first to take the characteristics of the pre-trained text feature extractor into account.

By integrating the multi-modal prototypes and the efficient two-stage training strategy with an advanced detector [10], we propose the multi-modal prototypical few-shot object detector (MP-FSDet) and achieve state-of-the-art few-shot object detection performance on DIOR and NWPU VHR-10.v2, which are two widely used datasets for few-shot object detection in RSIs. As shown in Figure 2, our method remarkably outperforms other state-of-the-art detectors, achieving a maximum performance improvement of 8.7% in the remote-sensing FSOD task.

In summary, the contributions of this paper are summarized as follows.

1. We propose a few-shot object detector based on multi-modal prototypes, aiming to enhance few-shot detection performance by combining visual prototypes and textual prototypes.

2. We propose a prototype aggregating method for the construction of multi-modal prototypes, which maintains the spatial details of visual prototypes and semantic prior of textual prototypes.

3. Based on comprehensive analyses and extensive experiments on different components of detectors, we propose an efficient two-stage training strategy (ETS) for FSOD on RSIs, which, to the best of our knowledge, is the first to take the characteristics of the pre-trained text feature extractor into account.

4. By integrating our proposed methods with the state-of-the-art detector [10], we achieve state-of-the-art detection performance on two widely used benchmarks, attaining a maximum performance improvement of 8.7% in the remote-sensing FSOD task.

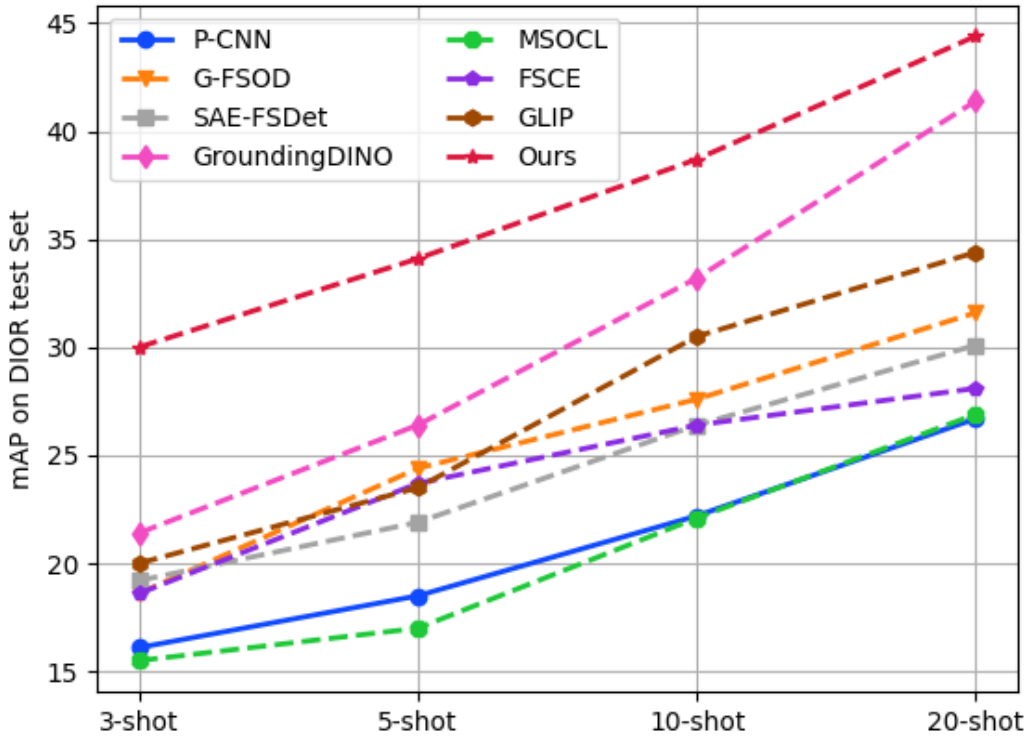

Figure 2. FSOD performance on DIOR test set at different shot numbers. We report the average AP50 on four splits, followed by [8]. As the figure shows, our proposed method is remarkably superior to other state-of-the-art approaches.

2. Related Work

In this chapter, we will introduce relevant works from two aspects. Firstly, we will introduce some generic object detection methods. Secondly, we will discuss various few-shot object detection approaches and analyze the shortcomings of previous research.

2.1. Generic Object Detection

Based on handcraft anchors or reference points, early convolution-based detectors are designed as either two-stage or one-stage models. Two-stage detectors [12,13] typically incorporate a region proposal network (RPN) to propose potential boxes and subsequently classify them in the second stage. One-stage detectors like YOLO v2 [14] and YOLO v3 [15] directly output offsets relative to predefined anchors and classification scores. Currently, a transformer-based end-to-end detector [16] has been proposed, which eliminates hand-designed components like anchors and NMS. To alleviate the computational burden, Deformable DETR [17] predicts 2D anchor offsets and incorporates a deformable attention module that focuses on a specific set around a reference point. By using a contrastive way for denoising training, a mixed query selection method, and a look-forward-twice scheme, DINO [18] further boosts the performance of transformer-based detectors.

Recently, the rapid advancement of natural language processing has enabled researchers to leverage textual information for object detection. Object detection approaches [9,10,19–21] via text queries have exhibited remarkable advancements within natural scenes. They are trained using existing bounding-box annotations and aim at detecting arbitrary classes with the help of generalized text queries. OVR-CNN [19] is a pioneering work that first pre-trains a visual projector on image–caption pairs to learn rich vocabularies and then fine-tunes a detector on the detection dataset. Then, by distilling the knowledge from the CLIP [20] model into an R-CNN-like detector, ViLD [21] further advances open-set object detection. GLIP [9] first formulates object detection as a grounding task and achieves an even stronger performance on fully supervised detection benchmarks without fine-tuning. GroundingDINO [10] utilizes the training strategy of

GLIP as a stronger detector DINO [18] and achieves a state-of-the-art open-set detection performance. The MQ-Det [22] utilizes both textual and visual queries for detection, which is similar to our work. However, it solely utilizes vectors as image queries, thereby failing to preserve the spatial details of the target. Furthermore, MQ-Det [22] is designed for natural scene images, which does not account for the domain gap between remote sensing and natural scene imagery. Since the models are only pre-trained on natural images and only use text embeddings as prototypes, the detection performance of these models is degraded on RSIs.

Along with the rapid development of detectors in natural scenes, object detection methods for aerial images have also achieved tremendous progress [23]. To address the challenge of rotation invariance for feature representation, Mei et al. [24] propose a novel cyclic polar coordinate convolutional layer. Considering the expansive spatial coverage of aerial images and diverse scales of objects, Deng et al. [25] propose a multi-scale object proposal network (MS-OPN), which comprises three proposal branches to predict multi-scale proposals. Due to the unique bird's-eye views of RSIs, objects in aerial images often exhibit arbitrary orientations. To achieve rotation-invariant detection, Cheng et al. [26] propose a Rotation-Invariant CNN (RICNN), which incorporates a novel rotation-invariant layer. Li et al. [27] propose a novel region proposal network that incorporates additional multi-angle anchors to detect objects with arbitrary orientations. Recent studies [28,29] have begun to explore the potential of detecting objects labeled with rotated bounding boxes.

Detectors, as mentioned above, have achieved remarkable performance. However, most of these are prone to overfitting when the labeled data become scarce and this significantly restricts the application of object detection for aerial images.

2.2. Few-Shot Object Detection

With the abomination of time-consuming or even impractical large-scale labeling, FSOD has attracted extensive attention.

Current FSOD methods can be divided into transfer-learning methods [1–3,30] and prototype-based methods [5–7,9,10,22,31]. Transfer-learning methods opt to fine-tune a pre-trained model on limited data for the few-shot task. Flagship works include LSTD [1], TFA [2], FSCE [3], and DeFRCN [30]. Prototype-based methods [5–7,9,10,22,31] aim to build a class prototype for each class and detect novel classes by matching candidate regions to class prototypes. They are trained across tasks for quick adaptation, and they are trained over multiple episodes. Early works [5–7,31,32] only employ few-shot support images as class prototypes. However, the visual prototypes produced by the few-shot images lack generalization. Inspired by the rapid development of natural language processing, the authors of [9,10,22] input class names into a pre-trained text encoder and utilize the resulting text features as class prototypes. As the text encoder is pre-trained on a large-scale textual corpus, textual prototypes generated from it exhibit strong generalization capabilities. Nevertheless, textual prototypes lack specific details about target classes.

Currently, FSOD for aerial images has attracted increasing attention. Li et al. [33] introduce a novel meta-learning approach with multi-scale detection capabilities tailored for the FSOD task in aerial images. Then, Cheng et al. [8] develop a framework named Prototype-CNN, which aims to leverage prototypes for recognizing objects with limited annotations. Similarly, Le et al. [34] also propose a novel prototypical network based on representation learning. Huang et al. [35] address the issue of significant data imbalance between the novel class and base classes by proposing a novel balanced fine-tuning approach and incorporating a shared attention module to exploit the abundant ground information within aerial images. DH-FSDet [36] presents an innovative annotation sampling and preprocessing methodology. To tackle the issue of limited scale diversity in aerial images, MSOCL [37] presents a multiscale object contrastive learning framework. TEMO [11] employs class descriptions as class prototypes. However, the text encoder it utilizes is training from scratch, thereby lacking generalization capabilities. In addition, single-modal prototypes in TEMO [11] cannot provide comprehensive information for target classes. Similarly,

UMFT [38] also proposes a multi-modal transformer to integrate the query features from ViT and the textual features extracted from BERT, aligning multi-modal representations in an end-to-end manner. To overcome the challenges of substantial scale and orientation variations of objects in RSIs, TINet [39] proposes a novel feature pyramid network (FPN) and utilizes prototype features to enhance query features. To rectify the issue of inconsistent label assignments across base training and fine-tuning stages, SAE-FSDet [40] proposes a novel label-consistent classifier named LCC and proposes a novel gradual rpn to enhance the localization performance of novel classes, inspired by the authors of [41,41–43].

Although prototype-based detection approaches have demonstrated remarkable performance in the field of FSOD, single-modality prototypes exhibit distinct strengths and limitations. To further boost the performance of FSOD, we propose a multi-modal prototype that can harness the strengths of both textual and visual prototypes. Moreover, the transferability of detectors based on multi-modal prototypes in the field of remote-sensing FSOD has not been explored. Through a comprehensive analysis, we find that the text encoder is domain-agnostic, while the visual encoder is domain-specific in RSIs, and we propose an efficient two-stage training strategy.

3. Proposed Approach

In this section, we will introduce the proposed MP-FSDet network in detail. Before introducing our approach, we will first present some preliminary knowledge in Section 3.1. Subsequently, we will provide a brief overview of the architecture of our proposed method in Section 3.2. To leverage both the prior knowledge from textual prototypes and the spatial details from visual prototypes, we present a multi-modal prototype construction method introduced in Section 3.3. Furthermore, we have found that the previous training strategy can harm the generalization capability of the text encoder. Thus, we propose an efficient two-stage training strategy in Section 3.4 to address the problem.

3.1. Preliminary Knowledge

3.1.1. Problem Setting

FSOD aims to train a detector on a base dataset D_b with abundant annotated data and a novel dataset D_n containing limited data. Classes in D_b are denoted as base classes C_b, and those in D_n are denoted as novel classes C_n. The base classes are fully annotated, containing numerous annotated object instances, while the novel classes have only K-shot annotations. Note that the base classes C_b and the novel classes C_n are non-overlapping, namely $C_b \cap C_n = \emptyset$. For the inference phase in FSOD, the test set contains both base and novel classes ($C_{test} = C_b + C_n$), and the detector is required to detect all objects. As a result, a balanced few-shot dataset D_{few} containing K-shot object annotations for both base and novel classes is constructed for few-shot fine-tuning.

3.1.2. Base Detector

Our model is grounded on GroundingDINO [10], which employs single-modal textual prototypes. It is originally designed for open-set object detection, a task that endeavors to detect arbitrary objects within human inputs, such as category names or referring expressions. However, owing to highly generalized language encoders and large-scale language–image alignment, it demonstrates remarkable performance in the domain of few-shot object detection [22]. The detector first obtains the respective feature maps via the image and text encoders. Subsequently, the textual prototypes generated from the text encoder and image features are fed into the transformer encoder and decoders to obtain the final detection results. The transformer encoder and decoder can be conceptualized as a matcher that matches the feature points on the feature map with the class prototypes of each class, utilizing the class with the highest confidence as the final classification result.

3.2. Multi-Modal Prototypical Few-Shot Object Detector

Textual prototypes that are pre-trained from a large-scale natural language corpus are highly generalizable, while visual prototypes contain detailed spatial information about specific objects. Therefore, we propose a detection framework that aims to combine the generalization properties of textual prototypes with the spatial details of objects in aerial images.

The architecture overview is illustrated in Figure 3. As the figure shows, the textual prototypes $P_t \in \mathbb{R}^{|C| \times D}$ are extracted from the class names by a pre-trained BERT [44] encoder, where $|C|$ is the number of classes. Benefiting from the pre-training of the BERT encoder on large-scale corpora, textual prototypes encompass extensive prior knowledge of targets and exhibit robust generalization capability. Meanwhile, the visual prototypes $P_v \in \mathbb{R}^{|C| \times H_s W_s \times D}$ are extracted from support images by a vision encoder and an RoI pooling layer, where H_s and W_s denote the height and width of support RoI features, respectively. Visual prototypes preserve spatial details of the target within the image. It is noteworthy that, unlike the previous approach [6], which sends the vector after splicing and resizing the support images and foreground masks to the vision encoder, we directly input the original support images into the vision encoder for processing. We argue that consistent pre-processing of support images and query images can benefit the alignment between query features and visual prototypes.

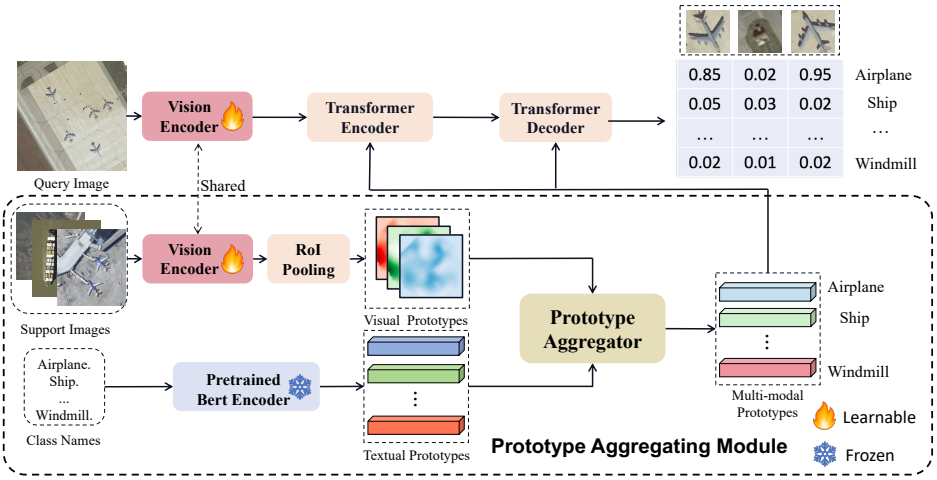

Figure 3. The architecture of our multi-modal prototypical few-shot object detector. It comprises two branches: a query image detection branch (**top**) and a multi-modal prototype construction branch (**bottom**). The upper branch employs the DETR architecture, which is responsible for detecting objects in the query image based on the multi-modal prototypes. The feature enhancer in the branch is a DETR encoder that employs a cross-attention mechanism to enhance the interaction between query features and multi-modal prototypes. The multi-modal decoder in the branch can be regarded as a matcher that matches each query feature with multi-modal prototypes. The bottom branch utilizes the prototype aggregating module (PAM) to construct multi-modal prototypes. The PAM feeds support images and class names into the visual encoder and BERT encoder to obtain visual prototypes and textual prototypes, respectively. The proposed prototype aggregator then aggregates the textual and visual prototypes into multi-modal prototypes. Note that the BERT encoder is frozen in the base training stage to maintain the generalization of textual prototypes.

Subsequently, the textual prototypes P_t and visual prototypes P_v are then fed into the prototype aggregator to obtain multi-modal prototypes $P \in \mathbb{R}^{|C| \times D}$. The prototype aggregator comprises a cross-modal attention mechanism and the details will be introduced in detail in Section 3.3. This module fuses semantic features from the textual prototypes with spatial details from the aerial images to construct multi-modal prototypes. After obtaining multi-modal prototypes, they will be fed into a transformer encoder introduced in GroundingDINO [10], along with the query features $F_q \in \mathbb{R}^{H_q W_q \times D}$. The transformer

encoder [10] comprises six mutual cross-attention layers to enhance the interactivity between multi-modal prototypes and query features. At last, the multi-modal prototypes and enhanced query features will be jointly input into the multi-modal decoder [10] to obtain the final detection outputs. The classification layer in the decoder will match the selected object queries with the multi-modal prototypes of each category one by one to obtain the final classification results. The decoder also incorporates several regression layers to refine the positions of object queries.

The training loss of our model is composed as follows:

$$L = L_{cls} + L_{loc} + L_{IoU} + L_{dn} \tag{1}$$

where L_{cls} and L_{loc} represent the classification loss and localization loss of the bounding boxes, respectively. Specifically, we employ the focal loss [45] for the classification loss and the smooth L1 loss [12] for the localization loss. The L_{IoU} is an additional loss employed to enhance the ability of detectors to handle objects with varying scales and shapes. Furthermore, in line with DINO [18], we also incorporate a denoising loss L_{dn} to address the low training efficiency of the DETR [16] model and enhance model convergence, as mentioned in [18,46].

3.3. Prototype Aggregating Module

Inspired by [9], we propose a prototype aggregating module to aggregate visual prototypes and textual prototypes. This module primarily comprises a prototype aggregator, designed to leverage both the spatial detail of visual prototypes and the generalization capability of textual prototypes.

Specifically, we employ a multi-head attention module in the prototype aggregator, where each head computes the cross-modal prototype by aggregating visual features into a text prototype, as shown in Figure 4. The prototype aggregator fuses the visual prototype and textual prototype for each class individually. Without loss of generality, the following description focuses on the construction of a single-class multi-modal prototype.

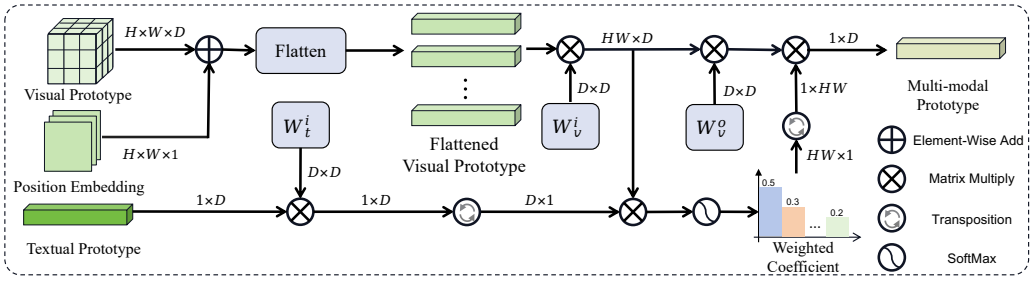

Figure 4. The architecture of the proposed prototype aggregator. The visual prototype, with positional embedding incorporated, is firstly reshaped into $P_v \in \mathbb{R}^{HW \times D}$. The flattened visual prototype and textual prototype are subsequently projected through the input projection weights W_v^i and W_t^i, respectively, to perform feature transformation. W_v^i and W_t^i are the transformation matrices for the input textual features and input visual features of the ith class, respectively. W_v^o represents the output transformation matrix of the visual features and the weighted visual prototypes will ultimately be produced as the multi-modal prototype. Then, the correlation between textual prototypes and visual prototypes is quantified by computing the normalized attention weights. At last, the weighted visual prototype is integrated into the text prototype, resulting in the multi-modal prototype.

The detailed formulas for a multi-modal prototype are presented as follows:

$$O_v = P_v W_v^i, O_t = P_t W_t^i, \tag{2}$$

$$A = \text{SoftMax}(\frac{O_t O_v^T}{\sqrt{D}}), \tag{3}$$

$$P = AP_vW_v^o, \tag{4}$$

where $P_v \in \mathbb{R}^{HW \times D}$, $P_t \in \mathbb{R}^{1 \times D}$ are flattened visual and textual prototypes, respectively, W_v^i and W_t^i are input projection weights, and W_v^o is the output project weights. The visual and textual prototypes are first projected into a high-dimension semantic space by input projection weights, as shown in Equation (2) with O_v and O_t. Then, the correlation matrix A between the semantic textual prototype and the detailed visual prototype is computed by calculating the normalized attention weights, as demonstrated in Equation (3). At last, the multi-modal prototype P is obtained by integrating weighted visual details into the text prototype, as illustrated in Equation (4).

While the proposed PAM appears similar to the feature fusion module in the GLIP [9] model, the two modules serve quite different roles in detectors. The fusion module in GLIP [9] fuses textual features and visual features of query images in the ATSS detection head, while PAM aggregates visual features of support images into textual features before the detection head. Therefore, the proposed PAM and GLIP [9] models are compatible, and the textual prototypes of GLIP [9] can be further enhanced. There are also some differences in the fusion methods between PAM and the fusion module in GLIP [9]. GLIP [9] adopts a bidirectional fusion approach based on early fusion, while our method uses a fusion based on late diffusion.

In addition, note that if each class has multiple samples, directly using the mean of visual samples as a visual prototype can lead to feature degradation due to potential misalignments in the spatial locations of different samples. Therefore, when the number of support images exceeds one, we employ the following formula to align support features from multiple shots. We select the first support feature S_0 of each class as the reference for spatial alignment, so all other support features should be aligned with it. We first compute the similarity between S_i and the S_0 at each spatial location, subsequently normalizing the values to obtain the $\delta(S_0, S_i)$. The detailed formula of $\delta(S_0, S_i)$ is presented as follows:

$$\begin{aligned} \delta(S_0, S_i) &= \sigma((S_0 - E(S_0))(S_0 - E(S_i))^T) \\ i &= 0, 1, ..., K-1, \end{aligned} \tag{5}$$

where $S_i \in \mathbb{R}^{HW \times D}, i = 0, 1, ..., K-1$ represent the feature of the i-th support image, K is the number of support images, $\sigma(.)$ is the SoftMax function, and $E(S_i)$ is the mean of S_i. Subsequently, we spatially align S_i with S_o by multiplying $\delta(S_0, S_i)$ with S_i, and then compute the mean of all aligned support features to obtain the final visual prototype P_v, as shown in Equation (6).

$$P_v = \frac{1}{K} \sum_{i=0}^{K-1} \delta(S_0, S_i) S_i \tag{6}$$

3.4. Efficient Two-Stage Training Strategy

Due to the domain gap between aerial images and natural images, detectors trained on natural scene datasets are not good at extracting the features of aerial targets. Therefore, previous few-shot object detectors [8,37] often adopt two-stage training strategies to enhance the feature extraction capabilities of detectors. However, the robustness of pre-trained text encoders such as BERT [44] in few-shot object detection for remote sensing images remains largely unexplored. To fill the blank, we have conducted comprehensive experiments and propose an efficient two-stage training strategy for multi-modal FSOD in RSIs.

In line with prior research, our two-stage training approach also comprises a base training stage followed by a few-shot fine-tuning stage. The base training stage aims to train feature extractors for aerial objects using the sufficiently annotated base dataset D_{base}, while the few-shot fine-tuning stage seeks to transfer the prior knowledge to enhance the

performance of few-shot novel classes. The overall framework of ETS is illustrated in Figure 5.

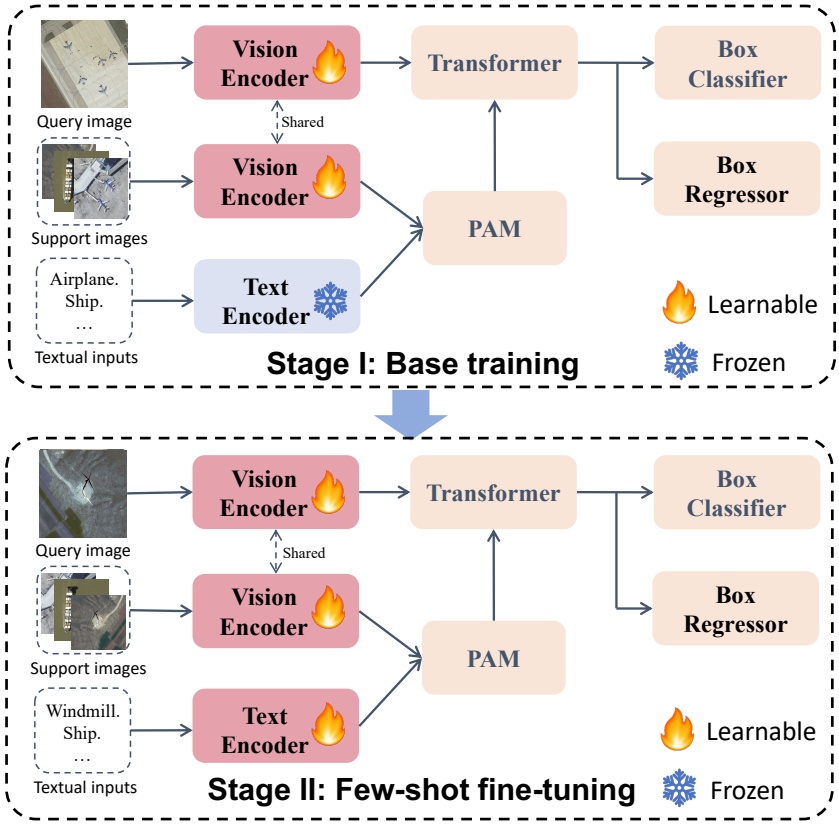

Figure 5. Proposed efficient two-stage training strategy. The modules in red indicate that the modules are learnable at this stage, while the modules in blue indicate that the module is frozen at this stage. Modules in orange are learnable at all stages.

Base training stage. Unlike previous methods [2–4] that train all learnable parameters during the base training stage, our approach trains only the image feature extractor and the DETR [16] encoder and decoder while keeping the text encoder frozen in the stage. The reason for the setting is that we surprisingly find that training the text encoder in the base training stage will harm the detection performance for few-shot novel classes. We argue that this may be because the trained text encoder has a preference for base classes. And we believe that the text encoder is domain-agnostic, implying that the semantic information encoded in the text encoder holds true in both aerial images and natural images. However, in the base training stage, the predominance of base class instances can bias the text encoder, compromising its generalization and diminishing the detection performance for few-shot novel classes. As a result, freezing the text encoder in the base training stage can maintain the generalization capability of the text encoder and can also conveniently reduce training time during large-scale base training.

Few-shot fine-tuning stage. In previous methods [3,4,11], only the image feature extractor is frozen, and all other components are learnable in the few-shot fine-tuning stage. However, our findings suggest that the image feature extractor trained on the base dataset is not optimally suited for the few-shot novel classes. Consequently, to enhance the feature extraction capability of the detector for novel classes, we fine-tune the image feature extractor of the detector with the balanced few-shot set. Since the pre-trained text encoder shows strong robustness, whether the text encoder is trained or not has little impact on the experimental results. Nonetheless, training the text encoder during the fine-tuning stage can still improve performance by about one point, so we unfroze the text encoder during this stage to further improve the detection performance on novel classes.

In summary, during the base training phase, our method maintains the text feature extractor in a frozen state since the textual encoder is domain-agnostic and has been well pre-trained on a large-scale natural language corpus. We are surprised to find that training the text encoder during the base training stage with relatively limited text data actually diminishes the generalization performance of the text encoder. In the fine-tuning stage, we fine-tune all components of the detector to obtain a more robust feature extractor.

4. Experiments and Results

To test the performance of our proposed MP-FSDet, a series of comparative and ablation experiments is conducted. Section 4.1 introduces the details of the experimental setup and evaluation metric. Section 4.2 presents implementation details of the experiments. In Section 4.3, we compare the proposed method with several state-of-the-art detectors, including both generic and few-shot detectors. At last, we carefully analyze the role of each module in the proposed method through ablation studies in Section 4.4.

4.1. Datasets and Evaluation Metrics

DIOR: The DIOR dataset is a large-scale object detection dataset for optical remote sensing images, released by the authors of [23]. In the DIOR dataset, images are sourced from Google Earth, comprising a total of 23,463 images and 192,472 objects distributed across 20 classes. The classes in the DIOR dataset consist of *airplane*, *airport*, *baseball field*, *basketball court*, *bridge*, *chimney*, *dam*, *expressway service area*, *expressway toll station*, *harbor*, *golf course*, *ground track field*, *overpass*, *ship*, *stadium*, *storage tank*, *tennis court*, *train station*, *vehicle*, and *windmill*. The dataset is partitioned into three subsets: a training set, a validation set, and a testing set, comprising 5682, 5863, and 11,738 images, respectively. The images in the DIOR dataset have a spatial resolution ranging from 0.5 to 30 m, with all images being 800 $\times$ 800 pixels in size.

NWPU VHR-10.v2: The NWPU VHR-10.v2 dataset is a high-resolution optical remote sensing dataset introduced by the authors of [47]. It includes 1173 optical images with the size of 400 $\times$ 400. The dataset comprises a total of ten geospatial object classes, which include *airplane*, *baseball diamond*, *basketball court*, *bridge*, *ground track field*, *harbor*, *ship*, *storage tank*, *tennis court*, and *vehicle*.

To evaluate the detection performance of our proposed method when there are only a few objects per class, we divide the classes of each dataset into two sets: base class sets and novel class sets. Each class within the base class sets comprises sufficient annotated objects, while each class within the novel class sets contains only K objects. Specifically, we select K = 3, 5, 10, and 20 objects for each class from both the training and validation sets in the K-shot detection task. For the DIOR dataset, we follow the FSOD setup outlined in [8,11] and utilize five different novel/base splits, each comprising 15 base classes and five novel classes as specified in Table 1. For the NWPU VHR-10.v2 dataset, following previous works [40], three categories, *airplane*, *baseball diamond*, and *tennis court*, are adopted as novel classes, while the others are considered as the base classes.

We evaluate the performance of detectors using mean Average Precision (mAP) introduced in the PASCAL VOC2007 benchmark [48]. The mAP is calculated by averaging 11 precision values as recall ranges from 0 to 1 in increments of 0.1, with an IoU threshold of 0.5.

Table 1. Five different base/novel split settings in the DIOR dataset.

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

4.2. Implementation Details

Our proposed method is based on GroundingDINO [10], and we adopt the pre-trained detector with the backbone of Swin-T [49] and BERT [44]. In the base training stage, the model is trained by 12 epochs, where the learning rate starts from 1.25×10^{-5} and decreases to 1.25×10^{-6} at 11 epochs. In the fine-tuning stage, the model is fine-tuned by 72 epochs with a learning rate of 1.25×10^{-5}. In both the base training and the fine-tuning stages, we use AdamW as the gradient backward method with a weight decay coefficient of 1×10^{-4}. The training batch size is set to 2 on one NVIDIA GeForce RTX 4090 GPU.

Given the maximum input length of the BERT encoder being 256 tokens, we do not use the class descriptions from the DIOR dataset in TEMO [11] as input text. Instead, we utilize the features of class names encoded by BERT as the textual prototypes. The separators "." between textual prototypes are eliminated during fusion, as we have found that separators lacking semantic information significantly interfere with the performance of prototype aggregation. Regarding visual prototypes, the cropped features of the target objects in support images serve as support features, and the mean features obtained after aligning are utilized as the visual prototypes, as detailed in Section 3.3.

4.3. Experimental Results and Comparisons

4.3.1. Results on the DIOR Dataset

To demonstrate the effectiveness of our proposed method, we compare our approach with several advanced detectors, including detectors based on fine-tuning [3,4,40], visual prototypes [8], and textual prototypes [9–11], as shown in Table 2. We also integrate our MP-FSDet with GLIP-T [9] and GroundingDINO-T [10], naming them MP-GLIP-T and MP-GroundingDINO-T, respectively. Given that GLIP-T [9] and GroundingDINO-T [10] are originally designed for detection in natural scenes, we fine-tune them on the balanced few-shot DIOR dataset to ensure a more equitable comparison.

As illustrated in Table 2, compared with the detectors based on visual prototypes [8] and transfer-learning [3,4,40], detectors based on textual prototypes [9–11] demonstrate superior performance on many data splits. We attribute this to the strong generalization of textual prototypes. The textual prototypes are generated by a massively pre-trained text encoder, thereby inheriting rich prior knowledge about targets from the pre-trained text corpus. The prior knowledge in textual prototypes helps to enhance the generalization of detectors and reduce the risk of overfitting in few-shot scenarios. However, single-modal textual prototypes are highly conceptual and lack specific spatial details of targets, which limits their performance in FSOD. To further enhance the performance of prototype-based few-shot detectors, we combine the semantic-rich textual prototypes with detail-rich visual prototypes through our proposed prototype aggregating module and efficient two-stage training strategy. As the table shows, our method achieves state-of-the-art detection performance on all data splits of the DIOR dataset. The results in the table demonstrate the effectiveness of our proposed MP-FSDet.

In addition to qualitative analysis, we also visualize the detection results of GLIP-T [9], GroundingDINO-T [10], and our proposed method in split1 under three-shot settings, as illustrated in Figure 6. As the figure shows, even when utilizing only three training samples, GLIP-T [9] and GroundingDINO-T [10] demonstrate the capability to detect objects belonging to novel classes. We argue that this ability stems from the fact that both GroundingDINO-T [9] and GLIP-T [10] employ textual prototypes, which possess rich prior knowledge. Yet, since textual prototypes lack spatial details about targets, these two advanced methods still misclassify some novel objects as background or base classes. Concurrently, the absence of spatial details in class prototypes also leads to poor object localization performance. By combining visual prototypes and textual prototypes, our method enhances the spatial information of class prototypes, consequently improving classification and localization accuracy.

Table 2. Few-shot object detection performance on the DIOR test set under 3-, 5-, 10-, and 20-shot settings. The best two performances are highlighted with red and blue. TEMO [11] is solely evaluated on split1, while others [3,4,8,40] are exclusively tested on the remaining four splits; missing data points are denoted by the − symbol.

Shots	Method	Year	Split1	Split2	Split3	Split4	Split5	Mean
3	P-CNN [8]	TGRS2021	-	18.0	14.5	16.5	15.2	16.1
	G-FSOD [4]	ISPRS2023	-	27.6	14.1	16.0	16.7	18.6
	TEMO [11]	TGRS2023	30.9	-	-	-	-	30.9
	SAE-FSDet [40]	TGRS2024	-	28.8	14.0	16.7	17.3	19.2
	FSCE [3]	CVPR2021	-	27.9	13.2	15.6	17.5	18.6
	GLIP-T [9]	CVPR2022	24.4	24.1	23.7	16.5	15.6	20.9
	GroundingDINO-T [10]	ECCV2024	27.8	30.8	24.5	15.2	14.8	22.6
	MP-GLIP-T (Ours)		30.0	25.5	31.0	24.3	27.0	27.6
	MP-GroundingDINO-T (Ours)		42.1	35.7	32.2	28.1	24.0	32.4
5	P-CNN [8]	TGRS2021	-	22.8	14.9	18.8	17.5	18.5
	G-FSOD [4]	ISPRS2023	-	37.5	15.8	23.3	21.0	24.4
	TEMO [11]	TGRS2023	37.2	-	-	-	-	37.2
	SAE-FSDet [40]	TGRS2024	-	32.4	15.6	19.1	20.5	21.9
	FSCE [3]	CVPR2021	-	28.6	14.1	16.2	20.4	19.8
	GLIP-T [9]	CVPR2022	28.2	26.7	28.1	20.3	19.0	20.5
	GroundingDINO-T [10]	ECCV2024	33.1	36.2	29.9	19.4	20.0	27.7
	MP-GLIP-T (Ours)		35.3	29.5	34.5	27.2	29.7	31.2
	MP-GroundingDINO-T (Ours)		46.4	37.3	36.8	30.9	31.2	36.5
10	P-CNN [8]	TGRS2021	-	27.6	18.9	23.3	18.9	22.2
	G-FSOD [4]	ISPRS2023	-	37.5	20.7	26.3	25.8	27.6
	TEMO [11]	TGRS2023	42.8	-	-	-	-	42.8
	SAE-FSDet [40]	TGRS2024	-	37.1	17.4	28.4	22.7	26.4
	FSCE [3]	CVPR2021	-	33.1	15.8	23.8	22.2	23.7
	GLIP-T [9]	CVPR2022	38.5	30.7	33.6	30.6	27.1	32.1
	GroundingDINO-T [10]	ECCV2024	42.4	40.4	33.5	31.2	27.8	35.1
	MP-GLIP-T (Ours)		45.9	32.6	41.0	31.5	34.4	37.1
	MP-GroundingDINO-T (Ours)		53.1	41.0	42.2	36.1	35.5	41.6
20	P-CNN [8]	TGRS2021	-	29.6	22.8	28.8	25.7	26.7
	G-FSOD [4]	ISPRS2023	-	39.8	22.7	32.1	31.8	31.6
	TEMO [11]	TGRS2023	-	-	-	-	-	-
	SAE-FSDet [40]	TGRS2024	-	42.5	21.3	29.9	26.8	30.1
	FSCE [3]	CVPR2021	-	37.6	20.9	28.9	25.0	28.1
	GLIP-T [9]	CVPR2022	42.1	36.3	36.6	33.0	31.5	35.9
	GroundingDINO-T [10]	ECCV2024	48.5	49.3	40.1	39.1	37.1	42.8
	MP-GLIP-T (Ours)		50.0	41.0	40.5	35.9	35.0	40.5
	MP-GroundingDINO-T (Ours)		55.4	51.5	45.5	40.9	39.8	46.6

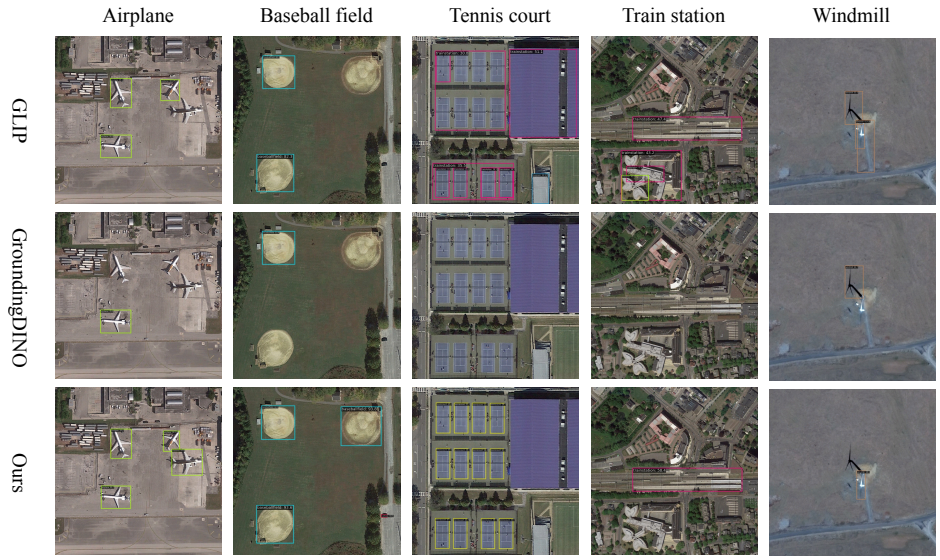

Figure 6. Novel class detection results in SPLIT 1 of the DIOR dataset under three-shot settings. Detection boxes with different colors represent different objects. The first row represents the results of GLIP-T [9], the second row represents the results of GroundingDINO-T [10], and the last row represents the detection results of our proposed method. Note that only the bounding boxes with a confidence greater than 0.3 are visualized.

4.3.2. Results on the NWPU VHR-10 Dataset

Table 3 presents the detection performance of our approach and other comparative methods on the NWPU VHR-10.v2 test set. Similar to the experiments on the DIOR dataset, we still combine our method with GLIP-T [9] and GroundingDINO-T [10]. Note that GLIP-T [9] and GroundingDINO-T [10] have been fine-tuned on the balanced few-shot dataset of NWPU VHR-10.v2.

Table 3. FSOD performance on the NWPU VHR-10.v2 test dataset under 3-, 5-, 10-, and 20-shot settings. The best two performances are highlighted with red and blue.

Method	3-Shot	5-Shot	10-Shot	20-Shot
P-CNN [8]	41.8	49.2	63.3	66.8
G-FSOD [4]	49.1	56.1	71.8	75.4
TEMO [11]	57.6	66.8	75.1	-
SAE-FSDet [40]	58.0	59.4	71.0	85.1
FSCE [3]	41.6	48.8	60.0	79.6
GLIP-T [9]	47.9	57.0	63.2	70.8
GroundingDINO-T [10]	59.1	62.1	84.3	88.1
MP-GLIP-T (Ours)	57.8	65.0	67.4	76.5
MP-GroundingDINO-T (Ours)	77.9	85.3	88.5	88.8

As Table 3 shows, with the combination of textual and visual prototypes and the efficient two-stage training strategy, MP-GroundingDINO-T achieves the mAP of 77.9% with just three samples, surpassing all other detectors by a minimum of 18.8%. This indicates that our approach performs effectively with extremely limited samples. Furthermore, we also observe that our approach converges to saturation performance more rapidly as the number of shots increases. This implies that our method exhibits a lower demand for training data and greater suitability for FSOD compared to other methods.

We visualize the detection results of GLIP-T [9], GroundingDINO-T [10], and our MP-GroundingDINO-T for three novel classes under the three-shot setting, as shown in Figure 7. The results indicate that GLIP-T [9] and GroundingDINO-T [10] suffer from significant misclassification. For instance, GLIP-T [9] misclassifies numerous background regions as foregrounds, while GroundingDINO [10] misclassifies numerous foreground regions as backgrounds. We attribute the poor detection performance to the inherent limitations of single-modal prototypes and the inadequacy of previous training strategies [4,11] to enhance the feature extraction capabilities of the vision encoder while preserving the generalization performance of the text encoder. However, by introducing the multi-modal prototypes and efficient two-stage training strategy, our method can leverage the advantages of detailed visual prototypes and generalized textual prototypes, thereby enhancing the detection performance for few-shot novel classes.

4.4. Ablation Studies

4.4.1. Ablation Study for the Components of Our Proposed Method

Table 4 shows the effectiveness of the proposed components in our approach. Given that GroundingDINO-T [10] is originally designed for detection in natural scenes, we fine-tune it on the balanced few-shot DIOR dataset and report the performance in the first row of Table 4. However, even though it utilizes the few-shot dataset to enhance the model's ability to detect the target classes, it still does not perform well on aerial images. Therefore, to introduce more prior knowledge about RSIs and enhance the feature extraction capability of remote-sensing objects, we employ a novel two-stage training strategy named the efficient two-stage training strategy (ETS) for the baseline model. As shown in the second row of Table 4, the incorporation of ETS significantly enhances the detection performance, especially when the number of shots is extremely low. Additionally, to leverage both the semantic information of textual prototypes and the spatial details of visual prototypes, the prototype aggregating module (PAM) is then employed on the detector, which improves the mAP by 9.3% on average. At last, we integrate our proposed ETS and PAM into the baseline detector, which further enhances the performance of FSOD

in RSIs. In summary, compared with the baseline detector, our proposed method improves the detection performance by 20.8%, 13.3%, 10.7%, and 6.9% in 3, 5, 10, and 20 shots, respectively, in the first split of the DIOR dataset. In addition, our method also improves the detection performance by 18.8%, 23.2%, 4.2%, and 0.7% in 3, 5, 10, and 20 shots, respectively, in the NWPU VHR-10.v2 dataset. We also find the saturation performance in the 20-shot NWPU VHR-10.v2 dataset, which indicates that with an increasing number of shots, the performance tends to converge. However, our method converges more rapidly, indicating a lower demand for training data.

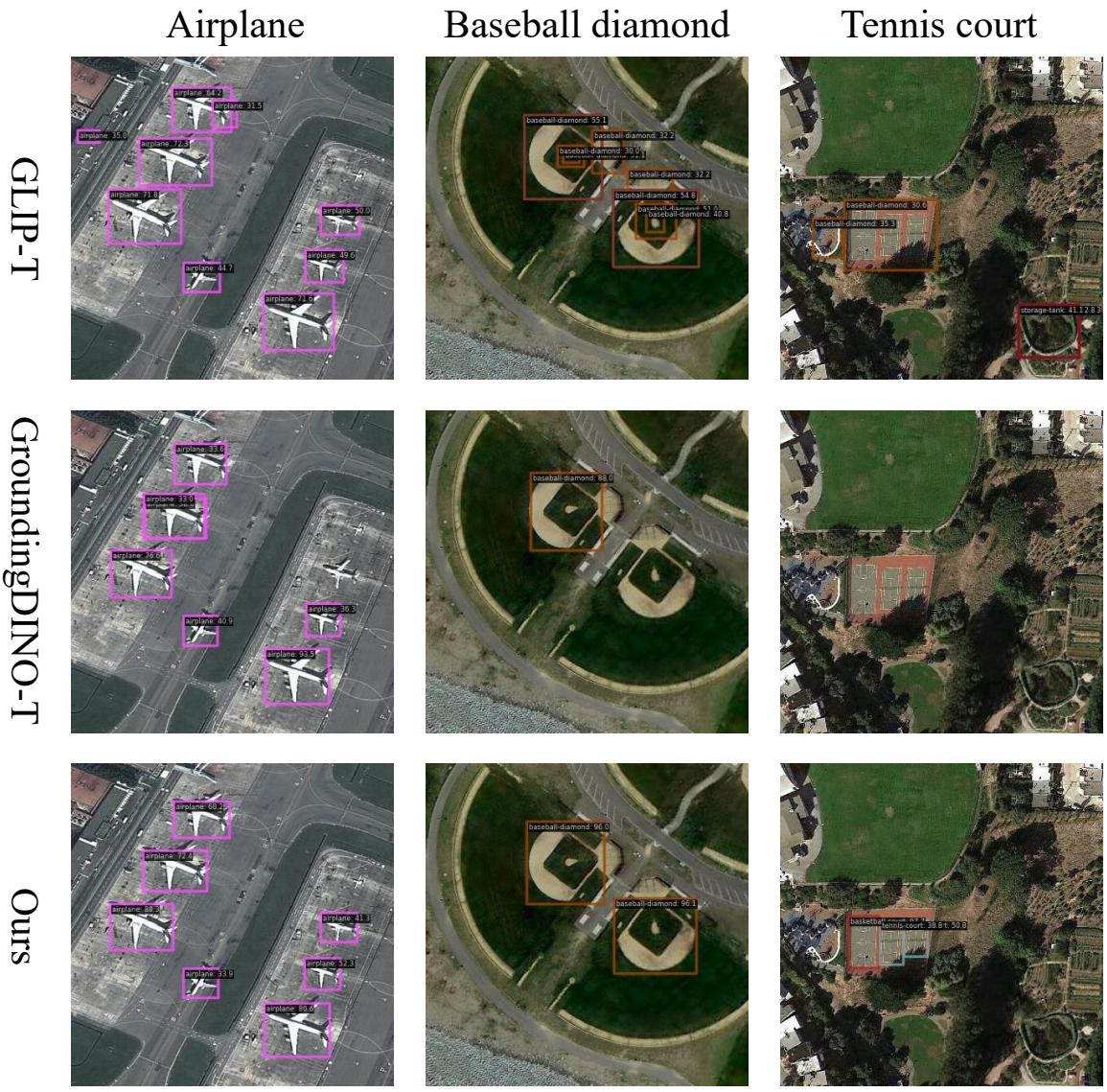

Figure 7. Novel class detection results of the NWPU-VHR10.v2 dataset under three-shot settings. Detection boxes with different colors represent different objects. The first row represents the results of GLIP-T [9], the second row represents the results of GroundingDINO-T [10], and the last row represents the detection results of our proposed method. Note that only the bounding boxes with a confidence greater than 0.3 are visualized.

To demonstrate the effectiveness of each proposed module in our approach, we also conduct extensive ablation studies. All experiments are performed on the DIOR test set in the first base/novel split or the NWPU VHR-10.v2 dataset.

Table 4. Ablation study of each component of our approach on the first split in the DIOR and NWPU VHR-10.v2 datasets.

Method	Components		DIOR				NWPU VHR-10			
	ETS	PAM	3	5	10	20	3	5	10	20
GroundingDINO			27.8	33.1	42.4	48.5	59.1	62.1	84.3	88.1
Ours	✓		38.8	44.7	50.1	53.6	75.2	85.5	86.9	88.4
Ours		✓	40.6	41.8	50.8	53.6	73.4	83.4	87.6	88.5
Ours	✓	✓	**42.1**	**46.4**	**53.1**	**55.4**	**77.9**	**85.3**	**88.5**	**88.8**

4.4.2. Ablation for the Performance of the Prototype Aggregating Module

To further analyze the properties of various prototypes, we conduct a comprehensive study on the first base/novel split of the DIOR dataset, as shown in Table 5. To enhance the feature extraction capability of detectors for aerial images, we have applied the ETS to all detectors in Table 5. As the table illustrates, on the one hand, in the condition of extremely limited shot number, the visual prototypes extracted by the model exhibited poor generalization capabilities, resulting in poor detection performance. Nonetheless, as the number of shots increases, the visual prototypes exhibit enhanced generalization capabilities, leading to a substantial improvement in the detection performance. On the other hand, owing to the robust generalization of textual prototypes, detectors based on textual prototypes can achieve high performance even with extremely limited training data. However, single-modal textual prototypes lack spatial details, which limits their detection performance. At last, by combining visual prototypes and textual prototypes, our proposed approach further enhances the detection performance, achieving a maximum improvement of 3.3%.

Table 5. Abalation study for the performance of different types of prototypes.

Visual Prototype	Textual Prototype	3-Shot	5-Shot	10-Shot	20-Shot
✓		21.3	27.0	42.6	50.6
	✓	38.8	44.7	50.1	53.6
✓	✓	**42.1**	**46.4**	**53.1**	**55.4**

In the above experiments, we find that the visual prototypes perform poorly when the number of training data points is extremely small, possibly due to the difficulty that vision encoders face in adequately extracting semantic features with limited training samples. Due to the rich prior knowledge of the text encoder, detectors based on textual prototypes demonstrate strong detection performance even with extremely limited training data. By combining the spatial information of visual prototypes into textual prototypes, we have achieved better detection performance. We have found that the spatial alignment between textual prototypes and visual prototypes is crucial. When visual prototypes do not retain spatial information or are not aligned with text prototypes, detectors often rely solely on textual prototypes to make predictions.

4.4.3. Ablation for the Performance of the Efficient Two-Stage Training Strategy

In order to analyze the importance of each component in the base training stage or fine-tuning stage, we conduct comprehensive ablation studies on the first base/novel set of the DIOR dataset, as demonstrated in Table 6.

As shown in the first row of Table 6, the detector designed for natural images exhibits poor detection performance in RSIs due to the large domain gap between natural images and aerial images, although it is claimed to be an open-set detector. Subsequently, we fine-tune the baseline detector on the few-shot dataset, and the detector's performance exhibits an improvement, as demonstrated in the second row of Table 6. However, in cases where the number of samples is extremely small, the performance of the detector is still not satisfactory. To further enhance the feature extraction capability of remote-sensing objects, in line with the previous training strategy [4,11], we performed a base training stage on a large-scale base dataset, and the improved detection performance is illustrated

in the third row of the table. It is noteworthy that no parameters are frozen during the base training phase. However, we argue that the text encoder is domain-agnostic, while the vision encoder is regarded as domain-specific. Simply training all parameters in the base training stage is inefficient and harms the generalization of the text encoder. Therefore, to further analyze the domain sensitivity of each component in the base training stage, we independently freeze the vision encoder and text encoder during the base training phase, as shown in the fourth and fifth rows of Table 6. Since the primary objective of base training is to enhance the feature extraction capability of the vision encoder, freezing the vision encoder during the base training stage will impair the detector's performance for remote sensing targets, as illustrated in the fourth row of the table. At last, by freezing the text encoder in the base training phase, our method can preserve the generalization of the text encoder and reduce the risk of overfitting to base classes, as demonstrated in the last row of Table 6. The experiment results also corroborate our perspective that the text encoder is domain-agnostic, whereas the vision encoder is domain-specific. Consequently, we believe that leveraging the extensive natural language corpora available on the Internet as prior knowledge can further enhance the capability of FSOD in RSIs.

Table 6. Ablation study for the performance of ETS. The ✗ indicates that the module is frozen during the training process, whereas the ✓ denotes that the module is unfrozen.

Method	Base Training Stage		Fine-Tuning Stage		3-Shot	5-Shot	10-Shot	20-Shot
	Vision Encoder	Text Encoder	Vision Encoder	Text Encoder				
Baseline	✗	✗	✗	✗	11.9	11.9	11.9	11.9
Baseline	✗	✗	✓	✓	21.3	33.1	42.4	48.5
Baseline	✓	✓	✓	✓	40.6	41.8	50.8	53.6
Baseline	✗	✓	✓	✓	35.9	39.2	49.2	51.6
Baseline	✓	✗	✗	✓	39.2	41.5	51.8	53.4
Baseline	✓	✗	✓	✗	41.7	45.7	51.1	54.4
Ours	✓	✗	✓	✓	**42.1**	**46.4**	**53.1**	**55.4**

In addition to conducting experiments during the base training stage, we have also experimented on whether it is necessary to freeze the text encoder and visual encoder during the fine-tuning stage. Although the vision encoder has undergone base training on base datasets, further fine-tuning the vision encoder during adaption stage can still yield some performance improvements during the fine-tuning stage, as shown in the fifth and last rows of Table 6. Since the text encoder already obtains sufficient prior semantic knowledge from a large-scale corpus, additional fine-tuning during this stage has only a negligible impact on performance.

4.4.4. Ablation for Efficiency

To further analyze the computational efficiency of our method, we perform comprehensive experiments on the DIOR dataset, as Table 7 illustrates. All experiments are conducted on a single NVIDIA GeForce RTX 4090 GPU and the training time includes the time of both the base training stage and the fine-tuning stage. It is worth noting that once the model training is complete, the fusion of visual and textual prototypes can be performed offline, thus exerting only a minimal impact on the FLOPs and FPS during inference. Furthermore, as the ETS does not introduce additional parameters and PAM employs a late fusion paradigm that only marginally increases the number of parameters, our method does not differ significantly in the scale of parameters from the baseline. Finally, through our proposed ETS, the model can achieve superior performance with reduced training time.

Table 7. Ablation study for efficiency. All experiments are conducted on one NVIDIA GeForce RTX 4090 GPU.

Method	ETS	PAM	Params	Training Time	FLOPs	FPS
Baseline			175 M	8.0 h	265 G	18.5
Ours	✓		175 M	7.5 h	265 G	18.5
Ours		✓	176 M	11.7 h	268 G	18.0
Ours	✓	✓	176 M	10.9 h	268 G	18.0

4.4.5. Ablation Study for the Construction of Visual Prototypes

We also conduct experiments on the construction method of visual prototypes, as shown in Table 8. We first randomly select a feature from the support features as the visual feature. To reduce randomness, we repeatedly conducted the experiment three times, selecting a different support feature each time and reporting the average results in the first row of Table 8. Subsequently, we present the experimental results obtained by directly employing the mean of all support features and utilizing the features aligned according to Equation (5) as visual prototypes. As the table demonstrates, the performance of randomly sampled visual prototypes is unstable since it is influenced by the quality of the sampled feature. In addition, spatial feature alignment also helps to improve the performance of the few-shot detector.

Table 8. Ablation study for the construction of visual prototypes.

Method	wo/align	w/align	3-Shot	5-Shot	10-Shot	20-Shot
random select	✓		41.8	45.6	52.9	54.5
mean	✓		42.0	45.6	52.6	54.9
Ours		✓	42.1	46.4	53.1	55.4

4.4.6. Failure Cases and Analysis

Although our proposed MP-FSDet has achieved better performance compared with previous methods, there is still room for improvement. We carefully studied the limitations of our method and elaborated on three possible reasons.

(a) The dense distribution of objects: As shown in the first column of Figure 8, the method fails to detect densely distributed airplanes in the airport. Due to the DETR-like methods [10,16,18] employing a fixed number of content queries for matching with feature maps, these approaches perform inadequately when detecting dense objects. To alleviate the failure, it is possible to consider integrating some iterative methods like [50] with our methods to progressively detect dense objects.

(b) Poor feature extraction capability for novel classes: Due to the limited number of objects for the novel classes, the feature extraction capability of the detector is insufficient to detect some hard examples. However, current feature extractors [51,52] exhibit robust feature extraction for remote-sensing objects. Utilizing a more robust feature extractor could potentially enhance the performance of few-shot novel classes.

(c) Large inter-class similarity between classes: As demonstrated by the third column of Figure 8, some classes in the dataset exhibit significant inter-class similarity. High inter-class similarity can result in some objects being misclassified as similar classes, which leads to performance degradation. Some potential works like [53] can be further utilized to address the issue.

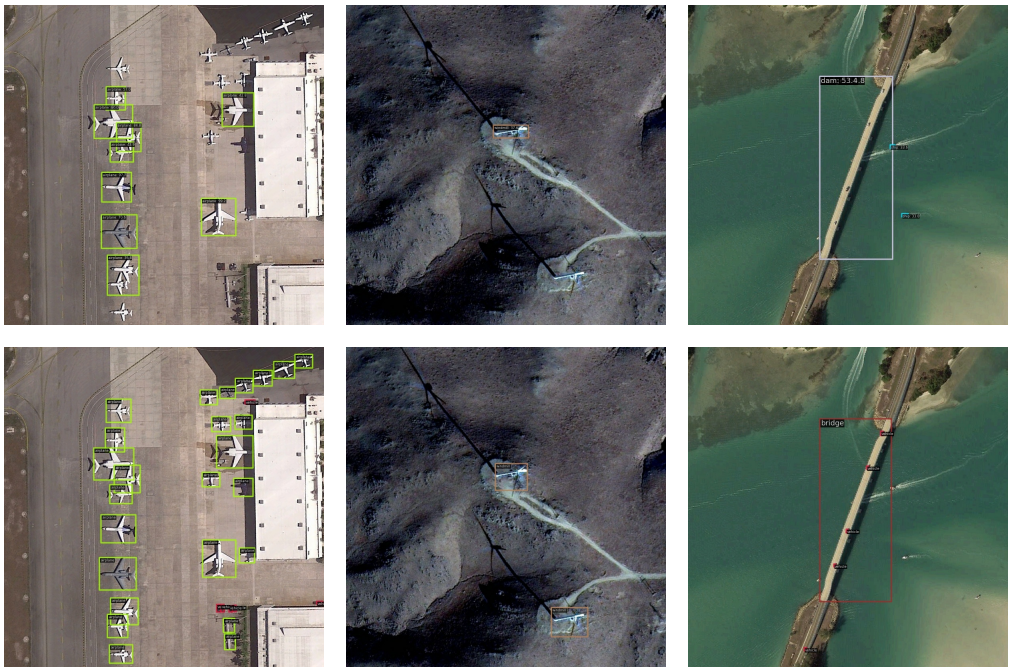

Figure 8. Failure cases in the DIOR test set for the first base/novel split under the 10-shot setting. The top row of the figure shows the predictions of our method, while the bottom row shows the ground truths.

5. Discussion

Textual prototypes generated from class names can inherit prior knowledge from a large-scale pre-trained language model. However, in RSIs, many of the distinctions between classes are very subtle. Consequently, textual prototypes that are generated directly from class names may not provide sufficient discriminative features. Therefore, our future work will attempt to use both class descriptions and class names to enhance the detection performance between fine-grained classes.

6. Conclusions

In this article, we propose a novel construction of multi-modal prototypes, termed the prototype aggregating module (PAM), and apply these multi-modal prototypes to a state-of-the-art detector. Multi-modal prototypes simultaneously leverage the advantages of text prototypes and visual prototypes by integrating the generalization properties of text prototypes with the spatial and detailed information of objects in aerial images. Furthermore, we investigate the transferability of pre-trained multi-modal object detectors in remote sensing scenarios. Our comprehensive experimental analyses reveal that the text encoder exhibits domain independence, while the vision coder demonstrates domain dependence. Consequently, we propose an efficient two-stage training strategy (ETS) that can mitigate the risk of overfitting in the text encoder while reducing the overall training time. The experimental results demonstrate that our proposed ETS achieves an average performance improvement of 2.55% compared to previous training strategies. By integrating multi-modal prototypes with the ETS, our proposed approach can further enhance the detection performance by an average of 2.5%.

Author Contributions: Conceptualization, Y.L. and Z.P.; methodology, Y.L., J.Y., and P.Z.; resources, Z.P. and B.Z.; writing—original draft preparation, Y.L. and P.Z.; and supervision, Z.P. and B.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Youth Innovation Promotion Association, CAS, under number 2022119.

Data Availability Statement: The data presented in this study are openly available in https://github.com/YanxingLiu/RS-Datasets/tree/main/DIOR (accessed on 3 November 2024) at https://doi.org/10.1016/j.isprsjprs.2019.11.023.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chen, H.; Wang, Y.; Wang, G.; Qiao, Y. Lstd: A low-shot transfer detector for object detection. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
2. Wang, X.; Huang, T.E.; Darrell, T.; Gonzalez, J.E.; Yu, F. Frustratingly simple few-shot object detection. *arXiv* **2020**, arXiv:2003.06957.
3. Sun, B.; Li, B.; Cai, S.; Yuan, Y.; Zhang, C. Fsce: Few-shot object detection via contrastive proposal encoding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 7352–7362.
4. Zhang, T.; Zhang, X.; Zhu, P.; Jia, X.; Tang, X.; Jiao, L. Generalized few-shot object detection in remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2023**, *195*, 353–364. [CrossRef]
5. Kang, B.; Liu, Z.; Wang, X.; Yu, F.; Feng, J.; Darrell, T. Few-shot object detection via feature reweighting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8420–8429.
6. Yan, X.; Chen, Z.; Xu, A.; Wang, X.; Liang, X.; Lin, L. Meta r-cnn: Towards general solver for instance-level low-shot learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9577–9586.
7. Xiao, Y.; Marlet, R. Few-Shot Object Detection and Viewpoint Estimation for Objects in the Wild. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020.
8. Cheng, G.; Yan, B.; Shi, P.; Li, K.; Yao, X.; Guo, L.; Han, J. Prototype-CNN for few-shot object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–10. [CrossRef]
9. Li, L.H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.N.; et al. Grounded language-image pre-training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 10965–10975.
10. Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv* **2023**, arXiv:2303.05499.
11. Lu, X.; Sun, X.; Diao, W.; Mao, Y.; Li, J.; Zhang, Y.; Wang, P.; Fu, K. Few-shot object detection in aerial imagery guided by text-modal knowledge. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–19. [CrossRef]
12. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1137–1149. [CrossRef]
13. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
14. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
15. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767. [CrossRef]
16. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
17. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
18. Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.M.; Shum, H.Y. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv* **2022**, arXiv:2203.03605.
19. Zareian, A.; Rosa, K.D.; Hu, D.H.; Chang, S.F. Open-vocabulary object detection using captions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 14393–14402.
20. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 8748–8763.
21. Gu, X.; Lin, T.Y.; Kuo, W.; Cui, Y. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv* **2021**, arXiv:2104.13921.
22. Xu, Y.; Zhang, M.; Fu, C.; Chen, P.; Yang, X.; Li, K.; Xu, C. Multi-modal queried object detection in the wild. *Adv. Neural Inf. Process. Syst.* **2024**, *36*.
23. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [CrossRef]
24. Mei, S.; Jiang, R.; Ma, M.; Song, C. Rotation-invariant feature learning via convolutional neural network with cyclic polar coordinates convolutional layer. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–13. [CrossRef]

25. Deng, Z.; Sun, H.; Zhou, S.; Zhao, J.; Lei, L.; Zou, H. Multi-scale object detection in remote sensing imagery with convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 3–22. [CrossRef]
26. Cheng, G.; Zhou, P.; Han, J. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 7405–7415. [CrossRef]
27. Li, Q.; Mou, L.; Liu, Q.; Wang, Y.; Zhu, X.X. HSF-Net: Multiscale deep feature embedding for ship detection in optical remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 7147–7161. [CrossRef]
28. Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; Fu, K. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8232–8241.
29. Xu, Y.; Fu, M.; Wang, Q.; Wang, Y.; Chen, K.; Xia, G.S.; Bai, X. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 1452–1459. [CrossRef]
30. Qiao, L.; Zhao, Y.; Li, Z.; Qiu, X.; Wu, J.; Zhang, C. Defrcn: Decoupled faster r-cnn for few-shot object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 8681–8690.
31. Fan, Q.; Zhuo, W.; Tang, C.K.; Tai, Y.W. Few-shot object detection with attention-RPN and multi-relation detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 4013–4022.
32. Lyu, Q.; Wang, W. Compositional prototypical networks for few-shot classification. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 9011–9019.
33. Li, X.; Deng, J.; Fang, Y. Few-shot object detection on remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14. [CrossRef]
34. Le Jeune, P.; Lebbah, M.; Mokraoui, A.; Azzag, H. Experience feedback using representation learning for few-shot object detection on aerial images. In Proceedings of the 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), Virtual, 13–16 December 2021; IEEE: New York, NY, USA, 2021; pp. 662–667.
35. Huang, Y.; Peng, J.; Sun, W.; Chen, N.; Du, Q.; Ning, Y.; Su, H. Two-Branch Attention Adversarial Domain Adaptation Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [CrossRef]
36. Wolf, S.; Meier, J.; Sommer, L.; Beyerer, J. Double Head Predictor based Few-Shot Object Detection for Aerial Imagery. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 721–731.
37. Chen, J.; Qin, D.; Hou, D.; Zhang, J.; Deng, M.; Sun, G. Multiscale Object Contrastive Learning-Derived Few-Shot Object Detection in VHR Imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
38. Azeem, A.; Li, Z.; Siddique, A.; Zhang, Y.; Zhou, S. Unified multimodal fusion transformer for few shot object detection for remote sensing images. *Inf. Fusion* **2024**, *111*, 102508. [CrossRef]
39. Liu, N.; Xu, X.; Celik, T.; Gan, Z.; Li, H.C. Transformation-invariant network for few-shot object detection in remote-sensing images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–14. [CrossRef]
40. Liu, Y.; Pan, Z.; Yang, J.; Zhang, B.; Zhou, G.; Hu, Y.; Ye, Q. Few-Shot Object Detection in Remote-Sensing Images via Label-Consistent Classifier and Gradual Regression. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–14. [CrossRef]
41. Zheng, Z.; Yang, L.; Wang, Y.; Zhang, M.; He, L.; Huang, G.; Li, F. Dynamic spatial focus for efficient compressed video action recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *34*, 695–708. [CrossRef]
42. Yang, L.; Zheng, Z.; Wang, J.; Song, S.; Huang, G.; Li, F. Adadet: An adaptive object detection system based on early-exit neural networks. *IEEE Trans. Cogn. Dev. Syst.* **2023**, *16*, 332–345. [CrossRef]
43. Yang, L.; Jiang, H.; Cai, R.; Wang, Y.; Song, S.; Huang, G.; Tian, Q. Condensenet v2: Sparse feature reactivation for deep networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 3569–3578.
44. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
45. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
46. Li, F.; Zhang, H.; Liu, S.; Guo, J.; Ni, L.M.; Zhang, L. Dn-detr: Accelerate detr training by introducing query denoising. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 13619–13627.
47. Cheng, G.; Han, J.; Zhou, P.; Guo, L. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS J. Photogramm. Remote Sens.* **2014**, *98*, 119–132. [CrossRef]
48. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
49. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
50. Rukhovich, D.; Sofiiuk, K.; Galeev, D.; Barinova, O.; Konushin, A. Iterdet: Iterative scheme for object detection in crowded environments. In Proceedings of the Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshops, S+ SSPR 2020, Padua, Italy, 21–22 January 2021; Proceedings; Springer: Berlin/Heidelberg, Germany, 2021; pp. 344–354.

51. Wang, D.; Zhang, J.; Xu, M.; Liu, L.; Wang, D.; Gao, E.; Han, C.; Guo, H.; Du, B.; Tao, D.; et al. MTP: Advancing remote sensing foundation model via multi-task pretraining. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**. [CrossRef]
52. Jiang, W.; Zhang, J.; Wang, D.; Zhang, Q.; Wang, Z.; Du, B. LeMeViT: Efficient vision transformer with learnable meta tokens for remote sensing image interpretation. *arXiv* **2024**, arXiv:2405.09789.
53. Wang, Z.; Yang, B.; Yue, H.; Ma, Z. Fine-Grained Prototypes Distillation for Few-Shot Object Detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 5859–5866.