remote sensing

MDPI

Review

Few-Shot Object Detection in Remote Sensing Image Interpretation: Opportunities and Challenges

Sixu Liu [1], Yanan You [1,*], Haozheng Su [1], Gang Meng [2], Wei Yang [3] and Fang Liu [1]

[1] School of Artifical Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2] Beijing Institute of Remote Sensing Information, Beijing 100192, China
[3] School of Electronics and Information Engineering, Beihang University, Beijing 100191, China
* Correspondence: youyanan@bupt.edu.cn

Abstract: Recent years have witnessed rapid development and remarkable achievements on deep learning object detection in remote sensing (RS) images. The growing improvement of the accuracy is inseparable from the increasingly complex deep convolutional neural network and the huge amount of sample data. However, the under-fitting neural network will damage the detection performance facing the difficulty of sample acquisition. Thus, it evolves into few-shot object detection (FSOD). In this article, we first briefly introduce the object detection task and its algorithms, to better understand the basic detection frameworks followed by FSOD. Then, FSOD design methods in RS images for three important aspects, such as sample, model, and learning strategy, are respectively discussed. In addition, some valuable research results of FSOD in computer vision field are also included. We advocate a wide research technique route, and some advice about feature enhancement and multi-modal fusion, semantics extraction and cross-domain mapping, fine-tune and meta-learning strategies, and so on, are provided. Based on our stated research route, a novel few-shot detector that focuses on contextual information is proposed. At the end of the paper, we summarize accuracy performance on experimental datasets to illustrate the achievements and shortcomings of the stated algorithms, and highlight the future opportunities and challenges of FSOD in RS image interpretation, in the hope of providing insights into future research.

Keywords: object detection; few-shot learning; few-shot object detection; remote sensing image interpretation

Citation: Liu, S.; You, Y.; Su, H.; Meng, G.; Yang, W.; Liu F. Few-Shot Object Detection in Remote Sensing Image Interpretation: Opportunities and Challenges. *Remote Sens.* **2022**, *14*, 4435. https://doi.org/10.3390/rs14184435

Academic Editor: Silvia Liberata Ullo

Received: 4 August 2022
Accepted: 2 September 2022
Published: 6 September 2022

1. Introduction

1.1. Motivation

Remote sensing (RS) images are an important data carrier for the human to realize space or earth observation. In this study, we focus on research of optical satellite remote sensing images, where the operating band of the sensor is limited to the visible band range (0.38–0.76 μm), with a high spatial resolution (usually less than 10 m). (Unless otherwise specified, the remote sensing images referred to in this paper are all high-resolution optical remote sensing images.) As a critical branch of remote sensing image interpretation, target (or object) detection task has manifested remarkable progress in recent years since the import of deep learning, that is, the most successful artificial intelligence (AI) paradigm today [1,2]. The brilliance of deep learning-based object detection methods is inseparable from sophisticated network structures and massive trainable parameters, requiring large-scale well-annotated datasets [1,3,4].

However, massive labeled datasets are not easily available, and inevitably, the deep learning-based object detector deteriorates by the sample limitation. Firstly, real scenes are more complicated than experimental datasets, and it takes a large amount of manpower, material, and financial resources to collect and calibrate samples to construct a dataset that can cover complete sample distribution. Secondly, a variety of targets present a long-tail

distribution, and some categories with rich samples often have much proportion, while it is difficult for other targets of interest to obtain enough samples. These are specifically divided into the following three scenarios:

(1) The actual number of targets is large but restricted by weather, illustration, viewing angle, resolution, and other factors, it is difficult to collect samples with high quality in diverse remote sensing images;
(2) The number of high-value targets is rare and it is difficult to obtain enough images;
(3) Low-value targets own a few historical images with the not-rich feature.

In the case of insufficient instances from candidate identifiable classes, the massive samples-driven object detector will face serious overfitting problems in supervised learning. So, few-shot learning (FSL) [5,6] attracts attention in that field. FSL aims at "learning to learn" and can make predictions based on a few samples, which is an imaginative and challenging learning paradigm. The pioneer FSL works concern the few-shot classification (FSC) task [7–9], which predicts image-level labels by using 1-, 2-, 3-, 5-, and 10-shot, which means training a classifier with a corresponding number of samples. Research on the more complex few-shot object detection (FSOD) task [10,11] that locates and predicts instance-level labels in an image flourishes later.

Although FSOD has developed rapidly, its application on RS images is still in its infancy. Recently, several pieces of literature have been published on FSOD in the CV field. Only the cost of sample annotation is considered in [12]. Thus, they consider that the combination of self-supervised object detection that can learn from unlabeled samples and FSOD can be a new research direction. Antonelli et al. [13] followed the FSL taxonomy proposed in [14] and categorized FSOD approaches into four families: (1) data augmentation; (2) transfer learning; (3) distance metric learning; (4) meta learning. Meanwhile, in remote sensing image interpretation, the authors of [15] overviewed the research progress in FSL and briefly introduced several FSOD works on optical RS images and Synthetic Aperture Radar (SAR) images. Different from the existing reviews, this paper comprehensively and carefully combs through the current published important literature both in FSOD and RS object detention fields. We aim to provide a structured review of recent advances in FSOD to help researchers obtain a holistic picture of this field and better understand the latest research findings. This paper sorts out a complete few-shot object detection technical route in the view of sample, model, and strategy, to discuss existing studies of FSOD on RS images and prospects for future development direction and application scenarios. It is our desire to help researchers realize the importance of FSOD task, and focus on key technical points and potential technical optimizations in FSOD in RS images interpretation.

1.2. Literature Retrieval and Analysis

To systematically analyze the tendency of few-shot object detection in remote sensing image interpretation for the past few years, the existing articles from Web of Science and Engineering Village are collected and analyzed in our work. Considering the interdisciplinarity between remote sensing (RS) image interpretation and computer vision (CV), more generally, artificial intelligence (AI), and an embryonic stage of FSOD in both fields, the fundamental work is elaborately demonstrated in this paper. Furthermore, articles on deep learning-based object detection and few-shot learning in RS are also included, to better understand the mechanism of FSOD. Finally, a total of 123 articles are selected, including 60 in RS and 63 in CV, and the details can be seen in Figure 1. It is worth mentioning that we only collected FSOD works in CV that are valuable for the RS field, and all published literature statistics until March 2022 can be found in Figures 2 and 3.

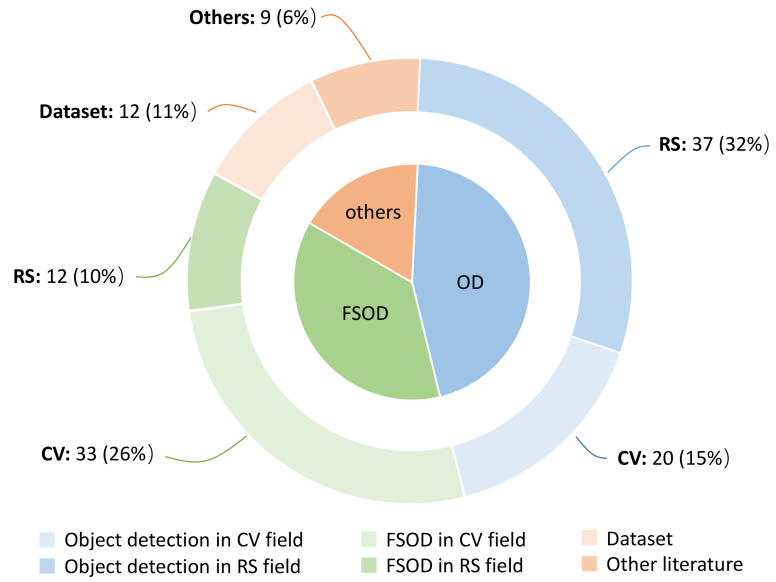

Figure 1. Literature cited in this paper. OD refers to object detection. Not all FSOD works in CV field are included because of their value in remote sensing image interpretation.

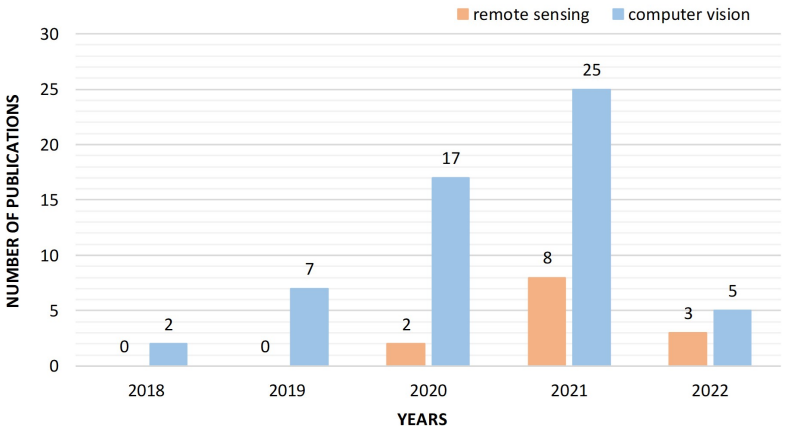

Figure 2. Published literature statistics of FSOD in both remote sensing image interpretation and computer vision fields. The statistics are counted according to the keywords "remote sensing" and "few-shot object detection" up to March 2022 in Web of Science and Engineering Village, acquiring a total of 69 articles.

As shown in Figure 2, the research on FSOD started in 2018, and was introduced into remote sensing in 2020, with the number of published articles increasing yearly. The statistics indicate that FSOD has become a new research hotspot in both algorithm and application practice. We also count the sources of literature in Figure 3, and find that the RS field prefers journals while CV field focuses on conferences, and CV field pays more attention to the improvement of algorithm accuracy and RS field places more emphasis on a complete application scenario. From the number of articles cited in this paper, the FSOD articles in RS field are fewer than those in the CV field. On the one hand, it takes a development process to convert an emerging method into a complete scheme. On the other hand, the application in professional field will also provide unique support for the improvement of FSOD method. Given the fact that works on CV are helpful to refine the studies of FSOD in RS, it is inevitable to present several important FSOD works of CV, to guide in addressing some specific problems in RS application scenarios.

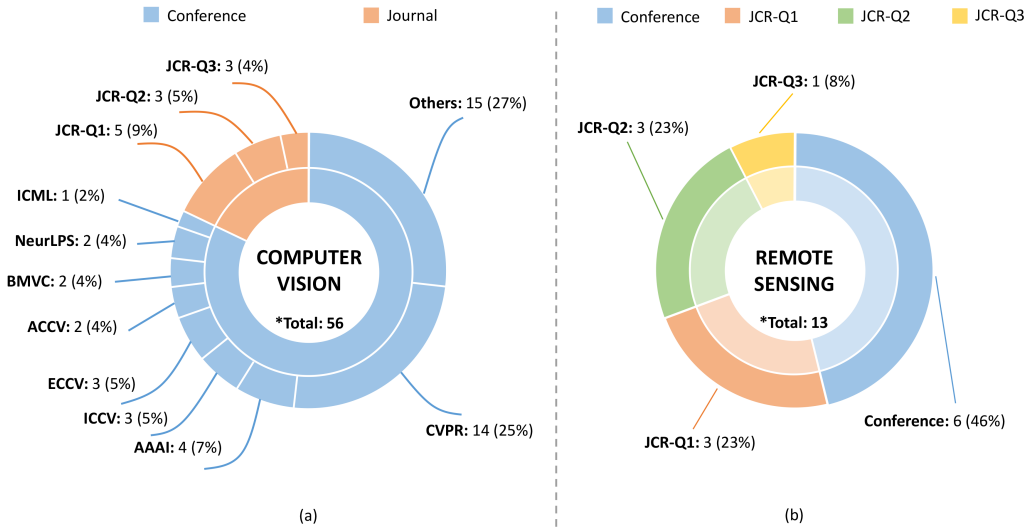

Figure 3. Publication sources of literature in the field of remote sensing image interpretation (**a**) and computer vision (**b**). The statistical result covers some journals and conferences. Journal Citation Reports (JCR) above is a partitioning method for journals retrieved by the Science Citation Index (SCI), which updates yearly and divides journals into four categories, from Q1 to Q4.

1.3. Problem Definition

AI aims to approach human intelligence, but the massive data-driven deep learning-based methods are different from common sense in that humans can learn novel concepts with a few samples. In the object detection task, few-shot object detection method attempts to simulate this human cognitive process.

Few-shot object detection (FSOD) is to achieve localization and classification abilities on novel classes using limited annotated instances. Given two sets of object categories, *base classes* $\mathcal{C}_{base}$ and *novel classes* $\mathcal{C}_{novel}$, where $\mathcal{C}_{base} \cap \mathcal{C}_{novel} = \varnothing$, FSOD methods aim to detect objects of $\mathcal{C}_{base} \cup \mathcal{C}_{novel}$ by learning from a *source domain* $D_s = \{(x_i, y_{i,loc}, y_{i,cls}) \in \mathcal{C}_{base}\}_{i=1}^{N_{base}}$ and a *target domain* $D_t = \{(x_i, y_{i,loc}, y_{i,cls}) \in \mathcal{C}_{novel}\}_{i=1}^{N_{novel}}$, where N_{base} refers to abundant annotated instances and N_{novel} refers to a few ones. x is an image with a category label y_{cls} and position label y_{loc}. Localization is to find the position of potential targets in an image, which is a regression problem, aiming to learn offsets between predicted boxes with their ground truth. Classification acts on target features, whose essence is to find a high-dimensional feature space (i.e., embedding space), to make target features (i.e., embedding vectors) cluster according to belonging categories and force clusters separable. Obviously, it is impossible to well train deep learning detectors on the target domain, subjecting to insufficient supervisory information. Instead, FSOD relies on prior knowledge learned from the source domain to adapt its detection ability in the target domain, and it has two types shown in Figure 4, according to how to extract the prior knowledge.

(1) **Transfer learning-based FSOD** learns from the source domain by supervised learning to obtain prior knowledge about targets, and the detector transfers from the source domain to the target domain. Thus, the training has two stages: pre-training on base classes and fine-tuning on novel classes in the target domain. These methods consider localization as category-irrelevant and classification as category-specific. The localization ability inherits from the training results of base classes in the source domain. So, when fine-tuning, nearly the entire trained parameters are frozen except the classifier.

(2) **Meta learning-based FSOD** regards a set of learning patterns found in the source domain by meta learning as the meta (prior) knowledge, which is going to be embedded into the detection of novel classes. It trains and tests with a series of episodes, consisting of support sets and query sets, to simulate few-shot situations, as shown in

Figure 5. The aim of training with these changeable episodes is to force the model to find out parameters that are adaptable after a few episodes from novel classes.

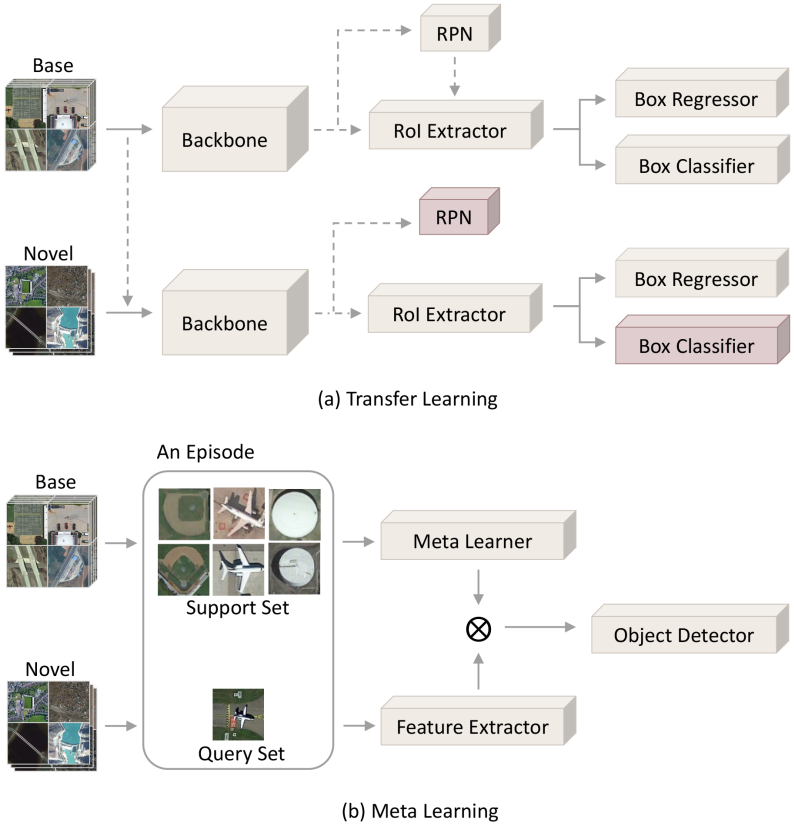

Figure 4. The illustration of two branches of few-shot object detection, transfer learning (**a**) and meta learning (**b**). Transfer learning follows standard supervised learning. The upper pipeline is the base training phase, while the lower pipeline is the few-shot fine-tuning phase. Meta learning episode setting is determined by novel classes. ⊗ means semantic fusion operation, such as depth-wise cross-correlation.

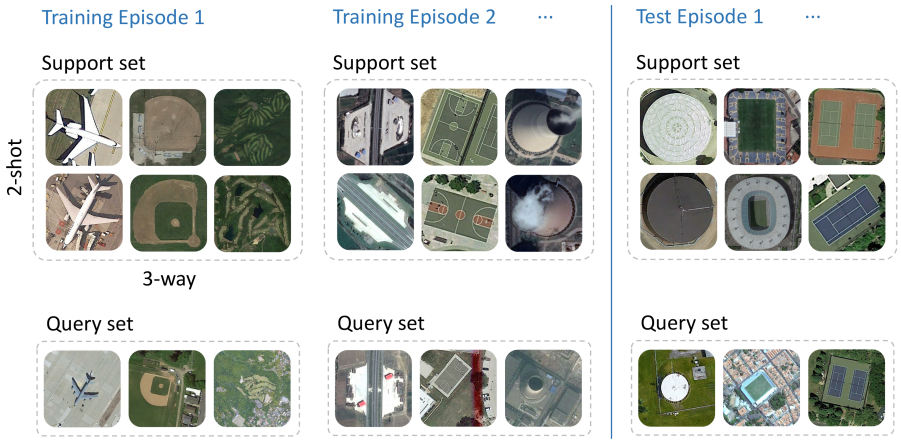

Figure 5. The illustration of 3-way 2-shot episodes in meta learning.

Currently, FSOD still follows the deep learning object detection frameworks in terms of implementation level, whose classic detection model, especially in the RS field, will be briefly summarized in Section 2.

Our view is that, similar to the current supervised learning paradigm, to analyze or design the FSOD framework or solution, there should be three considerations: limited and

representative ***samples***, optimal object detection ***models***, and few-shot learning ***strategies***. Samples are the core resources of RS image interpretation and their characteristics as the input conditions must be the first concern during the model design. Even under some extreme conditions, such as few-shot samples, the "intrinsic value" should be explored as much as possible to support the generalization ability of the model. Models are well-designed network architectures to achieve FSOD tasks, in which different components have different roles, aiming to implement cross-domain mapping of shared semantics. Strategies refer to learning strategies, which means how to conduct a network to implement FSOD, such as meta learning [10,11], transfer learning [16,17], etc. In this paper, we will follow the above clue to review and discuss the research status of few-shot object detection in remote sensing images and look forward to the development prospects.

The rest of this paper is organized as follows. In Section 2, classical deep learning-based object detection frameworks and remote sensing object detection methods are briefly introduced. The FSOD framework design concerning sample, model, and strategy are respectively demonstrated in Section 3. Section 4 lists the few-shot experimental performance on classical remote sensing image datasets and other natural image datasets. Section 5 presents future challenges to FSOD architecture improvements and has initial insight into the potential application scenarios in remote sensing image interpretation. The conclusion will be found in the final part.

2. Object Detection Review

In this section, we need to briefly introduce object detection tasks based on deep learning before FSOD. This is mainly for two considerations: (1) FSOD task is a continuation of the object detection task, and discussing the object detection methods will help researchers understand the detection mechanisms inherent in FSOD; (2) in the field of remote sensing image object detection, there are some studies on specific targets, which involve model specificity, i.e., how to design a detection model that has a preference for the specific targets, and that research direction may also have an impact on the research of FSOD.

2.1. Classical Deep Learning Object Detection

During the past decade, deep learning brings significant performance improvement to object detection and develops two mainstream technology routes: two-stage and one-stage, as shown in Figure 6. Two-stage object detectors, such as the R-CNN series [18–20], firstly generate region proposals that are candidate areas containing potential objects by selective search [18,19] or Region Proposal Network (RPN) [20], then project region proposals back to feature maps through a region of interest (RoI) pooling. Finally, the RoIs are processed to regress accurate bounding boxes and classify them into corresponding classes. One-stage object detectors such as YOLO series [21–24] and SSD [25], discarding the region proposal generation stage, directly regress offset of bounding boxes and predict class probabilities on feature maps. Relatively speaking, two-stage ones have advantages in accuracy while one-stage ones specialize in inference speed.

In addition to the categorization according to an explicit ROI feature extraction process, object detection methods via deep learning are split into two types based on the way to locate targets: anchor-based detectors and anchor-free detectors. Anchor-based approaches start from RPN [20], which regards each point in feature maps as an anchor point and generates a set of pre-set anchor boxes with different aspect ratios and scales on it. The localization is achieved by regressing from anchor boxes close to the ground truth. However, plenty of anchors lead to massive negative proposals and hyperparameters. Indeed, CornerNet [26] is the first attempt to achieve anchor-free, which predicts top-left and bottom-right points and groups them in an embedding space to localize an object. Thereafter, some other keypoint-based detectors [27–30], with different expressions of bounding boxes, are raised.

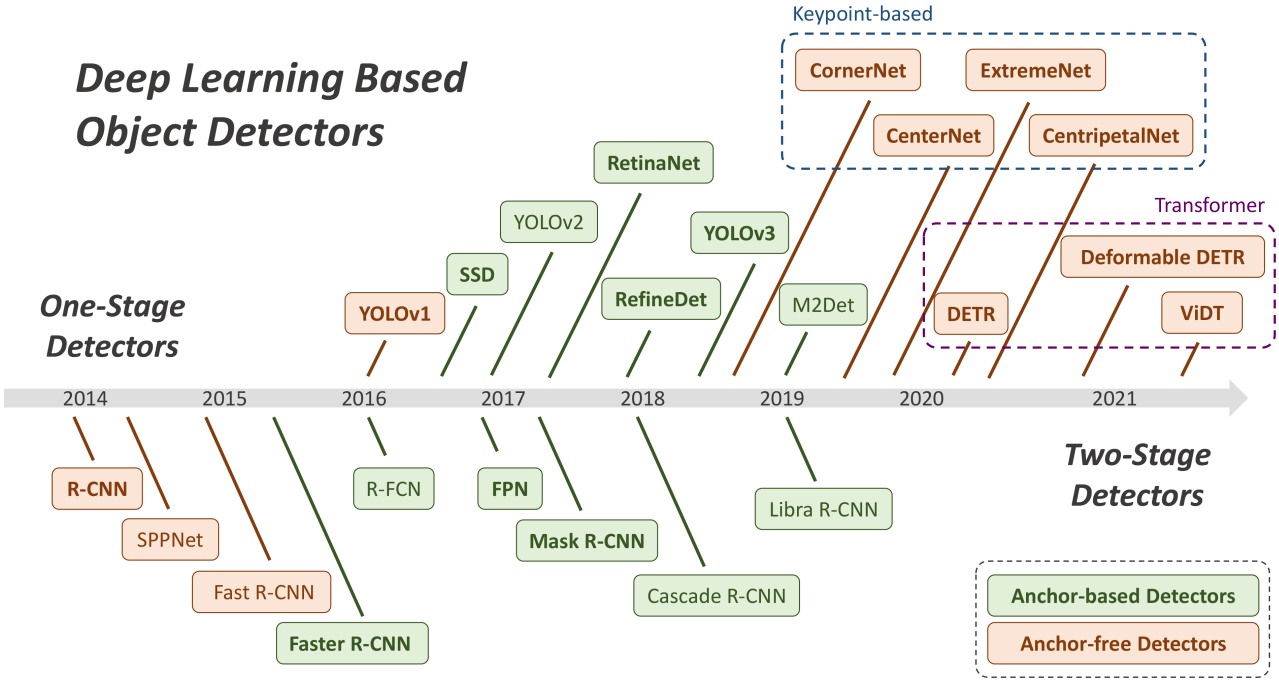

Figure 6. A road map of deep learning object detection. The models above the timeline are one-stage detectors, and the models below the timeline are two-stage detectors. The models in green are anchor-free ones while the models in orange are anchor-based ones. The models in the bond are milestones.

All the above methods are based on the images to carry out spatial encoding and decoding for target features. However, Transformer [31] is proposed as a powerful tool in machine translation tasks, and it has become the mainstream algorithm in natural language processing (NLP), which focuses on long-term information extraction by the designed self-attention mechanism. Because of the easier approach to get a bigger receptive field than stacked CNNs, vision transformer emerges constantly since 2020. The first attempt on object detection is DETR [32], which inputs flattened features after backbone into an encoder together with position information and predicts classes and locations by a decoder. Then, aiming at the long training cycle and insufficient ability to detect small targets of DETR, deformable DETR [33] proposes deformable attention to fix the sparsely distributed notable regions and considers the multi-scale issue. DETR series discard post-processing steps of non-maximum suppression (NMS) and prior knowledge and constraints such as anchors, simplifying the detection pipeline and realizing a truly end-to-end object detection network. The introduction of Transformer brings a new paradigm for object detection and becomes a research hotspot.

2.2. *Remote Sensing Object Detection*

RS images are captured in space or sky, and only the top view of observed targets can be seen, leading to different data characteristics from natural images. Taking some optical images as an example in Figure 7, more complicated background information, huger scale variance, more target directions, and densely small targets result in performance degradation if directly applying CV object detectors. Thus, according to these data issues, remote sensing object detection proposes adaptive improvements.

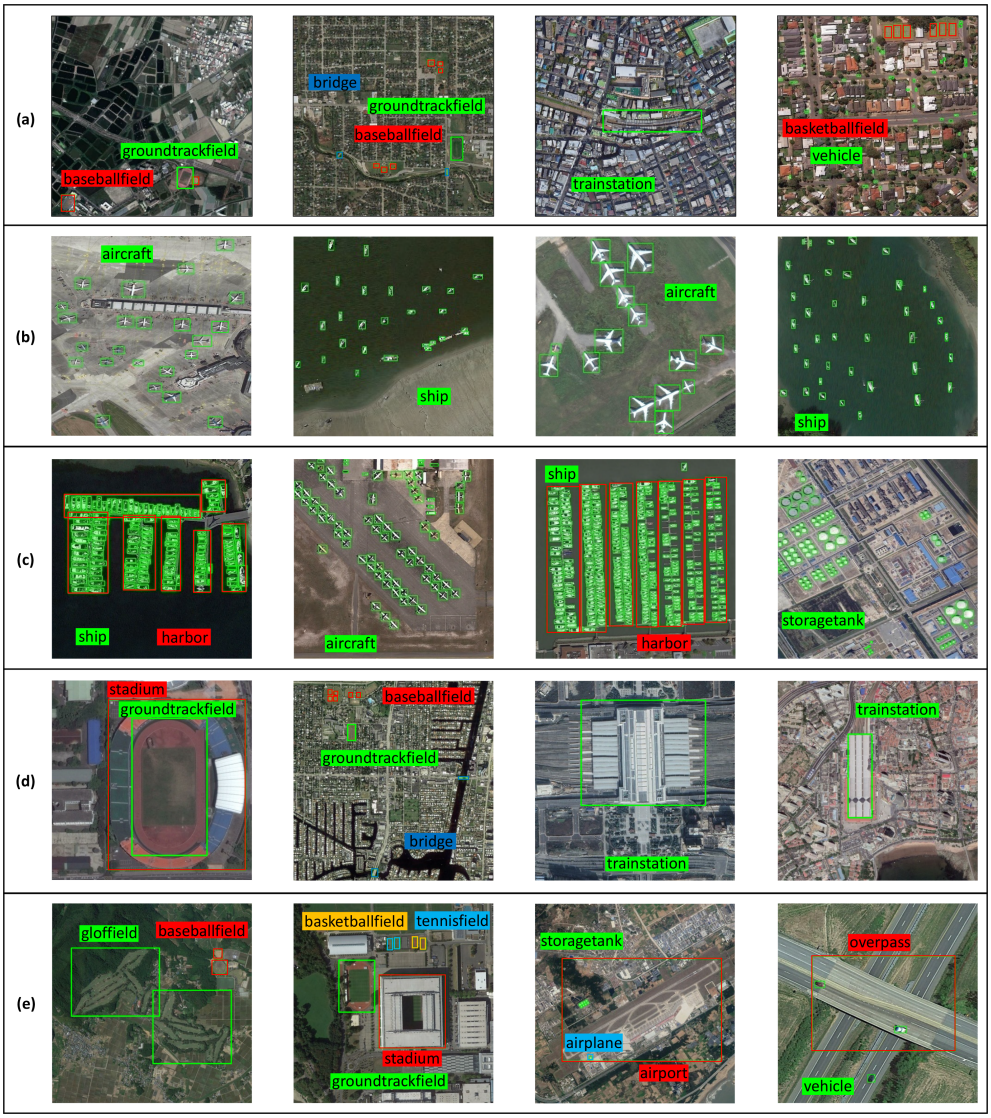

Figure 7. Data characteristics of remote sensing images collected from DIOR [1]: (**a**) complicated background information, (**b**) multiple target direction, (**c**) dense small targets, (**d**) inter-class scale diversity, (**e**) intra-class scale diversity.

(1) **Complicated background**: RS images own rich ground information, while the categories of interest are limited, so the background area occupies the majority, leading to the possible foreground-background misjudgment. Therefore, the solutions tend to suppress background information and highlight the target region features, where the attention mechanism [34–38] is effective. Moreover, our previous work [39] tried to divide the RS images into the target and non-target regions. The complex background information of non-target is extracted and suppressed by introducing a semantic segmentation task, which can effectively guide the correct selection of candidate regions. Similar ideas also appear in vehicle detection [40–43].

(2) **Multiple target directions**: In RS images, specific targets, such as ships and aircraft, are usually distributed in diversified angles on a single image. However, only certain rotation invariance of CNN makes extracted features sensitive to target directions. The intuitive solution is to rotate and expand the samples [44–48], but with limited effect. Thus, various rotation-insensitive modules are proposed for feature extraction, such as regularization before and after rotation [49], multi-angle anchors in RPN [50], rotation-invariance in frequency domain [51], target position coordination redefinition

in complex domain [52]. Direction prediction modules [44,47,53,54] in ship detection can also deal with the angle issue.

(3) **Densely small targets**: Targets in RS images are too densely distributed in some specific situations, such as vehicles in a parking lot, airplanes in an airport, ships in a harbor, etc. Dense distribution and small-scale targets are two considerations. The improvements on the former mainly focus on feature enhancement to improve the discrimination of a single target [44,53,55], while the latter depends on increasing the size of the feature maps [56] or introducing shallow layer feature maps to provide more information [44,47,57–59].

(4) **Huge scale variance**: As shown in Figure 7, scale variance occurs not only inter-class but also intra-class. Simultaneously keeping satisfactory detection accuracy on large- and small-scale targets is a hard issue. The current solution is to introduce a multi-scale mechanism, such as simply detecting on different scales [34,60], importing a feature pyramid network (FPN)[44,61], and designing or refining feature fusion networks [36,37,62–67]. However, the existing methods are still not effective for extreme scales, such as targets that only have a few pixels or almost occupy the whole image.

All algorithms mentioned in this section are designed for massive data on common object detection datasets, such as DOTA [3,68] with oriented annotations, or DIOR [1], NWPU VHR-10 [69] with horizontal annotations. However, in real applications, collecting adequate RS images and carefully annotating is time-costing and sometimes impossible. Then, in the face of insufficient labeled samples resource, massive data-driven methods will face serious overfitting problems. Thus, few-shot learning-based object detection emerges, that is, few-shot object detection.

3. FSOD Design for Sample, Model and Strategy

3.1. Sample: Feature Enhancement and Multimodal Fusion

Data augmentation is the naivest way to enlarge the sample size, under the premise of not introducing additional annotated data resources, which is one of the commonly used methods in deep learning and has also been studied in FSOD. It should be clear that data augmentation cannot completely solve the problem of sample scarcity, but it can relieve the pressure on that sample situation to some extent.

Traditional data augmentation methods, such as zooming, rotating, flipping, brightness adjustment, contrast enhancement, etc., have been utilized in remote sensing object detection [46–49,70]. These transformations of shallow features are easy to implement, but transformed samples own high similarity and are still original ones in essence. A method based on deep learning, which can imitate the characteristics of samples, i.e., generate adversarial network (GAN) [71,72], can be used for data enhancement. It generates fake samples between massive-annotated datasets and few-shot datasets in the embedding space, and distinguishes fake candidates from the true data distribution, finally achieving a Nash equilibrium that the generated samples can mix the spurious with the genuine. Chen et al. [71] innovatively introduced a GAN into its object detector under semi-supervision, where only part of the training samples is labeled. GAN here learns data distribution from unlabeled samples to better balance the classification boundary. The authors of [71] only paid attention to the difficulties of sample annotation, ignoring the learning method, while FSOD focuses on the situation that only a few labeled samples are available to build an effective detector. Thus, GAN in [72] focuses on domain shifting, transferring detection abilities from massive sample-driven models to few-shot models. GAN generates fake samples between massive datasets and few-shot datasets in the embedding space to gradually achieve sample distribution shifting. These GAN-based methods proved to be effective with insufficient annotated samples, but the training procedure is often erratic and time-consuming.

The above data augmentation methods based on GAN generate samples at the image level, while instance-level multi-scale positive sample augmentation [73–75] is also

widely adopted. Positive samples refer to foreground objects, which are the concerned parts in images. Noticing the scale sparsity problem of few-shot samples, MPSR [73] resizes the cropped input instances to refine prediction at various scales, while FSSP [74] further considers the background sparsity problem, not only scaling instances, but also putting the resized instance back to a random position in the mask image full of zero pixels. The introduction of multi-scale information effectively enhances instance-level samples and offers overall accuracy improvements. It is worth noting that remarkable performance gain is witnessed in the case of extremely few samples, i.e., less than or equal to 3 shots. The above two instance-level augmentation methods only include the situation that the possible location and scale of a single instance, and FSOD-SR [75] believes that the interdependence between objects can help cognition. It resizes the positive sample and joints the resized object back to its original position, increasing scale information and retaining location information. This consideration of interdependency leads to a 2∼8% improvement in detection accuracy relative to the previous two. Considering the characteristics of huge scale variance, RS images should pay more attention to the instance scale sparsity problem caused by limited samples. Although there is no attempt at positive sample enhancement on FSOD in RS images up to now, the visible achievements of the stated algorithms are enough attractive and worth to be considered.

Increasing positive instances is feasible, and so do utilize negative proposals [76]. Negative proposals are generated when regressing bounding boxes, which are incorrect proposals including partial foreground objects or total background. The principle of classification in detection pipeline is to achieve separability between embedding vectors of samples from different categories in a high-dimensional embedding space, where the embedding vectors are usually generated by positive proposals [77–79]. In addition, NP-Repmet [76] figures out that the introduction of negative proposals could enhance the separability of all classes in the embedding space. Proposals belonging to a certain category should be both close to the positive representatives and away from the negative representatives of that class. Thresholds are set to IoU between predicted boxes and ground truth to get positive and negative proposals, which are used to generate representatives. NP-Repmet significantly improves detection accuracy compared with Repmet that only embeds positive representatives, with the highest improvement up to 24%. In addition, the illustration of the above feature enhancement methods is shown in Figure 8.

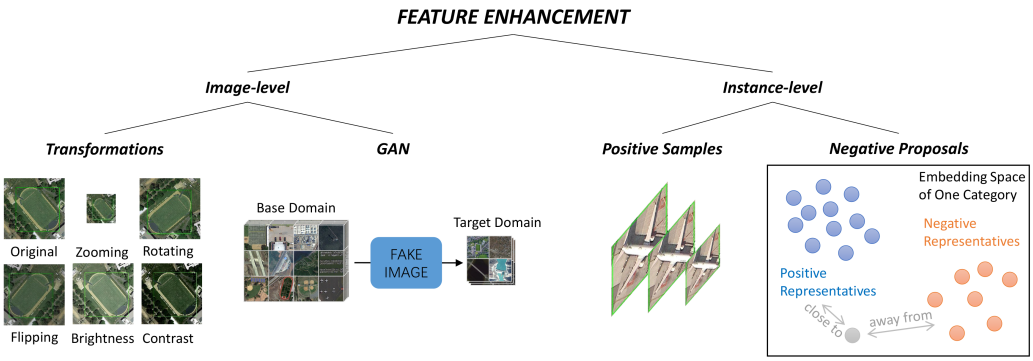

Figure 8. Principles of feature enhancement methods, including image-level and instance-level data augmentations. Details can be found in Section 3.1.

Objects are not independent of their surroundings, especially in images, just as humans build contextual associations of objects and backgrounds during visual recognition, and thus extracting the contextual information about the targets in images is also one of the ways for sample augmentation. So, Yang et al. [80] designed a plug-and-play module, called context-transformer, to adaptively construct a contextual field around the possible objects, as in Figure 9, which allows detectors to avoid the few-shot confusion problem. Logically, in the complicated and diverse remote sensing image, the target of interest has an obvious context relationship with its living environment; therefore, FSOD should

pay more attention to the context information of the target in the RS image, which is a beneficial revelation.

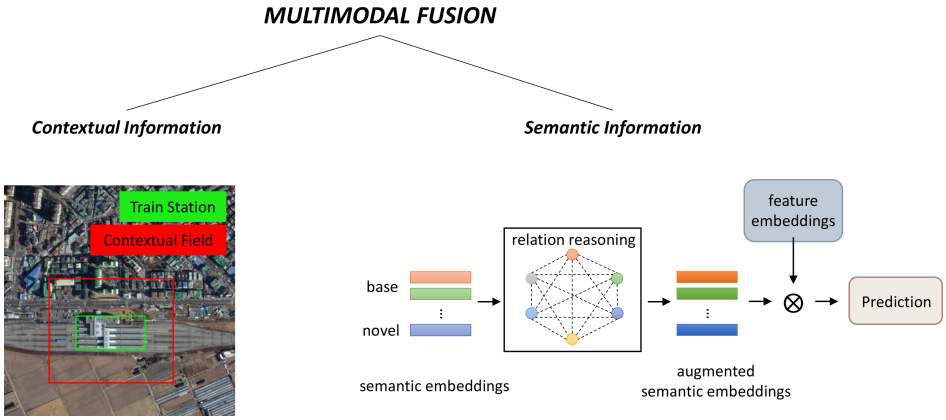

Figure 9. Principles of multi-modal fusion methods, including the introduction of contextual information and textual semantic information. Details can be found in Section 3.1.

In addition to visual information, SRR [81] firstly introduces textual semantic information. It considers that the semantic relation between the base classes and novel classes is constant regardless of data availability, especially, a word embedding is learned for each class from a large text corpus. The detection model projects and aligns visual representations with the corresponding semantics embedding by learning a dynamic relation graph in Figure 9. Experiments show the significant accuracy improvement in 1-shot, 2-shot, and 3-shot situations, which inspires us that when facing a serious sample scarce problem, the introduction of text semantics can assist few-shot learning. However, for RS images, typical textual semantic datasets such as Word2Vec [82] are too shallow, and it is difficult to provide enough semantic information to assist RS image target detection. However, the import of cross-modal fusion is worth learning from. Furthermore, the cross-modal FSOD can contain textual semantics for certain RS images, or different types of RS images, such as optical RS images, infrared images, or synthetic aperture radar (SAR) images. However, in this research direction, it is necessary to consider the semantic confusion caused by the inconsistency of multi-modal RS image target features. In short, it is worthy of in-depth study.

3.2. Model: Semantics Extraction and Cross-Domain Mapping

FSOD model framework needs to be established following the cognition process of "source vision–semantics–semantics–target vision", the model provides the function from visual task to semantic extraction in the source domain or target domain, and two domains are connected by some high-level semantics, or to say, meta knowledge. Then, how to extract and share meta knowledge is the key consideration in FSOD model design.

To our best knowledge, instead of designing a distinctly new detection framework from scratch, FSOD appends the process of meta knowledge extraction and sharing on classic deep learning object detection baselines, such as two-stage Faster R-CNN [11,16,17,72,73,78,83–98], one-stage YOLO [10,74,99], CenterNet [100,101], and Vision Transformer [102]. The numerous published reports indicate that the two-stage network is more favored, and the two-stage network has more advantages because of its higher detection accuracy, more interpretive, and extensible network structure.

No matter what kind of object detection baseline is used, the meta knowledge-related modules act the same for the few-shot situation. Therefore, we now summarize and put forward a research technical route of FSOD in Figure 10. In the source domain, base object information is obtained from massive-annotated datasets, and FSOD distills target semantics from this extracted information, shares semantics across domains, and finally executes prediction on target objects. We tease out the FSOD pipeline of target information

extraction, target semantics sharing, cross-domain mapping mechanism, and few-shot prediction head, which will be introduced in detail later.

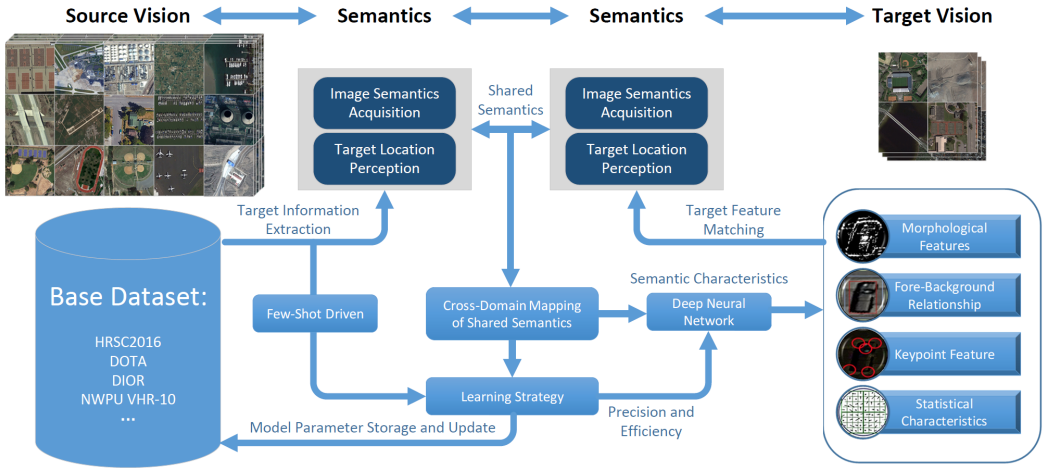

Figure 10. The technical route of few-shot object detection in remote sensing image. FSOD model design follows the cognition process of "source vision–semantics–semantics–target vision". The shared semantics extracted from samples are transmitted from the source domain to the target domain. The extraction and cross-domain mapping of shared semantics is the key to achieving FSOD. The base dataset in the source domain offers target information for semantics generation. Semantics shared to the target domain achieve feature fusion by target feature matching. Fused features are used for further detection, and the few-shot strategy supports the whole training procedure.

3.2.1. Target Information Extraction

In FSOD pipeline, target information extraction is achieved by the feature extraction backbone, which is a frontend module in the whole network to generate feature maps and the foundation for further feature processing. Backbone here refers to the stack of convolution layer, pooling layer combining with some performance enhancement methods, such as dropout, activation functions, batch normalization, residual block, etc. It can be self-designed but in terms of detection accuracy, usually well-designed mature backbones, such as ResNet [103] and DarkNet [22,23], are ideal for FSOD.

Limited by massive data-driven approaches, information including target category and background will be centralized and one-way transmitted to further feature processing modules, such as the region proposal network in a traditional two-stage network, leading to insufficient mining to object characteristics. Once the backbone network completes the forward mapping of features, the connection with low-level semantics will be cut off. Due to the scarce training data, support samples will suffer from unforeseeable background noise interference, so that the detection network cannot realize the foreground location of the query samples and will be in a disorder or over-fitting state, resulting in serious detection errors.

Therefore, it is necessary to build a robust information extraction network for reliable training and prediction, which should include network functional units as shown in Figure 11, such as high- and low-level semantic feature extraction of targets, background interference suppression, and target attention mechanism, and obtain multi-scale features in the embedded space for object classification.

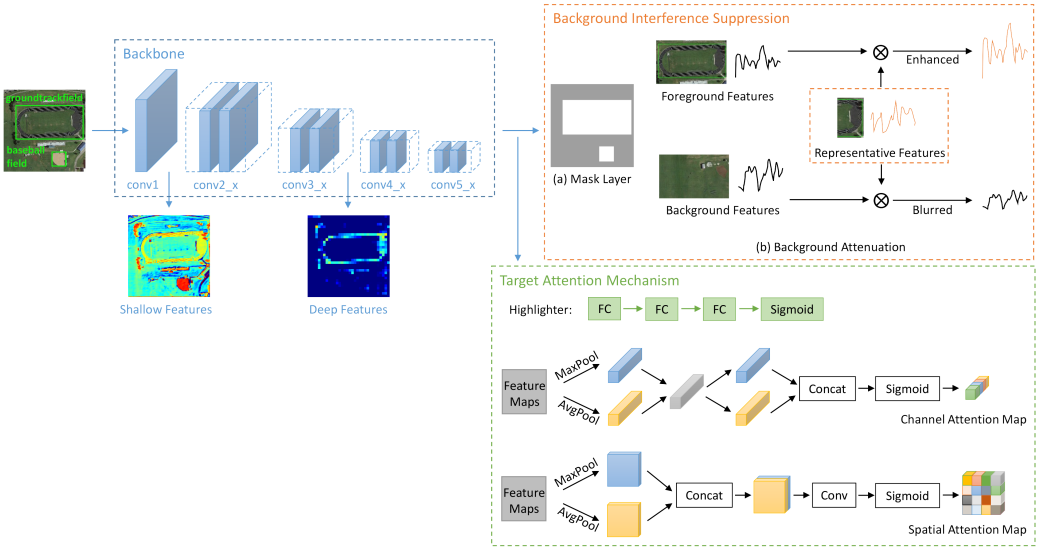

Figure 11. The illustration of high- and low-level semantic feature extraction, background interference suppression, and target attention mechanism. Details can be found in Section 3.2.1.

(1) High- and Low-Level Semantic Feature Extraction:

In a feature extraction backbone, the receptive field of the neural network expands with the stacking of convolutional layers. Shallow feature maps are suitable for detecting small-scale objects while deep features are for large-scale, according to the size of reception field. Huge scale variance of targets in RS images leads to the consideration of multi-scale problems. To extract features simultaneously from both high-level and low-level, FSODM [10] connects Darknet-53 [23] with a feature pyramid network (FPN) [104] to form query feature maps of three different scales, where FPN can enlarge the reception field of low-level features and provide multi-scale target perception. It also introduces another lightweight CNN from FR [99] to generate multi-scale support reweighting vectors to activate and match queries. To avoid matching and detecting multiple times at different scales, deconvolution and element-wise product are used to merge different-level semantics, keeping multi-scale information and saving calculation [105].

Multi-scale feature extraction is always a research interest in remote sensing object detection, while some attempts have not been introduced to FSOD but are valuable. Wang et al. [34] designed a skip-connected encoder-decoder model to extract multi-scale features, while [62,106] made improvements on FPN. The authors of [106] introduced a selectively refined module in FPN to selectively integrate feature maps from various scales. Fu et al. [62] added a bottom-up pathway on the top of top-down pathway as in FPN to further enhance multi-scale information.

(2) Background Interference Suppression:

The rich features in RS images will cause serious interference to the judgment of foreground or background, especially in the few-shot situation. The most easily implementable method is a binary mask channel, where 0 indicates background and 1 refers to foreground. Borrowed from FR [99], an additional mask is concatenated to support images to bring localization information and highlight boundaries between foreground and background [10,95] in RS images.

Turning to FSOD in natural images, suppression can also be achieved by extracting background features and setting a regularization for minimization [85]. With the help of the principle of interference phenomenon in physics, i.e., in-phase enhancement and inverse-phase attenuation, DAnA [107] superimposed the representative feature vectors extracted by the attention mechanism back to the original picture, thus achieving the effect of highlighting the foreground and suppressing the background.

Background depression regularization is suitable to transfer a well-trained model on base classes to novel ones with a few shots. The interference principle simultaneously depresses background and enhances discriminative foreground, which is connected to the support branch before the generation of semantics. These attempts are proved valuable and effective in natural scenes and afford experiences to RS images that merit attention.

(3) Target Attention Mechanism:

Background suppression and target attention work on different parts in images but for the same purpose, which is to help the FSOD models accurately separate the foreground and background of objects even under a few-shot setting. The attentional mechanism is essentially a feature re-weighting process, increasing the weight of the features in the concerned categories while attenuating that of the other categories and background information. Attention modules are inserted after the backbone and before the detection backend, which are trained and fine-tuned with the entire network together. Xiao et al. [83] designed a light-weighted feature attention highlight module that contains a cascaded coarse highlighter and a fine highlighter, making the general features to be specific in a serial way. In addition, in [17], a more complicated shared attention module for channel and spatial information is offered. Channel-wise attention focuses on the importance of different channels in the feature cube, while spatial-wise attention prefers the importance between patches. The above two attention modules both focus on optical RS images in open-source datasets [1,69,108]. The work [83] comes up earlier as one of the first methods to introduce FSOD into RS images. Its baseline is the classical Faster R-CNN framework combined with the transfer learning training method. The simple but effective attention module gets large or small performance gain on nearly all categories under any shot. Recently, a new attempt has came up [17]. Relative to the classical FSOD attempt in RS, i.e., FSODM [10], the authors of [17] offered a significant and average 15% improvement in detection accuracy. Ablation experiments also confirmed the performance of its shared attention module.

As for FSOD in natural images, the attention mechanism has other thoughts. Noticing that for generic object detection, one single sample is not such remarkable because the features of massive samples usually tend to be a constant [74], but FSOD does not. Limited samples lead to inaccurate feature representation of the category to which they belong. Thus, FSSP [74] introduces a self-attention module (SAM) into the backbone, which forces objects to homogenize with others of the same category. The self-attention learns the weight distribution of image features and applies it back to original images to enhance key features and ignore others. Moreover, to improve the network representation ability, HOSENet inserts a hyper-order semantic enhancement module into the backbone [87]. SAM and HOSE are both designed to be plug-and-play modules, thus are easy to deploy and worth attempting for FSOD in RS images.

3.2.2. Target Semantics Sharing

FSOD frameworks output enhanced feature maps with potential objects after target information extraction, according to the FSOD technical route. Then, the feature maps will be transmitted to further region proposal generation (option) and detection head in generic object detection. There exists an extraction and mapping of shared semantics from the source domain into the target domain before further detection. In brain-like intelligence [109], image semantic meaning is divided into three layers: the bottom features layer, the object layer, and the concept layer. The bottom features are like objects' outlines, edges, colors, textures, shapes, etc., which are shallow and have no image high-level semantic information. The object layer mainly considers the objects and objects' space in images, that is, the state of an object at a given moment. The conceptual layer is the closest thing an image can represent to human understanding, usually involves abstract attributes of images, and is the shared semantics we desire. The shared semantics here refers to some

abstract general category-independent knowledge that guides the recognition process, as in Figure 12, which is an imitation of the human cognitive process.

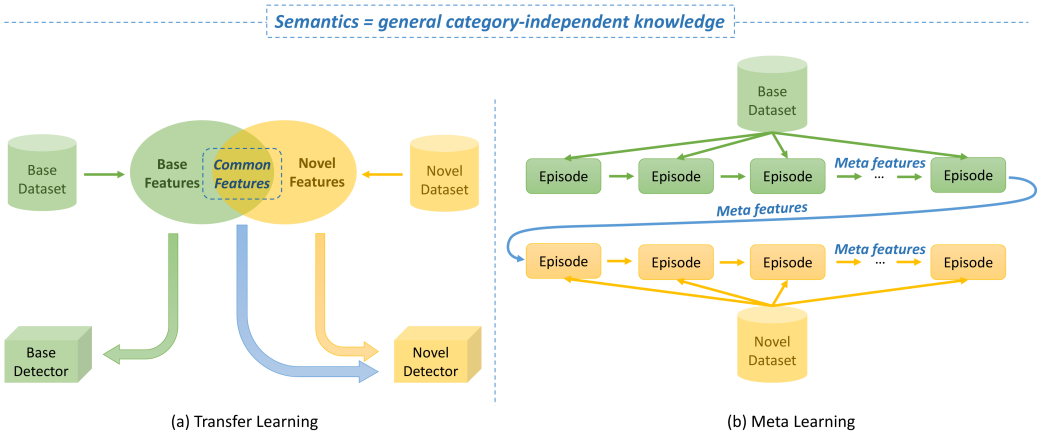

Figure 12. The semantics refer to in transfer few-shot detectors and meta few-shot detectors.

Transfer learning-based FSOD follows traditional supervised learning, firstly training on base classes to acquire the abilities to extract meaningful target information, accurately regress pre-set anchor boxes to ground truth, and construct a representative space for classification, then fine-tuning on novel classes. **The semantics shared between base and novel classes are the abstract object knowledge about feature extraction, bounding box regression, and existing classification representative space.** Extraction and regression are always regarded as category-independent. At the implementation level, the weights of corresponding modules are frozen when fine-tuning. Limited by a few shots from novel classes, it is impossible to completely reconstruct a representative space, which will lead to a serious overfitting problem. Thus, the novel classification space is built on the basis of base representation space [16,17,78,98].

Meta learning-based FSOD trains and fine-tunes with episodes. The design of episodes is to imitate a few-shot situation, to force the model to learn a new learning pattern, that is, achieving detection on query sets with the shared semantics from support sets. **The semantics here is the abstract knowledge about the discriminative characteristics of support objects, which will be used to measure similarities with query features to achieve the justification of category.** The generation of support semantics can be simply achieved by global max pooling [10] or global average pooling [84] on support feature maps from the target information extraction module. When down-sampling, average pooling focuses on the overall information of a dimension and the location of objects while max-pooling concerns the distinguishable feature. So, the combination of both pooling forms more representative semantics [17]. Shared fully-connected layers can also be used to extract support features [11,83,97].

3.2.3. Cross-Domain Mapping of Shared Semantics

According to the FSOD technical route in Figure 10, semantics generated from the source domain should be mapped to the target domain in order to help object detection with a few shots. For transfer learning-based FSOD, the responses of the base detector are going to be merged with target features via parameter fine-tuning [16,98] or feature fusion [17]. Parameter fine-tuning only aims at prediction head [16,98] and RPN [16], while the feature extraction backbone is frozen. Beyond this, Huang et al. [17] designed two attention extractors in channel-wise and spatial-wise connecting base and novel detectors, merging features generated from both. In addition, for meta learning-based FSOD, support semantics through depth-wise cross-correlation [83,105], channel-wise multiplication [10,84], element-wise multiplication [17], or Relation-GRU [11] to finish feature fusion with query region proposals, where the principle is shown in Figure 13. Alternatively, between parameter

iterations, cross-domain mapping is finished by gradient descent on the detection result of novel classes from the base detector [16]. In addition, distance measurement with source prototype clusters also represents one mapping method [78].

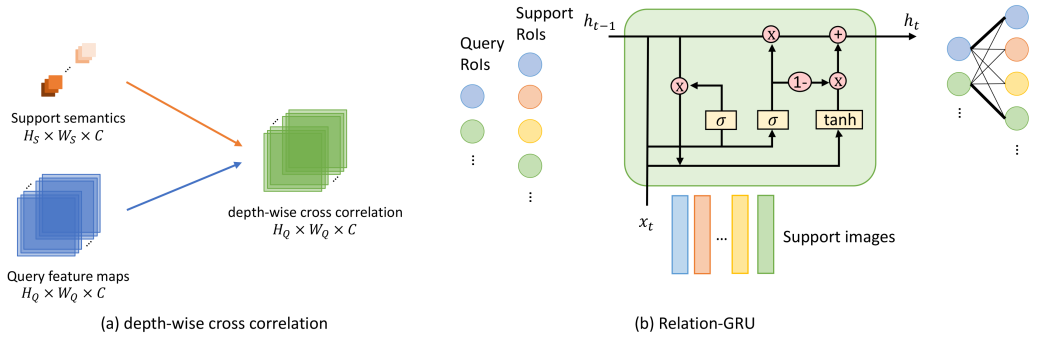

Figure 13. The principle of deep-wise cross correlation and relation-GRU. (**a**) For every channel, a convolution procedure is implemented where the convolutional kernel is the corresponding channel of the support semantics. Deep-wise cross-correlation is equivalent to channel-wise multiplication when $\mathcal{H}_S = \mathcal{W}_S = 1$, and is also equivalent to element-wise multiplication when $\mathcal{H}_S = \mathcal{H}_Q, \mathcal{W}_S = \mathcal{W}_Q$. (**b**) The green block refers to relation-GRU. h is the RoIs from detection framework, the left column belongs to query images and the right column refers to support images.

3.3. Few-Shot Prediction Head

From an implementation perspective, the current FSOD methods in RS images usually directly inherit the prediction backend in a generic object detection framework without structural modifications [10,11,16,17,83,84,97,98,105], which is composed of fully-connected layers and trained both on the source domain and the target domain, as shown in Figure 14a. The other attempt for the prediction head is inspired by FSC solutions, retaining localization layers and replacing classification layers with metric learning ideas [78]. Metric learning aims to train a network to output a representation for each instance in the embedding space, where these representations can cluster according to the belonging categories and different clusters are separable from each other. The authors of [78] respectively constructed embedding spaces for the two classification layers in the Faster R-CNN framework, which are foreground-background classification in RPN and the final category classification. The classification result is determined by the distances between the embedding vector and the centers of clusters in the embedding space.

About the construction of the embedding space for classification, FSOD in natural scenes has some inspirable considerations in Figure 14. CME [79] considers that there exists a contradiction between the representation and classification of novel classes. To better classify novel classes, the distribution of base classes in embedding space should be far away from each other, that is, max-margin, while to better represent novel classes, the distribution of base classes should be close to each other, referring to min-margin. Thus, CME presents an adversarial learning strategy to find a balance between the requirement of max-min margin. In addition, FSCE [88] discovers that freezing RPN, FPN, and RoI extraction modules when fine-tuning will lead to misjudgment from the novel class to the base class, and introduces contrastive learning in the embedding space to reduce the similarity between similar objects in different categories.

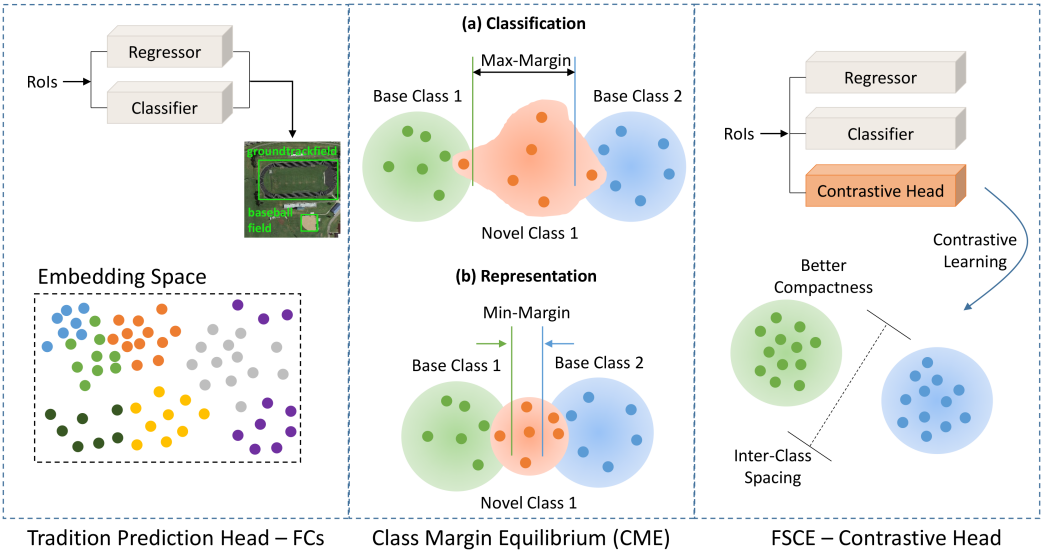

Figure 14. The principle of three kinds of few-shot prediction heads. FCs indicate fully-connected layers.

For RS images, because of the rich information, special attention should be paid to the target information extraction module. High- and low-level semantic feature extraction can simultaneously consider multiple target sizes, background suppression can reduce the interference of irrelevant information to detection, and the target attention mechanism is to help the network find the part that should be paid attention to, i.e., the concern foreground targets. After the feature extraction is finished, semantic information should be generated and cross-domain shared. Then, the fused feature maps are transmitted to the few-shot prediction head, thus achieving few-shot object detection.

3.4. Strategy: Fine-Tuning and Learning to Learn

Learning strategy is critical for deep learning-based algorithms. A well-designed and appropriate learning strategy can quickly converge to the optimal. As mentioned in Section 1.3, the dataset of FSOD task is split into base classes and novel classes, thus requiring two-stage training separately. Both stages are trained with the same few-shot strategy but in different categories. The current training strategy for the FSOD algorithm in RS images has two types in Figure 15: transfer learning [16,17,98] and meta learning [10,11,78,83,84,105].

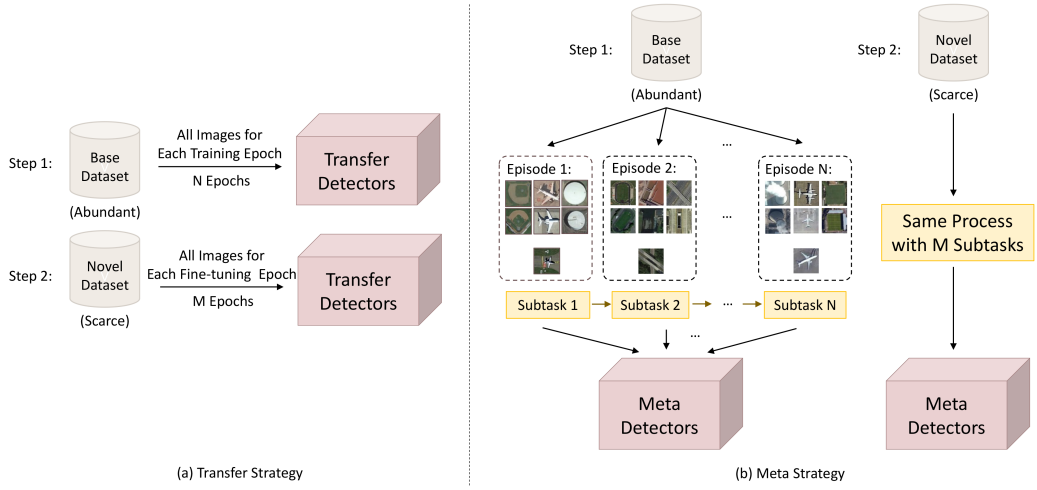

Figure 15. The difference between (**a**) transfer strategy and (**b**) meta detectors.

Transfer learning aims to find a function $\mathcal{F}$ to map from sample distribution X to label distribution Y, it achieves FSOD task by fine-tuning parameters with few-shot novel classes

on the massive data-driven base detector, and concerns sample space and optimizes only a single task, while that of meta learning is to learn a function $\mathcal{F}$ to map from a task set D to the set of optimal functions $f(x)$ for each task; it forms plenty of few-shot sub-tasks from the base dataset by imitating some few-shot datasets, expecting the network to learn how to learn from limited information, and executes on both sample space and task space and trains on multiple tasks. Relatively speaking, the training costs and difficulties of transfer learning is lower, but the meta learning strategy performs better on an unknown task because of the training mechanism that learns how to quickly adapt between tasks.

As mentioned before, transfer learning is first trained on the source domain as in deep learning object detection. Then, the parameters of the well-trained feature extraction backbone are frozen and the remaining parts of regression and classification will be fine-tuned on the target domain. The creativities usually happen in the fine-tuning phase. The two-stage fine-tuning approach proposed by TFA [86] is an early and classical training strategy for transfer learning-based FSOD algorithms, where at the fine-tuning stage, the trainable parameters of almost the entire network are fixed except the last layer of classifier and regressor in the prediction head. PAMS-Det [98] follows the training strategy in TFA to train the network, while DH-FSDet [16] considers that the foreground-background binary classification is also category-relevant and unfreezes RPN to be trainable at the fine-tuning stage. The fine-tuned RPN is proved to be effective at detection accuracy compared with the simple TFA strategy.

Furthermore, Huang et al. [17] noticed that the imbalance problem between massive base samples and few novel samples leads to generally poor confidence scores of novel classes. Because the maximum number of proposals generated by RPN is fixed, proposals of novel classes with relatively low confidence scores are mostly filtered out, resulting in insufficient foreground objects of novel classes. To solve the problem, the number of pre-set proposal generation in RPN is doubled. In addition, the sample imbalance also affects the calculation of the regression loss function. Under the control of a smooth L1 loss function, all the classes share the same weight. Due to quantity variance, the loss of base classes will be significantly higher than that of novel classes, leading to accurate localization but incorrect classification. Thus, a balanced L1 loss is proposed to increase the influence of novel classes on loss calculation. The above RPN adjustment and loss function improvement together form a balanced fine-tuning strategy, which solves the class imbalance problem under few-shot settings and improves detection accuracy.

Reviewing transfer learning-based FSOD in natural images, some valuable and novel attempts are worth discussing here. LSTD [85] is the first work that introduces transfer learning into the FSOD task, which proposes regularization, background depression, and transfer knowledge, and only retrains the last layer of the final classifier at the fine-tuning stage. Then, TFA [86] fine-tunes classifier and regressor in the prediction head. FSCE [88] only freezes the backbone and RoI pooling module and fine-tunes base parameters on the rest network with the balanced training datasets, K-shot from each base, and novel class. Retentive R-CNN [89] keeps base RPN and RoI prediction head, to train additional ones for novel classes, and regularizes fine-tuning to avoid base class detection performance deterioration. According to the above methods, the more the fine-tuned category-relevant modules, the better the detection accuracy performance. In addition, the operation of not forgetting detection capabilities on base classes is more important for FSOD in RS images.

Meta learning strategy aims at learning to learn, constructing a set of sub-tasks, i.e., episodes, for training and fine-tuning. Training episodes are formed entirely from base classes and fine-tuning episodes are constructed from the combination of base and novel or completely from novel classes. In each episode, query images are randomly selected, which determine the categories, and support instances are extracted according to these categories under few-shot settings. The episode-based training method is also known as meta strategy. The early work FR [99] on one-stage detection baseline and Meta R-CNN [91] on two-stage detection baseline for natural images are the benchmark

for current FSOD in RS images. FSODM [10] replaced the YOLOv2 detection baseline with the more advanced YOLOv3 in FR, introducing multi-scale meta training modules for the characteristics of huge scale variances in RS images. Experiments confirm the advancement of learning (convergence) speed and detection accuracy under the few-shot situation. Then, the works [11,83,84,97,105] follow the few-shot setting in Meta R-CNN. On this foundation, Gao et al. [105] imported the FPN network and multi-scale processing, while Chen et al. [83] designed a feature attention module. The authors of [84] improved the semantics extraction to generate class-aware prototypes, while in [11] a graph neural network to guide the feature fusion between supports and queries is applied. In addition, Zhang et al. [97] proposed an aggregator and an encoder to refine semantics generation and cross-domain mapping procedure. These methods respectively adjust on the baseline Meta R-CNN and achieve performance gain in comparison to this baseline.

Turning to meta learning-based FSOD in natural images, there are some classical and effective detection algorithms that have not been introduced into RS images. For example, AttentionRPN [92] designs a multi-relation head on the top of a few-shot pipeline in Meta R-CNN, which measures the similarities between support instances and query proposals at three different levels, i.e., global relation, local relation, and patch relation. Global relation learns an embedding for global matching on the entire feature maps, while local relation and patch relation respectively focus on pixel-wise and patch-wise. Its ablation experiments confirm the success of the design of the three relations. The great breakthrough in computer vision for object detection is Transformer [32], which utilizes the attention mechanism to obtain better feature representation. Meta-DETR [102] is the first work to combine Transformer and meta strategy under few-shot settings, and becomes the SOTA algorithms when our survey is prepared. Meta-DETR works at the image-level task, thus avoiding the weakness of few-shot proposal generation and with the help of an encoder-decoder structure, it can concern multiple support classes to better capture the inter-class correlation and reduce the misclassification. The creative network structure and the SOTA accuracy are bound to have a positive impact on FSOD in RS images.

4. Performance Evaluation

The deployment of FSOD algorithms relies on the computing power resources of the graphics processing unit (GPU). The most commonly used GPUs are the NVIDIA RTX series and Tesla series. In terms of software, the FSOD algorithms are typically constructed using the open-source programming framework, i.e., Tensorflow, Pytorch, or others. With the continuous development of object detection, some deep learning object detection toolboxes are released successively, such as Detectron [110] and mmDetection [111], and so on, which contain the implementations of the classical algorithms and are of great help to further research.

4.1. Experimental Datasets

The selection of datasets is critical and fundamental for the training of deep learning models. Dataset is a collection of samples with different types of annotations designed based on demands. For the object detection task, the annotations contain information about the bounding box locations and belonging categories of targets within images. The difference between few-shot object detection and generic object detection is essential whether the sample size is sufficient instead of the annotation type. Thus, the current FSOD attempts on RS images follow and reorganize the following object detection datasets. The detailed information about these datasets can be found in Figure 16 and the specific category partitions are shown in Table 1.

Table 1. Dataset selection and partitioning for FSOD algorithms on remote sensing images. DOTA and HRSC2016 are OBB annotated datasets but here they are converted to HBB. HRSC2016 is a fine-grained dataset with different types of ships.

Dataset	# cls	Base Classes	Novel Classes	Papers
NWPU VHR-10	10	Remaining classes	3 classes: airplane, baseball diamond, and tennis court	[10,17,97,98]
		Remaining classes	Random 1 class except for storage tank and harbor	[11]
DIOR	20	Remaining classes	5 classes: airplane, baseball field, tennis court, train station, windmill	[10,17,97,98]
		Remaining classes	Split 1: baseball field, basketball court, bridge, chimney, ship Split 2: airplane, airport, expressway toll station, harbor, ground track field Split 3: dam, golf course, storage tank, tennis court, vehicle Split 4: expressway service area, overpass, stadium, train station, windmill	[84]
RSOD	4	Remaining classes	Random 1 class	[11,83]
iSAID	15	Remaining classes	Split 1: Helicopter, Ship, Plane, Large Vehicle Split 2: Baseball Diamond, Soccer Ball Field, Roundabout Split 3: Ground Track Field, Helicopter, Baseball Diamond Roundabout, Soccer Ball Field, Basketball Court	[16]
DOTAv1.5	16	Remaining classes	Split 1: plane, ship, tennis court Split 2: harbor, helicopter, soccer ball field	[78]
DAN	15	Remaining classes	3 classes: vehicle, storage tank, plane	[105]

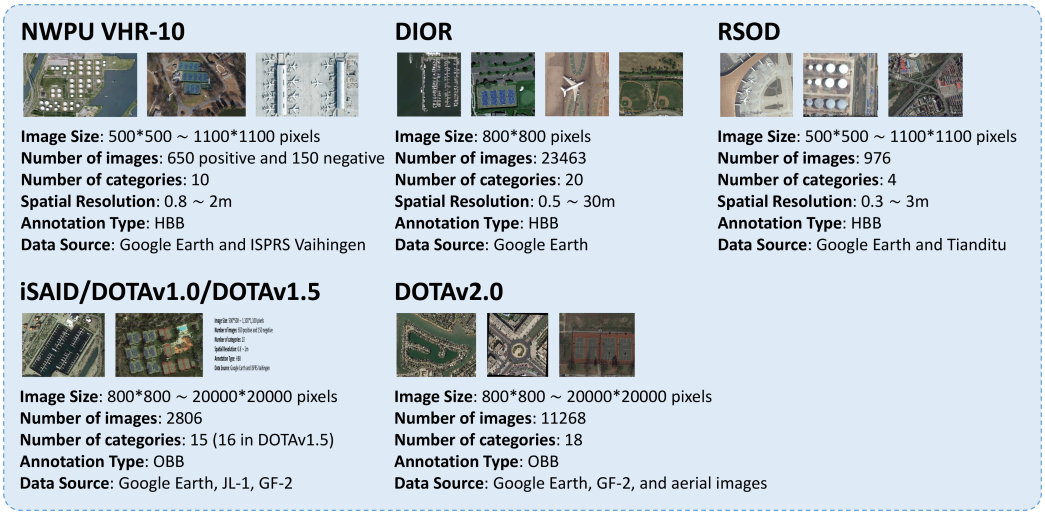

Figure 16. Detailed information of RS object detection datasets.

(1) *NWPU VHR-10* [69]: With 10 annotated categories, the works [10,17,97,98] share the same partition, where three classes are chosen as novel classes and the rest as base ones. All the images that do not contain any novel object are selected for base-training and a very small set that contains both base and novel classes with only k annotated boxes for each are constructed for fine-tuning, where k equals 1, 2, 3, 5, and 10. Then, the test set includes all the images containing at least one novel object except those for training, whose quantity is about 300. Xiao et al. [11] excluded storage tank and harbor because the number of these objects in an image is much larger than 10, which violates few-shot settings. In addition, the other designs are the same as it does on RSOD [112].

(2) *DIOR* [1]: As a large-scale publicly available dataset, DIOR has 20 categories with more than 23K images and 190K instances, in which five categories are selected to be

novel classes while keeping the remaining 15 classes as the base ones. The authors of [17,97,98] followed the few-shot setting of [10], which takes all the images of base classes in training set for base-training and extracts k annotated boxes for each novel class in training set for fine-tuning, where k belongs to 5, 10, and 20. Then, the performance is tested on the entire DIOR's validation set. When training, Cheng et al. [84] had k object instances for each novel class randomly from training set and validation set, where k equals to 3, 5, 10, 20, and 30, and performed evaluation on a testing set with both base and novel classes.

(3) *RSOD* [112]: There are four categories in RSOD. In [83], one class is randomly selected as the novel class and the remaining three are base classes. Four different base/novel splits are separately evaluated for objectivity. In the base training phase, for each base class, 60% of the samples are for training, and the rest for testing. In addition, in the novel fine-tuning phase, only k annotated boxes are selected in novel classes, where k equals to 1, 2, 3, 5, and 10.

(4) *iSAID* [113]: Instead of selecting a fixed number of novel classes, the authors of [16] designed three base/novel splits with a different number of categories according to data characteristics, as shown in Table 1. The novel categories in the first split mainly contain small-sized objects with a large variation in appearance, while the second split holds objects with relatively large dimensions and small variances in appearance. In addition, the third split takes the top-6 categories with the fewest instances as novel ones, which are regarded as the lowest occurrence. The base training phase takes all objects from base classes. When fine-tuning, the number of annotated boxes per class is respectively set to 10, 50, and 100.

(5) *DOTA* [3,68]: As an open source and constantly updated dataset, DOTA now has released versions 1.0, 1.5, and 2.0, whose number of categories increased from 15 to 18, and the number of instances expands nearly tenfold to 1.79 million. Jeune et al. [78] took DOTAv1.5 as the experimental dataset, and randomly chose two base/novel classes splits, 3 for novel and the remaining 13 for base. When constructing episodes, the number of shots for novel classes is 1, 3, 5, and 10.

(6) *DAN*: To enrich the samples, Gao et al. [105] constructed a dataset named DAN as a collection of DOTA and NWPU VHR-10, which consists of 15 representative categories. In addition, three categories are chosen to be novel ones, keeping the rest as base classes. The number of shots used for evaluation is not stated.

In addition to investigating the research status of FSOD in RS, this paper also focuses on the remarkable and inspiring works of FSOD in CV. So, the experimental datasets of natural scenarios and their few-shot settings will be introduced, including widely-used object detection benchmarks VOC 2007 [114], VOC 2012 [115], MS-COCO [116], and the first FSOD datasets called FSOD [92]. Different from the individualized dataset selection and splitting in RS, the CV field follows the designs in FR [99].

(1) *PASCAL VOC*: This is initially a large-scale well-annotated dataset for competition with two versions, VOC 2007 and VOC 2012. Following the common practice in classical object detection works [20,22], the VOC 07 test set is used for testing while VOC 07 and 12 train/val sets are for training. A total of 5 classes are randomly selected as novel ones, while keeping the remaining 15 ones as the base. In addition, 3 different base/novel splits are evaluated for objectiveness. When base-training, only annotations of base classes are available. In the novel fine-tuning phase, a very small set is constructed that includes k annotated boxes for each base and novel category, where $k \in \{1, 2, 3, 5, 10\}$.

(2) *MS-COCO*: COCO is a much larger dataset than VOC, which contains 80 categories with 1.5 million object instances. A total of 5000 images from the validation set are used for evaluation and the remaining images in train/val set are for training. The 20 categories overlapped with VOC are selected as novel classes, and the remaining 60 categories are used as base classes, so as to perform the cross-dataset learning

problem, that is, from COCO to PASCAL. Here, the number of annotated boxes with fine-tuning is 10 and 30.

(3) *FSOD*: In [92], it is indicated that the key to FSL is the generalization ability of the model when facing new categories. Thus, it proposes a highly-diverse dataset specifically designed for the FSOD task with a large number of object categories, called FSOD, which contains 1000 categories, 800 base classes for training, and 200 novel classes for testing. The average number of images per category is far less than that of the datasets designed for generic object detection, while the number of categories is far more than them, as shown in Table 2. Here, $k \in \{1, 5\}$.

Table 2. Dataset summary for natural images. Generic object detection datasets place emphasis on image diversity, while FSOD datasets focus on category diversity.

Dataset	# Class	# Image	# Instance	Avg. # Img/Cls
PASCAL VOC 07+12	20	21,493	52,090	2604.5
MS-COCO	80	330 K	1.5 million	4125
FSOD	1000	66,502	182,951	182.951

4.2. Evaluation Criteria

Similar to generic object detection, mean Average Precision (mAP) is also a standard evaluation metric in FSOD. There exist two calculations in PASCAL VOC and MS-COCO, whose difference is the threshold selection of IoU. The threshold is fixed to be 0.5 in VOC, while multiple thresholds are between 0.5 and 0.95 in COCO. In addition, COCO also provides the mAP calculation of objects in different sizes and the calculation of Average Recall. The current FSOD methods for RS images all follow the calculation of VOC, while CV field evaluates the both. The detailed calculation can be found in Appendix A.

4.3. Accuracy Performance

The current FSOD approaches on RS images are at an early stage. The benchmarks are still under exploration and a remote sensing dataset designed for few-shot scenario is missing. In Tables 3–9, we present a performance comparison of the existing approaches on each mentioned dataset.

Table 3. FSOD performance (AP50 in %) on the base and novel classes of the **NWPU VHR-10 dataset**. Novel classes are airplane, baseball diamond, and tennis court. Base classes are the remaining ones. Red and blue indicate the state-of-the-art and the second-best performance, respectively.

Method	Type	Novel mAP					Base mAP
		1-Shot	2-Shot	3-Shot	5-Shot	10-Shot	
FSODM [10]	Meta learning	-	-	32	53	65	**78**
SAM&BFS [17]	Transfer learning	**16.4**	**36.3**	**47**	**61.6**	**74.9**	-
OFA [97]	Meta learning	-	**34**	**43.2**	**60.4**	**66.7**	-
PAMS-Det [98]	Transfer learning	-	-	37	55	66	**88**

Table 4. FSOD performance (AP50 in %) on the novel classes of the **NWPU VHR-10 dataset**. Random 1 class is chosen to be the novel one, keeping the rest classes as base ones. Red and blue indicate the 1st and 2nd highest accuracy, respectively. SAAN is a meta learning-based few-shot detector.

Method	Novel Class	1-Shot	2-Shot	3-Shot	5-Shot	10-Shot
SAAN [11]	Airplane	9.09	15.72	25.71	34.68	38.78
	Ship	28.77	**52.78**	**57.95**	78.05	81.14
	Baseball diamond	**32.6**	**63.41**	**80.26**	**81.2**	**89.75**
	Tennis court	18.32	23.73	29.45	47.68	54.11
	Basketball court	21.33	37.74	42.43	54.78	80.16
	Ground track field	**31.29**	42.24	52.12	**78.88**	**89.57**
	Bridge	3.08	9.31	16.94	18.68	43.94
	Vehicle	9.37	13.15	22.77	29.6	36.24

Table 5. FSOD performance (AP50 in %) on the base and novel classes of the **DIOR dataset.** Novel classes are airplane, baseball field, tennis court, train station and windmill. Red and blue in dicate the SOTA and the second best, respectively.

Method	Type	Novel mAP					Base mAP
		2-Shot	3-Shot	5-Shot	10-Shot	20-Shot	
FSODM [10]	Meta learning	-	-	25	32	**36**	**54**
SAM&BFS [17]	Transfer learning	-	-	**38.3**	**47.3**	**50.9**	-
OFA [97]	Meta learning	**27.6**	**32.8**	**37.9**	**40.7**	-	-
PAMS-Det [98]	Transfer learning	-	**28**	33	38	-	**65**

Table 6. FSOD performance (AP50 in %) on the base and novel classes of the **DIOR dataset**. The base/novel splitting can be found in Table 1. Red and blue indicate the 1st and 2nd highest accuracy, respectively. P-CNN is a meta learning-based few-shot detector.

Method		Novel mAP					Base mAP				
		3-Shot	5-Shot	10-Shot	20-Shot	30-Shot	3-Shot	5-Shot	10-Shot	20-Shot	30-Shot
P-CNN [84]	Split 1	**18**	**22.8**	**27.6**	**29.6**	**32.6**	47	48.4	50.9	52.2	**53.3**
	Split 2	14.5	14.9	18.9	22.8	25.7	48.9	**49.1**	**52.5**	51.6	53
	Split 3	**16.5**	**18.8**	**23.3**	**28.8**	**32.1**	**49.5**	**49.9**	**52.1**	**53.1**	**53.6**
	Split 4	15.2	17.5	18.9	25.7	28.7	**49.8**	**49.9**	51.7	**52.3**	**53.6**

Table 7. FSOD performance (AP50 in %) on the novel classes of the **RSOD dataset**. Random one class is selected as the novel class and the remaining classes are base ones. Red and blue refer to the state-of-the-art and the second best. SAAN and FAHM are both meta-learning-based few-shot detectors. Our methods are based on transfer learning.

Method/Shot	Split 1					Split 2					Split 3					Split 4				
	1	2	3	5	10	1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
SAAN [11]	9.09	4.79	8.02	10.15	17.54	4.44	**17.69**	21.8	38.91	41.95	1.14	2.9	2.31	4.55	11.54	20.21	27.3	41.16	38.9	64.72
FAHM [83]	**9.1**	**10.6**	**15**	**20.2**	**43.5**	**16.54**	**34.8**	**51.6**	**60.92**	**71.27**	**9.09**	**11.36**	**21.01**	**41.4**	**59.63**	**33.54**	**50.5**	**61.6**	**77.1**	**88.98**
Ours	**12.9**	**16.2**	**22.5**	**27.7**	**32.1**	**9.09**	15.5	**37.4**	**58.4**	**63.2**	**11.5**	**15.2**	**20.8**	**42.7**	**66.9**	**37.6**	**68.1**	**84.8**	**90.3**	**96.6**

Table 8. FSOD performance (AP50 in %) on the base and novel classes of the **iSAID dataset**. The base/novel splitting can be found in Table 1. Red indicates the top highest accuracy. DH-FSDet is a transfer learning-based few-shot detector.

Method		Novel mAP			Base mAP		
		10-Shot	50-Shot	100-Shot	10-Shot	50-Shot	100-Shot
DH-FsDet [16]	Split 1	5.2 ± 0.8	12.8 ± 0.8	16.7 ± 1.7	65.0 ± 0.2	65.1 ± 0.1	65.2 ± 0.1
	Split 2	**14.5 ± 1.7**	**28.9 ± 3.4**	**36.0 ± 1.7**	64.5 ± 0.1	64.7 ± 0.1	64.8 ± 0.1
	Split 3	9.7 ± 2.2	19.6 ± 2.4	23.1 ± 0.9	**67.8 ± 0.1**	**68.0 ± 0.1**	**68.1 ± 0.1**

Table 9. FSOD performance (mAP) on the novel classes of the **DOTA dataset**. The base/novel splitting can be found in Table 1. Prototypical FRCN is a meta-learning-based few-shot detector.

Method		Novel mAP				Base mAP			
		1-Shot	2-Shot	5-Shot	10-Shot	1-Shot	2-Shot	5-Shot	10-Shot
prototypical FRCN [78]	Split 1	0.047 ± 0.02	0.024 ± 0.01	0.038 ± 0.01	0.041 ± 0.01	0.275 ± 0.01	0.352 ± 0.02	0.390 ± 0.01	0.384 ± 0.02
	Split 2	0.08 ± 0.01	0.101 ± 0.02	0.121 ± 0.01	0.101 ± 0.02	0.415 ± 0.03	0.392 ± 0.03	0.434 ± 0.02	0.414 ± 0.03

From the tables, we can find that detection accuracy increases significantly with the number of shots. More samples provide more information, which helps the learning procedure of few-shot detectors. In the existing FSOD approaches, transfer detectors account for a few [16,17,98] while meta detectors are more [10,11,78,83,84,97,105,117]. Each FSOD method has its own dataset organization and splitting. NWPU VHR-10 and DIOR are two relatively commonly used datasets, on which transfer detectors usually perform better. The superiority of one technical route over the other cannot be judged in the other datasets, because only transfer or meta detectors are adopted.

(1) **Experimental results on NWPU VHR-10.** As shown in Table 3, in terms of predictive performance on novel classes, the transfer detector [17] achieves to be state-of-the-art (SOTA) due to its shared attention module and the balanced fine-tuning strategy. The former calculates multi-attention maps from the rich base samples to share with the novel fine-tuning stage as prior knowledge, while the latter alleviates the training imbalance problem caused by sample sizes between base and novel classes. Meanwhile, migrating the detection capability to the target domain often sacrifices the performance on base classes, in which respect, transfer learning also impacts less. A different base/novel splitting is adopted in Table 4. After discarding the categories that violate the few-shot scenario, in each experimental setting, the remaining classes are respectively treated to be the novel ones. These results demonstrate that the difficulty of detecting novel classes is distinct from class to class. For baseball, diamond, and ground track field, 10 shots are adequate to achieve satisfactory results, while for airplane or vehicle, it is just not enough. Thus, [11] came to the conclusion that the number of objects that are defined as few shots should vary among categories.

(2) **Experiments results on DIOR.** In Table 5, the SOTA on DIOR when 5, 10, and 20 shots is illustrated to be the transfer detector [17], which is also the SOTA on NWPU VHR-10; it profited from its two proposed attention modules and training mechanism. Hence, when the sample size of novel objects is relatively large, the effect of the transfer detector is better. In addition, when only very few shots are provided, i.e., under a 2-shot or 3-shot setting, the meta detector [97] owns higher precision, which benefits from the meta-learning mechanism that can quickly learn from limited samples by training with a series of few-shot episodes. For the base accuracy attenuation problem, the new transfer-learning attempt PAMS-Det [98] is more able to handle it than the classical meta-learning approach FSODM [10], which also confirms the aforementioned conclusion that fine-tuning-based methods sacrifice less performance on the base classes than meta-based ones. Table 6 shows a different dataset reorganization thinking. The novel classes of the four splittings are non-overlapped. It is obvious that different base/novel splits own different detection accuracies. The reason is that, limited by the insufficient categories, the base training phase may not be able to extract totally category-irrelevant features. When there is a category-insufficient situation on both base and novel classes, the meta knowledge extraction from source domain and matching in target domain may result in contingency, thus influencing detection accuracy.

(3) **Experiments on RSOD.** As shown in Table 7, our two-stage multi-scale context-aware method has significantly better performance in most cases, which proves the effectiveness of our method in introducing the context mechanism into the two-stage multi-scale detection framework. The authors of [83] reached SOTA in the rest of the cases, which proposed a feature highlight module that can extract meta knowledge from coarse to fine. The creativity of the slightly inferior method [11] is a relation GRU for the matching between support knowledge and query instances. Thus, the accuracy comparison manifests that the import of contextual information does help and the extraction of meta knowledge is even more important than the matching procedure. Only the discriminative and representative features are founded, and the similarity measurement between support and query is meaningful.

(4) **Experiments on iSAID.** As stated before, the dataset splits in [16] are carefully designed, instead of random selection. The first split contains objects with large variations in appearance, while the second split is on the contrary. The accuracy comparison between the two in Table 8 confirms an assumption that 10 shots may be not enough for diversified objects. Although the third split also varies less in appearance, it detects worse than the second one, which indicates that considering more novel classes impedes the few-shot learning procedure. Contrasting between the second and third splits, an assumption that fewer base classes and more novel classes yield worse performance is made and verified by another set of experiments. Hence, we can conclude that higher diversity of classes and instances can generate less class-specific features that can benefit the novel fine-tuning phase, and considering more novel classes has the opposite effect.

(5) **Experiments on DOTA.** Generally speaking, more samples can bring better performance. However, this is not always observed in [78]. For instance, in the second split, the accuracy is stable with respect to the number of shots. In addition, Table 9 also shows that the accuracy gain is not as significant from 5 to 10 shots as that between 1, 3, and 5 shots. A relatively low number of samples may correctly approximate the class prototype and increasing the number might do damage. These are all because few samples lead to unsatisfactory representative space construction.

Considering the interdisciplinarity between RS and CV, some instructive FSOD works in CV are introduced in the paper. Hence, the few-shot detection performance of the classical, creative, well-performed algorithms on natural image datasets is listed in Tables A2–A5 in Appendix B.

In conclusion, the detection accuracy of few-shot object detection algorithms has made great progress in the recent years. Models based on transfer learning, meta learning perform equally well. To talk about the shots, obviously, the detection performance increases with the number of provided references. In addition, for transfer detectors, it is worth noting that there exists a base class performance degradation problem, which means that after achieving the detection of novel classes, the ability to detect base classes is damaged.

5. Opportunities and Challenges

Given the difficulties of RS sample acquisition, the demand for few-shot detection is increasingly prominent. So far, the development of FSOD in RS is still in its infancy, and many problems bother, such as inconsistent dataset selection and splitting, lacking specialized datasets, chaotic evaluation criteria, unsatisfactory detection performance, unstable training procedure, and so on. Thus, in this section, we introduce the opportunities and challenges in future research.

5.1. High-Diversity FSOD Dataset of RS Images

Although various large-scale well-annotated remote sensing datasets have been proposed for generic object detection and the existing FSOD works are also evaluated on the redistributed version of them, however, there exists a different emphasis on samples between the few-shot and enough-shot scenarios pointed out by [92]. Enough-shot datasets focus on the sample richness of each category and rely on massive data to provide sufficient prior knowledge for the model to achieve a high-precision detection effect of the known categories, while few-shot ones pay attention to category diversity instead of heavily redundant samples, in order to extract category-irrelevant meta knowledge that can be transferred from source domain to target domain that makes the few-shot detection models (FSDMs) quickly adapt to the detection of new categories. The RS datasets currently in use mentioned in Section 4 only have 20 categories at most, with 5 as a novel group and 15 as a base, which is far from perfect. The more base classes, the better for the FSDMs to extract category-independent knowledge, no matter whether transfer learning or meta learning, due to the mechanism of FSL. In addition, the more novel classes, the more robust and effective the FSDMs can be proved.

5.2. Few-Shot Sample Acquisition Principle

In addition to category diversity, the sample size is another consideration. To be a few-shot RS dataset, it is necessary to establish a multi-mode target acquisition rule with limited samples. Based on filtering and analysis on enough-shot base datasets, the pattern completeness should be guaranteed as much as possible while reducing the number of sampled objects. The pattern here means how targets exist in images, containing background diversity, object scale, and form variety, as shown in Figure 17. It is worth noting that owing to the single viewing perspective and the high resolution of RS images, there may exist a situation where one single image contains hundreds of objects, such as the densely distributed ships, airplanes, and vehicles, which is the expectation of enough-shot datasets, however violating the few-shot setting. Thus, the control of the sample amount needs deleting images with plenty of objects, and ensuring the separability between base classes and novel classes. All the potential objects should be annotated and belong to base or novel classes. If an image owns both base and novel object classes, it should not be selected. Because just removing partial annotations in an image will confuse the few-shot detectors.

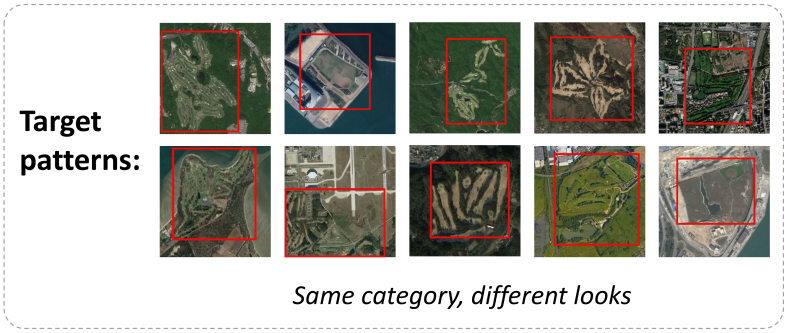

Figure 17. The illustration of target patterns of golf track fields.

5.3. Representation Abilities of Context Information

The target and its living environment together constitute the basic features for object detection. Under the condition of few-shot, the feature relationship between the target and its background is helpful to improve the classification accuracy in the detection process, which is proved in a pioneer one-stage SSD-style work named context-transformer [80]. Following our proposed roadmap, based on the context mechanism, we propose **multi-scale context-aware R-CNN** (the work has been accepted in IEEE International Geoscience and Remote Sensing Symposium (IGARSS) 2022) for FSOD in remote sensing images. The framework of our proposed method is shown in Figure 18. We adopt a two-stage training strategy, that is, source training and target fine-tuning. The **context-aware module** only assists in the second phase, and its main components are described below.

Context-aware Area Construction. Human attention usually focuses on the contextual information around the object rather than every detail. Thus, max-pooling is used to build the context-aware area of the feature maps pyramid,

$$A_k = MaxPooling(F_k), k = 1, 2, 3, 4, \tag{1}$$

where k denotes the different scales of the feature maps pyramid, F_k is the feature maps in different scales, A_k is the context-aware areas in different scales. The proposals on context-aware areas are generated by the proposals on the feature map pyramids, which are expanded to keep the size of the region on the feature maps unchanged according to the max-pooling step.

Relevance matrix. ROI features vectors P of the proposals and corresponding contextual information vectors Q are obtained by ROI align and FC layers transformation. Then, a relevance matrix R between the two is obtained by the dot product kernel,

$$R = FC(P) \times FC(Q^T), \tag{2}$$

where full connection layer FC can increase the learning flexibility of kernel computing.

Contextual information combination. To combine contextual information into ROI features as a kind of relational attention, we firstly use softmax in each row of the relevance matrix to obtain the importance of each contextual information. Then, it is used to obtain a weighted contextual vector V,

$$V = softmax(R) \times FC(Q), \tag{3}$$

Finally, the weighted contextual vectors V are restructured through a full connection layer, and combined into the ROI features P to obtain the feature vector P' for object classification in the target domain,

$$P' = P + FC(V), \tag{4}$$

Then, this vector is fed into a fully connected to obtain object scores in target domain.

Classification Acquisition. We use the BOX classifier at the last layer of the pre-training model and fine-tune it to obtain object scores of the source domain. Finally, object scores of the source domain and object scores of the target domain are concatenated as the final output of the classifier.

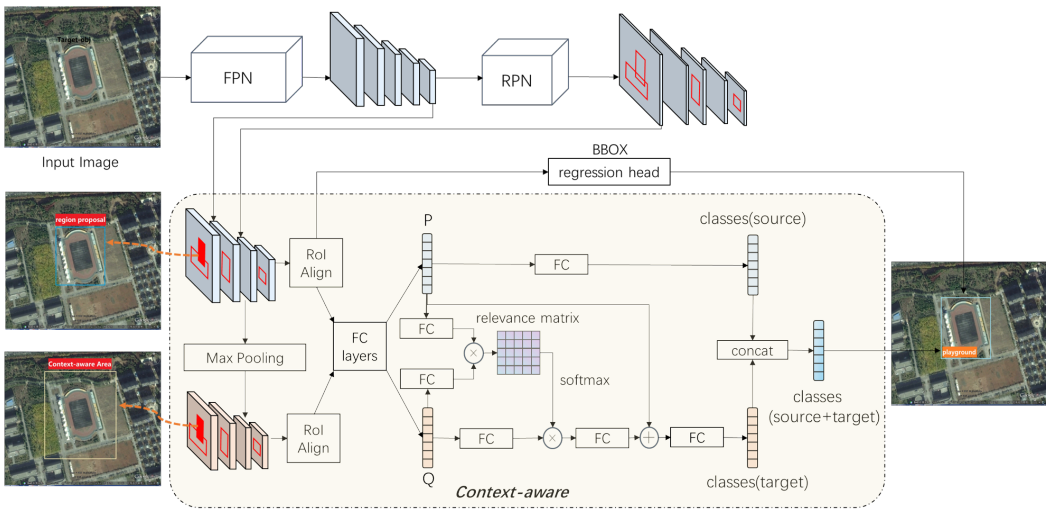

Figure 18. The framework of our proposed method. We regard base classes and novel classes as the source domain and the target domain, respectively. The context-aware module explores the contextual information to guide the classification in the target domain.

5.4. Effective and Reliable Training Strategy

With the continuous evolution of deep neural networks, especially for those oriented to special targets, the scale of the whole network is enormous, but the parameter between network nodes are still updated by conventional forward and backpropagation. So, the meta-learning strategy with multiple subtasks will lead to low operation efficiency of FSDMs, while a transfer learning strategy can reduce such risks to a certain extent. Although transfer technology provides theoretical and algorithm foundation of the computing mechanism, in the case that a regression task is difficult to migrate, how to design an efficient and reliable machine learning strategy to guarantee the robustness of network training and prediction is an urgent challenge.

This embodies two aspects: first, to acquire multi-mode target feature knowledge based on single-task online learning; second, to integrate the learning results of multiple tasks, and convert them into target-aggregated feature knowledge suitable for the current learning task. Single task online learning can ensure the update efficiency of FSDM and is the key to controlling the scale of the network and the high detection efficiency of specific targets. Based on this, with the continuous increase of learning tasks on detection networks,

it is necessary to integrate multi-mode features of learning targets from different training models to improve the predictive performance of the detection algorithm.

5.5. Model Ensemble for Industrial Application

Allen et al. [118] indicated that in a dataset, an object usually has multi-view samples, while multi-view samples in remote sensing images are usually caused by background differences, chromatic aberrations caused by light, differences in scales or directions, and varying degrees of cloud and fog occlusion. Because of the randomness in the learning process, one FSDM will quickly learn one subset of these view features and mark the remaining few samples that cannot be classified. Thus, to achieve higher application accuracy, it is a natural idea to integrate the detection abilities of multiple trained models, realized by a model ensemble or knowledge distillation. Model ensemble averages the outputs of independently trained FSDMs with or without weights, while knowledge distillation needs an extra model to remove irrelevant parameters from multiple FSDMs. By fusing these multiple trained models, all the learnable view features are selected, which significantly benefits the detection ability of few-shot detectors learned from a small number of samples.

6. Conclusions

Few-shot object detection is a burgeoning research area that can achieve localization and classification of possible targets within limited supervised information. It facilitates the achievement of object detection in the case of sample acquisition difficulties. Firstly, we review the milestones and state-of-the-art algorithms in deep learning-based object detection for remote sensing images. Then, we demonstrate some few-shot object detection design methods for sample, model, and strategy, respectively. In addition to the small amount of existing remote sensing literature, we also present valuable and remarkable works in computer vision. Based on our understanding, the current design of the few-shot object detection framework is inseparable from three vital aspects:

(i) Design few-shot data acquisition rules for remote sensing images.
Through the analysis of existing remote sensing object detection datasets, we find that, on the one hand, these datasets do not control the number of instances of the same category in one image. When the number is far more than the few-shot condition, such as 1, 3, or 5-shot, and thus, the data cannot be used for few-shot detectors. On the other hand, current annotations do not consider contextual information. However, some work suggests that the characteristics of the target's environment benefit few-shot tasks. Therefore, how to screen, supplement and label datasets that meet the requirement of few-shot object detection is the primary problem to be solved.

(ii) Build a deep neural network model close to the essence of human cognition.
The existing few-shot detectors focus on extracting information or detection abilities from base classes and how to apply these into novel classes, which follow our stated research route. The semantics shared between source and target domains need to obey the cognition of "visual-semantic-semantic-visual", and mining-shared features are the key to network model design.

(iii) Establish a learning strategy that can aggregate multi-modal knowledge and improve network efficiency.
Although there is still a performance gap between few-shot detectors and deep learning detectors, related studies show that the few-shot ones have shown acceptable detection ability for some kinds of targets with a limited number of shots. So, it is necessary to further focus on the integration and training methods, which can drive and build the practical application capabilities of few-shot detectors in these categories.

Given the above three problems, future research directions should focus on:

Sample: Different from the acquisition rules of deep learning, few-shot datasets require filtering from some massive datasets to reduce sample size while ensuring target pattern completeness. Meanwhile, the view discrepancy should also be amplified by highlighting

target modes and features of support images and enriching the background information of query images.

Model: Human cognition highly relies on prior knowledge, so the few-shot model focuses on semantic extraction to achieve the same effect. Therefore, a robust few-shot object detection network should contain background inference suppression, attention mechanism, and high- and low-level semantic extraction units.

Strategy: Meta strategy makes the network inefficient, but transfer strategy is hard to migrate regression tasks. Thus, it is necessary to design an efficient and reliable machine learning strategy that obtains multi-mode features from single-task online learning and integrates multiple learning tasks, to ensure the robustness of training and prediction.

Currently, this "***sample + model + strategy***" is the major research route, and it is also suitable for few-shot object detection algorithms in remote sensing images with complicated backgrounds, scale variance, multiple directions, and dense small target issues. Following that, we summarize the experimental datasets, evaluation criteria, and accuracy performance to illustrate the achievements and shortcomings of the stated algorithms. Finally, we highlight the opportunities and challenges of few-shot object detection in remote sensing images. In addition, based on our stated research route, novel research focused on contextual information has been proposed. In conclusion, this paper aims to present a clear technical route for readers and provide valuable references for the research methods in this field.

Author Contributions: Conceptualization, S.L. and Y.Y.; methodology, S.L., Y.Y. and H.S.; software, H.S.; validation, S.L., Y.Y. and H.S.; formal analysis, S.L. and Y.Y.; investigation, S.L., Y.Y. and H.S.; resources, Y.Y.; data curation, S.L.; writing—original draft preparation, S.L. and Y.Y.; writing—review and editing, S.L., Y.Y. and H.S.; visualization, S.L.; supervision, Y.Y., G.M., W.Y. and F.L.; project administration, Y.Y.; funding acquisition, Y.Y. and F.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China (62101060).

Data Availability Statement: Not applicable.

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

Abbreviations

The following abbreviations are used in this manuscript:

FSOD	Few-Shot Object Detection
FSL	Few-shot learning
FSC	Few-shot Classification

Appendix A. mAP Calculation

The purpose of FSOD is to achieve localization and classification of possible targets in an image by learning from a few samples. Thus, the model outputs a set of predicted boxes with belonging categories. To evaluate each prediction box, its *intersection over Union (IoU)* between ground truth is firstly calculated, which is the ratio of the intersection area of the prediction box and ground truth to the area of their union. Then, a threshold is given to judge the IoU values, which is fixed to be 0.5 in VOC, while multiple thresholds from 0.5 to 0.95 in COCO.

Then, three judgments come are obtained: *true positive (TP)*, *false positive (FP)*, and *false negative (FN)*. TP is the correct prediction, which indicates the number of prediction boxes whose IoU with ground truth is larger than a predefined threshold. FP means incorrect prediction, that is, there is no object here but it is predicted, which points to the number of prediction boxes whose IoU with ground truth is less than a predefined threshold. FN represents missing prediction, that is, there are objects here but they are not detected, which refers to the number of ground truth that is not being detected.

With the above three, *average precision (AP)* that evaluates the performance of FSOD models can be calculated by *precision* and *recall*.

$$precision = TP/(TP + FP), \tag{A1}$$

which refers to the percentage of correctly detected objects in the total number of detected objects.

$$recall = TP/(TP + FN), \tag{A2}$$

which indicates the percentage of correctly detected objects in the total number of objects in the test set. Then, the AP is the area of P-R curve formed with recall as the horizontal axis and Precision as the vertical axis. In addition, *mAP*, stated in Section 4, refers to the average AP value of each category.

The explanation of evaluation criteria in MS-COCO dataset is shown in Table A1. mAP in VOC datasets is equivalent to AP50 in COCO.

Table A1. Object detection evaluation criteria. AP, AP50, AP75, APS, APM, APL, AR1, AR10, AR100, ARS, ARM, ARL for MS-COCO dataset.

Average Precision (AP):		Average Recall (AR):	
AP	AP at IoU = :50 : :05 : :95	AR1	AR given 1 detection per image
AP50	AP at IoU = :50	AR10	AR given 10 detections per image
AP75	AP at IoU = :75	AR100	AR given 100 detections per image
AP Across Scales:		AR Across Scales:	
APS	AP for small objects: area < 322	ARS	AR for small objects: area < 322
APM	AP for medium objects: 322 < area < 962	ARM	AR for medium objects: 322 < area < 962
APL	AP for large objects: area > 962	ARL	AR for large objects: area > 962

Appendix B. FSOD Performance in Natural Image Datasets

Table A2. FSOD performance (AP50 in %) on the **PASCAL VOC 2007 test set**. The upper half is transfer learning-based FSOD approaches, while the lower half refers to meta learning-based FSOD methods. Red, and blue indicate the state-of-the-art and the second-best, respectively.

Method/Shot	**Novel Set 1**					**Novel Set 2**					**Novel Set 3**				
	1	**2**	**3**	**5**	**10**	**1**	**2**	**3**	**5**	**10**	**1**	**2**	**3**	**5**	**10**
LSTD(YOLO)[85]	8.2	1	12.4	29.1	38.5	11.4	3.8	5	15.7	31	12.6	8.5	15	27.3	36.3
RepMet [77]	26.1	32.9	34.4	38.6	41.3	17.2	22.1	23.4	28.3	35.8	27.5	31.1	31.5	34.4	37.2
TFA w/ fc [86]	36.8	29.1	43.6	55.7	57	18.2	29	33.4	35.5	39	27.7	33.6	42.5	48.7	50.2
TFA w/ cos [86]	39.8	36.1	44.7	55.7	56	23.5	26.9	34.1	35.1	39.1	30.8	34.8	42.8	49.5	49.8
HOSENet [87]	32.9	-	47.4	52.6	54.9	-	-	-	-	-	-	-	-	-	-
LSCN [119]	30.7	43.1	43.7	53.4	59.1	22.3	25.7	34.8	41.6	**50.3**	21.9	23.4	30.7	43.1	55.6
Retentive RCNN [89]	42.4	45.8	45.9	53.7	56.1	21.7	27.8	35.2	37	40.3	30.2	37.6	43	49.7	50.1
FSSP [74]	41.6	-	49.1	54.2	56.5	30.5	-	39.5	41.4	45.1	36.7	-	45.3	49.4	51.3
MPSR [73]	41.7	43.1	51.4	55.2	61.8	24.4	29.5	39.2	39.9	47.8	35.6	40.6	42.3	48	49.7
TFA + Halluc [120]	45.1	44	44.7	55	55.9	23.2	27.5	35.1	34.9	39	30.5	35.1	41.4	49	49.3
CoRPNs+ Halluc [120]	47	44.9	46.5	54.7	54.7	26.3	31.8	37.4	37.4	41.2	40.4	42.1	43.3	51.4	49.6
FSCE [88]	44.2	43.8	51.4	61.9	63.4	27.3	29.5	43.5	44.2	50.2	37.2	41.9	47.5	54.6	58.5
SRR-FSD [81]	47.8	50.5	51.3	55.2	56.8	32.5	35.3	39.1	40.8	43.8	40.1	41.5	44.3	46.9	46.4
FSOD-SR [75]	**50.1**	**54.4**	56.2	**60**	62.4	29.5	**39.9**	43.5	44.6	48.1	**43.6**	**46.6**	**53.4**	53.4	**59.5**
DeFRCN [121]	**53.6**	**57.5**	**61.5**	**64.1**	60.8	30.1	38.1	**47**	**53.3**	47.9	**48.4**	**50.9**	52.3	**54.9**	57.4
FSRW [99]	14.8	15.5	26.7	33.9	47.2	15.7	15.3	22.7	30.1	40.5	21.3	25.6	28.4	42.8	45.9
MetaDet(YOLO) [90]	17.1	19.1	28.9	35	48.8	18.2	20.6	25.9	30.6	41.5	20.1	22.3	27.9	41.9	42.9
MetaDet(FRCN) [90]	18.9	20.6	30.2	36.8	49.6	21.8	23.1	27.8	31.7	43	20.6	23.9	29.4	43.9	44.1
Meta R-CNN [91]	19.9	25.5	35	45.7	51.5	10.4	19.4	29.6	34.8	45.4	14.3	18.2	27.5	41.2	48.1
FsDetView [122]	24.2	35.3	42.2	49.1	57.4	21.6	24.6	31.9	37	45.7	21.2	30	37.2	43.8	49.6
TIP [123]	27.7	36.5	43.3	50.2	59.6	22.7	30.1	33.8	40.9	46.9	21.7	30.6	38.1	44.5	50.9
DCNet [95]	33.9	37.4	43.7	51.1	59.6	23.2	24.8	30.6	36.7	46.6	32.3	34.9	39.7	42.6	50.7
NP-RepMet [76]	37.8	40.3	41.7	47.3	49.4	**41.6**	**43**	43.3	47.4	49.1	33.3	38	39.8	41.5	44.8
GenDet [94]	38.5	47.1	52.2	57.7	**63.5**	26.8	34	37.3	42.8	48.3	33.4	40	44.3	51.2	56.5
Meta-DETR [102]	40.6	51.4	**58**	59.2	**63.6**	**37**	36.6	**43.7**	**49.1**	**54.6**	41.6	45.9	**52.7**	**58.9**	**60.6**
CME [79]	41.5	47.5	50.4	58.2	60.9	27.2	30.2	41.4	42.5	46.8	34.3	39.6	45.1	48.3	51.5

Table A3. Few-shot detection performance (AP (%) and AR (%)) of novel classes on the **COCO dataset when 10-shot**. Red, and blue indicate the state-of-the-art and the second-best, respectively.

#Shots	Method	Average Precision						Average Recall					
		AP	AP50	AP75	APS	APM	APL	AR1	AR10	AR100	ARS	ARM	APL
10	LSTD(YOLO) [85]	3.2	8.1	2.1	0.9	2	6.5	7.8	10.4	10.4	1.1	5.6	19.6
	TFA w/ fc [86]	9.1	17.3	8.5	-	-	-	-	-	-	-	-	-
	TFA w/ cos [86]	9.1	17.1	8.8	-	-	-	-	-	-	-	-	-
	HOSENet [87]	10	-	9.1	-	-	-	-	-	-	-	-	-
	MPSR [73]	9.8	17.9	9.7	3.3	9.2	16.1	15.7	21.2	21.2	4.6	19.6	34.3
	FSOD-SR [75]	11.6	12.7	10.4	4.6	10.5	17.2	16.4	23.9	24.1	9.3	21.8	37.7
	FSCE [88]	11.9	-	10.5	-	-	-	-	-	-	-	-	-
	FSSP [74]	9.9	20.4	9.6	-	-	-	-	-	-	-	-	-
	SRR-FSD [81]	11.3	23	9.8	-	-	-	-	-	-	-	-	
	LSCN [119]	12.4	26.3	7.57	-	-	-	-	-	-	-	-	
	FSRW [99]	5.6	12.3	4.6	0.9	3.5	10.5	10.1	14.3	14.4	1.5	8.4	28.2
	MetaDet [90]	7.1	14.6	6.1	1	4.1	12.2	11.9	15.1	15.5	1.7	9.7	30.1
	Meta R-CNN [91]	8.7	19.1	6.6	2.3	7.7	14	12.6	17.8	17.9	7.8	15.6	27.2
	GenDet [94]	9.9	18.8	9.6	3.6	8.4	15.4	-	-	-	-	-	-
	FsDetView [122]	12.5	27.3	9.8	2.5	13.8	19.9	20	25.5	25.7	7.5	27.6	38.9
	DCNet [95]	12.8	23.4	11.2	4.3	13.8	21	18.1	26.7	25.6	7.9	24.5	36.7
	CME [79]	15.1	24.6	16.4	4.6	16.6	26	16.3	22.6	22.8	6.6	24.7	39.7
	Meta-DETR [102]	19	30.5	19.7	-	-	-	-	-	-	-	-	-

Table A4. Few-shot detection performance (AP (%) and AR (%)) of novel classes on the **COCO dataset when 30-shot**. Red, and blue indicate the state-of-the-art and the second-best, respectively.

#Shots	Method	Average Precision						Average Recall					
		AP	AP50	AP75	APS	APM	APL	AR1	AR10	AR100	ARS	ARM	APL
30	LSTD(YOLO) [85]	6.7	15.8	5.1	0.4	2.9	12.3	10.9	14.3	14.3	0.9	7.1	27
	TFA w/ fc [86]	12	22.2	11.8	-	-	-	-	-	-	-	-	-
	TFA w/ cos [86]	12.1	22	12	-	-	-	-	-	-	-	-	-
	HOSENet [87]	14	-	14	-	-	-	-	-	-	-	-	-
	MPSR [73]	14.1	25.4	14.2	4	12.9	23	17.7	24.2	24.3	5.5	21	39.3
	FSOD-SR [75]	15.2	27.5	14.6	6.1	14.5	24.7	18.4	27.1	27.3	9.8	25.1	42.6
	FSCE [88]	16.4	-	16.2	-	-	-	-	-	-	-	-	
	FSSP [74]	14.2	25	13.9	-	-	-	-	-	-	-	-	-
	SRR-FSD [81]	14.7	29.2	13.5	-	-	-	-	-	-	-	-	
	LSCN [119]	13.9	30.9	9.96	-	-	-	-	-	-	-	-	
	FSRW [99]	9.1	19	7.6	0.8	4.9	16.8	13.2	17.7	17.8	1.5	10.4	33.5
	MetaDet [90]	11.3	21.7	8.1	1.1	6.2	17.3	14.5	18.9	19.2	1.8	11.1	34.4
	Meta RCNN [91]	12.4	25.3	10.8	2.8	11.6	19	15	21.4	21.7	8.6	20	32.1
	GenDet [94]	14.3	27.5	13.8	4.8	13	24.2	-	-	-	-	-	-
	FsDetView [122]	14.7	30.6	12.2	3.2	15.2	23.8	22	28.2	28.4	8.3	30.3	42.1
	DCNet [95]	18.6	32.6	17.5	6.9	16.5	27.4	22.8	27.6	28.6	8.4	25.6	43.4
	CME [79]	16.9	28	17.8	4.6	18	29.2	17.5	23.8	24	6	24.6	42.5
	Meta-DETR [102]	22.2	35	22.8	-	-	-	-	-	-	-	-	-

Table A5. Base-forgetting comparisons on PASCAL VOC Novel set 1. Before fine-tuning, the base AP50 in base training is 80.8. Red, and blue indicate the state-of-the-art and the second-best, respectively.

Method	Base AP50			Novel AP50		
	1	3	5	1	3	5
FSRW [99]	62.9	61.2	60.7	14.2	29.8	36.5
TFA w/ fc [86]	77.1	76	75.1	22.9	40.4	46.7
TFA w/ cos [86]	77.6	77.3	77.4	25.3	42.1	47.9
MPSR [73]	59.4	66.2	67.9	25.5	41.1	49.6
HOSENet [87]	79.2	77.9	77.8	32.9	47.4	52.9
FSCE [88]	78.9	74.1	76.6	44.2	51.4	61.9
FSSP [74]	-	73.5	-	-	49.1	-
FSOD-SR [75]	-	77.4	-	-	56.2	-

References

1. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [CrossRef]
2. Cheng, G.; Han, J. A survey on object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28. [CrossRef]

3. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983.
4. Lam, D.; Kuzma, R.; McGee, K.; Dooley, S.; Laielli, M.; Klaric, M.; Bulatov, Y.; McCord, B. xview: Objects in context in overhead imagery. *arXiv* **2018**, arXiv:1802.07856.
5. Fei-Fei, L.; Fergus, R.; Perona, P. One-shot learning of object categories. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 594–611. [CrossRef] [PubMed]
6. Fink, M. Object classification from a single example utilizing class relevance metrics. In *Advances in Neural Information Processing Systems*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2004; Volume 17.
7. Shi, J.; Jiang, Z.; Zhang, H. Few-shot ship classification in optical remote sensing images using nearest neighbor prototype representation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 3581–3590. [CrossRef]
8. Yao, X.; Yang, L.; Cheng, G.; Han, J.; Guo, L. Scene classification of high resolution remote sensing images via self-paced deep learning. In Proceedings of the IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 521–524.
9. Kim, J.; Chi, M. SAFFNet: Self-attention-based feature fusion network for remote sensing few-shot scene classification. *Remote Sens.* **2021**, *13*, 2532. [CrossRef]
10. Li, X.; Deng, J.; Fang, Y. Few-Shot Object Detection on Remote Sensing Images. *IEEE Trans. Geosci. Remote. Sens.* **2021**, *60*, 1–14. [CrossRef]
11. Xiao, Z.; Qi, J.; Xue, W.; Zhong, P. Few-Shot Object Detection With Self-Adaptive Attention Network for Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 4854–4865. [CrossRef]
12. Huang, G.; Laradji, I.; Vazquez, D.; Lacoste-Julien, S.; Rodriguez, P. A survey of self-supervised and few-shot object detection. *arXiv* **2021**, arXiv:2110.14711.
13. Antonelli, S.; Avola, D.; Cinque, L.; Crisostomi, D.; Foresti, G.L.; Galasso, F.; Marini, M.R.; Mecca, A.; Pannone, D. Few-shot object detection: A survey. *ACM Comput. Surv. (CSUR)* **2021**. [CrossRef]
14. Wang, Y.; Yao, Q.; Kwok, J.T.; Ni, L.M. Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv. (CSUR)* **2020**, *53*, 1–34. [CrossRef]
15. Sun, X.; Wang, B.; Wang, Z.; Li, H.; Li, H.; Fu, K. Research progress on few-shot learning for remote sensing image interpretation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 2387–2402. [CrossRef]
16. Wolf, S.; Meier, J.; Sommer, L.; Beyerer, J. Double Head Predictor based Few-Shot Object Detection for Aerial Imagery. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 721–731.
17. Huang, X.; He, B.; Tong, M.; Wang, D.; He, C. Few-Shot Object Detection on Remote Sensing Images via Shared Attention Module and Balanced Fine-Tuning Strategy. *Remote Sens.* **2021**, *13*, 3816. [CrossRef]
18. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
19. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
20. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [CrossRef]
21. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788. [CrossRef]
22. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
23. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
24. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
25. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
26. Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 734–750.
27. Dong, Z.; Li, G.; Liao, Y.; Wang, F.; Ren, P.; Qian, C. Centripetalnet: Pursuing high-quality keypoint pairs for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10519–10528.
28. Zhou, X.; Zhuo, J.; Krahenbuhl, P. Bottom-up object detection by grouping extreme and center points. In Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 850–859.
29. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850.

30. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. Centernet: Keypoint triplets for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 6569–6578.
31. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2017; pp. 5998–6008.
32. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
33. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
34. Wang, C.; Bai, X.; Wang, S.; Zhou, J.; Ren, P. Multiscale visual attention networks for object detection in VHR remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2018**, *16*, 310–314. [CrossRef]
35. Li, Q.; Mou, L.; Jiang, K.; Liu, Q.; Wang, Y.; Zhu, X.X. Hierarchical region based convolution neural network for multiscale object detection in remote sensing images. In Proceedings of the IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 4355–4358.
36. Li, C.; Xu, C.; Cui, Z.; Wang, D.; Zhang, T.; Yang, J. Feature-attentioned object detection in remote sensing imagery. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 3886–3890.
37. Wang, J.; Ding, J.; Guo, H.; Cheng, W.; Pan, T.; Yang, W. Mask OBB: A semantic attention-based mask oriented bounding box representation for multi-category object detection in aerial images. *Remote Sens.* **2019**, *11*, 2930. [CrossRef]
38. Yang, F.; Li, W.; Hu, H.; Li, W.; Wang, P. Multi-scale feature integrated attention-based rotation network for object detection in VHR aerial images. *Sensors* **2020**, *20*, 1686. [CrossRef] [PubMed]
39. You, Y.; Cao, J.; Zhang, Y.; Liu, F.; Zhou, W. Nearshore ship detection on high-resolution remote sensing image via scene-mask R-CNN. *IEEE Access* **2019**, *7*, 128431–128444. [CrossRef]
40. Deng, Z.; Sun, H.; Zhou, S.; Zhao, J.; Zou, H. Toward fast and accurate vehicle detection in aerial images using coupled region-based convolutional neural networks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 3652–3664. [CrossRef]
41. Mou, L.; Zhu, X.X. Vehicle instance segmentation from aerial image and video using a multitask learning residual fully convolutional network. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 6699–6711. [CrossRef]
42. Li, Q.; Mou, L.; Xu, Q.; Zhang, Y.; Zhu, X.X. R^3-net: A deep network for multi-oriented vehicle detection in aerial images and videos. *arXiv* **2018**, arXiv:1808.05560.
43. Li, X.; Men, F.; Lv, S.; Jiang, X.; Pan, M.; Ma, Q.; Yu, H. Vehicle Detection in Very-High-Resolution Remote Sensing Images Based on an Anchor-Free Detection Model with a More Precise Foveal Area. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 549. [CrossRef]
44. Yang, X.; Sun, H.; Sun, X.; Yan, M.; Guo, Z.; Fu, K. Position detection and direction prediction for arbitrary-oriented ships via multitask rotation region convolutional neural network. *IEEE Access* **2018**, *6*, 50839–50849. [CrossRef]
45. He, Y.; Sun, X.; Gao, L.; Zhang, B. Ship detection without sea-land segmentation for large-scale high-resolution optical satellite images. In Proceedings of the IGARSS 2018–2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 717–720.
46. Fu, Y.; Wu, F.; Zhao, J. Context-aware and depthwise-based detection on orbit for remote sensing image. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 1725–1730.
47. Li, M.; Guo, W.; Zhang, Z.; Yu, W.; Zhang, T. Rotated region based fully convolutional network for ship detection. In Proceedings of the IGARSS 2018–2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 673–676.
48. Schilling, H.; Bulatov, D.; Niessner, R.; Middelmann, W.; Soergel, U. Detection of vehicles in multisensor data via multibranch convolutional neural networks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 4299–4316. [CrossRef]
49. Cheng, G.; Zhou, P.; Han, J. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 7405–7415. [CrossRef]
50. Li, K.; Cheng, G.; Bu, S.; You, X. Rotation-insensitive and context-augmented object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2017**, *56*, 2337–2348. [CrossRef]
51. Wu, X.; Hong, D.; Tian, J.; Chanussot, J.; Li, W.; Tao, R. ORSIm detector: A novel object detection framework in optical remote sensing imagery using spatial-frequency channel features. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5146–5158. [CrossRef]
52. You, Y.; Ran, B.; Meng, G.; Li, Z.; Liu, F.; Li, Z. OPD-Net: Prow detection based on feature enhancement and improved regression model in optical remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 6121–6137. [CrossRef]
53. Liu, Z.; Hu, J.; Weng, L.; Yang, Y. Rotated region based CNN for ship detection. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 900–904.
54. Liu, W.; Ma, L.; Chen, H. Arbitrary-oriented ship detection framework in optical remote-sensing images. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 937–941. [CrossRef]
55. Zhao, W.; Ma, W.; Jiao, L.; Chen, P.; Yang, S.; Hou, B. Multi-scale image block-level F-CNN for remote sensing images object detection. *IEEE Access* **2019**, *7*, 43607–43621. [CrossRef]
56. Zhang, W.; Wang, S.; Thachan, S.; Chen, J.; Qian, Y. Deconv R-CNN for small object detection on remote sensing images. In Proceedings of the IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 2483–2486.

57. Van Etten, A. You only look twice: Rapid multi-scale object detection in satellite imagery. *arXiv* **2018**, arXiv:1805.09512.
58. Ma, Y.; Wei, J.; Zhou, F.; Zhu, Y.; Liu, J.; Lei, M. Balanced learning-based method for remote sensing aircraft detection. In Proceedings of the 2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP), Chongqing, China, 11–13 December 2019; pp. 1–4.
59. Yu, L.; Hu, H.; Zhong, Z.; Wu, H.; Deng, Q. GLF-Net: A Target Detection Method Based on Global and Local Multiscale Feature Fusion of Remote Sensing Aircraft Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 4021505. [CrossRef]
60. Zhang, S.; He, G.; Chen, H.B.; Jing, N.; Wang, Q. Scale adaptive proposal network for object detection in remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 864–868. [CrossRef]
61. Zou, F.; Xiao, W.; Ji, W.; He, K.; Yang, Z.; Song, J.; Zhou, H.; Li, K. Arbitrary-oriented object detection via dense feature fusion and attention model for remote sensing super-resolution image. *Neural Comput. Appl.* **2020**, *32*, 14549–14562. [CrossRef]
62. Fu, K.; Chang, Z.; Zhang, Y.; Xu, G.; Zhang, K.; Sun, X. Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2020**, *161*, 294–308. [CrossRef]
63. Azimi, S.M.; Vig, E.; Bahmanyar, R.; Körner, M.; Reinartz, P. Towards multi-class object detection in unconstrained remote sensing imagery. In Proceedings of the Asian Conference on Computer Vision, Perth, Australia, 2–6 December 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 150–165.
64. Fu, K.; Chen, Z.; Zhang, Y.; Sun, X. Enhanced feature representation in detection for optical remote sensing images. *Remote Sens.* **2019**, *11*, 2095. [CrossRef]
65. Li, Y.; Huang, Q.; Pei, X.; Jiao, L.; Shang, R. RADet: Refine feature pyramid network and multi-layer attention network for arbitrary-oriented object detection of remote sensing images. *Remote Sens.* **2020**, *12*, 389. [CrossRef]
66. Zhang, G.; Lu, S.; Zhang, W. CAD-Net: A context-aware detection network for objects in remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 10015–10024. [CrossRef]
67. Xu, C.; Li, C.; Cui, Z.; Zhang, T.; Yang, J. Hierarchical semantic propagation for object detection in remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 4353–4364. [CrossRef]
68. Ding, J.; Xue, N.; Xia, G.S.; Bai, X.; Yang, W.; Yang, M.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; et al. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**. [CrossRef] [PubMed]
69. Cheng, G.; Han, J.; Zhou, P.; Guo, L. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS J. Photogramm. Remote Sens.* **2014**, *98*, 119–132. [CrossRef]
70. Li, X.; Wang, S. Object detection using convolutional neural networks in a coarse-to-fine manner. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 2037–2041. [CrossRef]
71. Chen, G.; Liu, L.; Hu, W.; Pan, Z. Semi-supervised object detection in remote sensing images using generative adversarial networks. In Proceedings of the IGARSS 2018–2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 2503–2506.
72. Wang, T.; Zhang, X.; Yuan, L.; Feng, J. Few-shot adaptive faster r-cnn. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7173–7182.
73. Wu, J.; Liu, S.; Huang, D.; Wang, Y. Multi-scale positive sample refinement for few-shot object detection. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 456–472.
74. Xu, H.; Wang, X.; Shao, F.; Duan, B.; Zhang, P. Few-shot object detection via sample processing. *IEEE Access* **2021**, *9*, 29207–29221. [CrossRef]
75. Kim, G.; Jung, H.G.; Lee, S.W. Spatial reasoning for few-shot object detection. *Pattern Recognit.* **2021**, *120*, 108118. [CrossRef]
76. Yang, Y.; Wei, F.; Shi, M.; Li, G. Restoring negative information in few-shot object detection. *arXiv* **2020**, arXiv:2010.11714.
77. Karlinsky, L.; Shtok, J.; Harary, S.; Schwartz, E.; Aides, A.; Feris, R.; Giryes, R.; Bronstein, A.M. Repmet: Representative-based metric learning for classification and few-shot object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5197–5206.
78. Jeune, P.L.; Lebbah, M.; Mokraoui, A.; Azzag, H. Experience feedback using Representation Learning for Few-Shot Object Detection on Aerial Images. *arXiv* **2021**, arXiv:2109.13027.
79. Li, B.; Yang, B.; Liu, C.; Liu, F.; Ji, R.; Ye, Q. Beyond Max-Margin: Class Margin Equilibrium for Few-shot Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 7363–7372.
80. Yang, Z.; Wang, Y.; Chen, X.; Liu, J.; Qiao, Y. Context-transformer: Tackling object confusion for few-shot detection. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12653–12660.
81. Zhu, C.; Chen, F.; Ahmed, U.; Shen, Z.; Savvides, M. Semantic relation reasoning for shot-stable few-shot object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8782–8791.
82. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*; Morgan Kaufmann Publishers Inc.:San Francisco, CA, USA, 2013 ; Volume 26.
83. Xiao, Z.; Zhong, P.; Quan, Y.; Yin, X.; Xue, W. Few-shot object detection with feature attention highlight module in remote sensing images. In *International Society for Optics and Photonics, Proceedings of the 2020 International Conference on Image, Video Processing and*

Artificial Intelligence, Shanghai, China, 21–23 August 2020; International Society for Optics and Photonics: Bellingham, WA, USA, 2020 ; Volume 11584, p. 115840Z.
84. Cheng, G.; Yan, B.; Shi, P.; Li, K.; Yao, X.; Guo, L.; Han, J. Prototype-cnn for few-shot object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5604610. [CrossRef]
85. Chen, H.; Wang, Y.; Wang, G.; Qiao, Y. Lstd: A low-shot transfer detector for object detection. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
86. Wang, X.; Huang, T.E.; Darrell, T.; Gonzalez, J.E.; Yu, F. Frustratingly simple few-shot object detection. *arXiv* **2020**, arXiv:2003.06957.
87. Zhang, L.; Chen, K.J.; Zhou, X. HOSENet: Higher-Order Semantic Enhancement for Few-Shot Object Detection. In Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision (PRCV), Seattle, WA, USA, 13–19 June 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 175–186.
88. Sun, B.; Li, B.; Cai, S.; Yuan, Y.; Zhang, C. FSCE: Few-shot object detection via contrastive proposal encoding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 7352–7362.
89. Fan, Z.; Ma, Y.; Li, Z.; Sun, J. Generalized Few-Shot Object Detection without Forgetting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4527–4536.
90. Wang, Y.X.; Ramanan, D.; Hebert, M. Meta-learning to detect rare objects. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 9925–9934.
91. Yan, X.; Chen, Z.; Xu, A.; Wang, X.; Liang, X.; Lin, L. Meta r-cnn: Towards general solver for instance-level low-shot learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 9577–9586.
92. Fan, Q.; Zhuo, W.; Tang, C.K.; Tai, Y.W. Few-shot object detection with attention-RPN and multi-relation detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 4013–4022.
93. Zhang, S.; Luo, D.; Wang, L.; Koniusz, P. Few-shot object detection by second-order pooling. In Proceedings of the Asian Conference on Computer Vision, Kyoto, Japan, 30 November–4 December 2020.
94. Liu, L.; Wang, B.; Kuang, Z.; Xue, J.H.; Chen, Y.; Yang, W.; Liao, Q.; Zhang, W. GenDet: Meta Learning to Generate Detectors From Few Shots. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 3448–3460. [CrossRef]
95. Hu, H.; Bai, S.; Li, A.; Cui, J.; Wang, L. Dense Relation Distillation with Context-aware Aggregation for Few-Shot Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 10185–10194.
96. Karlinsky, L.; Shtok, J.; Alfassy, A.; Lichtenstein, M.; Harary, S.; Schwartz, E.; Doveh, S.; Sattigeri, P.; Feris, R.; Bronstein, A.; et al. StarNet: Towards Weakly Supervised Few-Shot Object Detection. *arXiv* **2020**, arXiv:2003.06798.
97. Zhang, Z.; Hao, J.; Pan, C.; Ji, G. Oriented Feature Augmentation for Few-Shot Object Detection in Remote Sensing Images. In Proceedings of the 2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI), Fuzhou, China, 24–26 September 2021; pp. 359–366.
98. Zhao, Z.; Tang, P.; Zhao, L.; Zhang, Z. Few-Shot Object Detection of Remote Sensing Images via Two-Stage Fine-Tuning. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
99. Kang, B.; Liu, Z.; Wang, X.; Yu, F.; Feng, J.; Darrell, T. Few-shot object detection via feature reweighting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 8420–8429.
100. Perez-Rua, J.M.; Zhu, X.; Hospedales, T.M.; Xiang, T. Incremental few-shot object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13846–13855.
101. Li, F.; Yuan, J.; Feng, S.; Cai, X.; Gao, H. Center heatmap attention for few-shot object detection. In Proceedings of theInternational Symposium on Artificial Intelligence and Robotics 2021, Pretoria, South Africa, 10–11 November 2021; Springer: Berlin/Heidelberg, Germany, 2021; Volume 11884, pp. 230–241.
102. Zhang, G.; Luo, Z.; Cui, K.; Lu, S. Meta-detr: Few-shot object detection via unified image-level meta-learning. *arXiv* **2021**, arXiv:2103.11731.
103. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
104. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
105. Gao, Y.; Hou, R.; Gao, Q.; Hou, Y. A Fast and Accurate Few-Shot Detector for Objects with Fewer Pixels in Drone Image. *Electronics* **2021**, *10*, 783. [CrossRef]
106. Liu, Y.; Li, Q.; Yuan, Y.; Du, Q.; Wang, Q. ABNet: Adaptive Balanced Network for Multiscale Object Detection in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14. [CrossRef]
107. Chen, T.I.; Liu, Y.C.; Su, H.T.; Chang, Y.C.; Lin, Y.H.; Yeh, J.F.; Chen, W.C.; Hsu, W. Dual-awareness attention for few-shot object detection. *IEEE Trans. Multimed. arXiv* **2021**, arXiv:2102.12152.
108. Xiao, Z.; Liu, Q.; Tang, G.; Zhai, X. Elliptic Fourier transformation-based histograms of oriented gradients for rotationally invariant object detection in remote-sensing images. *Int. J. Remote Sens.* **2015**, *36*, 618–644. [CrossRef]
109. Shi, Z. Chapter 14—Brain-like intelligence. In *Intelligence Science*; Shi, Z., Ed.; Elsevier: Amsterdam, The Netherlands, 2021; pp. 537–593. [CrossRef]

110. Girshick, R.; Radosavovic, I.; Gkioxari, G.; Dollár, P.; He, K. Detectron. 2018. Available online: https://github.com/facebookresearch/detectron (accessed on 23 August 2022).
111. Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. MMDetection: Open MMLab Detection Toolbox and Benchmark. *arXiv* **2019**, arXiv:1906.07155.
112. Long, Y.; Gong, Y.; Xiao, Z.; Liu, Q. Accurate object localization in remote sensing images based on convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2486–2498. [CrossRef]
113. Waqas Zamir, S.; Arora, A.; Gupta, A.; Khan, S.; Sun, G.; Shahbaz Khan, F.; Zhu, F.; Shao, L.; Xia, G.S.; Bai, X. isaid: A large-scale dataset for instance segmentation in aerial images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 15–20 June 2019; pp. 28–37.
114. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
115. Everingham, M.; Eslami, S.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.* **2015**, *111*, 98–136. [CrossRef]
116. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
117. Zhang, X.; Zhang, H.; Jiang, Z. Few shot object detection in remote sensing images. In *International Society for Optics and Photonics, Proceedings of the Image and Signal Processing for Remote Sensing XXVII, Online, 13–17 September 2021*; Bruzzone, L., Bovolo, F., Eds.; SPIE: 2021; Volume 11862, pp. 76–81. [CrossRef]
118. Allen-Zhu, Z.; Li, Y. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv* **2020**, arXiv:2012.09816.
119. Li, Y.; Cheng, Y.; Liu, L.; Tian, S.; Zhu, H.; Xiang, C.; Vadakkepat, P.; Teo, C.; Lee, T. Low-shot Object Detection via Classification Refinement. *arXiv* **2020**, arXiv:2005.02641.
120. Zhang, W.; Wang, Y.X. Hallucination improves few-shot object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13008–13017.
121. Qiao, L.; Zhao, Y.; Li, Z.; Qiu, X.; Wu, J.; Zhang, C. Defrcn: Decoupled faster r-cnn for few-shot object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 20–25 June 2021; pp. 8681–8690.
122. Xiao, Y.; Marlet, R. Few-shot object detection and viewpoint estimation for objects in the wild. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 192–210.
123. Li, A.; Li, Z. Transformation invariant few-shot object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 3094–3102.