# Boosting Object Detectors via Strong-Classification Weak-Localization Pretraining in Remote Sensing Imagery

Cong Zhang, *Graduate Student Member, IEEE*, Tianshan Liu, Jun Xiao, Kin-Man Lam, *Senior Member, IEEE*, and Qi Wang, *Senior Member, IEEE*

***Abstract*— Deep learning-based object detectors in remote sensing (RS) scenarios typically follow the paradigm of pretraining and fine-tuning to alleviate the limitation of insufficient downstream data. Despite the improved performance, existing pretraining paradigms are suboptimal due to three deficiencies: 1) inconsistent domains, i.e., pretraining on natural scenes and fine-tuning for RS scenes; 2) mismatched task objectives, i.e., classification-oriented pretraining while detection-oriented fine-tuning; and 3) misaligned architectures, i.e., pretraining only one bare backbone yet neglecting other vital detection components. Against these issues, this article proposes a novel pretraining paradigm specifically for the task of RS object detection, namely, RS strong-classification weak-localization (SCWL) pretraining. Unlike conventional classification pretraining, such as the widely used ImageNet pretraining, our pretraining strategy can adaptively perform bounding box generation on a reconstructed large-scale RS classification-style dataset. These pseudobounding boxes are integrated with the original accurate class labels as location- and category-related supervisions, respectively, to pretrain the entire RS detectors. The proposed RS SCWL pretraining paradigm is able to significantly improve downstream detection performance and outperforms classification pretraining methods, including ImageNet pretraining. Extensive experiments on different object detection datasets demonstrate its effectiveness and superiority in boosting various RS detectors.**

***Index Terms*— Object detection, pretraining paradigms, remote sensing (RS) imagery, scene classification, weakly supervised object localization (WSOL).**

## I. Introduction

With the rapid development of satellite and unmanned aerial vehicle (UAV) systems, a plentiful amount of remote sensing (RS) images is available, which facilitates the research of various vision tasks, such as scene classification, object detection, and semantic segmentation [1], [2], [3], [4]. Among them, RS object detection [5], [6], [7] is a fundamental task in the field of Earth observation and measurement. It aims to accurately distinguish the categories of specific regions and locate geospatial objects, i.e., performing two subtasks: classification and localization.

The last decade has witnessed dramatic advances in generic object detection in natural scenes (NSs). Deep learning-based methods [8], [9], [10], [11], [12] lead this field due to their promising detection performance, which further motivates an increasing number of convolutional neural network (CNN)-based algorithms that focus on object detection in RS imagery [13], [14], [15], [16], [17], [18], [19]. Upstream pretraining and downstream fine-tuning play an important role in the success of CNN-based methods [20], [21], [22]. Typically, generic object detectors are first pretrained on large-scale datasets, for example, ImageNet [23] for classification, and then fine-tuned on downstream detection datasets, such as Pascal VOC [24] and COCO [25]. It has been introduced in [26] that pretraining benefits downstream tasks by learning low-level feature representations, such as boundaries, textures, and shapes, and collecting semantic relations in advance. At the same time, this presumption has been challenged in recent years. In [27], it was reported that a standard model trained on the COCO dataset from random initialization, i.e., from scratch, with sufficient training data and long training time, is still able to achieve detection performance comparable to ImageNet pretraining. This empirical study shows that ImageNet pretraining can be probably eliminated as long as the downstream detection datasets provide sufficient training samples.

However, compared to the COCO dataset, which has a substantial amount of data with more than 300 000 images, there is no such large-scale dataset for RS object detection. Limited by the characteristics of geospatial objects, such as small scale, dense, and cluttered arrangements, annotating RS instances is dramatically labor-intensive and time-consuming. The largest existing RS detection dataset, the DIOR dataset [28], contains only about 20 000 images, much smaller than the COCO dataset. Therefore, in recent years, the pretraining and fine-tuning paradigm still dominates RS object detection. In particular, most RS object detection algorithms [28] directly initialize the detection backbone using the pretrained models on ImageNet and then fine-tune the entire detector on a specific downstream RS detection dataset, such as DOTA [29], DIOR [28], and NWPU VHR-10 [30]. Nevertheless, in terms of

Manuscript received 26 June 2023; revised 18 August 2023; accepted 28 August 2023. Date of publication 26 September 2023; date of current version 2 October 2023. The Associate Editor coordinating the review process was Xiangchen Qian. *(Corresponding author: Cong Zhang.)*

Cong Zhang, Tianshan Liu, Jun Xiao, and Kin-Man Lam are with the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong (e-mail: cong-clarence.zhang@connect.polyu.hk).

Qi Wang is with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China.

Digital Object Identifier 10.1109/TIM.2023.3315392

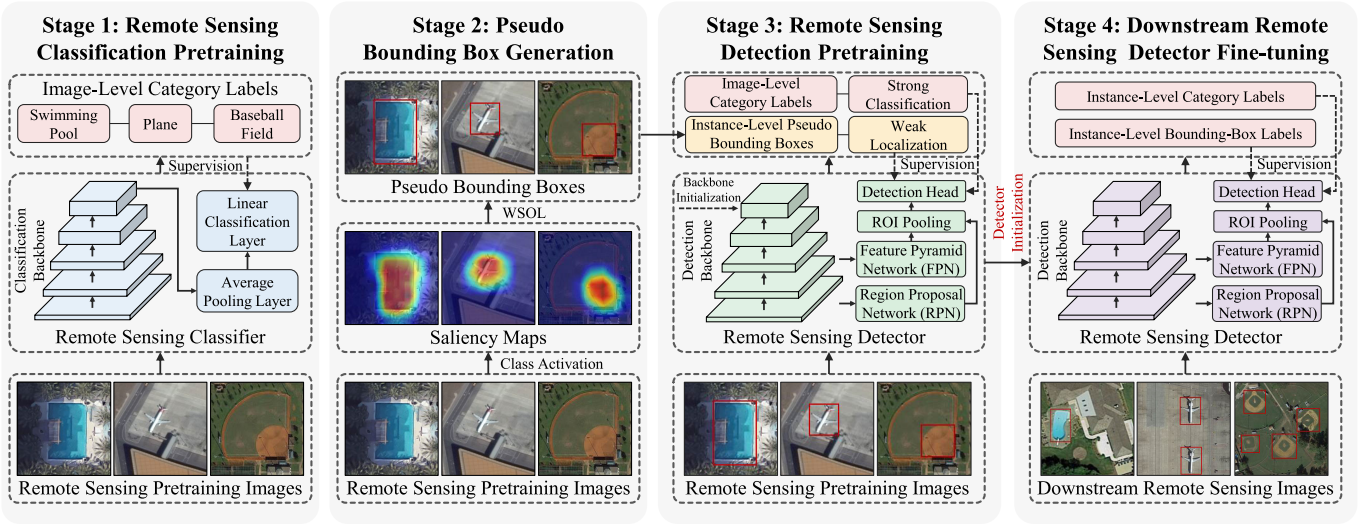


Fig. 1. Overview of our pretraining and fine-tuning pipeline for RS object detection, which consists of four stages: RS classification pretraining, pseudobounding box generation, RS detection pretraining, and downstream RS detector fine-tuning. The first three stages constitute the proposed RS SCWL pretraining, while retaining only Stage 1 forms its degraded version, named RS classification pretraining in this work. For the RS SCWL pretraining, the pretrained weights in Stage 3 are employed to initialize the entire downstream detector (e.g., the detection backbone, RPN, FPN, and the detection head in Faster R-CNN) in Stage 4.

RS object detection, this paradigm of ImageNet pretraining is questionable and suboptimal, and its empirical gain will reduce when fine-tuning RS detection datasets. The potential reasons can be summarized as follows.

1) *Inconsistent Domains:* In contrast to NS images, RS images, including satellite and aerial images, are usually captured from a bird's-eye view. In these two different scenarios, the appearances of objects belonging to the same category still vary considerably, leading to bias in feature learning. The semantic and discriminative information gleaned from ImageNet pretraining cannot appropriately contribute to downstream RS object detection and may even cause interference.
2) *Mismatched Task Objectives:* Detection and classification have different learning goals. However, during the classification pretraining phase, the objective of the downstream detection task is unknown. Generally, detection is sensitive to object scale and position, involving both classification and regression losses at the instance level. Nevertheless, employing only a single image-level classification loss during pretraining will adversely affect the performance of object detection with translation and scale variations. This mismatch caused by the difference in task objectives may diminish the benefits of the classification pretraining paradigm.
3) *Misaligned Architectures:* Existing pretraining networks typically consist of a bare backbone, an average pooling layer, and a linear classification layer. In contrast, object detection networks are more complicated and usually involve diverse additional components, such as the region proposal network (RPN) [10], the feature pyramid network (FPN) [31], the classification head, and the regression head [10]. In downstream detector fine-tuning, all these additional modules will simply be initialized randomly without being pretrained. This architectural misalignment between the pretraining and fine-tuning networks can result in performance degradation, especially in the case of low-data object detection.

The above three issues severely reduce the benefit of pretraining for RS object detection. Therefore, our goal is to bridge this gap by designing a novel pretraining paradigm tailored to the task of RS object detection. It is worth noting that no new detection modules, components, or architectures will be introduced in this work, and instead, with high flexibility, this pretraining paradigm should be able to extensively adapt to various existing detection frameworks. This is because pretraining fundamentally caters to specific downstream tasks, making it task-oriented rather than framework-constrained. In the context of RS object detection, pretraining is expected to showcase two indispensable properties.

1) *Strong-Classification Perception:* The classification capability should be pretrained locally and accurately within the RS domain, instead of globally and roughly on cross-domain NS images.
2) *Weak-Localization Awareness:* The pretrained representations should be aware of object localization knowledge, whose potential for bounding box prediction can be further unleashed after fine-tuning.

Based on the aforementioned design principles, in this article, we propose a more advanced pretraining paradigm, namely, strong-classification weak-localization (SCWL) pretraining, which can consistently enhance the performance of diverse RS object detectors. Explicitly, RS object detection still follows the common workflow of *first upstream pretraining* and *then downstream fine-tuning* yet respecting the proposed RS SCWL pretraining and inheriting the standard fine-tuning. As illustrated in Fig. 1, the whole pipeline is composed of four stages: RS classification pretraining, pseudobounding

box generation, RS detection pretraining, and downstream RS detector fine-tuning, where the first three stages constitute our proposed RS SCWL pretraining. Specifically, the first stage pretrains an accurate classifier on a large-scale RS scene classification dataset, while, in the second stage, pseudobounding boxes containing location information are generated based on the pretrained classifier and the concept of weakly supervised object localization (WSOL). The third stage treats the pseudobounding boxes as instance-level annotations to pretrain the entire detection framework. Finally, this object detector is further fine-tuned on a specific RS detection dataset by initializing with the pretrained weights. It is worth noting that, without requiring additional annotation effort, the proposed SCWL pretraining leverages both *image-level* strong-classification supervision and *instance-level* weak-localization supervision. Furthermore, unlike the conventional classification pretraining (e.g., ImageNet pretraining) that only pretrains the bare backbone, it performs explicit pretraining for all detection architectures in RS scenarios. Extensive experiments verify that our RS SCWL pretraining can consistently yield remarkable performance improvements for different RS detection frameworks across a wide variety of settings.

The main contributions can be summarized as follows.

1) A novel pretraining paradigm, namely, SCWL, is devised specifically for the task of RS object detection, which has the capability to efficiently pretrain the entire detection framework instead of only the bare backbone. To the best of our knowledge, we are the first to demonstrate that, by deviating from conventional NS-domain pretraining, pretraining the whole detector within only the RS domain can significantly boost downstream RS detection accuracy. Remarkably, the proposed SCWL pretraining does not impose specific detection architectures. On the contrary, it embodies a versatile paradigm and can be extensively instantiated into various frameworks for consistent downstream performance improvements.
2) The RS in-domain classification pretraining is also introduced to serve as a portable alternative to SCWL pretraining, which still respects the traditional paradigm of only pretraining the backbone for universality. It can generalize well on various frameworks with higher flexibility. More importantly, previous cross-domain pretraining inevitably brings adverse representation interference for downstream RS object detection, while this pipeline eliminates this issue in a straightforward and effective manner, i.e., unifying pretraining and fine-tuning with a single domain. Even with fewer pretraining samples, our proposed RS classification pretraining still outperforms ImageNet pretraining across different detection scenarios.
3) To enable the aforementioned two pretraining paradigms specialized for RS object detection, we reconstruct a large-scale RS dataset, including 276 300 samples, by aggregating and cleaning the 14 previous RS image classification datasets. Notably, based on our pseudobounding box generation strategy, numerous bounding boxes have been automatically created, together with the original category annotations to form a complete RS-domain pretraining dataset with instance-level labels.

The rest of this article is organized as follows. Section II briefly reviews related work. The reconstructed large-scale RS pretraining dataset is described in Section III. Section IV presents the proposed SCWL pretraining paradigm in detail, and Section V analyzes comprehensive experimental results. Finally, the conclusion is drawn in Section VI.

## II. Related Work

In this section, we first briefly review object detection in RS images and then investigate two highly related issues, pretraining paradigms and WSOL.

### A. Object Detection in Remote Sensing Images

As an important and classic vision task, object detection has extracted considerable attention in both the image processing and Earth observation communities. Traditional object detection methods [32], [33], [34] usually employ handcrafted features to represent different objects according to their intrinsic properties, such as color, texture, shape, and scale.

Due to significant progress in deep learning, most of the recent modern object detectors are based on CNNs, which exhibit more promising detection performance, compared to traditional algorithms. CNN-based object detectors can be divided into two categories: *two-stage* methods and *one-stage* methods. The former is represented by the well-known R-CNN family [8], [9], [10]. For instance, Faster R-CNN [10] generates category-agnostic region candidates using the RPN in the first stage. With the deep features extracted from these region proposals, classifiers and regressors are adopted for class prediction and bounding box prediction, respectively, in the second stage. Then, several necessary postprocessing procedures, such as nonmaximum suppression (NMS), are used to generate the final detection results. Most other two-stage detectors [35], [36] are also region-based methods, achieving leading detection accuracy in this field. In terms of object detection in RS images, there are some unique challenges to be solved, such as scale variation, arbitrary orientation, and dense and cluttered arrangements. Consequently, many improved two-stage algorithms [37], [38], [39] have been proposed toward RS object detection in recent years. For example, inspired by Faster R-CNN, Li et al. [37] adapted the original RPN to multiscale and multiangle anchors to match the scale-variant rotated geospatial objects. By removing the RPN, one-stage methods [11], [12], [40] directly perform object detection on the backbone features instead of region proposals, forming a more efficient pipeline but yielding worse accuracy. Among them, RetinaNet [12] achieved a decent speed–accuracy tradeoff, which has been widely utilized and improved for RS object detection [41], [42]. This work does not contribute new detection algorithms but, instead, focuses on a more effective RS detection-specific pretraining paradigm, which can consistently and broadly benefit the existing two- and one-stage detectors.

### B. Pretraining Paradigms

The pretraining and fine-tuning paradigm has played a vital role in the advancement of deep learning-based applications, including computer vision [10], natural language processing (NLP) [43], and speech processing [44]. As mentioned previously, in the past few years, the most common pretraining paradigm for object detection is to first utilize a pretrained network for classification on the ImageNet dataset as the detector backbone and then fine-tune the complete detector on downstream datasets. Theoretically, Shinya et al. [45] claimed that the pretrained models are able to form a narrower eigenspectrum than those trained from scratch. However, the performance gain achieved by this classification pretraining is still debated and has been extensively studied in recent years. For example, in [46], the impact of pretraining data was explored, such as the comparison of ImageNet pretraining with JFT-300M [47] pretraining, and the results demonstrate that large-scale data drastically enhances discriminative feature representation for multiple downstream tasks, such as detection, segmentation, and pose estimation. Moreover, He et al. [27] observed that, compared to large-scale fine-tuning datasets, pretraining is relatively more indispensable when downstream data are limited. Surprisingly, some modern detectors [48], [49] still achieve competitive performance only trained from scratch, suggesting that classification pretraining is less beneficial for detection or even hinders localization on larger datasets. An increasing amount of research [50], [51], [52], [53], [54] indicates that the traditional ImageNet pretraining is suboptimal and delicate, and thus, there is an urgent need to develop more effective and efficient pretraining paradigms for specific downstream tasks. With regard to RS scenarios, on the one hand, constructing an extremely large-scale object detection dataset that can be used to train a deep detector from scratch is considerably resource-intensive, so pretraining is still necessary. On the other hand, it is reasonable to improve and adapt the existing ImageNet pretraining to the RS scenes, resulting in a more advanced paradigm. This is the reason for the emergence of this research.

### C. Weakly Supervised Object Localization

WSOL refers to a class of object localization methods based only on image-level annotations as weak supervision [68], [69], [70], [71]. Class activation maps (CAMs) [68] can be regarded as pioneering work in this field, which is able to highlight the discriminative regions with the strongest activation. In this way, given the class of an image, the potential objects in the image can be roughly localized. Moreover, if extended to multilabeled images (one image having multiple categories), weakly supervised object detection (WSOD) [72], [73], highly related to WSOL, can be considered to detect multiple instances with different classes in an image. Although some recent WSOD algorithms [74], [75], [76], [77] also exploit weakly annotated samples to train object detectors, there are two crucial differences between them and our method. First, the purposes of generating pseudo-ground truths are significantly different. Our method aims to introduce in-domain weak-localization knowledge by pseudobounding boxes in advance during only pretraining, while its downstream fine-tuning still respects the standard training scheme of fully supervised object detection. In contrast, the detectors in [74], [75], [76], and [77] still follow the conventional ImageNet pretraining based on image-level supervision yet derive pseudo-ground truths as weak instance-level supervision during only downstream fine-tuning, which, in essence, is determined by the weakly supervised pipeline of WSOD. Consequently, pseudolabels are generated and utilized for upstream pretraining in our pipeline and for downstream training (i.e., fine-tuning) in WSOD methods. In practice, leveraging weak supervision for pretraining, instead of fine-tuning, can considerably alleviate the dependence on high-quality pseudolabels. Second, weak supervision is constructed in different ways. Specifically, in our proposed pretraining paradigm, the pseudoannotations can be simply produced by a lightweight well-trained classifier, while they are typically selected from massive proposals precomputed by a weak object detector in the WSOD algorithms [74], [75], [76]. Therefore, in our pipeline, pseudobounding boxes can be developed based on WSOL, rather than the more complicated WSOD requiring heavy computational load. This advantage relieves the requirement for available large-scale multilabeled training samples, which are difficult to acquire in the context of RS.

## III. Large-Scale Remote Sensing Pretraining Dataset Reconstruction

In order to improve the generalization and robustness of pretrained models, pretraining is usually built on large-scale datasets, such as ImageNet. However, as introduced in Section I, [21], and [27], traditional ImageNet pretraining may be suboptimal and can negatively impact downstream tasks, such as RS object detection. The ImageNet pretraining paradigm forces downstream detectors to perform transfer learning [78] from NS to RS scenarios during fine-tuning. To avoid this learning bias and break the dominance of ImageNet pretraining, it is necessary to establish a large-scale RS pretraining dataset, which should satisfy the following two properties. First, as an improved alternative to ImageNet specifically serving RS object detection, the dataset should be of large scale and wide data diversity to guarantee the effectiveness and generalization of pretraining. Second, each image is equipped with at least image-level supervision, such as scene category information for RS classification pretraining.

However, building such a large-scale pretraining dataset from scratch is still labor-intensive and time-consuming. In response to this issue, we reconstruct this dataset based on the existing RS datasets instead of collecting new data in the wild. Specifically, motivated by the promising performance of deep learning and its high demand for enormous data for training, many RS scene classification datasets have been publicly released in recent years. Moreover, in contrast to detection-specific datasets, RS classification datasets usually enjoy larger scale and higher accessibility due to the low requirement for image-level annotations. Table I illustrates

TABLE I
COMPARISON OF DIFFERENT RS SCENE CLASSIFICATION DATASETS AND OUR RECONSTRUCTED PRETRAINING DATASET

| Datasets | Year | Number of Categories | Number of Images Per Category | Image Size | Total Number of Images |
|---|---|---|---|---|---|
| WHU-RS19 [55] | 2012 | 19 | 50 ∼ 61 | 600×600 | 1,005 |
| RSC11 [56] | 2016 | 11 | 80 ∼ 142 | 512×512 | 1,232 |
| OPTIMAL-31 [57] | 2019 | 31 | 60 | 256×256 | 1,860 |
| UC-Merced [58] | 2010 | 21 | 100 | 256×256 | 2,100 |
| SIRI-WHU [59] | 2016 | 12 | 200 | 200×200 | 2,400 |
| RSSCN7 [60] | 2015 | 7 | 400 | 400×400 | 2,800 |
| MASATI v2 [61] | 2018 | 7 | 304 ∼ 1,789 | 512×512 | 7,389 |
| AID [62] | 2017 | 30 | 220 ∼ 420 | 600×600 | 10,000 |
| CLRS [63] | 2020 | 25 | 600 | 256×256 | 15,000 |
| RSI-CB256 [64] | 2017 | 35 | 198 ∼ 1,331 | 256×256 | 24,000 |
| PatternNet [65] | 2018 | 38 | 800 | 256×256 | 30,400 |
| NWPU-RESISC45 [66] | 2016 | 45 | 700 | 256×256 | 31,500 |
| RSI-CB128 [64] | 2017 | 45 | 173-1,550 | 128×128 | 36,000 |
| MLRSNet [67] | 2020 | 46 | 1,500 ∼ 3,000 | 256×256 | 109,161 |
| RS Pretraining | 2022 | 70 | 700 ∼ 13,405 | 128×128 ∼ 600×600 | 276,300 |

14 commonly adopted RS classification datasets and their data characteristics, including the release time, number of categories, number of images per category, image size, and total number of labeled images. In this work, by analyzing and comparing the characteristics of different datasets, we carefully aggregated, streamlined, and cleaned these 14 datasets to reshape a new pretraining dataset, namely, the RS pretraining dataset. It is worth noting that, considering that our goal is to boost downstream RS detectors through more advanced pretraining, several complicated scene categories are simplified or removed to avoid confusing downstream detection fine-tuning, and each sample corresponds to only one class. In this way, as shown in the last row of Table I, the reconstructed RS pretraining dataset contains 276 300 images of various resolutions and 70 different categories, covering all the classes in the widely used RS detection datasets including DOTA [29], DIOR [28], and NWPU VHR-10 [30]. In fact, this property can significantly benefit downstream RS detectors after pretraining, especially for low-data settings, which will be demonstrated in the experiments. Furthermore, this large-scale pretraining dataset appropriately meets the two above-mentioned requirements and can be utilized to replace ImageNet, facilitating the proposed SCWL pretraining paradigm that will be presented in detail in Section IV.

## IV. STRONG-CLASSIFICATION WEAK-LOCALIZATION PRETRAINING

Aimed at consistently and flexibly boosting downstream RS detectors by improving the discrimination of pretrained models, the proposed SCWL pretraining can be regarded as an advanced pretraining paradigm that is not constrained to a specific detection framework. Fig. 1 depicts the workflow of our proposed pretraining and fine-tuning paradigm for the task of RS object detection, whose four stages will be introduced in detail. Algorithm 1 summarizes these four stages concisely for clear understanding.

### A. Remote Sensing Classification Pretraining

The first stage is to train a deep learning-based classifier on the proposed large-scale upstream pretraining dataset, i.e., the RS pretraining dataset. Unlike the widely adopted ImageNet pretraining, this foremost step focuses on achieving in-domain classification pretraining, thereby making downstream fine-tuning free from transfer learning across different domains. In particular, following the standard practice for image classification [79], we construct the classifier as a deep backbone followed by an average pooling layer and a liner classification layer. As shown in Fig. 1, the classifier is lean and typical, but more sophisticated architectures can also work well, which will be verified in the experiments. Then, the classifier can be simply trained with the cross-entropy loss on the proposed RS pretraining dataset under image-level category supervision. Specifically, the classification loss is formulated as follows:

$$\mathcal{L}_{\text{cls}} = -\sum_{c=1}^{K} \gamma_c \log(p_c) \tag{1}$$

where $K$ is the number of classes ($K = 70$ for the used RS pretraining dataset) and $c$ is the category prediction. $\gamma_c$ denotes the binary indicator, where $\gamma_c = 1$ if class $c$ is the correct classification; otherwise, $\gamma_c = 0$. $p_c$ denotes the predicted probability of class $c$.

Therefore, this classifier has been pretrained on the large-scale RS pretraining dataset, and the weights and the backbone network can be directly inherited by downstream detectors for fine-tuning in the fourth stage. As shown in Fig. 1, by skipping the second and third stages, the pretraining approach is named RS classification pretraining, which essentially only pretrains the backbone, similar to the traditional ImageNet pretraining, but constrained to the RS domain with reduced feature disturbance from the NS. In other words, without the second and third stages, RS classification pretraining can be viewed as a weakened method of SCWL pretraining, which maintains high flexibility and efficiency. Moreover, it still appropriately achieves empirical performance gains over ImageNet pretraining.

### B. Pseudobounding Box Generation

*1) Motivations:* As illustrated in Fig. 1, the first stage and its corresponding RS classification pretraining approach

**Algorithm 1** RS SCWL Pretraining and Fine-Tuning

**Stage 1: RS Classification Pretraining**

**Input:** The pretraining image $\mathcal{I}$ from the large-scale RS pretraining dataset

**Output:** The category prediction $c$ and its predicted probability $p_c$

**Initialization:** From scratch

**Supervision:** Image-level category labels

**Pretraining:** Training the RS classifier using Equation (1)

**Stage 2: Pseudo Bounding Box Generation**

**Input:** The pretraining image $\mathcal{I}$ and the well-trained classifier in Stage 1

**Output:** Pseudo bounding boxes as localization ground-truths for Stage 3

**Steps:**

1. Feed the RS image $\mathcal{I}$ to the well-trained classifier
2. Generate its coarse CAM $\mathcal{S}$ and the normalized $\tilde{\mathcal{S}}$ using Equation (2)
3. Generate the cleaned saliency map $\mathcal{Z}$ using Equation (3) for WSOL
4. Yield the pseudo bounding boxes using Equation (4) and Equation (5)

**Stage 3: RS Detection Pretraining**

**Input:** The pretraining image $\mathcal{I}$ from the large-scale RS pretraining dataset

**Output:** The category prediction $c$, the corresponding classification probability $p_c$, and the predicted bounding boxes $t^u$

**Initialization:** Initialize the detector backbone by the pretrained weights in Stage 1 and randomly initialize other detector components

**Supervision:** Image-level category labels as strong-classification supervision and pseudo bounding boxes as weak-localization supervision

**Pretraining:** Training the RS detector using Equation (1), Equation (6) and Equation (7)

**Stage 4: Downstream RS Detector Fine-tuning**

**Input:** The image from the downstream RS detection dataset

**Output:** The category prediction, the corresponding classification probability, and the predicted bounding boxes

**Initialization:** Initialize the entire detector by the pretrained weights in Stage 3 except the last classification layer

**Supervision:** Instance-level category and bounding box labels

**Fine-tuning:** Training the RS detector using Equation (1), Equation (6) and Equation (7)

still follow the conventional pretraining paradigm since the backbone network is mainly pretrained as initialization for the downstream detector. In contrast, other detector components, including RPN, FPN, and detection head, are randomly initialized. However, as mentioned above, pretraining the entire detector, instead of only the backbone, is more reasonable and beneficial to downstream fine-tuning. To this end, it is necessary to construct a pretraining dataset with instance-level

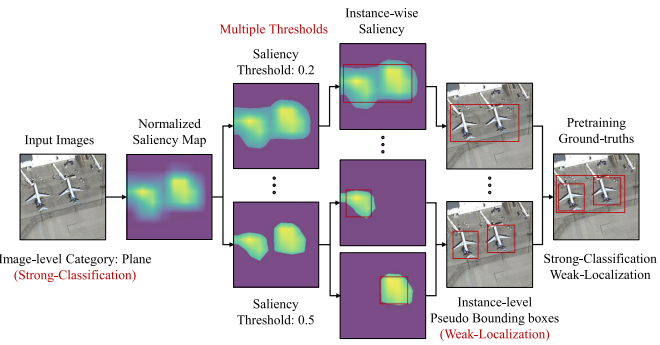


Fig. 2. Procedure of pseudobounding box generation. For each connected salient region, a bounding box is created to enclose potential object instances. Moreover, a multithreshold strategy is adopted to increase the recall, whose outputs are merged to form the final pseudoresults. These category-box pairs are treated as the ground truths for RS SCWL pretraining. This demonstrates that noisy SCWL labels are satisfactory for our method, which can achieve remarkable performance gains.

annotations to serve our proposed detection-specific pretraining paradigm. Naturally, based on the reconstructed RS pretraining dataset, it is reasonable to generate instance-level pseudobounding boxes and then integrate them with existing image-level category labels to form paired class-bounding box ground truths as SCWL supervision. Since each image in the RS pretraining dataset has been assigned an accurate ground-truth class, the problem can be transformed into *how to obtain available massive bounding box supervision automatically, instead of manually*.

*2) Salient Region Positioning:* Fig. 2 illustrates the procedure of our proposed pseudobounding box generation. Inspired by the WSOL methods [68], [69], [70], [80], with a well-trained RS scene classifier as described in Section IV-A, the salient regions containing potential geospatial objects can be roughly inferred and located by bounding boxes. For instance, the pioneer WSOL work [68], namely, CAM, first removes the activation function and global average pooling and then replaces the final linear classification layer corresponding to the ground-truth class with a $1 \times 1$ convolution, resulting in a CAM that highlights salient regions. Thus, in this work, we adopt this method to extract discriminative regions for pseudobounding box generation. Although many improved versions of CAM have been developed recently, such as Grad-CAM [81], Grad-CAM++ [82], Score-CAM [83], and LayerCAM [84], our method is still based on the original CAM due to its simplicity and ubiquity, which has already shown satisfactory results. In addition, data augmentation, such as multiple scales, horizontal flips, and other transformations, can be applied to the input images by averaging to enhance the quality of CAMs. Given an input image $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$ from the RS pretraining dataset in Fig. 2, its CAM is generated based on the well-trained classifier, denoted as $\mathcal{S} \in \mathbb{R}^{H\times W}$, and each pixel in CAM can then be normalized to [0, 1] to compress the background noise as follows:

$$\tilde{\mathcal{S}} = \frac{\mathcal{S} - \min \mathcal{S}}{\max \mathcal{S} - \min \mathcal{S}}. \tag{2}$$

To adaptively locate potential geospatial objects with their coordinates, we employ a threshold $\alpha \in [0, 1]$ to excite only

salient regions while suppressing the background, formulated as follows:

$$\mathcal{Z} = \mathbb{1}\{\tilde{\mathcal{S}} > \alpha\} \odot \tilde{\mathcal{S}} \tag{3}$$

where $\mathbb{1}$ is a binary indicator function and $\odot$ denotes elementwise multiplication. As illustrated in Fig. 2, for some specific categories like planes (PLs), a lower saliency threshold $\alpha$ tends to keep a larger area and, finally, derives a coarser bounding box, while a higher $\alpha$ may imply a more accurate region. However, for other categories, simply using only high thresholds may suppress the most salient pixels, leading to no pseudobounding boxes being generated. Therefore, to increase the recall across all classes, as depicted in Fig. 2, multiple saliency thresholds $\{\alpha_i \mid i = 1, 2, \ldots, \}$ can be utilized to generate a sequence of 2-D saliency maps $\{\mathcal{Z}_i\}$ (corresponding to multiple pseudobounding boxes) for each potential target. By default, we simultaneously apply four different thresholds, i.e., $i = 1, 2, 3, 4$, and the advantages of this multithreshold strategy will also be verified in the experiments. In addition, our method can flexibly handle multiple objects belonging to the same category appearing in an image. To filter out noisy small regions, within each $\mathcal{Z}_i$, any components whose area is less than half of the largest component's area are suppressed, while only large contiguous components will be connected to generate the final pseudobounding boxes in the following procedure.

*3) Bounding Box Determination:* With these secured salient components, pseudobounding boxes can be determined as follows. For the sake of generality, suppose that $\mathcal{P}_j$ denotes the point set of the $j$th salient component in $\mathcal{Z}_i$, and $p = (p_x, p_y)$ is a point in $\mathcal{P}_j$, where $p_x$ and $p_y$ are the horizontal and vertical coordinates, respectively. Consequently, the goal is to generate a pseudobounding box represented by $(x_j, y_j, w_j, h_j)$ to properly cover $\mathcal{P}_j$, where $(x_j, y_j)$, $w_j$, and $h_j$ are the coordinates of the center, width, and height of the bounding box, respectively. In practice, this bounding box can be calculated by matching the first and second moments of $\mathcal{P}_j$ with a potential rectangle. Here, the first and second moments are defined as the mean and variance, respectively, formulated as follows:

$$x_j = \frac{\sum_{p\in\mathcal{P}_j} p_x \ \mathcal{Z}_i(p)}{\sum_{p\in\mathcal{P}_j} \mathcal{Z}_i(p)}, \quad y_j = \frac{\sum_{p\in\mathcal{P}_j} p_y \ \mathcal{Z}_i(p)}{\sum_{p\in\mathcal{P}_j} \mathcal{Z}_i(p)} \tag{4}$$

$$w_j = \sqrt{12\frac{\sum_{p\in\mathcal{P}_j}\left(p_x - x_j\right)^2 \mathcal{Z}_i(p)}{\sum_{p\in\mathcal{P}_j} \mathcal{Z}_i(p)}}$$

$$h_j = \sqrt{12\frac{\sum_{p\in\mathcal{P}_j}(p_y - y_j)^2 \mathcal{Z}_i(p)}{\sum_{p\in\mathcal{P}_j} \mathcal{Z}_i(p)}}. \tag{5}$$

In this way, the bounding box for the point set $\mathcal{P}_j$ can be constructed. Notably, for different saliency thresholds, the above procedure for pseudobounding box determination should be repeated, which can prompt high recall. As illustrated in (3), for a specific object instance, higher thresholds generally favor activating fewer pixels, resulting in more compact bounding boxes. Conversely, lower thresholds tend to produce larger bounding boxes, which possibly involves more background. Finally, based on a predefined intersection-over-union (IoU) threshold $\beta$, these adjacent bounding boxes will be suppressed through NMS to improve the precision, whose optimal setting will be determined in the experiments. Using a higher IoU threshold typically results in more pseudobounding boxes for supervision, since only a small number of bounding boxes can be merged by NMS, and vice versa. In this way, the proposed RS pretraining dataset is smoothly equipped with instance-level pseudobounding box labels for localization while still inheriting image-level category annotations. Although some of the generated pseudobounding boxes may be noisy or even imprecise without any human correction, the original accurate (*strong*) class ground truths can still properly cooperate with these pseudo (*weak*) bounding box ground truths in the following detection-specific pretraining stages. Furthermore, this weak-localization supervision only takes effect during pretraining, and fine-tuning in the last stage will provide more accurate supervision as compensation. Hence, this approach has shown remarkable performance gains for downstream RS detection in experiments.

### C. Remote Sensing Detection Pretraining

The second stage, pseudobounding box generation, can smoothly transform our RS classification pretraining dataset into a detection dataset with both classification and localization annotations, making it possible to train a detector conveniently in an end-to-end manner in the third stage. It should be noted that this procedure for training detectors on our detection dataset is considered a crucial component of pretraining instead of downstream fine-tuning. The goal of this stage is to effectively introduce RS-domain detection-specific knowledge into the whole detector in advance, based on our available large-scale RS detection dataset, which covers almost all categories and scenarios of downstream RS detection datasets. In particular, the backbone is initialized with the pretrained classification model in the first stage, while other components serving detection, such as the detection head, FPN, and RPN, are initialized randomly. Then, the entire detector will be trained following the standard detection training [10], [12], [85]. Typically, the multitask loss for training object detectors can be unified as follows:

$$\mathcal{L}_{\text{pre}} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{loc}}\mathcal{L}_{\text{loc}} + \mathcal{L}_{\text{ext}} \tag{6}$$

where $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{loc}}$ denote the classification loss and the localization loss, respectively, and $\lambda_{\text{cls}}$ and $\lambda_{\text{loc}}$ are their corresponding balance weights. In practice, different detectors employ different classification and localization losses inside $\mathcal{L}_{\text{pre}}$. In addition, $\mathcal{L}_{\text{ext}}$ represents the specific loss for training extra components, like RPN in the detector. Taking the most representative two-stage detector Faster R-CNN [10] as an example, we adopt the cross-entropy loss as $\mathcal{L}_{\text{cls}}$ [same as (1)] and smoothed $L_1$ loss as $\mathcal{L}_{\text{loc}}$, formulated as follows:

$$\mathcal{L}_{\text{loc}} = \sum_{j\in\{x,y,w,h\}} \text{smooth-L}_1\left(t_j^u - v_j\right) \tag{7}$$

$$\text{smooth-L}_1(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases} \tag{8}$$

where $t^u = (t_x^u, t_y^u, t_w^u, t_h^u)$ and $v = (v_x, v_y, v_w, v_h)$ represent the 4-D coordinate tuples for the predicted bounding boxes and its ground truths, respectively. In this case, $\mathcal{L}_{\text{ext}}$ denotes the RPN loss, the same as defined in [10]. In contrast, for one-stage detectors, such as RetinaNet, $\mathcal{L}_{\text{ext}}$ is fixed to 0, and $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{loc}}$ follow their original definitions in [12].

Since the entire detector will be fully trained in an end-to-end manner, all detection components should have the capability to generalize discriminative semantics in various raw RS images, especially for downstream datasets with limited categories. Thus, the proposed pretraining approach in this stage is concise but actually overturns the traditional classification pretraining, including ImageNet pretraining and our RS classification pretraining in Section IV-A. Moreover, this SCWL pretraining paradigm is not constrained to a specific detection framework. Instead, it can be flexibly leveraged for various detectors, such as anchor-free and anchor-based detectors, with consistent performance improvement, which will be verified in experiments. Meanwhile, SCWL pretraining can work well with existing training strategies, including complicated data augmentation, ensuring its effectiveness for recent advanced Transformer-based RS detectors.

### D. Downstream Remote Sensing Detector Fine-Tuning

Aiming at fine-tuning downstream RS detection datasets, the last stage directly initializes the detector with the pretrained weights from the third stage. Unlike the traditional initialization strategies based on classification pretraining, the proposed SCWL pretraining paradigm supports the initialization of all detector layers, except for the last classification layer, which is determined by the number of geospatial object categories of the downstream detection dataset. After that, we fine-tune the entire detector following the normal training objective, the same as in (6), i.e., $\mathcal{L}_{\text{fine}} = \mathcal{L}_{\text{pre}}$.

## V. Experiments and Analysis

In this section, we conducted comprehensive experiments to compare the proposed SCWL pretraining paradigm with other pretraining approaches, and the results verify its effectiveness and superiority. First, the downstream RS detection datasets and evaluation metrics are briefly introduced. Then, various experimental setups are adopted for ablation studies. Finally, we present some quantitative and qualitative detection results, boosted by our pretraining method, on three widely used RS detection datasets.

### A. Dataset Description

In order to extensively evaluate the proposed pretraining methods, we leverage our large-scale RS pretraining dataset for upstream pretraining and three public RS object detection datasets for downstream fine-tuning, namely, DOTA [29], DIOR [28], and NWPU VHR-10 [14], [30] in the experiments.

*1) DOTA Dataset:* DOTA is a very representative RS dataset [29], consisting of 2806 optical RS images with a total of 188 282 manually annotated object instances. Notably, various complex RS scenarios, such as diverse object scales and appearances, unpredictable imaging conditions, and different spatial resolutions from 800 × 800 to 4000 × 4000 pixels, make this dataset more challenging and closer to real-world applications. Moreover, with both horizontal and rotated bounding box annotations, this dataset can be extensively utilized for both object detection and oriented object detection in the RS domain. We also conducted ablation studies on this dataset to demonstrate the effectiveness of the proposed pretraining paradigms. The DOTA dataset covers 15 different categories, almost all of which are included in our RS pretraining dataset, defined as PL, baseball diamond (BD), bridge (BR), ground field track (GFT), small vehicle (SV), large vehicle (LV), ship (SH), tennis court (TC), basketball court (BC), storage tank (ST), soccer ball field (SBF), roundabout (RA), harbor (HA), swimming pool (SP), and helicopter (HC). In the experiments, considering the huge computational load of large-resolution images, we uniformly split them into subimages of size 600 × 600. The ratio of the number of images in the training set, the validation set, and the test set is 3:1:2.

*2) DIOR Dataset:* DIOR [28] is the largest publicly available dataset for object detection in RS images. It is composed of 23 463 images, all scaled to the same resolution, i.e., 800 × 800 pixels. This dataset involves 20 different RS object categories, denoted as c1–c20 for brevity: airplane (c1), airport (c2), baseball field (c3), BC (c4), BR (c5), chimney (c6), dam (c7), expressway service area (c8), expressway toll station (c9), golf course (c10), ground track field (c11), HA (c12), overpass (c13), SH (c14), stadium (c15), ST (c16), TC (c17), train station (c18), vehicle (c19), and windmill (c20). Furthermore, most of the categories have been included in our reconstructed RS pretraining dataset. In the experiments, we usually partition the DIOR dataset into two equal subsets: the training set and the test set.

*3) NWPU VHR-10 Dataset:* NWPU VHR-10 is another widely used dataset for RS object detection, released by Cheng et al. [30], which contains a total of 800 very-high-resolution (VHR) optical RS images. Of these, 650 are positive samples, and the remaining 150 images are negative samples, i.e., with background only. The NWPU VHR-10 dataset contains ten different classes, namely, airplanes, SHs, STs, BDs, TCs, BCs, ground track fields, HAs, BRs, and vehicles. In our experiments, 75% of the positive images are randomly selected as the training set, while the others are used for testing.

### B. Evaluation Metrics and Implementation Setup

To quantify the performance of different detectors with different pretraining strategies, we employ several widely used evaluation metrics in the experiments. In general, all detection results can be divided into four types: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). Precision and recall can be calculated by counting the number of samples per case as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{9}$$

TABLE II

COMPARISON OF ARCHITECTURE DETAILS IN DIFFERENT STAGES, TAKING FASTER R-CNN AS EXAMPLE. "DET," "CLS," AND "REG" ARE ABBREVIATIONS FOR "DETECTION," "CLASSIFICATION," AND "REGRESSION," RESPECTIVELY. "GAP" MEANS THE GLOBAL AVERAGE POOLING OPERATION, WHILE "FC" IS THE FULLY CONNECTED LAYER. "SHARED" INDICATES THAT THESE LAYERS ARE SHARED BY THE CLASSIFICATION BRANCH AND THE REGRESSION BRANCH

| Component Types | Stage 1 | Stage 3 | Stage 4 |
|---|---|---|---|
| Backbone | ResNet50: | (7×7, 64, Stride 2), (3×3, Max Pooing, Stride 2), $\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$, $\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 6$, | $\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 4$, $\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$ |
| Neck | GAP | FPN: $\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \end{bmatrix} \times 4$ | |
| Head | CLS Head: 70-d FC | RPN: Shared (3 × 3, 256), CLS (1 × 1, 3), REG (1 × 1, 12) | |
| | | DET Head: Shared 1024-d FC×2, CLS 71-d FC, REG 280-d FC | CLS 16-d FC, REG 60-d FC |

TABLE III

TRAINING DETAILS FOR DIFFERENT STAGES. † REPRESENTS THAT, UNLESS OTHERWISE SPECIFIED, ALL DETECTORS ARE FINE-TUNED FOR 12 EPOCHS ON A SPECIFIC DOWNSTREAM DATASET. ‡ MEANS THAT THE TRAINING TIME IS CALCULATED BASED ON THE ARCHITECTURES IN TABLE II

| Training Details | Stage 1 | Stage 3 | Stage 4 |
|---|---|---|---|
| Training Dataset | Our Reconstructed RS Pretraining Dataset | | DOTA, DIOR, or NWPU VHR-10 |
| Total Epochs | 100 Epochs | 12 Epochs | 12 Epochs† |
| Optimizer | Momentum Stochastic Gradient Descent (SGD) | | |
| Initial Learning Rate | 0.1 | 0.01 | |
| Momentum | 0.9 | | |
| Weight Decay | 0.0001 | | |
| Learning Rate Decay Policy | Decreased by 0.1 on the 30th, 60th, and 90th epochs | Decreased by 0.1 on the 8th and 11st epochs | |
| Anchor Settings of RPN (If Applicable) | – | Base size: $[4, 8, 16, 32, 64]$, Strides: $[4, 8, 16, 32, 64]$, Ratios: $[0.5, 1.0, 2.0]$ | |
| Training Time‡ | 21.6 hours | 24.5 hours | 4.2 hours for DOTA, 6.4 hours for DIOR, 12 minutes for NWPU VHR-10 |

Then, by applying different thresholds to examine multiple precision and recall values, the precision–recall (PR) curve can be plotted, while average precision (AP) is defined as the area under the PR curve. In multiclass object detection, mAP denotes the mean value of APs of all object categories, computed as follows:

$$\text{mAP} = \frac{1}{\mathcal{C}} \sum_{k=1}^{\mathcal{C}} \text{AP}_k \tag{10}$$

where $\mathcal{C}$ is the total number of categories and $k$ represents the category index. Generally, a higher mAP implies better detection performance. In our experiments, unless otherwise specified, mAP (or $\text{mAP}_{50}$) means an IoU threshold of 0.5, while $\text{mAP}_{75}$ means an IoU threshold of 0.75. In particular, $\text{mAP}_{50:95}$ represents the multithreshold mAP averaged with ten different IoU thresholds, i.e., 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, and 0.95.

The proposed methods in this work are independent of the detector architecture, and its extensive effectiveness on various frameworks will be validated in the experiments. Unless otherwise specified, we adopt Faster R-CNN [10] with the ResNet50 [79] backbone as the base detector due to their widespread application in RS object detection. Table II presents and compares the detailed architectures of different stages, where all stages share the same backbone, while Stage 3 and Stage 4 share the same FPN, RPN, and part of the detection head. As mentioned above, this work does not contribute to the detection architectures. Thus, for fairness and generality, the backbone ResNet50, the neck FPN, RPN, and the detection head also follow the same structures in their original papers [10], [31], [79], respectively. For example, as illustrated in Table II, 7 × 7 refers to that the kernel size of the convolutional layer is 7, and 64 represents its number of output channel. 70-d means that the dimension of the output channels of the FC layer is 70. In addition, Table III summarizes, in detail, all the training settings at different stages, and each model is trained and evaluated on an NVIDIA Tesla V100 GPU. Four different pretraining strategies will be extensively analyzed in the experiments, i.e., training from scratch (without any pretraining procedures on the upstream datasets), ImageNet pretraining, RS classification pretraining, and RS SCWL pretraining.

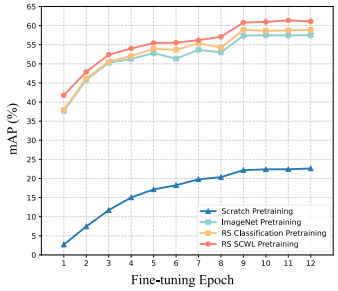


Fig. 3. Comparison of the learning curves during fine-tuning for downstream RS object detection, which are initialized with four different pretrained models. The RS SCWL pretraining paradigm achieves the best accuracy throughout the fine-tuning procedure.

### C. Ablation Studies and Analysis

In this section, we carry out comprehensive ablation studies on the DOTA dataset to substantiate the effectiveness of the proposed pretraining approaches under various experimental settings and then thoroughly compare and analyze the results.

*1) Effect of Different Pretraining Strategies on RS Object Detection:* Two new pretraining methods are developed in this work, namely, RS classification pretraining and RS SCWL pretraining. The latter essentially serves as a pioneering detection-specific pretraining paradigm. As mentioned above, in addition to them, two previous pretraining methods, i.e., training from scratch and ImageNet pretraining, are

TABLE IV
PERFORMANCE EVALUATION AND COMPARISON OF FOUR DIFFERENT PRETRAINING PARADIGMS

| Methods | PL<br>BC | BD<br>ST | BR<br>SBF | GTF<br>RA | SV<br>HA | LV<br>SP | SH<br>HC | TC | $mAP_{50}$ (%) | $\Delta mAP_{50}$ (%) | $mAP_{75}$ (%) | $mAP_{50:95}$ (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| From Scratch | 55.8<br>0.5 | 2.3<br>28.7 | 0.0<br>0.4 | 0.1<br>0.0 | 39.8<br>15.7 | 61.1<br>31.9 | 53.9<br>8.5 | 40.9 | 22.6 | – | 9.5 | 11.2 |
| ImageNet Pretraining | 78.9<br>43.9 | 60.5<br>55.4 | 34.5<br>47.2 | 51.1<br>51.7 | 58.4<br>67.5 | 78.4<br>43.4 | 74.6<br>28.0 | 89.2 | 57.5 | +34.9 | 33.8 | 33.8 |
| RS Classification Pretraining | 79.7<br>**53.1** | **61.5**<br>60.5 | 35.5<br>46.0 | 47.4<br>53.7 | 57.9<br>65.8 | 79.2<br>45.5 | 74.7<br>**34.0** | 89.1 | 58.9 | +36.3 | 35.6 | 34.0 |
| RS SCWL Pretraining | **82.3**<br>52.7 | 58.2<br>**62.2** | **38.6**<br>**53.8** | **53.5**<br>**58.5** | **59.1**<br>**70.1** | **81.0**<br>**47.9** | **76.3**<br>30.7 | **91.4** | **61.1** | **+38.5** | **40.5** | **37.8** |

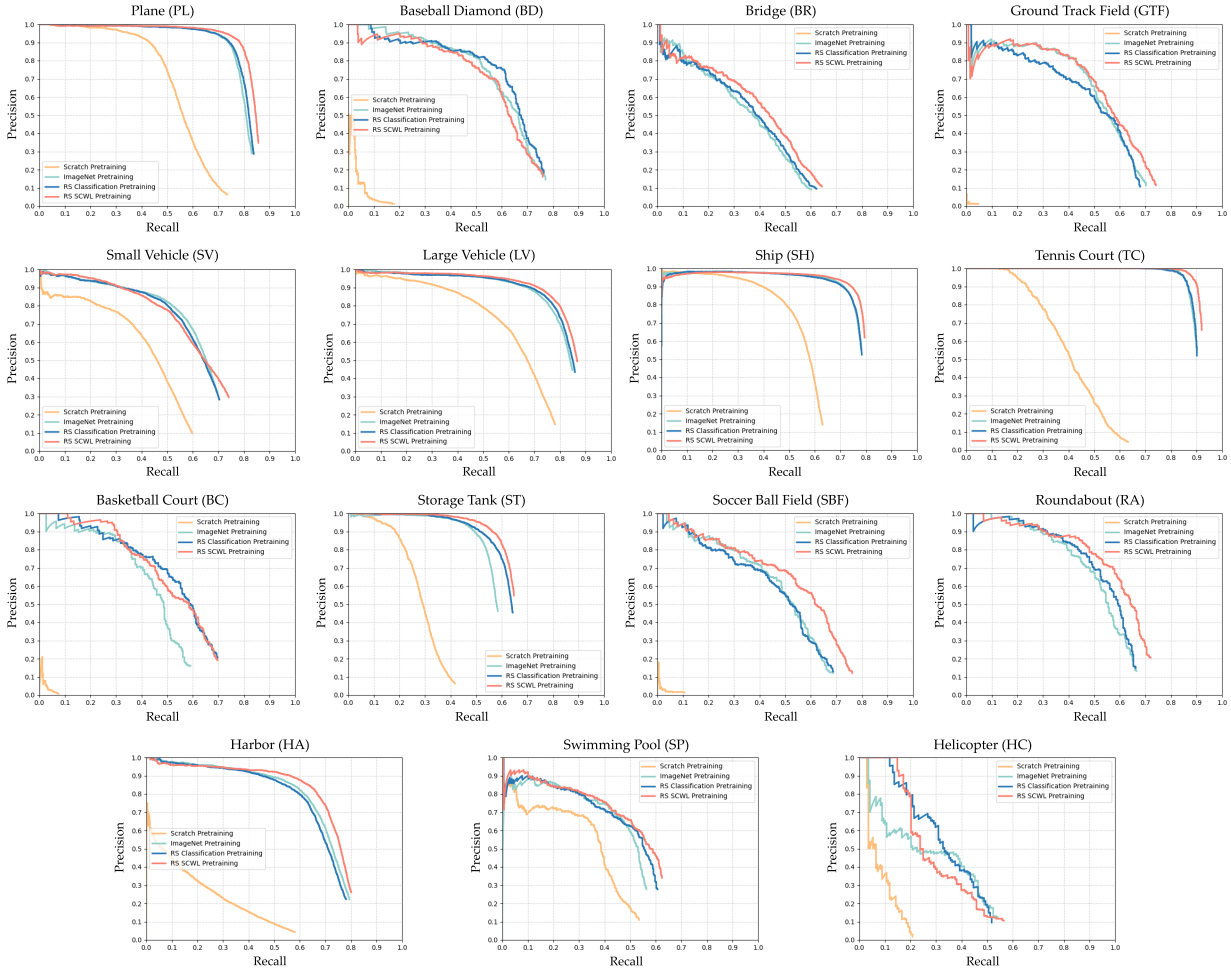


Fig. 4. PR curves of four different pretraining paradigms on the DOTA dataset.

also investigated in the ablation experiments for comparison. As shown in Table IV, the detector trained from scratch can only achieve 22.6% $mAP_{50}$, dramatically lower than 57.5% for traditional ImageNet pretraining. This result implies the necessity of pretraining for the task of RS object detection, as the related downstream datasets usually have limited training data, which is insufficient to train a modern detector well from scratch. When applying RS classification pretraining and RS SCWL pretraining, the detection results are 58.9% and 61.1%, respectively, which are 1.4% and 3.6% higher than ImageNet pretraining. For the other two evaluation metrics, $mAP_{75}$ and $mAP_{50:95}$, similar performance gains can be obtained, as shown in Table IV. Moreover, in terms of categorywise detection performance, our proposed RS SCWL pretraining consistently brings considerable benefits to almost every geospatial class, which proves its generalizability in different scenes.

Fig. 3 presents the learning curves during downstream fine-tuning for epochwise performance comparison. It can be noticed that these four methods follow similar convergence trends under a given training schedule, while RS SCWL pretraining consistently provides the most noticeable accuracy boost in each fine-tuning epoch. In addition, we plot the PR curves of these four pretraining methods in Fig. 4, where RS SCWL pretraining also generally outperforms the other methods on most RS object categories.

TABLE V

ABLATION EXPERIMENTS ON USING DIFFERENT AMOUNTS OF RS PRETRAINING DATASET

| Methods | Proportion of the Used RS Pretraining Data | | | | Average mAP |
|---|---|---|---|---|---|
| | 1/4 | 1/2 | 3/4 | 1 | |
| RS Classification Pretraining | 57.7 | 57.9 | 58.4 | **58.9** | 58.2 |
| RS SCWL Pretraining | 58.4 | 59.9 | 60.8 | **61.1** | 60.1 |

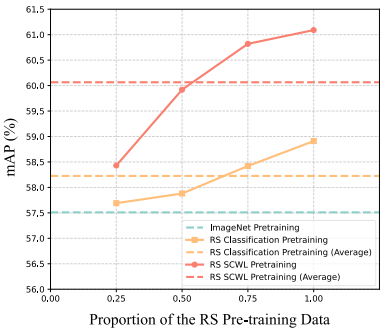


Fig. 5. Downstream fine-tuning results yielded by the same RS detector but pretrained on different proportions of the upstream pretraining data. On the horizontal axis, 0.25, 0.50, 0.75, and 1.00 mean that one-quarter, half, three-quarters, and all of the samples from the pretraining dataset are used to pretrain the detector, respectively. The yellow and red dashed lines represent the averaged final detection results of the two proposed pretraining strategies, while the green line represents the traditional ImageNet pretraining.

*2) Effect of the RS Pretraining Dataset Size:* As described above, we reconstruct a large-scale pretraining dataset in the RS scenes, namely, the RS pretraining dataset, with both classification and localization supervision. As mentioned above, most existing RS detection datasets only include limited training samples, so the size of the pretraining data plays an important role in improving the final detection results. Table V shows that both proposed RS pretraining strategies prefer larger pretraining datasets, and downstream detection performance improves steadily with increasing pretraining data. Our RS SCWL pretraining outperforms RS classification pretraining at all different sizes, indicating the advancement of the novel pretraining paradigm. When all images are utilized, RS SCWL pretraining can achieve the best performance. As shown in Fig. 5, they both outperform the previous ImageNet pretraining with average mAP gains of 0.7% and 2.6%, which demonstrates the necessity of building a large-scale pretraining dataset specifically for the RS scenarios.

*3) Effect of Freezing the Pretrained Detection Backbone:* Both ImageNet pretraining and RS classification pretraining still follow the paradigm of only pretraining backbone networks, albeit in different domains, whereas RS SCWL pretraining aims to pretrain the entire detector. To determine whether the improvement only comes from pretraining the additional detection components, we conducted an ablation experiment, and the results are shown in Table VI. Specifically, during the detector pretraining stage (the third stage in Section IV-C), the weights of the backbone are frozen, while only the remaining components, i.e., the RPN, FPN, and detection heads in Faster R-CNN, are pretrained. As reported in the last row of Table VI, despite a gain of +1.8% $mAP_{50}$ over ImageNet pretraining, freezing the backbone produces worse detection results than our proposed normal RS SCWL pretraining strategy in the third row, i.e., 59.3% versus 61.1% $mAP_{50}$. Therefore, the experimental results suggest that RS SCWL pretraining enhances the representation capabilities of all layers (including its backbone) and adapts the whole detector to RS-specific object detection, rather than merely pretraining additional layers.

TABLE VI

PERFORMANCE COMPARISON OF FREEZING THE BACKBONE AND ONLY PRETRAINING THE ADDITIONAL COMPONENTS BASED ON THE PROPOSED RS SCWL PRETRAINING STRATEGY

| Methods | $mAP_{50}$ (%) | $\Delta mAP_{50}$ (%) | $mAP_{75}$ (%) | $mAP_{50:95}$ (%) |
|---|---|---|---|---|
| ImageNet Pretraining | 57.5 | – | 33.8 | 33.8 |
| RS Classification Pretraining | 58.9 | +1.4 | 35.6 | 34.0 |
| RS SCWL Pretraining | **61.1** | **+3.6** | **40.5** | **37.8** |
| RS SCWL Pretraining (Freeze) | 59.3 | +1.8 | 39.1 | 37.2 |

TABLE VII

PERFORMANCE COMPARISON OF THE FOUR DIFFERENT PRETRAINING METHODS WITH LONGER FINE-TUNING TIME

| Methods | 1× | | 2× | | 3× | |
|---|---|---|---|---|---|---|
| | mAP (%) | ΔmAP (%) | mAP (%) | ΔmAP (%) | mAP (%) | ΔmAP (%) |
| From Scratch | 22.6 | – | 25.2 | – | 25.2 | – |
| ImageNet Pretraining | 57.5 | +34.9 | 58.8 | +33.6 | 58.8 | +33.6 |
| RS Classification Pretraining | 58.9 | +36.3 | 59.6 | +34.4 | 59.8 | +34.6 |
| RS SCWL Pretraining | **61.1** | **+38.5** | **61.9** | **+36.7** | **61.8** | **+36.6** |

*4) Effect of a Longer Fine-tuning Schedule:* The default training strategy utilizes a 12-epoch (1×) schedule. We experiment to explore the effect of longer training schedules, such as 24 epochs (2×) and 36 epochs (3×), and Table VII tabulates the detection results of the four pretraining methods. Note that the training schedule here refers to fine-tuning on the downstream RS detection dataset rather than pretraining, and in practice, three different schedules exploit the same pretrained model as initialization for fine-tuning. It can be seen from Table VII that extending the fine-tuning time cannot compensate for the performance differences brought by different pretraining strategies. Although the accuracy gaps are reduced marginally for longer schedules, the proposed RS SCWL pretraining paradigm still leads its counterparts, achieving the best result of 61.9% mAP when trained with 2×. The benefits of our state-of-the-art pretraining method can be maintained accordingly, which also validates its effectiveness for long training schedules.

*5) Effect of Different Pretraining Paradigms on Low-Data Object Detection:* The proposed pretraining paradigm is capable of benefiting low-data object detection, which is of significant practical value and alleviates annotation labor. Table VIII tabulates the downstream detection results with different pretraining strategies, where we actually perform low-shot ($k$-shot) object detection by sampling a fixed number of training samples per class on the downstream dataset. It has

TABLE VIII

PERFORMANCE COMPARISON OF $k$-SHOT LOW-DATA RS OBJECT DETECTION WITH THREE DIFFERENT PRETRAINING STRATEGIES. $k$ REFERS TO ONLY $k$ TRAINING SAMPLES ACTIVATED FOR EACH RS CATEGORY

| Methods | $k = 50$ | | $k = 100$ | | $k = 200$ | | $k = 500$ | | Average |
|---|---|---|---|---|---|---|---|---|---|
| | mAP (%) | ΔmAP (%) | mAP (%) | ΔmAP (%) | mAP (%) | ΔmAP (%) | mAP (%) | ΔmAP (%) | mAP (%) |
| ImageNet Pretraining | 1.3 | – | 10.2 | – | 17.0 | – | 31.2 | – | 14.9 |
| RS Classification Pretraining | 1.9 | +0.6 | 11.1 | +0.9 | 20.1 | +3.1 | 32.3 | +1.1 | 16.4 |
| RS SCWL Pretraining | **4.0** | **+2.7** | **13.3** | **+3.1** | **25.6** | **+8.6** | **37.0** | **+5.8** | **20.0** |

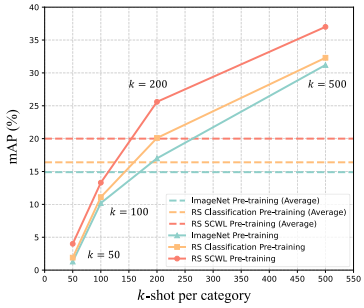


Fig. 6. Low-shot detection results following the same downstream fine-tuning mechanism but using three different pretrained models. By varying the number of training samples per category on the horizontal axis, different performance gaps can be observed.

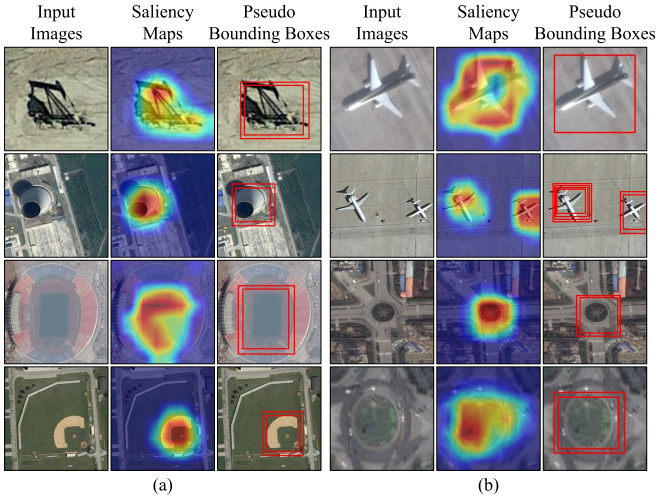


Fig. 7. Visualization of some original input images, their saliency maps, and pseudobounding boxes. All example images are from the reconstructed RS pretraining dataset. (a) Results for different geospatial categories. (b) Results for different object scales with the same categories.

been reported that, compared to classification pretraining, including the typical ImageNet pretraining and the proposed RS classification pretraining, our detector-specific RS SCWL pretraining consistently improves detection accuracy by large margins under diverse low-data settings. Fig. 6 visually compares the fine-tuning accuracy differences based on the three pretraining methods. In the 200-shot case, it achieves the most noticeable performance gain, about +8.6% mAP over the ImageNet pretraining. Moreover, for low-data detection, RS SCWL pretraining generally achieves more remarkable improvements than the above full-data experiments. These improvements during fine-tuning are likely attributed to the efficient utilization of rich RS in-domain data in the pretrained models, which also demonstrates the effectiveness of our proposed pretraining paradigm.

*6) Visualization of Pseudobounding Boxes Based on RS SCWL Pretraining:* In Fig. 7, we visualize some examples of original images, the salient regions (i.e., the corresponding CAMs), and the generated pseudobounding boxes in the second stage of the proposed RS SCWL pretraining (see Section IV-B). Through the visualization, it can be observed that, given the categories, most foreground geospatial object instances are roughly localized by computing and cleaning the CAMs, despite the complicated background. Furthermore, in multi-instance scenarios, the proposed method can distinguish multiple neighboring instances belonging to the same category. As shown in Fig. 7, compared to manual annotations, the produced pseudobounding boxes are relatively noisy (weak), such as incomplete, redundant, and missing boxes, yet still significantly contribute to RS detection. This procedure provides discriminative knowledge from a large number of RS in-domain images, making our detection-specific pretraining (see Section IV-C) effective. Finally, during fine-tuning, the detector is able to inherit favorable knowledge in almost all layers for more efficient learning on only a limited amount of downstream samples.

*7) Effect of Saliency Thresholds on Pseudobounding Box Generation:* As elaborated on in Section IV-B, we adopt a set of saliency thresholds $\{\alpha_i\}$ to generate four CAMs simultaneously, where $\{\alpha_1, \alpha_2, \alpha_3, \alpha_4\} = \{0.2, 0.3, 0.4, 0.5\}$ in practice. Then, salient regions are localized and extracted for pseudobounding box generation based on multiple CAMs, instead of a single CAM to enhance recall and accuracy. Table IX tabulates the downstream detection results with different thresholds, which validates the effectiveness of this multithreshold strategy. Specifically, multiple thresholds will result in diverse bounding box characteristics. As shown in Table IX, as the threshold $\alpha$ increases, the average number of boxes per image slightly increases, while the enclosed area decreases significantly, indicating that only the most discriminative parts of potential objects are focused. Notably, our proposed multithreshold strategy encourages more bounding boxes, on average 3.24 boxes per image, dramatically more than the number generated by utilizing a single threshold. Moreover, regarding downstream RS object detection, it outperforms the other four in three evaluation metrics. In fact, due to large variations in the scale of RS objects, it is reasonable

TABLE IX

EFFECTS OF DIFFERENT SALIENCY THRESHOLDS ON THE CHARACTERISTICS OF THE PSEUDOBOUNDING BOXES AND THE PERFORMANCE OF DOWNSTREAM RS DETECTION. THE FIRST FIVE ROWS PROVIDE STATISTICAL PROPERTIES OF THE GENERATED BOUNDING BOXES, WHILE THE LAST THREE ROWS REPORT THE CORRESPONDING DETECTION ACCURACY. "MULTIPLE THRESHOLDS" DENOTE THAT THE MULTITHRESHOLD STRATEGY IS UTILIZED, WHERE THE RESULTS FROM ALL FOUR DIFFERENT THRESHOLDS ARE AGGREGATED

| Saliency Threshold | $\alpha = 0.2$ | $\alpha = 0.3$ | $\alpha = 0.4$ | $\alpha = 0.5$ | Multiple Thresholds |
|---|---|---|---|---|---|
| Number of Boxes / Image | 1.12 | 1.13 | 1.15 | 1.15 | 3.24 |
| Box Center Coordinates $x$ | 142.67 | 143.89 | 144.99 | 145.85 | 143.17 |
| Box Center Coordinates $y$ | 142.66 | 143.68 | 144.50 | 145.09 | 142.73 |
| Bounding Box Width | 135.18 | 118.78 | 103.23 | 88.04 | 106.28 |
| Bounding Box Height | 131.80 | 115.56 | 100.23 | 85.29 | 103.24 |
| $mAP_{50}$ | 59.6 | 60.0 | 60.7 | 60.5 | **61.1** |
| $mAP_{75}$ | 36.5 | 36.9 | 38.1 | 37.8 | **40.5** |
| $mAP_{50:95}$ | 35.1 | 35.3 | 36.8 | 36.0 | **37.8** |

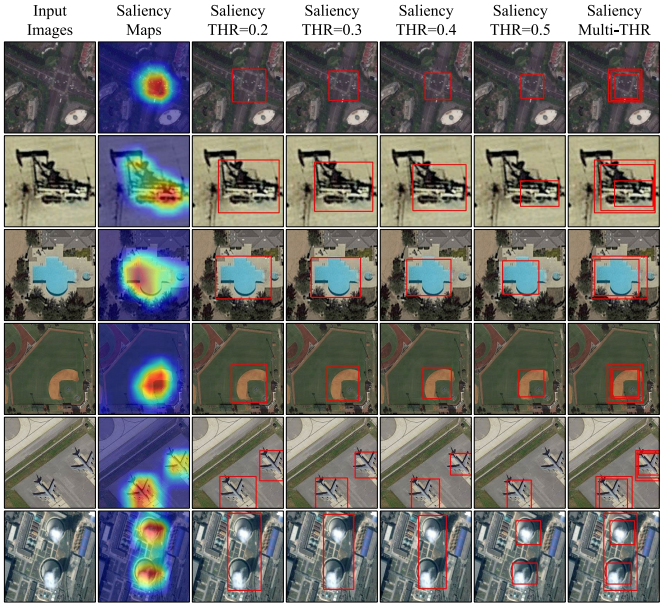


Fig. 8. Visualization of the pseudobounding boxes produced by different saliency thresholds. It can be noticed that using multiple thresholds tends to produce more bounding boxes. "Multi-THR" means the simultaneous use of four thresholds, i.e., 0.2, 0.3, 0.4, and 0.5.

and necessary to employ multiple thresholds to aggregate salient regions of different sizes, with the benefits verified in Table IX. Fig. 8 presents some pseudobounding boxes with different saliency thresholds. As illustrated in the top five rows, utilizing only high thresholds such as 0.5 may lead to visibly incorrect results that only part of instances are enclosed or even no bounding box survives since, in these cases, the saliency may not highlight the entire foreground objects, causing high thresholds oversuppressing some border regions. Interestingly, using high thresholds can also generate more accurate results in the last row, where the two instances with the same category are separately distinguished. However, the opposite patterns can be observed for other lower thresholds that perform better on other examples except the last one. Therefore, it can be inferred that different objects in different contexts prefer

TABLE X

EFFECTS OF DIFFERENT NMS IOU THRESHOLDS ON THE CHARACTERISTICS OF THE PSEUDOBOUNDING BOXES AND THE PERFORMANCE OF DOWNSTREAM RS OBJECT DETECTION. THE THRESHOLD VARIES FROM 0.5 TO 1.0, WHILE SETTING $\beta$ TO 0.8 YIELDS THE BEST RESULTS

| NMS IoU Threshold | $\beta = 0.5$ | $\beta = 0.6$ | $\beta = 0.7$ | $\beta = 0.8$ | $\beta = 0.9$ | $\beta = 1.0$ |
|---|---|---|---|---|---|---|
| Number of Boxes / Image | 1.89 | 2.20 | 2.57 | 3.24 | 4.26 | 4.54 |
| Box Center Coordinates x | 143.13 | 143.05 | 143.24 | 143.17 | 143.93 | 144.36 |
| Box Center Coordinates y | 142.84 | 142.73 | 142.91 | 142.73 | 143.53 | 143.99 |
| Bounding Box Width | 112.65 | 109.81 | 108.42 | 106.28 | 108.00 | 111.12 |
| Bounding Box Height | 109.70 | 106.81 | 105.41 | 103.24 | 104.91 | 108.04 |
| $mAP_{50}$ | 60.6 | 60.6 | 60.8 | **61.1** | **61.1** | 60.8 |
| $mAP_{75}$ | 38.4 | 40.0 | 39.8 | **40.5** | 40.3 | 38.5 |
| $mAP_{50:95}$ | 36.7 | 37.4 | 37.4 | **37.8** | 37.7 | 36.7 |

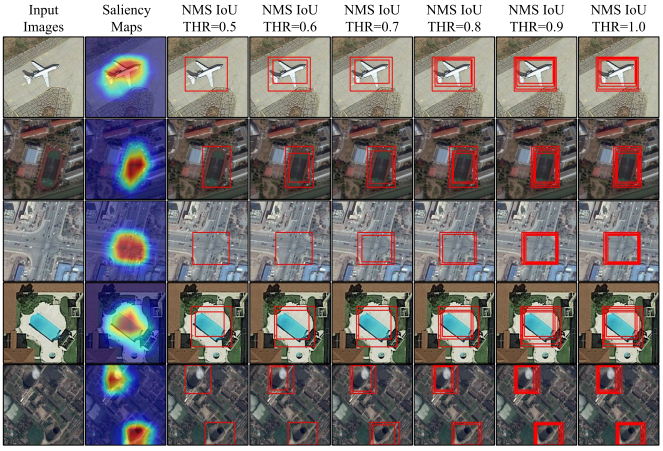


Fig. 9. Visualization of pseudobounding boxes with different NMS IoU thresholds. Utilizing higher thresholds tends to keep more boxes, and 0.8 is the default setting in our experiments.

different saliency thresholds, and the proposed multithreshold strategy favorably handles this challenge, achieving a high recall of pseudobounding boxes across various cases.

*8) Effect of the NMS IoU Threshold on Pseudobounding Box Generation:* After applying the multithreshold strategy, several replicated boxes are generated for each instance, all of which will be suppressed by NMS to form the final pseudobounding boxes. As with the standard NMS postprocessing in the object detection pipeline, there is a hyperparameter, the NMS IoU threshold, to be predefined with the range from 0 to 1. Thus, to further explore its effect, we preset different NMS IoU thresholds in the pseudobounding boxes generation stage, and the corresponding bounding box statistics and downstream detection results are shown in Table X. It is suggested that, as the threshold $\beta$ increases, the average number of boxes increases steadily, whereas the position, shape, and size of the bounding boxes only vary marginally. Empirically, when $\beta = 0.8$, the best detection accuracy can be obtained. Therefore, it is considered the default setting in our experiments. Fig. 9 visualizes some example pseudobounding boxes with different NMS IoU thresholds.

*9) Effect of Different Pretraining Paradigms on Oriented Object Detection:* The proposed pretraining paradigms can be properly extended to oriented object detection, which is a representative task in the field of RS with promising

TABLE XI
PERFORMANCE COMPARISON OF DIFFERENT PRETRAINING PARADIGMS ON ORIENTED OBJECT DETECTION IN RS SCENARIOS

| Methods | $mAP_{50}$ (%) | $\Delta mAP_{50}$ (%) | $mAP_{75}$ (%) | $mAP_{50:95}$ (%) |
|---|---|---|---|---|
| From Scratch | 22.9 | – | 5.5 | 9.2 |
| ImageNet Pretraining | 53.9 | +31.0 | 17.2 | 24.5 |
| RS Classification Pretraining | 55.3 | +32.4 | 20.1 | 25.9 |
| RS SCWL Pretraining | **60.1** | **+37.2** | **25.9** | **30.9** |

practical value. Table XI tabulates the downstream oriented object detection results of different pretraining paradigms, where the oriented Faster R-CNN is developed as the basic detector with rotated bounding box (instead of horizontal bounding box) representations. It can be observed that pretraining still plays an indispensable role in this task, and the proposed RS SCWL pretraining substantially enhances the downstream-oriented object detector to produce the best results across different evaluation criteria. Specifically, our RS SCWL pretraining achieves 60.1% $mAP_{50}$, surpassing 53.9% $mAP_{50}$ of the traditional ImageNet pretraining. Regarding the other two more challenging metrics $mAP_{75}$ and $mAP_{50:95}$, our strategy also consistently outperforms other pretraining paradigms by a large margin. The above experimental results have demonstrated that the detector can be preliminarily empowered with localization awareness of geospatial objects by the proposed RS SCWL pretraining, thereby benefiting downstream-oriented object detection.

*10) Effect of Various Backbone Networks:* To study the adaptability of the proposed pretraining paradigm to different backbones, we conducted extensive experiments and compared their downstream RS detection performance in Table XII. Concretely, ResNet [79] and Swin Transformer [86] are adopted as the current mainstream representative backbones, CNN-style and Transformer-style, respectively. For all four pretraining methods, Swin Transformer performs better than ResNet but usually requires more parameters and higher computational loads. In terms of training from scratch, gradually replacing ResNet18 with a larger ResNe101 leads to a significant accuracy drop, from 29.7% to 21.8% mAP, which implies that the RS detector cannot be fully trained with only the downstream data, and numerous parameters are suboptimal in such a large backbone. In contrast, for the same backbones (ResNet18 to ResNet101), the proposed RS SCWL pretraining is capable of improving the detection performance from 59.3% to 63.0% mAP. Furthermore, our novel pretraining paradigm adapts well to the Transformer-style backbones, achieving the best absolute accuracy at 67.2% mAP with the Swin-Large backbone.

Generally, as reported in Table XII, advanced or sophisticated backbones deliver improved detection performance, but, more importantly, it can be observed that our proposed RS SCWL pretraining surpasses all other pretraining strategies, consistently with various backbones networks, which demonstrates its superiority.

*11) Effect of Various Object Detector Architectures:* In theory, the proposed pretraining methods are independent of the detection architecture and can be transferred to all existing detectors. We exploit Faster R-CNN [10], a two-stage anchor-based detector, by default, in most experiments. Furthermore, we introduce two famous one-stage frameworks, RetinaNet [12] and FCOS [85]. The latter serves as a recently advanced anchor-free detector, as opposed to traditional anchor-based detectors. Table XIII compares the detection performance of the four pretraining strategies with three different types of architectures. Two types of backbones with the same-level model size, ResNet50 and Swin-Tiny, are also introduced for an extensive comparison. With a simple pipeline, RetinaNet requires less computational effort but reduces detection accuracy. Noticeably, the advanced anchor-free detector FCOS is not only computationally efficient but also achieves results comparable to the sophisticated Faster R-CNN, e.g., 63.8% versus 63.5% mAP with the Swin-Tiny backbone. Moreover, it can be concluded from Table XIII that our RS SCWL pretraining paradigm is able to generalize effectively to various detection architectures and consistently outperforms other strategies.

Fig. 10. Qualitative downstream detection results of faster R-CNN with the proposed RS SCWL pretraining paradigm on the NWPU VHR-10 dataset.

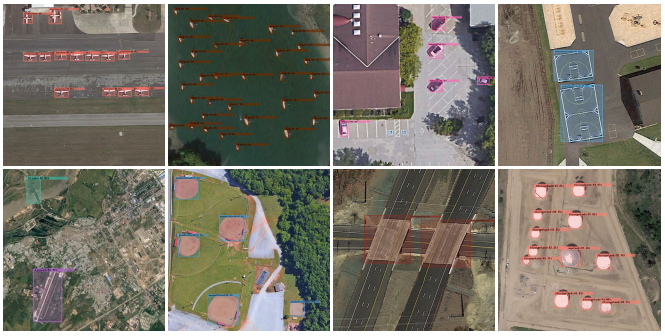

Fig. 11. Qualitative downstream detection examples of faster R-CNN with the proposed RS SCWL pretraining paradigm on the DIOR dataset.

### D. Comparison on Different RS Object Detection Datasets

This section compares the detection results improved by our proposed RS SCWL pretraining paradigm with other competitive algorithms on three popular RS object detection datasets, namely, NWPU VHR-10, DIOR, and DOTA.

*1) Results on NWPU VHR-10:* Table XIV tabulates the detection results of different RS object detectors on the NWPU VHR-10 dataset, all of which are based on deep learning and generally require pretraining due to the limited number of

TABLE XII

COMPARISON OF DOWNSTREAM RS DETECTION PERFORMANCE OF VARIOUS BACKBONE NETWORKS PRETRAINED BY FOUR DIFFERENT STRATEGIES. LARGER FLOPS MEANS HIGHER COMPUTATIONAL LOADS

| Backbones | | Pretraining Strategies | Params (M) | FLOPs (G) | mAP (%) | ΔmAP (%) |
|---|---|---|---|---|---|---|
| ResNet | ResNet18 | From Scratch | 28.19 | 64.50 | 29.7 | – |
| | | ImageNet Pretraining | | | 55.5 | +25.8 |
| | | RS Classification Pretraining | | | 57.4 | +27.7 |
| | | RS SCWL Pretraining | | | 59.3 | +29.6 |
| | ResNet34 | From Scratch | 38.23 | 78.16 | 29.7 | – |
| | | ImageNet Pretraining | | | 54.7 | +25.0 |
| | | RS Classification Pretraining | | | 57.0 | +27.3 |
| | | RS SCWL Pretraining | | | 60.1 | +30.4 |
| | ResNet50 | From Scratch | 41.20 | 83.56 | 22.6 | – |
| | | ImageNet Pretraining | | | 57.5 | +34.9 |
| | | RS Classification Pretraining | | | 58.9 | +36.3 |
| | | RS SCWL Pretraining | | | 61.1 | +38.5 |
| | ResNet101 | From Scratch | 60.19 | 111.03 | 21.8 | – |
| | | ImageNet Pretraining | | | 58.8 | +37.0 |
| | | RS Classification Pretraining | | | 61.0 | +39.2 |
| | | RS SCWL Pretraining | | | 63.0 | +41.2 |
| Swin Transformer | Swin-Tiny | From Scratch | 44.82 | 85.15 | 33.9 | – |
| | | ImageNet Pretraining | | | 58.5 | +24.6 |
| | | RS Classification Pretraining | | | 60.9 | +27.0 |
| | | RS SCWL Pretraining | | | 63.5 | +29.6 |
| | Swin-Small | From Scratch | 66.14 | 118.10 | 36.8 | – |
| | | ImageNet Pretraining | | | 61.3 | +24.5 |
| | | RS Classification Pretraining | | | 62.4 | +25.6 |
| | | RS SCWL Pretraining | | | 64.2 | +27.4 |
| | Swin-Base | From Scratch | 104.17 | 170.22 | 38.1 | – |
| | | ImageNet Pretraining | | | 62.9 | +24.8 |
| | | RS Classification Pretraining | | | 63.7 | +25.6 |
| | | RS SCWL Pretraining | | | 64.1 | +26.0 |
| | Swin-Large | From Scratch | 212.66 | 318.76 | 42.8 | – |
| | | ImageNet Pretraining | | | 64.3 | +21.5 |
| | | RS Classification Pretraining | | | 64.6 | +21.8 |
| | | RS SCWL Pretraining | | | 67.2 | +24.4 |

TABLE XIII

DOWNSTREAM DETECTION RESULTS OF VARIOUS RS DETECTORS PRETRAINED FOLLOWING FOUR DIFFERENT PRETRAINING STRATEGIES. LARGER FLOPS INDICATE HIGHER COMPUTATIONAL LOADS

| Detector Architectures | Types | Backbones | Pretraining Strategies | Params (M) | FLOPs (G) | mAP (%) | mAP (%) |
|---|---|---|---|---|---|---|---|
| Faster R-CNN | Two-stage & Anchor-based | +ResNet50 | From Scratch | 41.20 | 83.56 | 22.6 | – |
| | | | ImageNet Pretraining | | | 57.5 | +34.9 |
| | | | RS Classification Pretraining | | | 58.9 | +36.3 |
| | | | RS SCWL Pretraining | | | 61.1 | +38.5 |
| | | +Swin-Tiny | From Scratch | 44.82 | 85.15 | 33.9 | – |
| | | | ImageNet Pretraining | | | 58.5 | +24.6 |
| | | | RS Classification Pretraining | | | 60.9 | +27.0 |
| | | | RS SCWL Pretraining | | | 63.5 | +29.6 |
| RetinaNet | One-stage & Anchor-based | +ResNet50 | From Scratch | 36.39 | 76.07 | 25.8 | – |
| | | | ImageNet Pretraining | | | 57.1 | +31.3 |
| | | | RS Classification Pretraining | | | 57.6 | +31.8 |
| | | | RS SCWL Pretraining | | | 58.8 | +33.0 |
| | | +Swin-Tiny | From Scratch | 37.11 | 78.30 | 34.9 | – |
| | | | ImageNet Pretraining | | | 57.0 | +22.1 |
| | | | RS Classification Pretraining | | | 57.7 | +22.8 |
| | | | RS SCWL Pretraining | | | 61.7 | +26.8 |
| FCOS | One-stage & Anchor-free | +ResNet50 | From Scratch | 31.87 | 71.29 | 32.0 | – |
| | | | ImageNet Pretraining | | | 53.6 | +21.6 |
| | | | RS Classification Pretraining | | | 57.4 | +25.4 |
| | | | RS SCWL Pretraining | | | 59.4 | +27.4 |
| | | +Swin-Tiny | From Scratch | 36.77 | 75.65 | 36.1 | – |
| | | | ImageNet Pretraining | | | 61.1 | +25.0 |
| | | | RS Classification Pretraining | | | 61.9 | +25.8 |
| | | | RS SCWL Pretraining | | | 63.8 | +27.7 |

training samples. Without additional modules and tricks, Faster R-CNN equipped with our RS SCWL pretraining paradigm attains 90.6% mAP, which is higher than the results of most previous detectors. This indicates that, with more advanced pretraining strategies, such as our proposed RS SCWL pretraining, the basic detector Faster R-CNN is able to catch up with its improved versions and other sophisticated detectors for the task of RS object detection. Moreover, our detector only adopts the lightweight backbone ResNet50 instead of the larger ResNet101, as shown in Table XIV, which demonstrates the effectiveness of RS SCWL pretraining in boosting model efficiency. By introducing a more advanced Transformer-style backbone, Swin-Tiny, the detection accuracy can be further improved to 91.9% mAP. We visualize some qualitative detection results in Fig. 10.

*2) Results on DIOR:* For extensive evaluation and comparison, we report the detection results of different methods on the DIOR dataset in Table XV, all based on CNN. Many recent RS object detectors promote performance by introducing additional auxiliary modules or layers into existing detection frameworks, such as Faster R-CNN [10] and RetinaNet [12]. For example, GLNet proposed in 2022 [98] is a two-stage

TABLE XIV

DOWNSTREAM DETECTION PERFORMANCE COMPARISON WITH OTHER METHODS ON THE NWPU VHR-10 DATASET. AP IS EMPLOYED TO EVALUATE THE DETECTION PERFORMANCE IN EACH CATEGORY, WHILE MAP INDICATES PERFORMANCE ACROSS ALL CATEGORIES. "IMAGENET" AND "RS SCWL" REFER TO USING THE IMAGENET PRETRAINING OR OUR PROPOSED RS SWCL PRETRAINING STRATEGIES, RESPECTIVELY. THE TOP TWO BEST RESULTS IN EACH COLUMN ARE BOLDED

| **Method** | Backbone | Pretraining Strategy | Airplane | Ship | Storage Tank | Baseball Diamond | Tennis Court | Basketball Court | Ground Track Field | Harbor | Bridge | Vehicle | mAP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RICNN [14] | AlexNet | ImageNet | 88.4 | 77.3 | 85.3 | 88.1 | 40.8 | 58.5 | 86.7 | 68.6 | 61.5 | 71.1 | 72.6 |
| YOLOv2 [87] | DarkNet19 | ImageNet | 83.0 | 84.4 | 81.9 | 84.3 | 85.0 | 53.5 | 62.8 | 78.7 | 85.0 | 70.0 | 76.8 |
| SSD [40] | VGG16 | ImageNet | 90.6 | 83.7 | 77.4 | 97.4 | 87.6 | 69.3 | **100.0** | 88.2 | **98.2** | 38.4 | 83.1 |
| MSCA [88] | VGG16 | ImageNet | 99.5 | 80.0 | 90.4 | 90.5 | 90.6 | 77.3 | **100.0** | 76.1 | 65.9 | 80.6 | 85.1 |
| HRBM [37] | ZF CNN | ImageNet | 99.7 | 90.8 | 90.6 | 92.9 | 90.3 | 80.1 | 90.8 | 80.3 | 68.5 | 87.1 | 87.1 |
| SAPNet [89] | ResNet101 | ImageNet | 97.8 | 87.6 | 67.2 | 94.8 | **99.5** | **99.5** | 95.9 | **96.8** | 68.0 | 85.1 | 89.2 |
| FMSSD [90] | VGG16 | ImageNet | 99.7 | 89.9 | 90.3 | 98.2 | 86.0 | **96.8** | 99.6 | 75.6 | 80.1 | 88.2 | 89.2 |
| Cascade R-CNN [91] | ResNet101 | ImageNet | 99.4 | **95.7** | 66.4 | 96.7 | 93.1 | 94.3 | 99.7 | 87.8 | 84.6 | 85.3 | 90.3 |
| R-FCN [35] | ResNet101 | ImageNet | **99.9** | **95.8** | 66.9 | 97.7 | 93.6 | 90.7 | 98.7 | 89.8 | 75.9 | 84.0 | 90.4 |
| CAD-Net [92] | ResNet101 | ImageNet | 97.0 | 77.9 | **95.6** | 93.6 | 87.6 | 87.1 | 99.6 | **100.0** | **86.2** | **89.9** | 91.5 |
| MEDNet [93] | ResNet101 | ImageNet | 99.2 | 94.4 | 82.2 | 98.5 | **95.4** | 95.2 | 98.3 | 88.1 | 75.1 | 89.3 | **91.6** |
| Faster R-CNN (Ours) | ResNet50 | RS SCWL | 99.2 | 92.9 | 88.3 | **99.3** | 91.3 | 84.1 | 98.9 | 80.7 | 83.0 | 88.0 | 90.6 |
| Faster R-CNN (Ours) | Swin-Tiny | RS SCWL | **100.0** | 91.9 | **91.8** | **99.3** | 92.6 | 85.7 | 98.7 | 86.3 | 82.5 | **89.8** | **91.9** |

TABLE XV

DOWNSTREAM DETECTION PERFORMANCE COMPARISON WITH OTHER METHODS ON THE DIOR DATASET. AP IS EMPLOYED TO EVALUATE THE DETECTION PERFORMANCE IN EACH CATEGORY, WHILE MAP INDICATES PERFORMANCE ACROSS ALL CATEGORIES. "IMAGENET" AND "RS SCWL" REFER TO USING THE IMAGENET PRETRAINING AND OUR PROPOSED RS SWCL PRETRAINING STRATEGIES, RESPECTIVELY. THE TOP TWO BEST RESULTS IN EACH COLUMN ARE BOLDED

| **Method** | Backbone | Pretraining Strategy | c1 | c2 | c3 | c4 | c5 | c6 | c7 | c8 | c9 | c10 | c11 | c12 | c13 | c14 | c15 | c16 | c17 | c18 | c19 | c20 | mAP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SSD [40] | VGG16 | ImageNet | 59.5 | 72.7 | 72.4 | 75.7 | 29.7 | 65.8 | 56.6 | 63.5 | 53.1 | 65.3 | 68.6 | 49.4 | 48.1 | 59.2 | 61.0 | 46.6 | 76.3 | 55.1 | 27.4 | 65.7 | 58.6 |
| CornerNet [94] | Hourglass104 | ImageNet | 58.8 | **84.2** | 72.0 | 80.8 | 46.4 | 75.3 | 64.3 | 81.6 | **76.3** | 79.5 | 79.5 | 26.1 | 60.6 | 37.6 | 70.7 | 45.2 | 84.0 | 57.1 | 43.0 | 75.9 | 64.9 |
| Mask R-CNN [36] | ResNet101 | ImageNet | 53.9 | 76.6 | 63.2 | 80.9 | 40.2 | 72.5 | 60.4 | 76.3 | 62.5 | 76.0 | 75.9 | 46.5 | 57.4 | 71.8 | 68.3 | 53.7 | 81.0 | 62.3 | 43.0 | 81.0 | 65.2 |
| RetinaNet [12] | ResNet101 | ImageNet | 53.3 | 77.0 | 69.3 | 85.0 | 44.1 | 73.2 | 62.4 | 78.6 | 62.8 | 78.6 | 76.6 | 49.9 | 59.6 | 71.1 | 68.4 | 45.8 | 81.3 | 55.2 | 44.4 | **85.5** | 66.1 |
| PANet [95] | ResNet101 | ImageNet | 60.2 | 72.0 | 70.6 | 80.5 | 43.6 | 72.3 | 61.4 | 72.1 | 66.7 | 72.0 | 73.4 | 45.3 | 56.9 | 71.7 | 70.4 | 62.0 | 80.9 | 57.0 | **47.2** | 84.5 | 66.1 |
| CSFF [96] | ResNet101 | ImageNet | 57.2 | 79.6 | 70.1 | 87.4 | 46.1 | 76.6 | 62.7 | **82.6** | **73.2** | 78.2 | 81.6 | 50.7 | 59.5 | 73.3 | 63.4 | 58.5 | 85.9 | 61.9 | 42.9 | 86.9 | 68.0 |
| SCRDet++ [41] | ResNet50 | ImageNet | 64.3 | 79.0 | 73.2 | 85.7 | 45.8 | 76.0 | **68.4** | 79.3 | 68.9 | 77.7 | 77.9 | **56.7** | **62.2** | 70.4 | 67.7 | 60.4 | 80.9 | **63.7** | 44.4 | 84.6 | 69.4 |
| MSFC-Net [97] | ResNeSt101 | ImageNet | **85.8** | 76.2 | 74.4 | **90.1** | 44.2 | 78.1 | 55.5 | 60.9 | 59.5 | 76.9 | 73.7 | 49.6 | 57.2 | **89.6** | 69.2 | **76.5** | 86.7 | 51.8 | **55.2** | 84.3 | 70.1 |
| GLNet [98] | ResNet101 | ImageNet | 62.9 | **83.2** | 72.0 | 81.1 | **50.5** | **79.3** | **67.4** | **86.2** | 70.9 | **81.8** | **83.0** | 51.8 | **62.6** | 72.0 | 75.3 | 53.7 | 81.3 | **65.5** | 43.4 | **89.2** | 70.7 |
| ASSD [99] | VGG16 | ImageNet | 85.6 | 82.4 | **75.8** | 89.5 | 40.7 | 77.6 | 64.7 | 67.1 | 61.7 | **80.8** | 78.6 | **62.0** | 58.0 | **84.9** | **76.7** | 65.3 | **87.9** | 62.4 | 44.5 | 76.3 | **71.1** |
| Faster R-CNN (Ours) | ResNet50 | RS SCWL | **91.2** | 72.8 | **92.5** | **90.5** | **46.2** | **84.2** | 56.6 | 70.2 | 67.9 | 74.5 | **89.3** | 41.6 | 57.9 | 75.8 | **92.7** | **66.2** | **93.7** | 49.9 | 42.5 | 84.7 | **72.1** |

TABLE XVI

DOWNSTREAM DETECTION PERFORMANCE COMPARISON WITH OTHER METHODS ON THE DOTA DATASET. AP IS EMPLOYED TO EVALUATE THE DETECTION PERFORMANCE IN EACH CATEGORY, WHILE MAP INDICATES PERFORMANCE ACROSS ALL CATEGORIES. "IMAGENET" AND "RS SCWL" REFER TO USING THE IMAGENET PRETRAINING OR OUR PROPOSED RS SCWL PRETRAINING STRATEGIES, RESPECTIVELY. THE TOP TWO BEST RESULTS IN EACH COLUMN ARE REPRESENTED IN BOLD

| **Method** | Backbone | Pretraining Strategy | PL | BD | BR | GTF | SV | LV | SH | TC | BC | ST | SBF | RA | HA | SP | HC | mAP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| YOLOv2 [11] | GoogLeNet | ImageNet | 76.9 | 33.9 | 22.7 | 34.9 | 38.7 | 32.0 | 52.4 | 61.7 | 48.5 | 33.9 | 29.3 | 36.8 | 36.4 | 38.3 | 11.6 | 39.2 |
| MSCA [88] | VGG16 | ImageNet | 78.1 | 67.7 | 28.3 | 42.2 | 24.0 | 62.2 | 48.3 | 82.6 | 45.0 | 38.4 | 40.5 | 36.4 | 69.2 | 38.5 | 36.1 | 49.2 |
| AF-SSD [5] | ResNet50 | ImageNet | 88.0 | 61.6 | 31.2 | 43.6 | 59.4 | 32.8 | 54.8 | 90.7 | 62.4 | 74.0 | 29.8 | 54.0 | 33.0 | 49.8 | 23.9 | 52.6 |
| R-FCN [35] | ResNet101 | ImageNet | 81.0 | 59.0 | 31.6 | 59.0 | 49.8 | 45.0 | 49.3 | 69.0 | 52.1 | 67.4 | 41.8 | 51.4 | 45.2 | 53.3 | 33.9 | 52.6 |
| FR-H [29] | ResNet101 | ImageNet | 80.3 | **77.6** | 32.9 | **68.1** | 53.7 | 52.5 | 50.0 | 90.4 | **75.1** | 59.6 | **57.0** | 49.8 | 61.7 | 56.5 | 41.9 | 60.5 |
| HSF-Net [100] | ResNet101 | ImageNet | 80.0 | 69.9 | 37.7 | 58.0 | 66.8 | 64.2 | 71.8 | 87.9 | 69.4 | 61.8 | 47.5 | 52.8 | 66.8 | 59.2 | 41.8 | 62.4 |
| RFB-Net [101] | VGG16 | ImageNet | 87.8 | 49.8 | 35.7 | 36.4 | **78.0** | **77.5** | **90.1** | **93.8** | 63.8 | **89.5** | 33.6 | **67.9** | 63.2 | **72.8** | 20.5 | 64.0 |
| GLNet [98] | ResNet101 | ImageNet | **89.4** | 71.5 | 43.1 | **68.4** | 31.9 | 52.5 | 53.9 | 79.9 | 66.9 | 77.2 | **74.1** | **64.6** | **73.1** | 59.2 | **74.0** | 65.3 |
| ASBL [102] | ResNet50 | ImageNet | **89.5** | 74.1 | **46.9** | 55.5 | **73.8** | **66.9** | **78.5** | **90.9** | 70.1 | 73.2 | 46.7 | 61.3 | **70.5** | 72.2 | 32.8 | **66.9** |
| Faster R-CNN (Ours) | ResNet50 | RS SCWL | 78.9 | **79.8** | **53.6** | 48.0 | 56.9 | 58.1 | 59.8 | 80.1 | **75.2** | **84.7** | 35.0 | 61.1 | 64.0 | **73.8** | **76.0** | **65.7** |

RS detector, which develops a Clip-LSTM module to explore the spatial correlation information between local geospatial regions and plugs the module between the feature extractor and detection head. Taking ResNet101 as the backbone, GLNet reaches an accuracy of 70.7% mAP. However, it is structurally redundant and relies on more computation during inference, which hinders its real-world applications. In contrast, following the proposed RS SCWL pretraining paradigm, our detector obtains 72.1% mAP, only based on ResNet50, achieving both higher accuracy and efficiency for inference, more friendly to practical applications. Moreover, as presented in Table XV, our method achieves state-of-the-art or competitive results across multiple categories, demonstrating the broad effectiveness of the proposed pretraining strategy. Some detection results are visualized in Fig. 11.

*3) Results on DOTA:* Involving various complex RS scenes, the DOTA dataset is another widely used dataset for downstream RS object detection, on which we have carried out ablation experiments. Table XVI tabulates the detection results of some representative CNN-based algorithms on this test set

Fig. 12. Qualitative downstream detection examples of faster R-CNN with the proposed RS SCWL pretraining paradigm on the DOTA dataset.

but many adopt the heavy ResNet101 as the backbone to handle challenging scenarios. Due to its inefficiency, especially in the inference stage, several studies have been devoted to making lightweight backbone-based detectors, which perform better on this dataset. For instance, based on the naive VGG16, in [103], RFB-Net [101] was leveraged to achieve comparable detection performance, 64.0% mAP. However, sophisticated backbones still dominate due to their high accuracy. Our advanced RS SCWL pretraining paradigm can appropriately bridge this gap, depending only on basic detectors, such as Faster R-CNN, and simple backbones, such as ResNet50. Without bells and whistles, our improved RS detector achieves a promising performance of 65.7% mAP, surpassing most of the results in Table XVI, which validates the benefits of our pretraining approach for RS object detection. Some qualitative results are presented in Fig. 12.

## VI. Conclusion

This article proposes a novel RS detection-specific pretraining paradigm, named RS SCWL pretraining, which can be easily adopted in various existing detectors to boost detection performance without additional parameters and steps in the inference stage. The deficiencies of previous pretraining strategies for RS detection are attributed to three issues: different scenarios, mismatched task objectives, and misaligned architectures. To pave the way, RS SCWL pretraining breaks the traditional classification pretraining paradigm by pretraining backbone networks only on natural images, such as the widely used ImageNet pretraining. Instead, it aims to provide detection-invariant discriminative knowledge from RS scenarios in advance by pretraining the entire detector. To enable this pretraining paradigm, we first reconstruct a large-scale RS dataset based on multiple existing RS classification datasets. Then, inheriting the accurate image-level category annotations as strong-classification supervision, our method additionally generates instance-level pseudobounding boxes as weak-localization supervision. This advanced pretraining procedure can be performed and leveraged for diverse detectors without changing their structures. Comprehensive experiments were conducted on three downstream datasets, DOTA, NWPU VHR-10, and DIOR, and the results have demonstrated the superiority and efficiency of our RS SCWL pretraining, which consistently boosts the detection performance of RS images.

In the future, we will focus on further improving the superiority of pretraining from the following two aspects. On the one hand, CAM can be regarded as a simple and straightforward object localization method, whose performance, however, has been surpassed by the recent WSOL algorithms. Thus, more advanced strategies are worth exploring to improve the efficiency and precision of pseudobounding box generation, which may benefit subsequent detection-specific pretraining. On the other hand, although the proposed method introduces the property of SCWL, significantly mitigating the requirements for bounding box annotations, unfortunately, it still relies on massive category labels. Consequently, it is reasonable to explore unsupervised pretraining strategies, which completely releases the limitations of manual annotations. Moreover, in the context of RS, with unsupervised paradigms, the scale of the pretraining dataset can be expanded to enhance downstream detection performance.

## References


[1] H. Wang et al., "BDR-Net: Bhattacharyya distance-based distribution metric modeling for rotating object detection in remote sensing," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–12, 2023.

[2] W. Zhao, Y. Kang, H. Chen, Z. Zhao, Y. Zhai, and P. Yang, "A target detection algorithm for remote sensing images based on a combination of feature fusion and improved anchor," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–8, 2022.

[3] S. Sun, Y. Yin, X. Wang, and D. Xu, "Robust visual detection and tracking strategies for autonomous aerial refueling of UAVs," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 12, pp. 4640–4652, Dec. 2019.

[4] X. Zeng, S. Wei, J. Shi, and X. Zhang, "A lightweight adaptive RoI extraction network for precise aerial image instance segmentation," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–17, 2021.

[5] X. Lu, J. Ji, Z. Xing, and Q. Miao, "Attention and feature fusion SSD for remote sensing object detection," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–9, 2021.

[6] T. Ye, W. Qin, Y. Li, S. Wang, J. Zhang, and Z. Zhao, "Dense and small object detection in UAV-vision based on a global-local feature enhanced network," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–13, 2022.

[7] Y. Wu, K. Zhang, J. Wang, Y. Wang, Q. Wang, and X. Li, "GCWNet: A global context-weaving network for object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5619912.

[8] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[13] J. Han, J. Ding, J. Li, and G.-S. Xia, "Align deep features for oriented object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5602511.

[14] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.

[15] W. Zhang, L. Jiao, Y. Li, Z. Huang, and H. Wang, "Laplacian feature pyramid network for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5604114.

[16] Q. Wang, Y. Liu, Z. Xiong, and Y. Yuan, "Hybrid feature aligned network for salient object detection in optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5624915.

[17] C. Zhang, K.-M. Lam, and Q. Wang, "CoF-Net: A progressive coarse-to-fine framework for object detection in remote-sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5600617.

[18] C. Zhang, J. Su, Y. Ju, K.-M. Lam, and Q. Wang, "Efficient inductive vision transformer for oriented object detection in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5616320.

[19] C. Zhang, T. Liu, and K.-M. Lam, "Angle tokenization guided multi-scale vision transformer for oriented object detection in remote sensing imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2022, pp. 3063–3066.

[20] D. Mahajan et al., "Exploring the limits of weakly supervised pretraining," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 181–196.

[21] C. Vasconcelos, V. Birodkar, and V. Dumoulin, "Proper reuse of image classification features improves object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13618–13627.

[22] Y. Zhong, J. Wang, L. Wang, J. Peng, Y.-X. Wang, and L. Zhang, "DAP: Detection-aware pre-training with weak supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4535–4544.

[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[24] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[25] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2014, pp. 740–755.

[26] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.

[27] K. He, R. Girshick, and P. Dollar, "Rethinking ImageNet pre-training," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4917–4926.

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

[29] G.-S. Xia et al., "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3974–3983.

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

[31] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

[32] J. Han et al., "Efficient, simultaneous detection of multi-class geospatial targets based on visual saliency modeling and discriminative learning of sparse coding," *ISPRS J. Photogramm. Remote Sens.*, vol. 89, pp. 37–48, Mar. 2014.

[33] C. Zhu, H. Zhou, R. Wang, and J. Guo, "A novel hierarchical method of ship detection from spaceborne optical image based on shape and texture features," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 9, pp. 3446–3456, Sep. 2010.

[34] Q. Li, G. Wang, J. Liu, and S. Chen, "Robust scale-invariant feature matching for remote sensing image registration," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 2, pp. 287–291, Apr. 2009.

[35] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016.

[36] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[37] K. Li, G. Cheng, S. Bu, and X. You, "Rotation-insensitive and context-augmented object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2337–2348, Apr. 2018.

[38] R. Qin, Q. Liu, G. Gao, D. Huang, and Y. Wang, "MRDet: A multihead network for accurate rotated object detection in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5608412.

[39] Y. Liu, Q. Li, Y. Yuan, Q. Du, and Q. Wang, "ABNet: Adaptive balanced network for multiscale object detection in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5614914.

[40] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Oct. 2016, pp. 21–37.

[41] X. Yang, J. Yan, W. Liao, X. Yang, J. Tang, and T. He, "SCRDet++: Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2384–2399, Feb. 2023.

[42] X. Yang and J. Yan, "On the arbitrary-oriented object detection: Classification based approaches revisited," *Int. J. Comput. Vis.*, vol. 130, no. 5, pp. 1340–1365, May 2022.

[43] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2019, pp. 4171–4186.

[44] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12449–12460.

[45] Y. Shinya, E. Simo-Serra, and T. Suzuki, "Understanding the effects of pre-training for object detectors via eigenspectrum," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1931–1941.

[46] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 843–852.

[47] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Proc. NIPS Deep Learn. Represent. Learn.*, 2015.

[48] Z. Shen, Z. Liu, J. Li, Y.-G. Jiang, Y. Chen, and X. Xue, "DSOD: Learning deeply supervised object detectors from scratch," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1937–1945.

[49] R. Zhu et al., "ScratchDet: Training single-shot object detectors from scratch," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2263–2272.

[50] S. Kornblith, J. Shlens, and Q. V. Le, "Do better ImageNet models transfer better?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2656–2666.

[51] D. Zhou, X. Zhou, H. Zhang, S. Yi, and W. Ouyang, "Cheaper pre-training lunch: An efficient paradigm for object detection," in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2020, pp. 258–274.

[52] F. Wei, Y. Gao, Z. Wu, H. Hu, and S. Lin, "Aligning pretraining for detection via object-level contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 22682–22694.

[53] F. Wang, H. Wang, C. Wei, A. Yuille, and W. Shen, "$CP^2$: Copy-paste contrastive pretraining for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 499–515.

[54] O. Mañas, A. Lacoste, X. Giró-i-Nieto, D. Vazquez, and P. Rodríguez, "Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9394–9403.

[55] G.-S. Xia, W. Yang, J. Delon, Y. Gousseau, H. Sun, and H. Maître, "Structural high-resolution satellite image indexing," in *Proc. ISPRS TC VII Symp.-100 Years*, vol. 38, 2010, pp. 298–303.

[56] L. Zhao, P. Tang, and L. Huo, "Feature significance-based multibag-of-visual-words model for remote sensing image scene classification," *J. Appl. Remote Sens.*, vol. 10, no. 3, Jul. 2016, Art. no. 035004.

[57] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155–1167, Feb. 2019.

[58] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, Nov. 2010, pp. 270–279.

[59] Q. Zhu, Y. Zhong, B. Zhao, G.-S. Xia, and L. Zhang, "Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 6, pp. 747–751, Jun. 2016.

[60] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2321–2325, Nov. 2015.

[61] A.-J. Gallego, A. Pertusa, and P. Gil, "Automatic ship classification from optical aerial images with convolutional neural networks," *Remote Sens.*, vol. 10, no. 4, p. 511, Mar. 2018.

[62] G.-S. Xia et al., "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.

[63] H. Li et al., "CLRS: Continual learning benchmark for remote sensing image scene classification," *Sensors*, vol. 20, no. 4, p. 1226, Feb. 2020.

[64] H. Li et al., "RSI-CB: A large-scale remote sensing image classification benchmark using crowdsourced data," *Sensors*, vol. 20, no. 6, p. 1594, Mar. 2020.
[65] W. Zhou, S. Newsam, C. Li, and Z. Shao, "PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 197–209, Nov. 2018.
[66] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.
[67] X. Qi et al., "MLRSNet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding," *ISPRS J. Photogramm. Remote Sens.*, vol. 169, pp. 337–350, Nov. 2020.
[68] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.
[69] X. Zhang, Y. Wei, G. Kang, Y. Yang, and T. Huang, "Self-produced guidance for weakly-supervised object localization," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 597–613.
[70] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. Huang, "Adversarial complementary learning for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1325–1334.
[71] J. Choe and H. Shim, "Attention-based dropout layer for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2214–2223.
[72] Z. Zeng, B. Liu, J. Fu, H. Chao, and L. Zhang, "WSOD2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8291–8299.
[73] Y. Wei et al., "TS2C: Tight box mining with surrounding segmentation context for weakly supervised object detection," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 434–450.
[74] B. Dong, Z. Huang, Y. Guo, Q. Wang, Z. Niu, and W. Zuo, "Boosting weakly supervised object detection via learning bounding box adjusters," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2856–2865.
[75] Y. Shen, R. Ji, Z. Chen, Y. Wu, and F. Huang, "UWSOD: Toward fully-supervised-level capacity weakly supervised object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 7005–7019.
[76] Z. Huang, Y. Bao, B. Dong, E. Zhou, and W. Zuo, "W2N: Switching from weak supervision to noisy supervision for object detection," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2022, pp. 708–724.
[77] H. Li, X. Pan, K. Yan, F. Tang, and W.-S. Zheng, "SIOD: Single instance annotated per category per image for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14177–14186.
[78] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.
[79] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[80] J. Choe, S. J. Oh, S. Lee, S. Chun, Z. Akata, and H. Shim, "Evaluating weakly supervised object localization methods right," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3130–3139.
[81] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
[82] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 839–847.
[83] H. Wang et al., "Score-CAM: Score-weighted visual explanations for convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 111–119.
[84] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, and Y. Wei, "LayerCAM: Exploring hierarchical class activation maps for localization," *IEEE Trans. Image Process.*, vol. 30, pp. 5875–5888, 2021.
[85] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9626–9635.
[86] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.
[87] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525.
[88] J. Chen, L. Wan, J. Zhu, G. Xu, and M. Deng, "Multi-scale spatial and channel-wise attention for improving object detection in remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 4, pp. 681–685, Apr. 2020.
[89] S. Zhang, G. He, H.-B. Chen, N. Jing, and Q. Wang, "Scale adaptive proposal network for object detection in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 6, pp. 864–868, Jun. 2019.
[90] 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.
[91] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.
[92] 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.
[93] Q. Lin, J. Zhao, B. Du, G. Fu, and Z. Yuan, "MEDNet: Multiexpert detection network with unsupervised clustering of training samples," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4703114.
[94] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 734–750.
[95] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768.
[96] G. Cheng, Y. Si, H. Hong, X. Yao, and L. Guo, "Cross-scale feature fusion for object detection in optical remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 3, pp. 431–435, Mar. 2021.
[97] T. Zhang, Y. Zhuang, G. Wang, S. Dong, H. Chen, and L. Li, "Multiscale semantic fusion-guided fractal convolutional object detection network for optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5608720.
[98] Z. Teng, Y. Duan, Y. Liu, B. Zhang, and J. Fan, "Global to local: Clip-LSTM-based object detection from remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5603113.
[99] T. Xu, X. Sun, W. Diao, L. Zhao, K. Fu, and H. Wang, "ASSD: Feature aligned single-shot detection for multiscale objects in aerial imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607117.
[100] Q. Li, L. Mou, Q. Liu, Y. Wang, and X. X. Zhu, "HSF-Net: Multiscale deep feature embedding for ship detection in optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 7147–7161, Dec. 2018.
[101] S. Liu and D. Huang, "Receptive field block net for accurate and fast object detection," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 385–400.
[102] P. Sun, G. Chen, and Y. Shang, "Adaptive saliency biased loss for object detection in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7154–7165, Oct. 2020.
[103] G. Wang et al., "FSoD-Net: Full-scale object detection from optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5602918.

**Cong Zhang** (Graduate Student Member, IEEE) received the B.E. degree from the School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China, in 2018, and the M.E. degree from the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, in 2021. He is currently pursuing the Ph.D. degree with the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong.

His current research interests include remote sensing and computer vision.

**Tianshan Liu** received the B.Sc. and M.Sc. degrees from the School of Internet of Things Engineering, Jiangnan University, Wuxi, China, in 2016 and 2019, respectively. He is currently pursuing the Ph.D. degree with the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong.

His research interests include computer vision, machine learning, and video understanding.

**Jun Xiao** received the B.Sc. degree in telecommunication engineering from the Guangdong University of Technology, Guangzhou, Guangdong, China, in 2016, and the M.Sc. degree in electronic and information engineering and the Ph.D. degree from the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong, in 2018 and 2022, respectively.

His current research includes Bayesian machine learning, image and video restoration, and quality enhancement.

**Kin-Man Lam** (Senior Member, IEEE) received the M.Sc. degree in communication engineering from the Department of Electrical Engineering, Imperial College London, London, U.K., in 1987, and the Ph.D. degree from the Department of Electrical Engineering, The University of Sydney, Camperdown, NSW, Australia, in 1996.

He received his associateship in electronic engineering with distinction from The Hong Kong Polytechnic University (formerly called Hong Kong Polytechnic), Hong Kong, in 1986. From 1990 to 1993, he was a Lecturer with the Department of Electronic Engineering, The Hong Kong Polytechnic University. He joined the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, again as an Assistant Professor, in October 1996, where he became an Associate Professor in 1999 and has been a Professor since 2010. He is currently an Associate Dean of the Faculty of Engineering, The Hong Kong Polytechnic University. He was actively involved in professional activities. His current research interests include image and video processing, computer vision, and human face analysis and recognition.

Dr. Lam has been a member of the organizing committee or program committee of many international conferences. He is the IEEE Signal Processing Society (SPS) VP-Membership and a Member-at-Large of the Asia-Pacific Signal and Information Processing Association (APSIPA). He was the Chairperson of the IEEE Hong Kong Chapter of Signal Processing from 2006 to 2008 and the Director-Student Services and the Director-Membership Services of the IEEE SPS from 2012 to 2014 and 2015 to 2017, respectively. He was the VP-Member Relations and Development and VP-Publications of APSIPA from 2014 to 2017 and 2017 to 2021, respectively. He also serves as a Senior Editorial Board Member of *APSIPA Transactions on Signal and Information Processing* and an Associate Editor for *EURASIP Journal on Image and Video Processing*. He was an Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING from 2009 to 2014 and *Digital Signal Processing* from 2014 to 2018. He was an Editor of *HKIE Transactions* from 2013 to 2018 and an Area Editor of the *IEEE Signal Processing Magazine* from 2015 to 2017.

**Qi Wang** (Senior Member, IEEE) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively.

He is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. He is also with the Key Laboratory of Intelligent Interaction and Applications, Ministry of Industry and Information Technology, Beijing, China. His research interests include computer vision and pattern recognition.