FSOD4RSI: Few-Shot Object Detection for Remote Sensing Images via Features Aggregation and Scale Attention

Honghao Gao, *Senior Member, IEEE*, Shuping Wu, Ye Wang, Jung Yoon Kim, and Yueshen Xu, *Member, IEEE*

***Abstract*—Due to the continuous development of few-shot learning, there have been notable advancements in methods for few-shot object detection in recent years. However, most existing methods in this domain primarily focus on natural images, neglecting the challenges posed by variations in object scales, which are usually encountered in remote sensing images. This article proposes a new few-shot object detection model designed to handle the issue of object scale variation in remote sensing images. Our developed model has two essential parts: a feature aggregation module (FAM) and a scale-aware attention module (SAM). Considering the few-shot features of remote sensing images, we designed the FAM to improve the support and query features through channel multiplication operations utilizing a feature pyramid network and a transformer encoder. The created FAM better extracts the global features of remote sensing images and enhances the significant feature representation of few-shot remote sensing objects. In addition, we design the SAM to address the scale variation problems that frequently occur in remote sensing images. By employing multiscale convolutions, the SAM enables the acquisition of contextual features while adapting to objects of varying scales. Extensive experiments were conducted on benchmark datasets, including NWPU VHR-10 and DIOR datasets, and the results show that our model indeed addresses the challenges posed by object scale variation and improves the applicability of few-shot object detection in the remote sensing domain.**

***Index Terms*—Attention mechanism, feature aggregation, few-shot learning, object detection, remote sensing images.**

I. Introduction

THE rapid development of remote sensing technology in recent decades has led to widespread application in various domains, such as environmental monitoring [1], traffic management [2], and urban planning [3]. Within this context, object detection has emerged as a crucial image-processing technique and remains a prominent research focus in remote sensing. Object detection in remote sensing images involves the automated identification and localization of specific targets of interest within the imagery. This technology has proven invaluable in diverse applications, such as natural disaster detection [4] and ship detection [5], [6]. Initially, object detection in remote sensing heavily relied on traditional methods [7], such as template matching [8], [9], expert knowledge [10], [11], and object-based image analysis [12]. These methods predominantly rely on manually crafted features. However, with the limitations of performance produced by manual feature-based approaches, there has been a shift toward the development of deep learning-based methods [13], [14]. Nonetheless, deep learning-based approaches are limited in two aspects. First, deep learning models often require substantial amounts of labeled data for effective training, resulting in the high cost associated with data acquisition. Second, the scale of remote sensing images is considerably smaller than that of natural images, which may lead to overfitting and performance degradation when employing traditional deep learning models. Therefore, exploring innovative methodologies that address these challenges and enhancing the applicability of deep learning-based techniques for remote sensing object detection is necessary.

Few-shot learning [15] has emerged as a crucial approach to improve the performance of object detection when training data are limited. Its fundamental technique involves training a detection model on a base dataset, which produces considerable labeled data, thus enabling the identification of novel image classes with minimal or no labeled data. The existing few-shot object detection (FSOD) methodologies primarily encompass metalearning-based methods [16], transfer learning-based methods [17], and metric learning-based methods [18]. Although these methods have demonstrated promising results in natural image scenarios, they fail to learn the distinctive features of remote sensing images. Consequently, the following factors prevent the use of these methods for remote sensing images.

1) *Object Scale:* Due to the wide variety of altitudes recorded in remote sensing images, the same target may appear at different size scales.

Manuscript received 8 October 2023; revised 17 December 2023 and 11 January 2024; accepted 30 January 2024. Date of publication 6 February 2024; date of current version 20 February 2024. This work was supported by the National Key Research and Development Program of China under Grant 2020YFB1006003. *(Corresponding authors: Yueshen Xu; Jung Yoon Kim.)*

Honghao Gao is with the School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China, and also with the College of Future Industry, Gachon University, Seongnam, 13120, South Korea (e-mail: gaohonghao@shu.edu.cn).

Shuping Wu and Ye Wang are with the School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China (e-mail: wushuping@shu.edu.cn; wangye1994@126.com).

Jung Yoon Kim is with the College of Future Industry, Gachon University, Seongnam, 13120, South Korea (e-mail: kjyoon@gachon.ac.kr).

Yueshen Xu is with the School of Computer Science and Technology, Xidian University, Xi'an 710126, China (e-mail: ysxu@xidian.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3362748

2) *Object Pixels:* Many objects are difficult to detect accurately due to the small number of pixels, resulting in severe missed detections. Images captured through remote sensing may show dense clusters of several classes of objects, and typical examples include cars, planes, and fuel tanks.
3) *Object Range:* Because remote sensing photographs typically have a wider field of view, object detection is more difficult than natural image detection.

Considering these challenges, this study proposes a new few-shot learning-based approach for object detection in remote sensing images. The proposed method comprises two stages: the first stage is the base class training stage, and the second stage is the novel class finetuning stage. Our model is trained on substantial annotated base class data during the base class training stage, enabling it to acquire substantial prior knowledge. Consequently, during the subsequent finetuning stage, our model requires training solely on the new class of data, which contains few annotations, allowing the model to swiftly converge. The contributions of this work can be summed up, as shown in the following.

1) First, we designed the feature aggregation module (FAM) to enhance the representation of salient features of few-shot objects in remote sensing images. In this module, query features and support features are encoded and channel-multiplied using the transformer encoder. On this basis, support features are embedded into query features to better learn salient features.
2) Second, we designed a scale-aware attention module (SAM) to enable the model to perceive the issue of scale variation in remote sensing images. The network is guided to concentrate on regions with more information at proper image feature scales by convolving with different scales to obtain and aggregate contextual features.
3) Third, the Soft-NMS (nonmaximum suppression) algorithm is introduced in the postprocessing stage of object detection to help detect dense features in remote sensing images. This algorithm can avoid deleting occluded objects in dense images and effectively address the issue of missed detections caused by other ordinary NMS algorithms. Our method achieved significant improvements in performance on the DIOR dataset and the NWPU VHR-10 dataset, especially in the few-shot scenario.

The rest of this article is organized as follows. Section II introduces related works. Section III introduces the proposed few-shot remote sensing image object detection model. Section IV describes the experiments and gives the analysis of experimental results. Section V discusses the performance of the proposed method on both base and new classes. Finally, Section VI concludes this article.

II. Related Work

A. Object Detection for Remote Sensing Images Based on Deep Learning

Object detection is a research hotspot in the remote sensing domain and has broad application prospects. The effectiveness of object detection for remote sensing images has recently improved due to the development of deep learning, particularly the potent feature extraction capabilities of convolutional neural networks (CNNs). A number of deep learning-based methods have also been developed. Most of early researches were based on the region-based convolutional neural network (RCNN) architecture for detecting remote sensing images, and after achieving success, researchers also developed many regression-based methods.

- *RCNN-based approaches:* Zhang et al. [19] proposed two models, i.e., the pyramid local context network and the global context network (Get), to help the neural network extract relevant contextual features on the object of interest information. Wu et al. [20] attempted to address the problem of the high cost of manual annotations for remote sensing images. The authors attempted to increase the response strength of low response regions in the shallow feature maps and to improve the feature distribution of the shallow feature maps. The authors added a divergent activation module and a similarity module to a neural network model. Pang et al. [21] developed a neural network model with autonomous enhancement consisting of a lightweight residual backbone as well as classifiers and detectors to enhance the network's capacity for small target detection and computational efficiency.
- *Regression-based approaches:* Cai et al. [22] designed an unanchored target detection framework for remote sensing images, including a cross-channel feature pyramid network (CFPN) and foreground attention detection heads (FDHs). CFPN can deal with a wide range of target sizes in remote sensing images, and FDHs can enhance foreground features in remote sensing images and reduce interference from complicated background information. Zhang et al. [23] proposed the contextual bidirectional enhancement approach to remove irrelevant background information from remote sensing images.

However, existing deep learning-based approaches for object detection in remote sensing images always depend on a large volume of training annotation samples; moreover, the acquisition of remote sensing images is usually difficult, and manual annotation typically requires professional expertise. Therefore, in the case of insufficient training data, the performance of standard deep learning-based object detection techniques can easily be hampered by overfitting.

B. Object Detection for Remote Sensing Images Based on Few-Shot Learning

FSOD has not been widely studied in remote sensing, in contrast to previous works on the subject of natural photographs. The existing FSOD methods for remote sensing images can be classified into one-stage approaches and two-stage approaches.

- *One-stage approaches:* Li et al. [24] proposed the first FSOD method in remote sensing images. Their approach focused on addressing the inherent scale diversity of remote sensing images by incorporating a multiscale mechanism. Zhou et al. [25] designed a lightweight feature extractor and

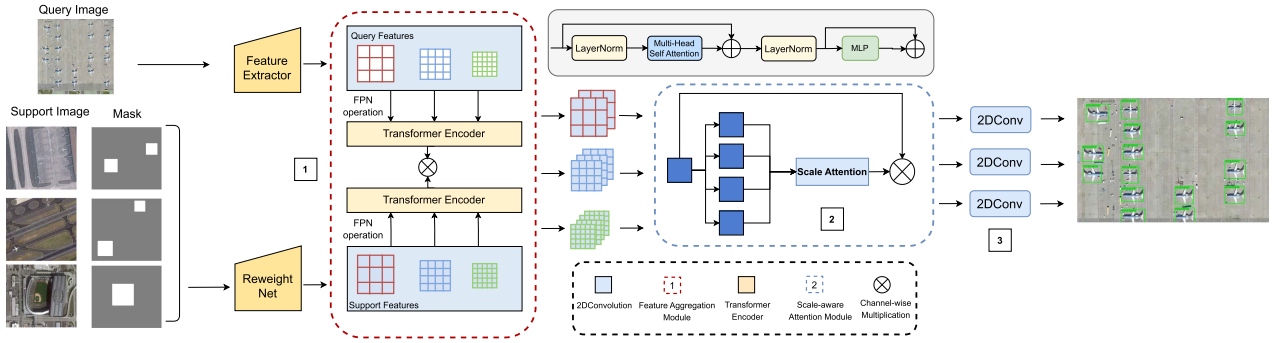

Fig. 1. Architecture of the developed FSOD model for images from remote sensing. The FAM represents the proposed transformer-based FAM, which aggregates query features and support features. The SAM represents the proposed scale-aware attention mechanism, which directs the neural network to concentrate on more information areas at the right feature scales.

a aggregation module and applied the model to synthetic aperture radar images.

- *Two-stage approaches:* Xiao et al. [26] proposed an FSOD approach based on adaptive attention networks and proposed a relational gate loop unit to obtain object-level relationships. Zhao et al. [27] constructed a path aggregation module and a feature pyramid based on a two-level finetuned object detection approach to solve the scale variation issue for remote sensing images. Cheng et al. [28] extended the meta-RCNN model [29] using the designed prototype-guided predictor-head remodeling network (P-G RPN) in place of the generic RPN, where the P-G RPN employed a class prototype with many fully connected layers. The output was used to determine the weight of the convolution layer to attach complementary classifiers. Zhou et al. [30] developed two modules: the context-aware pixel aggregation module, which employs convolutions at various scales to adjust to objects at various scales, and the context-aware feature aggregation, which employs the graph convolutional network to obtain more semantic information by enhancing context awareness. Zhang et al. [31] proposed the application of the self-adaptive global similarity (SAGS) and two-way foreground stimulator (TFS) modules in the FSOD model. SAGS computes the similarity between queries and supporting images while preserving the spatial information of the supporting images. TFS utilizes a bidirectional attention mechanism to mine the hidden information in the supporting image.

The majority of existing FSOD techniques ignore the significance of spatial and contextual information for remote sensing images. Furthermore, those methods that use a two-stage procedure may offer better accuracy than their single-stage counterparts; however, they also suffer from reduced detection speed, hampering their practicality in real-time applications.

III. Developed Approach

The objective of this study is to augment the salient feature representation of few-shot objects to utilize support features more effectively. In addition, we address the inherent multiscale characteristics of remote sensing images by designing a scale-aware attention mechanism. Fig. 1 presents the framework with two key components: the FAM and SAM. In this section, we first illustrate an overview of our methodology and then elaborate upon the designed components.

A. Overall Architecture

The model's overall architecture is composed of three parts, as shown in Fig. 1: Module 1 is the FAM, Module 2 is the SAM, and Module 3 is the target detection postprocessing part. The FAM is an integration module that combines a transformer [32] and a feature pyramid network (FPN) [33] to aggregate multiscale support features and query features. By employing the transformer encoder, the FAM jointly encodes support features and query features, leveraging channel multiplication to enhance the aggregation of these features. Extracting salient features, which are essential for object detection, is facilitated by this method. In the SAM, diverse context information is acquired through distinct convolution operations. This enables the aggregation of the obtained contextual details, which guides the network to adapt to objects of varying scales and emphasizes regions containing richer information. By adapting to different scale contexts, the SAM enhances the discriminative ability of the model. The object detection postprocessing step incorporates the Soft-NMS algorithm [34], which is tailored to address the characteristics of dense targets commonly encountered in remote sensing scenes. This algorithm mitigates the issue of missing target detection, thus improving both the detection and recognition rates of objects in remote sensing images. Overall, the developed method effectively combines the FAM, SAM, and postprocessing parts, enabling robust object detection for remote sensing images.

Given a query image Q_i as the input to feature extractor D to generate the query feature q_i and given a set of support images (with annotations) $S_i = (I_j, M_j)$, I_j represents the support image, and M_j represents the bounding box annotation. By highlighting the target area in white on the mask, the target can be precisely located, and its position and shape can be extracted. Simultaneously, masks can mask the areas of interest in an image. S_i is input to the reweight net M to obtain the

TABLE I
REWEIGHT NET STRUCTURE

Index	Type	Filters	Size	Stride	Output
1	Convolutional	32	3×3	1	512×512
2	Max-pooling		2×2	2	256×256
3	Convolutional	64	3×3	1	256×256
4	Max-pooling		2×2	2	128×128
5	Convolutional	128	3×3	1	128×128
6	Max-pooling		2×2	2	64×64
7	Convolutional	256	3×3	1	64×64
8	Max-pooling		2×2	2	32×32
9	Convolutional	512	3×3	1	32×32
10	Max-pooling		2×2	2	16×16
11	Convolutional	1024	3×3	1	16×16

support features v_i. The computation is shown in (1) and (2)

$$q_i = D(Q_i) \tag{1}$$

$$v_i = M(S_i). \tag{2}$$

Table I gives the network architecture of the reweight net M. "Convolutional" refers to a 2-D convolutional layer. "Filters" is the number of convolutional filters. "Size" represents a convolutional kernel's spatial dimensions as "kernel height × kernel width." "Stride" represents the step size of the convolutional kernel moving on the image; "Max-pooling" represents the max-pooling layer.

Then, the query features and support features are input to the FAM for feature aggregation to obtain the aggregated features F_A, where FAM$(\cdot)$ represents the FAM module

$$F_A = \text{FAM}(q_i, v_i). \tag{3}$$

The obtained aggregated features are subsequently subjected to the SAM for scale perception to obtain F'_A, where SAM$(\cdot)$ represents the SAM module

$$F'_A = \text{SAM}\left(F_A\right). \tag{4}$$

Finally, the prediction layer is the input for detection according to (5), where Det$(\cdot)$ represents the target classification and bounding box regression

$$O = \text{Det}\left(F'_A\right). \tag{5}$$

B. Feature Aggregation Module

In the context of FSOD for remote sensing images, the ability of models to identify novel classes is often compromised due to the lack of available labeled training data. Therefore, effectively harnessing the information contained within existing data becomes a crucial challenge. Traditional object detection methods typically rely on the feature map of the final convolutional layer for prediction, limiting their ability to detect smaller targets, which widely exist in remote sensing images.

To address the aforementioned challenges, we created a FAM based on an FPN and leveraged the Transformer encoder to aggregate support features and query features. Specifically, the query features correspond to the metafeatures extracted from the query image by the feature extractor. Support features consist of support image information extracted by the reweight net from the support images that contain labels and masks. Fig. 2 illustrates

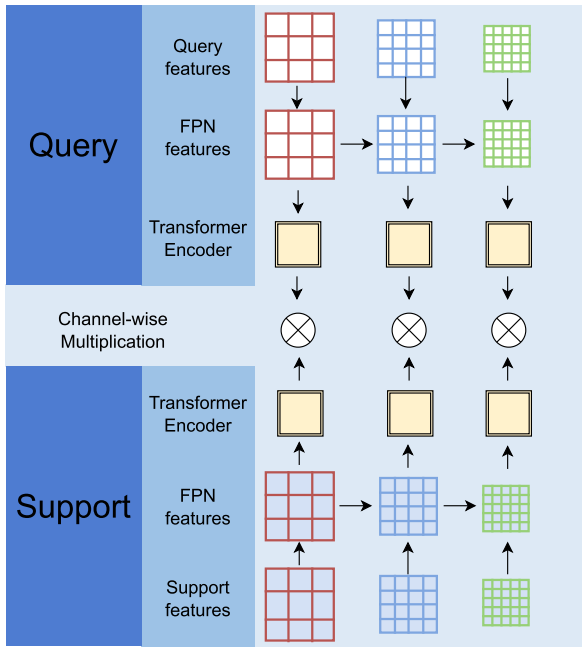

Fig. 2. Structure of the FAM.

the workflow of the feature aggregation process. Initially, the query features and support features acquired from the feature extractor undergo FPN processing. This step yields the support FPN features and query FPN features on three different scales. Subsequently, to enhance those features, the support features, query features, and post-FPN are sent to the transformer encoder for encoding. The structure of the transformer encoder consists of two sublayers: a multihead attention layer and a fully connected layer (i.e., a multilayer perceptron or MLP). Residual connections are used between each sublayer. The transformer encoder boosts the capacity to gather global information in addition to utilizing the self-attention method to tap the feature representation potential. Ultimately, the enhanced query features and support features are subject to a channel multiplication operation at each scale, facilitating the integration of information across the feature maps. This operation enables the effective fusion of query features and support features, leading to enhanced representations for subsequent tasks.

Algorithm 1 shows the outline of the FAM. Upon receiving the input support set and query set, the feature extraction procedure yields query features $q_i \epsilon R^{C\times H_q\times W_q}$ from the feature extractor, while the reweight net produces support features $v_i \epsilon R^{N\times C\times H_s\times W_s}$. Here, N represents the number of supported image object classes, where the supported image classes are determined by the selected base class. C signifies the number of channels; H_q and W_q represent the height and width of the query feature map, respectively; and H_s and W_s denote the height and width of the support feature map, respectively. In lines 2–8, the FPN [33] operation is demonstrated, where query features q_i and support features v_i from the FAM are the inputs. Multiscale feature fusion is performed through a feature pyramid to obtain q'_i and v'_i. In lines 9–10, the encoding operation is illustrated, utilizing the transformer encoder to

Algorithm 1: Feature Aggregation Module Algorithm.

input : Support category N of images, query meta features q, support features v_j, $j\epsilon\{1,2...,N\}$, feature map scale $i\epsilon\{1,2,3\}$, convolution layer $Conv()$, upsample layer $Upsample()$, Transformer encoder E_n;

output: Aggregated Features F_A;

1 **for** $i\epsilon\{1,2,3\}$ **do**
2 **if** i = *1* **then**
3 $q_i^{'} = Conv(q_i)$;
4 $v_i^{'} = Conv(v_i)$;
5 **else**
6 $q_i^{'} = Conv(q_i) + Upsampple(q_{i-1}^{'})$;
7 $v_i^{'} = Conv(v_i) + Upsampple(v_{i-1}^{'})$;
8 **end**
9 $q_i^E = E_n(q_i^{'})$;
10 $v_i^E = E_n(v_i^{'})$;
11 **for** $j\epsilon\{1,2...,N\}$ **do**
12 $F_A = q_i^E \bigotimes v_{ij}^E$
13 **end**
14 **end**

derive encoded query features q_i^E and encoded support features v_i^E. These encoded features are denoted as $q_i^E = E_n(q_i')$ and $v_i^E = E_n(v_i')$, respectively. The function E_n represents the transformer encoder, which applies a series of encoding transformations to the input features q_i' and v_i' obtained through multiscale feature fusion. The resulting encoded features q_i^E and v_i^E embody the enriched representations of query features and support features, respectively. In lines 11–13, the channel multiplication procedure is delineated. This step involves leveraging the encoded query features and support features to execute reweighting operations, resulting in the attainment of the aggregated feature F_A. The channel multiplication operation combines the enriched representations of query features and support features, facilitating the integration of feature information from the query and support images. By fusing these features through the reweighting process, the aggregated feature F_A is formed, which encapsulates the consolidated knowledge derived from the query and support sets.

$$F_A = q_i^E \bigotimes v_{ij}^E \quad , i = 1, 2, 3 \quad \text{and} \quad j = 1, 2, \ldots, N. \tag{6}$$

The symbol $\bigotimes$ denotes the channelwise multiplication operation, wherein the aggregated feature F_A is obtained through this operation. It should be noted that the encoded query features q_i^E and support features v_{ij}^E possess an equal number of channels. The feature extractor and the reweighting module (i.e., reweight net) can be jointly optimized as a result of the incorporation of the FAM into the training process. This simultaneous training scheme facilitates the acquisition of meaningful reweighted features. Moreover, the transformer encoder plays a pivotal role by leveraging the self-attention mechanism to explore potential feature representations and addressing the inherent limitation of CNNs in capturing global information. Thus, the transformer encoder enhances the overall feature extraction performance by incorporating global contextual information, augmenting the capacity to learn comprehensive and contextually rich representations.

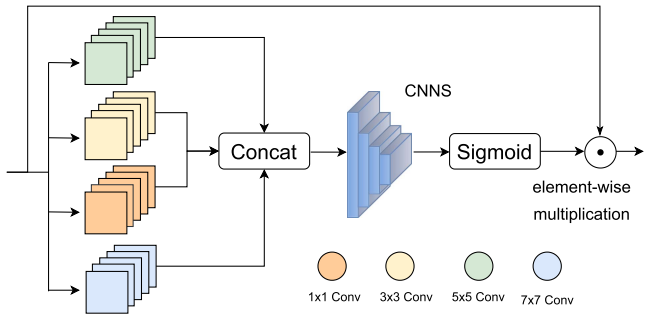

Fig. 3. SAM structure. The input features are concatenated after convolutions 1×1, 3×3, 5×5, and 7×7. The concatenated features are scaled for attention and connection to the initial features by the activation function.

C. Scale-Aware Attention Module

Visual attention has emerged as a valuable tool across various computer vision applications, including scene text recognition and image captioning. This concept draws inspiration from the remarkable capabilities of the human visual system, which adeptly scans an entire image, swiftly identifying areas that demand attention. By leveraging attention mechanisms, detailed information about the target can be obtained, effectively suppressing irrelevant information from surrounding regions. Pixel attention can assist in extracting additional semantic information, which helps to mitigate the issue of inadequate learnable knowledge resulting from insufficient data volume. However, because target objects in remote sensing images differ greatly in scale, standard pixel attention cannot accurately extract target information at all scales.

Inspired by the visual cortex of primates, Serre et al. [35] utilized convolutional kernels of different sizes to handle multiscale problems. In GoogLeNet [36], the authors also used the same strategy to design the network structure. Moreover, the author believes that this design concept also aligns with real-world visual perception, which involves processing visual information on different scales and then combining these processed results.

Therefore, we construct a SAM that allows the network to learn semantic information at multiple scales. By employing independent convolution kernels operating at different scales, the module extracts features from the input image. Fig. 3 provides an overview of the complete process implemented by the SAM. The input to the SAM consists of aggregated features $F_A \epsilon R^{H \times W \times C}$ that are obtained from the aggregation module FAM across three distinct scales. To enhance context awareness and accommodate varying levels of generalized semantic information, the features $F_{A_i}(i\epsilon[1,3])$ from each scale are processed through convolutional kernels of sizes 1×1, 3×3, 5×5, and 7×7. Subsequently, a concatenation operation merges the resulting feature maps, which are computed as follows:

$$X_i = \text{Conv}_{1\times1}(F_{A_i}) + \text{Conv}_{3\times3}(F_{A_i}) + \text{Conv}_{5\times5}(F_{A_i}) + \text{Conv}_{7\times7}(F_{A_i}) \tag{7}$$

where the input features are indicated by F_{A_i}, and the output features are indicated by X_i. The features $A_i \epsilon R^{1\times1\times C}$ are obtained by X_i after the attention map calculation and *sigmoid* operation. The original features are multiplied by the processed features A_i to acquire the features after scale-aware attention

$$A_i = \sigma[\varphi(X_i)] \tag{8}$$

$$S_i = A_i \bigodot F_{A_i}. \tag{9}$$

In this context, the *sigmoid* function $\sigma(\cdot)$ is employed for the transformation, while the elementwise multiplication is denoted by $\bigodot$. Computing the attention map $\varphi(\cdot)$ is performed by combining 2-D convolutional layers and ReLU activation functions to reduce the dimensionality of feature maps. Specifically, we choose the sigmoid function as the activation function for the output to enhance the information of interest while ignoring secondary information. Therefore, we expect to affect only the intensity of the output without changing its direction. Notably, a dedicated $\varphi(\cdot)$ is employed for each specific scale to compute the attention map corresponding to that scale. This approach can help the network prioritize pertinent regions of interest at appropriate scales, effectively suppressing irrelevant information. Moreover, utilizing multiscale convolution operations expands the network's receptive field. significantly Consequently, it can capture both local details and global background information, further enhancing the overall performance.

D. Postprocessing

In the context of remote sensing images, there are specific considerations related to the imaging angle of view. These images are typically taken from a top-down perspective, while the detection targets usually exhibit varying orientations. In addition, certain targets, such as airplanes, are often densely distributed within scenes. Consequently, the prediction frames generated by the detection algorithm tend to overlap. When employing the traditional NMS algorithm, the predicted frames with lower confidence scores are directly discarded. This approach, however, may lead to missed detections and a subsequent decrease in overall detection accuracy. To address this challenge and enhance the algorithm's ability to detect densely arranged targets, we propose substituting the conventional NMS algorithm with the Soft-NMS algorithm [34]. By employing Soft-NMS, we can mitigate the issues associated with overlapping predicted frames, allowing more refined and accurate detection.

The Soft-NMS algorithm [34] was developed to address the challenge of achieving accurate detection in scenarios where targets are occluded by one another. This algorithm incorporates an attenuation function that modifies the confidence scores of adjacent predicted frames based on their intersection over union (IoU) values. Instead of setting the confidence of bounding boxes with lower confidence to zero, the algorithm reduces confidence using a gradual attenuation approach. This approach preserves the detection frames, thus improving the algorithm's recall rate and mitigating instances of missing detection. The proposed computation to reduce confidence in Soft-NMS [34] is as follows:

$$\begin{cases} c_i, & \mathrm{IoU}(b_i, b_j) < T \\ c_i(1 - \mathrm{IoU}(b_i, b_j)), & \mathrm{IoU}(b_i, b_j) \geq T. \end{cases} \tag{10}$$

Among them, c_i reflects the model's confidence in the box containing objects and reflects its belief in the accuracy of box prediction. Formally, we define confidence as $P_r(\mathrm{Object}) \cdot \mathrm{IoU}_{\mathrm{pred}}^{\mathrm{truth}}$. If no object exists in that cell, the confidence score should be zero. Otherwise, we want the confidence score to be equal to the IoU between the predicted box and the ground truth.

At test time, we multiply the conditional class probabilities and the individual box confidence predictions

$$c_i = P_r(\mathrm{Class}|\mathrm{Object}) \cdot P_r(\mathrm{Object}) \cdot \mathrm{IoU}_{\mathrm{pred}}^{\mathrm{truth}}. \tag{11}$$

$P_r(\mathrm{Object})$ represents the probability of the presence of a target in the detection box, and $P_r(\mathrm{Class}|\mathrm{Object})$ represents the probability confidence formula of the target belonging to a certain category in a given detection box, which is used to calculate the probability of the presence of a target in the detection box and gives us class-specific confidence scores for each box.

For each bounding box b_i, if the IoU value between b_i and another bounding box b_j is greater than a predetermined threshold T, Soft-NMS maintains the correct result by lowering the confidence level of b_i. With this approach, bounding box redundancy can be avoided to some extent while increasing the object detection precision.

E. Training Scheme

To train the proposed model in few-shot scenarios effectively, a training data partitioning approach was employed. Specifically, the training data were divided into two distinct groups: the query set (Q) and the support set (S). A query set comprises query images along with the corresponding annotations (A)

$$Q = \{(I, A)\}. \tag{12}$$

The category of the support set is determined by the selected base class category. A support set consists of N support images, each from a different base class. Each support image I_j corresponds to a bounding box mask M_{I_j}, where $j = 1, 2, \ldots, N$ and the support set is defined as

$$S = \{(I_1, M_{I_1}), (I_2, M_{I_2}), \ldots, (I_N, M_{I_N})\}. \tag{13}$$

Each training set is defined as follows: The training set is separated into numerous sets; each training set consists of a query image and its annotations, and a set of supporting images and their masks

$$T_k = Q_k \bigcup S_k \tag{14}$$

where the query set images are input to the feature extractor (see Fig. 1), and the query set images are input to the recurrent net, which is also shown in Fig. 1.

To address the challenge of detecting objects with limited training samples, we use a two-stage training scheme that categorizes the whole dataset into base classes and novel classes. This training process offers improved performance in few-shot scenarios. The initial stage, known as basic training, focuses

Algorithm 2: Training Scheme.

input : A big dataset for base classes D_{base}, a small dataset for novel classes D_{novel}, the number of epochs in base training e_0, the number of epochs in fine-tuning e_1.
Initialize all parameters of model M;
output: Trained model parameters $\theta^{finetune}$;

1 **for** $i \to 0, 1, ..., e_0 - 1$ **do**
2 Construct a set of episodes T^{base} buit upon D_{base};
3 **for** T *in* T^{base} **do**
4 $l^{train} \to Loss^{train}(M, T)$;
5 Refresh all the trainable parameters θ^{base} of M by backprop;
6 **end**
7 **end**
8 **for** $i \to 0, 1, ..., e_1 - 1$ **do**
9 Construct a set of episodes $T^{finetune}$ buit upon $D_{base} \bigcup D_{novel}$;
10 **for** T *in* $T^{finetune}$ **do**
11 $l^{finetune} \to Loss^{finetune}(M, T)$;
12 Refresh all the trainable parameters $\theta^{finetune}$ of M by backprop;
13 **end**
14 **end**

on training the network's learning parameters using a comprehensive base class dataset. This stage typically requires a large amount of time to ensure effective learning. In the subsequent stage, referred to as finetuning, we leverage a novel class dataset to train the model built upon the knowledge gained during the first stage further. This finetuning process enables the model to attain superior performance with a relatively shorter training duration. Algorithm 2 describes the two-stage training procedure. First, we input the base class dataset D_{base}, the novel class dataset D_{novel}, and the model M. Lines 1–7 represent the base training procedure of the first training stage. Specifically, within each base training epoch, line 2 constructs a training dataset T^{base} based on the base class dataset D_{base}; line 3 represents to traverse each batch T of the training dataset T^{base}; line 4 calculates the training loss for model M and performs backpropagation, and line 5 refreshes the parameters θ^{base} based on backpropagation. Lines 8–14 explain the few-shot fine-tuning in the second training stage. Specifically, in each fine-tuning epoch, line 9 constructs the fine-tuning dataset T^{finetune} based on the base class dataset D_{base} and the novel class dataset D_{novel}; line 10 represents traversing each batch T of the fine-tuning dataset T^{finetune}; line 11 calculates the loss for model M and performs backpropagation; line 12 represents refreshing training parameters θ^{finetune} based on backpropagation.

IV. Experiments

This section presents a comprehensive overview of the performed experiments, encompassing two widely used public remote sensing object detection datasets, experimental configurations, evaluation metrics, experimental findings, and ablation analyses assessing the impact of individual components. The details are discussed in the following sections.

A. Dataset, Experimental Configuration, and Parameter Setting

Two publicly available datasets were employed to test the developed method as follows.

- *NWPU VHR-10*[1]*:* The high-resolution public dataset NWPU VHR-10 is used to identify objects in remote sensing images. The dataset contains 800 remote sensing images that were gathered from the ISPRS Vaihingen and Google Earth datasets. Within this dataset, 650 samples are classified as positive instances, each containing at least one discernible target object, while the remaining 150 samples are designated negative instances devoid of any target objects. This dataset encompasses annotations for ten distinct object categories, including baseball diamonds, tanks, basketball courts, tennis courts, ships, ground track fields, bridges, and harbors.
- *DIOR*[2]*:* DIOR is a sizable benchmark dataset for object detection in remote sensing images. This dataset includes 192 472 instances of 20 classes and 23 463 images taken from Google Earth. There are 20 object classes in this dataset: baseball field, airport, airplane, chimney, bridge, basketball court, expressway toll station, express service area, dam, ground track field, golf course, harbor, stadium, ship, overpass, train station, tennis court, storage tank, windmill, and vehicle. Each image within the dataset adheres to a consistent size of 800 $\times$ 800 pixels while providing a spatial resolution ranging from 0.5 to 30 m.

Each dataset was divided into two distinct subsets to assess the efficacy of the model in few-shot scenarios: the base class dataset and the novel class dataset. In the NWPU-VHR10 dataset, the novel class was defined by three specific classes, tennis court, airplane, and baseball diamond, while the remaining seven classes were designated the base class. For the DIOR dataset, a subset of five classes, comprising windmill, airplane, train station, tennis court, baseball field, and train station, was identified as the novel class, while the remaining classes were categorized as the base class.

All the experiments were implemented in PyTorch and trained on an Nvidia RTX 3090 GPU. For parameter settings, we used the Adam optimizer with momentum set to 0.9 during training. For the first base training stage, the initial learning rate was 0.001, while for the second finetuning stage, a continuous learning rate of 0.0001 was used. Meanwhile, the IoU threshold of the Soft NMS algorithm is set to 0.5. The reason is that we expect the algorithm could detect targets effectively and produces a high accuracy and stability when facing scenes with high overlapping between small targets in remote sensing images.

[1][Online]. Available: https://1drv.ms/u/s!AmgKYzARBl5cczaUNysmiFRH4eE

[2][Online]. Available: https://drive.google.com/drive/folders/1UdlgHk49iu6WpcJ5467iT-UqNPpx__CC

B. Evaluation Metrics

To validate the object detection results, we adopt visual object class metrics [37], including precision, recall, AP, and mAP

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{15}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{16}$$

where TP, FP, and FN represent true positives, false positives, and false negatives, respectively. The prediction box is positive when the IoU between it and the target box is linked to surpass 0.5. Otherwise, it is negative

$$\text{AP} = \int_0^1 p(r) d_r. \tag{17}$$

The coordinate system AP represents the region under the precision–recall curve with precision as the vertical coordinate and recall as the horizontal coordinate

$$\text{mAP} = \frac{\sum_{c=1}^{n} AP_c}{n} \tag{18}$$

where c denotes the category, n represents the number of categories, and mAP represents the average value of APs in multiple categories. The overall effectiveness of multicategory target detection can be well represented by mAP.

C. Baseline Methods

We compared the proposed few-shot detection method with the current one-stage FSOD approaches.

1) FSRW [16] is an FSOD approach that uses metalearning for initialization. Notably, FSRW introduces a reweighting module, which generates a collection of reweighting vectors from the support samples. These reweighting vectors recalibrate the metafeatures extracted from the feature extractor, thereby enabling effective detection in few-shot scenarios. By intelligently reweighting the metafeatures based on the support samples, the FSRW achieves good performance for FSOD.
2) The FSODM [24] was improved upon the FSRW to achieve multiscale object detection in few-shot scenarios. FSODM is designed to contain multiscale detection heads, achieving the first application of FSOD for remote sensing images.
3) YOLOv5 [38] stands out as one of the most prevalent single-stage object detection algorithms, garnering extensive utilization. Its distinguishing feature lies in its ability to detect all objects in an image with only a single forward propagation through the neural network. This unique characteristic endows YOLOv5 with advantages in terms of both detection speed and performance. By obviating multiple passes or complex mechanisms, YOLOv5 achieves a streamlined and efficient object detection process, making it an appealing choice for various applications.

D. Comparison With Baseline Methods

To showcase the efficacy of the approach in detecting objects via remote sensing in a few shots, we conducted a comprehensive comparison with the aforementioned baseline methods. Thus, we evaluated the developed approach in combination with baseline approaches using the DIOR dataset and the NWPU VHR-10 dataset in 3-shot, 5-shot, and 10-shot cases, as illustrated in Tables II and III. Notably, the results denoted with an asterisk (*) were originally reported in Wang's [39] paper; the results without an asterisk were obtained through rigorous testing in our experimental environment. This comparative analysis enables a robust assessment of the efficiency and performance of our suggested approach.

The performance comparison results obtained using the NWPU VHR-10 dataset are shown in Table II. Specifically, our model achieves 0.36 mAP in the 3-shot case, 0.52 mAP in the 5-shot case, and an impressive 0.65 mAP in the 10-shot case, outclassing YOLOv5 to a great extent by approximately 35% in all cases. Our model outperforms the FSRW in all the other cases by 15%. FSRW achieves 0.12 mAP in the 3-shot case, 0.24 mAP in the 5-shot case, and 0.40 mAP in the 10-shot case. Furthermore, compared to the FSODM, the proposed model outperforms the other models by 4% in the 3-shot case, by 2% in the 5-shot case, and by another 2% in the 10-shot case. FSODM achieves 0.32 mAP in the 3-shot case, 0.50 mAP in the 5-shot case, and 0.63 mAP in the 10-shot case. In stark contrast, the conventional YOLOv5 algorithm lags significantly, displaying inferior performance compared to the two few-shot-based methods. Note that the mAP of YOLOv5 reaches a mere 0.20 in the 10-shot case, which is a result of the impressive 0.36 mAP achieved in the 3-shot case.

These findings underscore the considerable effectiveness of our approach in FSOD scenarios for remote sensing images. As depicted in Table III, our approach exhibits commendable performance in the DIOR dataset, achieving mAPs of 0.25, 0.32, and 0.35 in the 5-shot, 10-shot, and 20-shot scenarios, respectively. Notably, the approach outperforms the FSODM by 1% in the 5-shot case and by 2% in both the 10-shot and 20-shot cases. These outcomes show the efficacy of our methodology in addressing potential challenges related to multiscale detection and the optimal utilization of support samples, which have been observed in previous models. Comparatively, the FAM enhances the representation of support features and query features, thereby maximizing the utilization of support samples. Similarly, the SAM module equips the network with adaptability to the diverse multiscale characteristics of remote sensing images, thus expanding the perceptual field. These improvements contributed to the overall performance enhancement produced by our approach in FSOD for remote sensing images.

E. Ablation Study

The effectiveness of each module employed in our research was carefully assessed through ablation experiments. Specifically, we evaluated the performance of each module in the NWPU VHR-10 dataset, which serves as a reliable benchmark. In an ablation study, we built upon the foundation of the

TABLE II
FEW-SHOT DETECTION RESULTS (MAP) COMPARISON IN THE NWPU VHR-10 DATASET (IOU OF 0.5)

	YOLOv5*			FSRW*			FSODM			Ours		
Class	3-shot	5-shot	10-shot	3-shot	5-shot	10-shot	3-shot	5-shot	10-shot	3-shot	5-shot	10-shot
airplane	0.06	0.10	0.18	0.13	0.24	0.20	0.15	0.33	0.40	**0.19**	**0.43**	**0.56**
baseball diamond	0.14	0.20	0.28	0.12	0.39	0.74	0.67	0.88	0.91	0.66	0.77	0.83
tennis court	0.12	0.15	0.15	0.11	0.11	0.26	0.15	0.29	0.57	0.22	**0.37**	**0.57**
mean	0.11	0.15	0.20	0.12	0.24	0.40	0.32	0.50	0.63	**0.36**	**0.52**	**0.65**

The bold values denote the best results.

TABLE III
FEW-SHOT DETECTION RESULTS (MAP) COMPARISON IN THE DIOR DATASET (IOU OF 0.5)

	YOLOv5*			FSRW*			FSODM			Ours		
Class	5-shot	10-shot	20-shot	5-shot	10-shot	20-shot	5-shot	10-shot	20-shot	5-shot	10-shot	20-shot
airplane	0.02	0.08	0.09	0.09	0.15	0.19	0.11	0.16	0.17	**0.14**	**0.17**	0.17
baseball field	0.09	0.27	0.30	0.33	0.45	0.52	0.37	0.44	0.47	0.30	**0.48**	0.51
tennis court	0.10	0.12	0.20	0.47	0.54	0.55	0.54	0.62	0.65	**0.62**	0.61	0.62
train station	0.00	0.00	0.02	0.09	0.07	0.18	0.06	0.11	0.12	0.10	**0.13**	0.15
wind mill	0.01	0.10	0.12	0.13	0.18	0.26	0.11	0.16	0.21	0.11	**0.22**	**0.31**
mean	0.04	0.11	0.15	0.22	0.28	0.34	0.24	0.30	0.32	**0.25**	**0.32**	**0.35**

The bold values denote the best results.

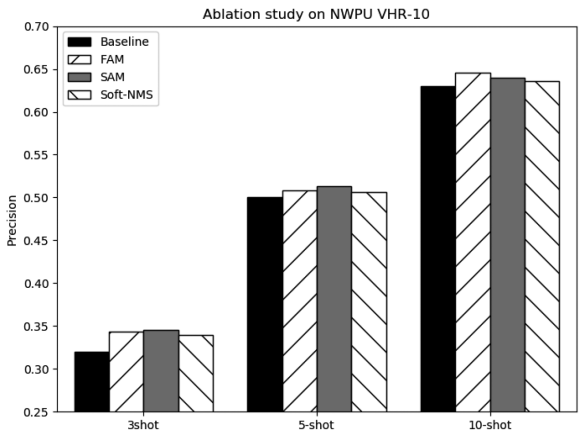

Fig. 4. Ablation study.

FSODM and systematically introduced three key parts: the FAM module, the SAM module, and the Soft-NMS algorithm. By incorporating these modules individually into the baseline, we carried out several experiments in few-shot scenarios, encompassing 3-shot, 5-shot, and 10-shot cases. This comprehensive evaluation allows us to thoroughly investigate and validate the efficacy of each component within the network, further contributing to a comprehensive understanding of the proposed approach.

The ablation results, as depicted in Fig. 4, showcase the effect of individual module additions. By incorporating the FAM alone, a notable improvement in performance of approximately 2% is observed. Specifically, the mAP increases from 0.3233 to 0.3432 in the 3-shot case, from 0.50 to 0.5086 in the 5-shot case, and from 0.6266 to 0.6459 in the 10-shot case. The integration of the FAM module, employing the transformer encoder, contributes to enhanced accuracy, albeit at the expense of computational speed. In addition, the inclusion of the SAM module alone yields a performance boost of approximately 2.5%. The mAP increases from 0.3233 to 0.3457 in the 3-shot case, from 0.50 to 0.5131 in the 5-shot case, and from 0.6266 to 0.6402 in the 10-shot case, underscoring the efficacy of the SAM module in addressing the multiscale challenges encountered in remote sensing images. Furthermore, employing the Soft-NMS algorithm as a standalone postprocessing technique on the baseline results in an improvement of approximately 1%. This improvement is evident in the 3-shot case, in which the mAP increases from 0.3233 to 0.3392; in the 5-shot case, from 0.50 to 0.5060; and in the 10-shot case, from 0.6266 to 0.6356.

Fig. 5 shows examples of the novel class detection outcomes of the proposed method in the NWPU VHR-10 dataset and the DIOR dataset. The majority of the novel class objects are correctly detected, as shown in Fig. 5, demonstrating the efficiency of our model. It is evident that the airplane class and the baseball diamond class are well-identified in both the NWPU dataset and the DIOR dataset. In Fig. 5, we can see that despite the tennis court's resemblance to the basketball court, the model correctly identifies the tennis court and does not misidentify the basketball court. Similarly, in the left of Fig. 5, despite the tennis court's small size and densely packed targets, the model still correctly identifies the majority of targets. Our model can effectively handle size variations and correctly recognize both large-scale train stations and small-scale wind turbines. However, it is important to acknowledge that certain challenges persist. For instance, the presence of cluttered backgrounds hampers the recognition of certain targets, such as airplanes. Moreover, complex backgrounds may result in the misclassification of background objects as novel class targets. In dense scenarios such as the tennis court, there may be instances where detection is missing. These observations provide valuable insights into the strengths and limitations of our proposed approach in object detection for remote sensing images, facilitating a comprehensive understanding of its performance.

V. DISCUSSION

The proposed method was evaluated by comparison with advanced FSOD methods in the experiment. The experimental

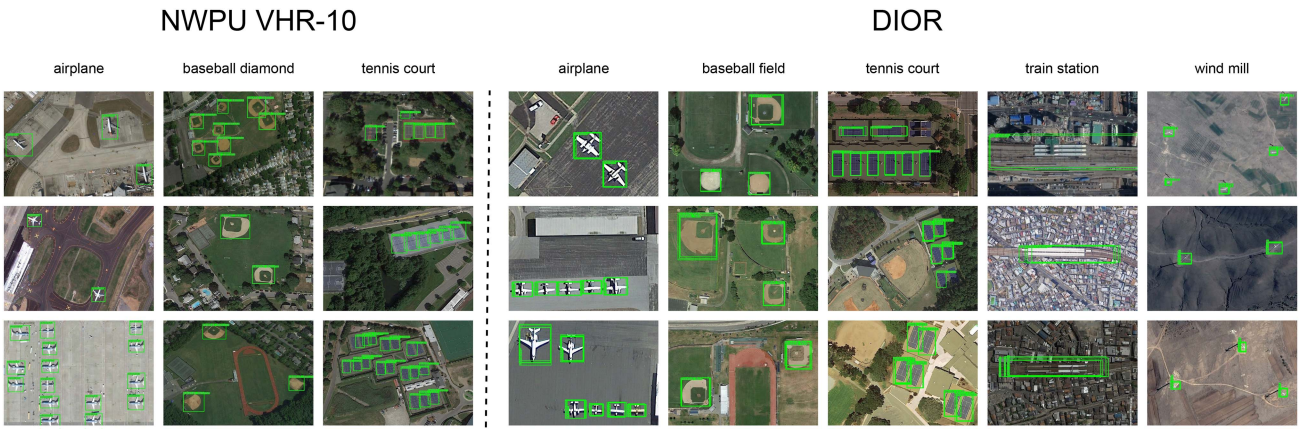

Fig. 5. Few examples of few-shot detection outcomes. (Left) Results of detection using a 10-shot setting on the novel classes in the NWPU VHR dataset. (Right) 20-shot detection results for the novel classes in the DIOR dataset.

results indicate the effectiveness of this method on the NWPU VHR-10 and DIOR datasets.

A. Performance on Novel Classes

According to Tables II and III, our proposed approach outperforms all the competing methods. The results of comparative experiments demonstrate the advantages of our proposed method, which will be discussed as follows. By analyzing the detection accuracy of each class, we can conclude that our method performs well on both datasets and outperforms FSODM [24] for targets with clear contours, such as airplanes, which are easy to distinguish between foreground and background. In addition, our approach performs well in detecting objects in tennis court categories where there is significant overlap. This is because our Soft-NMS algorithm effectively alleviates the problem of missed detections caused by overlapping targets.

Simultaneously, we discovered that our approach underperformed on the DIOR and NWPU VHR-10 datasets for the baseball field and diamond. We speculate that this might be due to the limited number of similar base class samples in the dataset, which is insufficient to enable the model to obtain a feature extractor with sufficient knowledge during base training.

Moreover, in object detection for remote sensing images, windmill detection is usually a difficult task, particularly for FSOD. Because windmills are often small, they can be readily mistaken for the background. Due to the problems of background interference and target scale, the accuracy of the FSODM in identifying different windmill groups is inadequate. However, our experimental findings suggest that the designed SAM can lessen the impact of irrelevant background data while assisting the model in better adapting to changes in object scale. Furthermore, the suggested FAMs can effectively enhance the ability to detect small targets.

B. Performance on Base Classes

A reliable FSOD model should excel in detecting novel classes and should also perform satisfactorily on base classes.

TABLE IV
PERFORMANCE COMPARISON FOR BASE CLASS DETECTION IN THE NWPU VHR-10 DATASET (IOU OF 0.5)

Class	YOLOv5*	FSRW*	FSODM	Ours
ship	0.80	0.77	0.81	**0.89**
storage tank	0.52	0.80	0.77	0.59
basketball court	0.58	0.51	0.65	**0.80**
ground track field	0.99	0.94	0.90	0.90
harbor	0.67	0.86	0.89	0.88
bridge	0.56	0.77	0.78	**0.88**
vehicle	0.70	0.68	0.58	**0.89**
mean	0.69	0.76	0.77	**0.83**

The bold values denote the best results.

TABLE V
PERFORMANCE COMPARISON FOR BASE CLASS DETECTION IN THE DIOR DATASET (IOU OF 0.5)

Class	YOLOv5*	FSRW*	FSODM	Ours
airport	0.59	0.59	0.48	0.57
basketball court	0.71	0.74	0.74	0.68
bridge	0.26	0.29	0.33	**0.33**
chimney	0.68	0.70	0.66	0.69
dam	0.40	0.52	0.30	0.41
expressway service area	0.55	0.63	0.55	**0.63**
express toll station	0.45	0.48	0.54	**0.59**
golf course	0.60	0.61	0.48	0.53
ground track field	0.65	0.54	0.63	**0.64**
harbor	0.31	0.52	0.31	0.43
overpass	0.46	0.49	0.47	0.47
ship	0.10	0.33	0.74	0.71
stadium	0.65	0.52	0.52	0.57
storage tank	0.21	0.26	0.57	0.52
vehicle	0.17	0.29	0.39	**0.41**
mean	0.45	0.50	0.51	**0.55**

The bold values denote the best results.

As Tables IV and V show, we carried out a detailed comparison of our approach's detection results on the base class to fully assess its capabilities. Remarkably, the model achieves an impressive base class mAP of 0.83 in the NWPU VHR-10 dataset, surpassing the base class mAPs of 0.75 and 0.77 obtained by FSRW and FSODM by 7% and 6%, respectively. In addition, the model outperforms YOLOv5 by a remarkable 14%. Similarly, in the DIOR dataset, the model achieves a commendable base class mAP of 0.55, surpassing the base class

mAPs of 0.50 and 0.51 obtained by FSRW and FSODM by 5% and 4%, respectively. Moreover, the model exhibited an impressive improvement of 10% over YOLOv5. These results highlight the outstanding performance and effectiveness of the suggested approach, particularly in maintaining high detection precision for both novel and base classes.

Most existing work adopts a two-stage detection scheme, such as the Fatser-RCNN, in the traditional object detection field. The inference speed after model deployment is usually slow and cannot meet real-time requirements. Our method adopts a one-stage detection architecture that can be easily applied to mobile devices such as drones and respond in a timely manner.

VI. Conclusion

This article develops a few-shot finetuning approach based on feature aggregation and scale attention for FSOD in remote sensing images. Our approach introduces the FAM, which leverages the transformer encoder and FPN architectures to effectively capture spatial position information and improve small object detection. In addition, we propose a SAM that effectively handles scale changes while adaptively directing attention to more unique regions of interest, reducing interference from background information. Conversely, to mitigate the issue of overlooked detection stemming from the overlapping of specific categories, we employed the Soft-NMS algorithm during the postprocessing phase to increase the detection precision. Experiments on real-world datasets show that our approach outperforms the FSODM and the well-known object detection algorithms YOLOv5 and FSRW, especially in few-shot scenarios. Our approach effectively utilizes support samples and small target information and adapts well to changes in remote sensing image scales and target overlap. Existing studies for FSOD usually focus on natural image scenes, and in contrast, our proposed approach contributes to the detection of few-shot objects for remote sensing images. Meanwhile, existing studies on FSOD for remote sensing images usually employ RCNN, which often consumes a large amount of running time and thus produces poor performance. Our proposed approach is equipped with a single-stage object detection, which could simultaneously achieve real-time performance and high accuracy. Moreover, the proposed approach is specially superior in the scenario where small targets overlap with each other.

Further analysis of the experimental results indicates that our approach also has enormous potential value in real-world applications. Our approach has brought new possibilities to the field of remote sensing image processing by addressing scale changes and few-shot problems. In particular, our approach can produce good results when dealing with small targets and overlapping scenes, such as in traffic management. The traffic scenes in remote sensing images usually involve a dense distribution of multiple targets, such as vehicles and pedestrians, and our approach can still accurately identify targets.

The discussion of experimental results indicates that our approach has some limitations in detecting novel classes when being faced with limited information from similar base class samples. To address this issue, we intend to investigate new strategies, such as expanding training samples, to produce a variety of samples to enhance detection performance. Possible methods include sophisticated generative methodologies, such as recognition flow [40], generative adversarial network [41], or autoencoder [42]. We also plan to study the object detection task for unmanned aerial vehicles. Through these future explorations, we attempt to explore the current performance of FSOD in remote sensing and expand its applications to real-world scenarios.

References

[1] Y. Himeur, B. Rimal, A. Tiwary, and A. Amira, "Using artificial intelligence and data fusion for environmental monitoring: A review and future perspectives," *Inf. Fusion*, vol. 86–87, pp. 44–75, 2022.

[2] A. Abdollahi, B. Pradhan, N. Shukla, S. Chakraborty, and A. Alamri, "Deep learning approaches applied to remote sensing datasets for road extraction: A state-of-the-art review," *Remote Sens.*, vol. 12, no. 9, 2020, Art. no. 1444.

[3] Z. Shao, N. S. Sumari, A. Portnov, F. Ujoh, W. Musakwa, and P. J. Mandela, "Urban sprawl and its impact on sustainable urban development: A combination of remote sensing and social media data," *Geo-spatial Inf. Sci.*, vol. 24, no. 2, pp. 241–255, 2021.

[4] A. Terentev, V. Dolzhenko, A. Fedotov, and D. Eremenko, "Current state of hyperspectral remote sensing for early plant disease detection: A review," *Sensors*, vol. 22, no. 3, pp. 1–31, 2022.

[5] X. Wang, D. Zhu, G. Li, X.-P. Zhang, and Y. He, "Proposal-copula-based fusion of spaceborne and airborne sar images for ship target detection**," *Inf. Fusion*, vol. 77, pp. 247–260, 2022.

[6] D. Zhu, X. Wang, G. Li, and X.-P. Zhang, "Vessel detection via multi-order saliency-based fuzzy fusion of spaceborne and airborne sar images," *Inf. Fusion*, vol. 89, pp. 473–485, 2023.

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

[8] K. Stankov and D.-C. He, "Detection of buildings in multispectral very high spatial resolution images using the percentage occupancy hit-or-miss transform," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 10, pp. 4069–4080, Oct. 2014.

[9] S. Leninisha and K. Vani, "Water flow based geometric active deformable model for road network," *ISPRS J. Photogrammetry Remote Sens.*, vol. 102, pp. 140–147, 2015.

[10] A. O. Ok, "Automated detection of buildings from single VHR multispectral images using shadow information and graph cuts," *ISPRS J. Photogrammetry Remote Sens.*, vol. 86, pp. 21–40, 2013.

[11] A. O. Ok, C. Senaras, and B. Yuksel, "Automated detection of arbitrarily shaped buildings in complex environments from monocular VHR optical satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 3, pp. 1701–1717, Mar. 2013.

[12] D. Contreras, T. Blaschke, D. Tiede, and M. Jilge, "Monitoring recovery after earthquakes through the integration of remote sensing, gis, and ground observations: The case of l'aquila (Italy)," *Cartogr. Geographic Inf. Sci.*, vol. 43, no. 2, pp. 115–133, 2016.

[13] Y. Xu et al., "Gliding vertex on the horizontal bounding box for multi-oriented object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1452–1459, Apr. 2021.

[14] P. Wang, X. Sun, W. Diao, and K. Fu, "FMSSD: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3377–3390, May 2020.

[15] B. M. Lake, R. R. Salakhutdinov, and J. Tenenbaum, "One-shot learning by inverting a compositional causal process," in *Proc. Conf. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013, pp. 2526–2534.

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

[17] B. Sun, B. Li, S. Cai, Y. Yuan, and C. Zhang, "FSCE: Few-shot object detection via contrastive proposal encoding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7348–7358.

[18] T.-I. Hsieh, Y.-C. Lo, H.-T. Chen, and T.-L. Liu, "One-shot object detection with co-attention and co-excitation," in *Proc. Conf. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 2725–2734.

[19] G. Zhang, S. Lu, and W. Zhang, "CAD-Net: A context-aware detection network for objects in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 10015–10024, Dec. 2019.
[20] Z.-Z. Wu, J. Xu, Y. Wang, F. Sun, M. Tan, and T. Weise, "Hierarchical fusion and divergent activation based weakly supervised learning for object detection from remote sensing images," *Inf. Fusion*, vol. 80, pp. 23–43, 2022.
[21] J. Pang, C. Li, J. Shi, Z. Xu, and H. Feng, "R^2-CNN: Fast tiny object detection in large-scale remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5512–5524, Aug. 2019.
[22] W. Cai, B. Zhang, and B. Wang, "Scale-aware anchor-free object detection via curriculum learning for remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 9946–9958, 2021.
[23] J. Zhang, C. Xie, X. Xu, Z. Shi, and B. Pan, "A contextual bidirectional enhancement method for remote sensing image object detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4518–4531, 2020.
[24] X. Li, J. Deng, and Y. Fang, "Few-shot object detection on remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.
[25] Z. Zhou et al., "FSODS: A lightweight metalearning method for few-shot object detection on SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2022.
[26] Z. Xiao, J. Qi, W. Xue, and P. Zhong, "Few-shot object detection with self-adaptive attention network for remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4854–4865, 2021.
[27] Z. Zhao, P. Tang, L. Zhao, and Z. Zhang, "Few-shot object detection of remote sensing images via two-stage fine-tuning," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[28] G. Cheng et al., "Prototype-CNN for few-shot object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–10, 2022.
[29] X. Wu, D. Sahoo, and S. Hoi, "Meta-RCNN: Meta learning for few-shot object detection," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 1679–1687.
[30] Y. Zhou, H. Hu, J. Zhao, H. Zhu, R. Yao, and W.-L. Du, "Few-shot object detection via context-aware aggregation for remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[31] Y. Zhang, B. Zhang, and B. Wang, "Few-shot object detection with self-adaptive global similarity and two-way foreground stimulator in remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 7263–7276, 2022.
[32] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," in *Proc. Conf. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 15908–15919.
[33] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 936–944.
[34] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS–improving object detection with one line of code," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5562–5570.
[35] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust object recognition with cortex-like mechanisms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 411–426, Mar. 2007.
[36] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.
[37] R. Padilla, S. L. Netto, and E. A. Da Silva, "A survey on performance metrics for object-detection algorithms," in *Proc. IEEE Int. Conf. Syst., Signals Image Process.*, 2020, pp. 237–242.
[38] G. Jocher et al., "Yolov5: V3. 1-bug fixes and performance improvements," 2020. Accessed: May 10, 2023. [Online]. Available: https://gitlab.com/ultralytics/yolov5/-/releases/v3.1
[39] Y. Wang, C. Xu, C. Liu, and Z. Li, "Context information refinement for few-shot object detection in remote sensing images," *Remote Sens.*, vol. 14, no. 14, 2022, Art. no. 3255.
[40] Y. Shen, J. Qin, L. Huang, L. Liu, F. Zhu, and L. Shao, "Invertible zero-shot recognition flows," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 614–631.
[41] K. Li, Y. Zhang, K. Li, and Y. Fu, "Adversarial feature hallucination networks for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13467–13476.
[42] E. Schwartz et al., "Delta-encoder: An effective sample synthesis method for few-shot object recognition," in *Proc. Conf. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 2850–2860.

Honghao Gao (Senior Member, IEEE) received his Ph.D degree in Computer Science and Technology in 2012 from Shanghai University, Shanghai, China.

He is currently with the School of Computer Engineering and Science, Shanghai University, Shanghai, China. He is also a Professor at the College of Future Industry, Gachon University, Seongnam, South Korea. Prior to that, he was a Research Fellow with the Software Engineering Information Technology Institute at Central Michigan University, Mt. Pleasant, MI, USA, and was an Adjunct Professor with Hangzhou Dianzi University, Hangzhou, China. He has publications in IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS (IEEE T-ITS), IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON SERVICES COMPUTING, IEEE TRANSACTIONS ON CLOUD COMPUTING, IEEE TRANSACTIONS ON FUZZY SYSTEMS, IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING, IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT, IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING, IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING, IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS, IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE, IEEE TRANSACTIONS ON CONSUMER ELECTRONICS, IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS. His research interests include software intelligence, cloud/edge computing, and intelligent data processing.

Dr. Gao was the 2022 recipient of Highly Cited Chinese Researchers by Elsevier, the 2023 recipient of Highly Cited Researcher by Clarivate, and is recognized as World's Top 2% Scientists 2021–2023. He is a Fellow of the Institution of Engineering and Technology (IET), a Fellow of the British Computer Society (BCS), and a Member of the European Academy of Sciences and Arts (EASA). He is the Editor-in-Chief for *International Journal of Web Information Systems* (IJWIS), Editor for *Wireless Network* (WINE), *The Computer Journal* (COMPJ), and *IET Wireless Sensor Systems* (IET WSS), and Associate Editor for IEEE T-ITS, *IET Intelligent Transport Systems (IET ITS)*, *IET Software*, *International Journal of Communication Systems (IJCS)*, *Journal of Internet Technology (JIT)*, and *Engineering Reports (EngReports)*. Moreover, he has broad working experience in cooperative industry-university-research. He is a European Union Institutions-appointed external expert for reviewing and monitoring EU Project, is a Member of the EPSRC Peer Review Associate College for U.K. Research and Innovation in the U.K., and a Founding Member of the IEEE Computer Society Smart Manufacturing Standards Committee.

Shuping Wu is currently working toward the M.S. degree in computer science with the School of Computer Engineering and Science, Shanghai University, Shanghai, China.

Her research interests include computer vision and object detection.

Ye Wang received the M.S. degree in control theory and control engineering from Shanghai Maritime University, Shanghai, China, in 2022. He is currently working toward the Ph.D. degree in Computer Science and Technology with the College of Computer Engineering and Science, Shanghai University, Shanghai, China.

His research interests include the Internet of Things and edge computing.

Jung Yoon Kim received the Ph.D. degree in game engineering from the Graduate School of Advanced Imaging Science, Multimedia and Film, Chung-Ang University, Seoul, South Korea, in 2013.

He is currently an Associate Professor with the Graduate School of Game, Gachon University, Seongnam, South Korea, and the Department of Game Media, College of Future Industry. He is also the Center Director of the Start-Up Education Center, Gachon University. From 2015 to April 2018, he was the Vice President of Korea Game Developer Association, Seoul, South Korea. His research interests include computer gaming, AI, virtual reality technology, and interactive technology.

Dr. Kim has been the Editor-in-Chief for the Korea Computer Game Association since 2016.

Yueshen Xu (Member, IEEE) received the Ph.D. degree in Computer Science and Technology from Zhejiang University, Hangzhou, China, in 2016, and was a cotraining Ph.D. candidate with the University of Illinois at Chicago, Chicago, IL, USA, from 2014 to 2015.

He is currently an Associate Professor with the School of Computer Science and Technology, Xidian University, Xi'an, China. He has authored or coauthored more than more than 70 papers in international conferences or journals. His research interests include software engineering, information retrieval, and multimodal learning.

Dr. Xu is the Reviewer of many journals and a PC Member at many international conferences.