remote sensing

MDPI

Review

Deep Learning-Based Object Detection Techniques for Remote Sensing Images: A Survey

Zheng Li [1,2], Yongcheng Wang [1,*], Ning Zhang [1,2], Yuxi Zhang [1,2], Zhikang Zhao [1,2], Dongdong Xu [1], Guangli Ben [1,2] and Yunxiao Gao [1,2]

[1] Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China; lizheng20@mails.ucas.ac.cn (Z.L.); zhangning171@mails.ucas.ac.cn (N.Z.); zhangyuxi18@mails.ucas.ac.cn (Y.Z.); zhaozhikang20@mails.ucas.ac.cn (Z.Z.); xudongdong@ciomp.ac.cn (D.X.); benguangli@ciomp.ac.cn (G.B.); gaoyunxiao19@mails.ucas.ac.cn (Y.G.)

[2] University of Chinese Academy of Sciences, Beijing 100049, China

* Correspondence: wangyc@ciomp.ac.cn

Abstract: Object detection in remote sensing images (RSIs) requires the locating and classifying of objects of interest, which is a hot topic in RSI analysis research. With the development of deep learning (DL) technology, which has accelerated in recent years, numerous intelligent and efficient detection algorithms have been proposed. Meanwhile, the performance of remote sensing imaging hardware has also evolved significantly. The detection technology used with high-resolution RSIs has been pushed to unprecedented heights, making important contributions in practical applications such as urban detection, building planning, and disaster prediction. However, although some scholars have authored reviews on DL-based object detection systems, the leading DL-based object detection improvement strategies have never been summarized in detail. In this paper, we first briefly review the recent history of remote sensing object detection (RSOD) techniques, including traditional methods as well as DL-based methods. Then, we systematically summarize the procedures used in DL-based detection algorithms. Most importantly, starting from the problems of complex object features, complex background information, tedious sample annotation that will be faced by high-resolution RSI object detection, we introduce a taxonomy based on various detection methods, which focuses on summarizing and classifying the existing attention mechanisms, multi-scale feature fusion, super-resolution and other major improvement strategies. We also introduce recognized open-source remote sensing detection benchmarks and evaluation metrics. Finally, based on the current state of the technology, we conclude by discussing the challenges and potential trends in the field of RSOD in order to provide a reference for researchers who have just entered the field.

Keywords: object detection; deep learning; remote sensing; neural network; weakly supervised learning

check for updates

Citation: Li, Z.; Wang, Y.; Zhang, N.; Zhang, Y.; Zhao, Z.; Xu, D.; Ben, G.; Gao, Y. Deep Learning-Based Object Detection Techniques for Remote Sensing Images: A Survey. *Remote Sens.* **2022**, *14*, 2385. https://doi.org/10.3390/rs14102385

Academic Editors: M. Jamal Deen, Subhas Mukhopadhyay, Yangquan Chen, Simone Morais, Nunzio Cennamo and Junseop Lee

Received: 7 April 2022
Accepted: 11 May 2022
Published: 16 May 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In recent years, considerable effort has been devoted to overcoming the challenge of object detection in computer vision. Unlike image classification, object detection [1] inherited from classification tasks not only needs to identify the category to which an object of interest belongs, but also to locate the position of the object using a bounding box (BBox), which makes the task more difficult and increases the requirements of the algorithm [2].

Large quantities of remote sensing data have been obtained from imaging optical sensors on artificial Earth satellites and aerial platforms; such approaches have the advantages of being realistic and obtainable in real time. According to different imaging spectral ranges, the data can be classified as visible, infrared, ultraviolet, multispectral, hyperspectral, or SAR images [3,4]. These images make different contributions to the Earth Observation System, promoting our understanding of the environment and facilitating people's activities. Recently, thanks to the rapid development of remote sensing platforms and sensors, the fact that the quantity and quality of remote sensing data are improving

has raised a new problem, i.e., how to effectively use existing data and maximize their application value. Thus, object detection using RSIs, as a basic image analysis application, is receiving more and more attention from researchers [5].

DL, which originated from research on artificial neural networks, can extract beneficial information by stacking extremely deep network layers autonomously, thereby imitating the learning mechanisms of the human brain. DL has undergone rapid development, attracting a great deal attention due to its powerful data mining and analysis capabilities. Compared with DL-based image classification [6,7], DL-based object detection was proposed later but developed faster. Excellent object detection algorithms have emerged one after another and have gradually extended to the remote sensing domain.

Due to the remote imaging of the Earth's surface, RSIs have the characteristics of large size, high viewpoint, and low spatial resolution compared with images acquired by ground cameras. The major challenges for RSOD are as follows:

1. **Complex object characteristics:** First, the wide coverage of RSIs leads to the frequent appearance of objects with large-scale variations, such as the coexistence of ships and harbors in the scenarios. A top-down imaging view often causes objects to present a disorderly directional arrangement, as shown in Figure 1a. Therefore, the detection model not only has to be sensitive to the scale but must also be perceptive in terms of orientation [8,9]. Second, the object size of some species may be small or even occupy only a few pixels, as illustrated in Figure 1b. Such objects make up only a very small part of the whole image and make extracting features from fewer pixels more arduous [10]. Third, a high degree of similarity may occur among objects in RSIs that are intensely similar [11], such as tennis courts and baseball fields, or roads and bridges, as pictured in Figure 1c. The extracted similar features may confuse the detector, resulting in incorrect judgments. Finally, RSIs may contain special categories such as mountain roads and cross-sea bridges with extreme aspect ratios, such as in Figure 1d; the slender appearance of such objects makes it challenging for the detector to identify features accurately [12].
2. **Complex image background:** A major characteristic of RSIs is that the background will occupy the majority of the scene. On the one hand, the extensive background may overwhelm the object regions, causing the detector to fail to outline the object effectively. On the other hand, the scene in which the image was taken can be relatively cluttered and noisy, which can affect the detector's ability to efficiently extract features and correctly locate objects [13]. Therefore, searching and positioning objects from highly complex scenes such as the one shown in Figure 1e turns out to be quite demanding.
3. **Complex instance annotation:** DL-based models rely heavily on accurately labeled training data. In general, a rich and high-quality dataset is more likely to provide relatively satisfactory results in terms of training. Accurately annotating RSIs that often present small and densely distributed objects is a time-consuming, labor-intensive chore, and inaccurate labeling degrades the performance of the model [14]. Therefore, complex sample annotations also inadvertently increase the complexity of detection implementation.

The above-mentioned possible complications hinder the development of the technical path of RSOD. Urban monitoring, building planning, port management, disaster prediction, post-disaster reconstruction, and various livelihood and military applications all require accurate and efficient object detection technology. However, most existing reviews of RSOD focus on the introduction of general-purpose detection algorithms and do not meticulously summarize or review the significant strategies proposed to address the adaptation problems arising from their application in the remote sensing field. For example, Ref. [15] was published too early to capture the current trends. In [16], the authors focused on the introduction of algorithms in the field of computer vision and proposed the DIOR data set. The authors of [17] mainly researched aircraft detection algorithms with regard to remote sensing. The development of detection algorithms based on DL was also described

in [18]. Unlike these references, in this paper, we mainly summarize the current DL-based improvement strategies concerning the characteristics of RSIs. At the same time, we classify various types of improvement algorithms according to their characteristics, which include the attention mechanism strategy, multiscale feature fusion strategy, mining contextual information strategy, refined anchor strategy, direction prediction strategy, super-resolution reconstruction detection strategy, transformer-based strategy, semi-supervised learning detection strategy, weakly supervised learning detection strategy, and others. We also divide these strategies into subcategories in order to form a complete classification system.

(**a**) (**b**) (**c**) (**d**) (**e**)

Figure 1. Main challenges in detecting remote sensing objects. (**a**) Chaotic direction of remote sensing objects. (**b**) Small objects. (**c**) High degree of similarity. (**d**) Extreme aspect ratio. (**e**) Objects are surrounded by complex background.

The main purpose of this review is to extract the core knowledge of DL-based object detection by collating recent studies, thereby helping researchers understand RSOD more thoroughly. At the same time, we provide some research suggestions by analyzing current state of the art technologies. The main contributions of this paper are as follows:

- We provide a comprehensive review of RSI object detection techniques based on DL, including representative methods, implemented processes, benchmark datasets, performance metrics, performance comparisons, etc.
- We systematically summarize the improved strategies proposed in recent years to address the complex challenges facing remote sensing, and classify them into a taxonomy in a hierarchical manner according to their characteristics.
- We discuss existing issues and provide a reference for potential future research directions.

A structure diagram of this paper is shown in Figure 2. In this paper, we describe traditional detection methods such as template matching, priori information, and machine learning methods, as well as presenting DL-based detection methods such as one- and two-stage families. The DL-based detection process is organized into five steps: data pre-processing, feature extraction and processing, BBox generation, detection, and post-processing. We exhaustively introduce improved strategies designed for the remote sensing domain, which include the attention mechanism, multi-scale feature fusion, mining contextual information, the refined anchor mechanism, direction prediction strategies, super-resolution strategies, Transformer-based methods, semi-supervised learning, weakly supervised learning, among other methods, and construct a taxonomy. We summarize the foundation of performance evaluations—benchmark datasets and performance metrics—and compare the performance of multiple models.

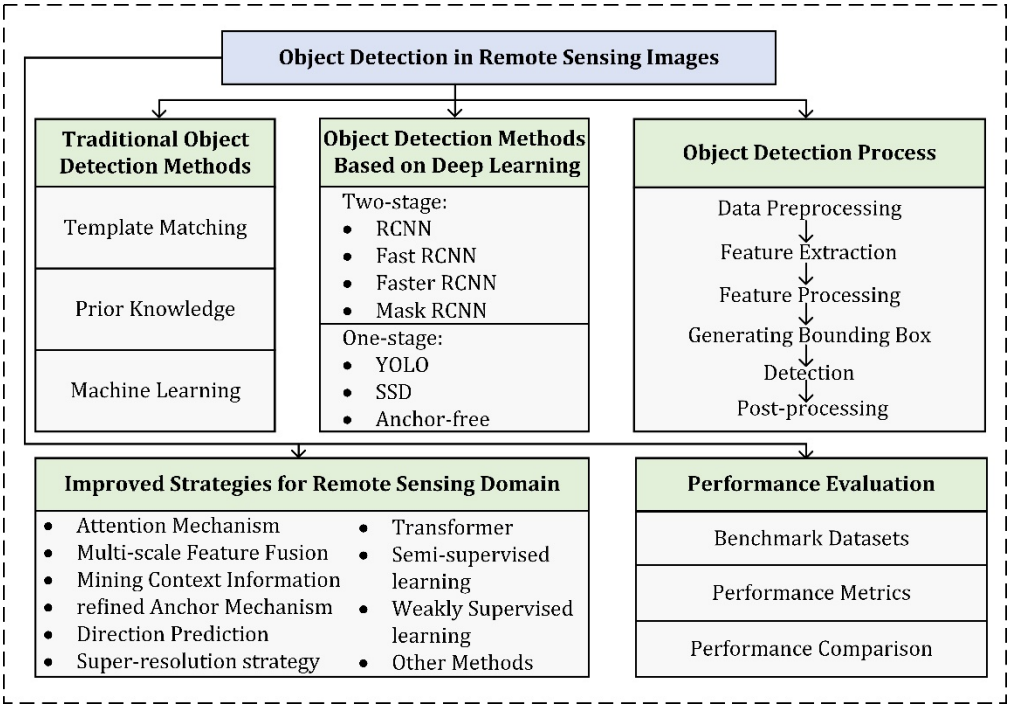

Figure 2. Structure of the article.

The rest of this paper is organized as follows. Section 2 briefly reviews representative methods and presents the implementation process of object detection algorithms. We also exhaustively review improvement strategies for RSOD. In Section 3, performance evaluations are carried out. Section 4 discusses potential problems and predicts promising directions for the future. Section 5 presents our conclusions.

2. Methods

2.1. Review of Object Detection Algorithms

2.1.1. Traditional Remote Sensing Object Detection Methods

Template matching [19,20] is the original RSOD method. The process includes template generation and similarity calculation. The first step creates a template, i.e., a patch containing only objects, which is handcrafted to detect objects in RSIs. The similarity calculation slide searches potential regions and calculates the similarity between the region and the template in order to locate the objects. This method has the benefits of simplicity (in principle) and concise processing, but is not intelligent, since it requires hand-crafted templates, and sliding has high computational complexity.

Prior knowledge [21,22] uses both geometry knowledge and context knowledge to search for objects. Geometry knowledge utilizes the appearance information of objects to design models. Context knowledge creates a special spatial constraint between the objects and the background and transforms implicit knowledge of objects into explicit detection rules to search for satisfactory objects according to a set of rules. The key issue with this approach is the accuracy of the priori knowledge, which depends more on human subjectivity; as such, excessive human intervention may lead to unstable results.

Machine learning [23] was the dominant approach among researchers until the advent of DL techniques. This traditional machine learning approach treats object detection as a classification problem, where the model will first search for possible object regions in an image and extract the histogram of gradients (HOG) features [24], bag of words (BoW) features [25], texture features, contextual features, and other information in potential regions. It then uses an independent classifier to discriminate among object categories to determine whether the sub-region contains objects. Subject to the drawbacks that the feature extractor and classifier cannot be trained in an end-to-end manner, high computational

overhead, and an inaccurate positioning function, the machine learning approach has been gradually replaced by DL and its use is being phased out.

2.1.2. Object Detection Methods Based on Deep Learning

At the ImageNet competition in 2012, the convolutional neural network (CNN)-based AlexNet, designed by Krizhevsky et al. [26], won first prize by a huge margin over machine learning-based SVM, signaling the arrival of the DL era, which has led to the evolution of various image analysis techniques, including object detection. Mainstream object detection algorithms can be roughly divided into two categories, i.e., one-stage and two-stage, with the main difference being whether they include a step for the proposed region. Two-stage algorithms add a step to the total process, which is equivalent to an additional screening process on top of the original one, and therefore has advantages in terms of accuracy. One-stage algorithms only perform a single detection; the absence of the proposal region generation step is advantageous in terms of speed. The RSOD mostly applies detection algorithms for natural images as the underlying framework and designs improvement strategies for the adaptation problems arising from the migration of source domains, which has gradually become a research hot spot in the field of image analysis and application.

As the name suggests, two-stage algorithms divide the implementation process into two steps. The first step generates a series of possible proposals, and the second refines the proposals to output the final results. The R-CNN family is a collection of algorithms which are representative of this technique. Girshick et al. [27] applied a CNN to detection tasks for the first time, giving rise to the creation of R-CNN. R-CNN generates nearly 2000 proposals by the Selective Search (SS) algorithm [28]. The CNN is then used to extract features. Finally, the SVM classifier and regressor are used to obtain the final detection results. To counter the disadvantage of R-CNN, i.e., the need to train the classifier separately, fast R-CNN [29] proposes a multi-task loss function and uses the Softmax classifier. Therefore, the network can perform in an end-to-end fashion. For proposals of different sizes, ROI Pooling was proposed to output feature maps into a fixed size, as required in the subsequent fully connected network. Faster R-CNN [30] abandoned the SS algorithm and designed the RPN subnet to obtain proposals from the anchor mechanism. The network directly inputs the whole image, which reduces computing time. Faster R-CNN is also frequently employed as the bottom network in RSOD. He et al. [31] proposed Mask R-CNN, which was the first system to integrate the detection task with the segmentation task. The mask branch was designed in parallel with the classification and detection branches, and the FPN [32] structure was introduced to enrich the features of shallow layers. At the same time, the network designed ROI Align, instead of ROI Pooling, to reduce the quantization error. Thus, the detection precision was improved.

One-stage algorithms abandon the time-consuming step of generating proposals and regard the detection task as a regress problem. Representative algorithms include the YOLO [33–36], the SSD [37–46], and the anchor-free families [47–50]. YOLOv3 [35] and SSD [37] have structural similarities, in that they both use multi-scale detection heads for objects of different sizes, and similar anchor mechanisms. They have also been more studied in RSOD. Liu et al. [44] proposed a new loss function named focal loss, which efficiently reduces the weight of easy samples and increases the weight of hard samples to make networks focus on them. Focal loss is also studied in remote sensing. The anchor-free approach discards the anchor mechanism completely, proposing instead a new method based on key point detection. CornerNet [47] proposed by Law et al. and CenterNet [49] proposed by Duan et al. generate BBoxes by detecting the corner points or center points of objects to further accelerate speed detection. These concepts have become a new research hotspot in remote sensing.

2.1.3. Summary

In this subsection, we will review traditional object detection methods and object detection algorithms based on DL technology. With the development of technology and

improvements in application field requirements, traditional methods that rely more on human intervention can no longer meet the current needs of intelligent technology. Instead, the combination of DL, as a new technology in artificial intelligence, and detection tasks has promoted the development of image location recognition tasks. This approach has rapidly grown to occupy the main position in the field of object detection. Two- and one-stage algorithms, as two benchmarks of object detection, are widely employed in remote sensing as the bottom framework. As shown in Figure 3, Faster R-CNN, which adopts the two-step regression process of rough and fine detection, has become the most widely available underlying framework. At the same time, as representatives of fast detectors, one-stage algorithm families such as YOLO and SSD pursue high detection efficiency at the cost of a little accuracy. As such, two- and one-stage algorithms have made remarkable contributions to RSOD.

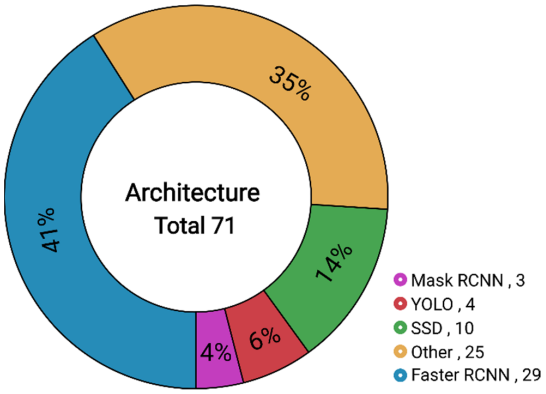

Figure 3. Statistical diagram of the underlying framework. The bottom right corner indicates the framework and the number of frameworks employed in the statistical paper, respectively.

2.2. Remote Sensing Object Detection Based on Deep Learning

The pipeline of RSOD can be broadly divided into five parts: (1) data pre-processing, (2) feature extraction and processing, (3) the generation of a BBox, (5) detection, and (5) post-processing. RSIs are first pre-processed to meet the input requirements of the detection network and feature extraction network in order to extract the rough features of the object, which are not conducive to the final detection and need to be enhanced via certain improvement strategies. On the basis of the features which have been processed to generate the BBox in order to outline the object, the subsequent head structures make predictions based on information regarding the features and box. Final post-processing filters out the useless detection information, yielding the final results. The overall pipeline is shown in Figure 4.

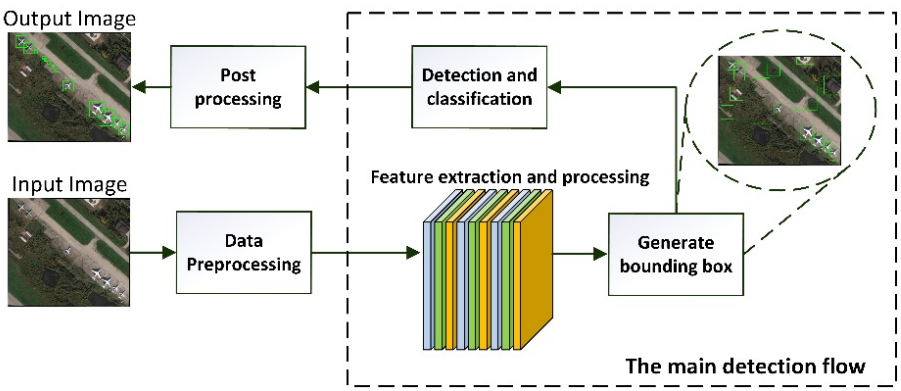

Figure 4. Detection flowchart of DL-based object detection network. The green box represents the BBox for locating the object.

2.2.1. Data Preprocessing

Data Augmentation: DL models typically have millions of parameters and require massive quantities of data to train. However, in practice, data are often insufficient or unbalanced, in which case the model will not be sufficiently trained, leading to poor generalization ability. Data augmentation [51–54], as a means of sample amplification using software generation without increasing the cost of data acquisition, has broad applications. The collection budget of remote sensing data, being more costly than other data, makes the sample expansion in remote sensing more reliant on data augmentation techniques. Currently, common data augmentation techniques include geometric transformation, color transformation, fuzzy transformation, etc.

Geometric transformation [51,53] comprises translation, rotation, flipping, scaling, etc. Translation transformation is undertaken to move the image in a particular direction by a certain distance, and is one of the simplest geometric transformation methods. The object appears in various locations in the image by simply moving it around, thereby increasing the amount of data while also enriching the object diversity. Rotation transformation describes the rotation of the object to a certain angle. Flipping, also known as mirror imaging, involves flipping the object along a symmetry axis (x-axis or y-axis). Due to a lack of training data, the detection network cannot adequately learn the orientation changes of the rotating objects in the training samples, resulting in the model being insensitive to the orientation. The rotation method can enrich the orientation information of the rotating objects so that the model can learn the object features of various orientations, effectively improving the generalization performance of the rotating objects. Scaling transformation involves enlarging or reducing the image size according to the scale factor, which effectively increases the scale diversity of the object. As described in the introduction, the scale of remote sensing objects varies widely, and there are large differences in the scales of the same class of objects. The scaling method can increase the scale diversity of an object so that the detector can effectively learn the object features at each scale and fully solve the limitation of the scale variation phenomenon of remote sensing objects.

Color transformation [54], usually referred to as HSV (hue, saturation, luminance) transformation, refers to the process of changing the image colors [55–57]. During the imaging of remote sensing sensors, the brightness of the acquired RSIs is different due to environmental factors such as lighting conditions, cloud cover, and atmospheric conditions. In these respects, differences can easily interfere with the accuracy of detections. HSV transformation changes the brightness of the image by adjusting the hue, saturation, and brightness channels to make the detection model more robust to RSIs with different colors. This transformation includes linear and nonlinear transformations, where the former changes the overall brightness of the image by randomly perturbing the pixels in the hue, saturation, and brightness channels, while the latter commonly comprises gamma transform, which enhances the gray values in the darker regions of the image through nonlinear forms to change the overall brightness of the image. This nonlinear transformation enhances the details of the image by brightening dark areas and reducing the brightness of light areas.

Blurred transformation [58], also called smoothing transformation, is a means of making images blurred. The main effect is that it reduces the otherness between pixels to smooth images and reduces the noise and the level of detail in the images to alleviate the network's reliance on image quality. Gaussian blur is one means to achieve the smooth effect by the weighted average of pixels in the form of normal distribution.

In addition to these conventional data augmentation methods, Sharma et al. [59] designed a new geometric data augment strategy which segments images into two rectangular parts and exchanges the locations thereof to acquire new images after splicing. Experimental results show that the effect is better than that of standard geometry and color transformations. Wu et al. [60] designed a new cropping method to avoid acquiring a low-quality image containing a large amount of background. In their approach, a large sub-block is first cropped, and then the final sub-image is obtained by performing further

cropping in the ±45° direction within this sub-block. The authors of [61,62] stretched the histograms of images by a certain factor, which not only augments the images but also effectively suppresses the noise in RSIs.

Image Clipping: Image cropping is a strategy for slicing an image into a series of small patches of preset sizes. Owing to a special imaging mechanism, RSIs have a large image size, i.e., a high-resolution RSI usually has millions of pixels, but existing systems cannot effectively process such huge images due to limitations in computing capacity. Image cropping involves sliding the original complete RSIs into a series of sub-blocks whose sizes need to meet the requirements of network input [63]. Then, the sub-blocks are merged and stitched back into a complete image to obtain the final detection result [56], which can effectively overcome the problem of the inefficient detection effects of large RSIs. The image is usually cut with a certain overlap area during processing to avoid the problem of the object at the segmentation boundary not being detected properly after it is split into parts. As such, choosing an optimal the overlap rate is key; a high overlap rate yields a relatively intact object but also creates more sub-blocks. Conversely, a low overlap rate yields fewer sub-blocks for faster detection, but the object may be incomplete.

2.2.2. Feature Extraction and Processing

Feature extraction is an indispensable step for various tasks including image recognition and object detection, and can be considered the foundation of DL technology. Clear features contribute to network forecasting and reduce the complexity of detection tasks. Current mainstream feature extraction networks include AlexNet [26], VGGNet [51], GoogleNet [64], ResNet [52], and DenseNet [65]. AlexNet [26] was proposed in 2012 and won the ImageNet competition that year. That network successfully brought DL into the limelight with its first design of a CNN as a deep network and use of GPUs to accelerate the process. Subsequently, VGGNet [51] attempted to create small processing kernels such as 3 × 3 small convolutional kernels and 2 ×2 pooling kernels to reduce the number of parameters. Instead of stacking deep layers, GoogleNet [64] improved performance by connecting several convolutional modules in parallel; this approach was the ImageNet champion in 2014. ResNet [52] is the most commonly used feature extraction network. In it, the residual module and skip connection structure effectively overcome the problem of gradient disappearance and gradient explosion, thereby successfully implementing an extremely deep network design. DenseNet [65] connects all layers by skip connection, which enables not only the transmission of information but also the full utilization thereof. In addition, the YOLOv2 [34] and YOLOv3 [35] models proposed for object detection tasks independently created a specific backbone network, DarkNet, which borrowed from the idea of residual networks. The utilization of the backbones is illustrated in Figure 5. The most widely used backbones are ResNet and VGGNet, as these can adequately extract the features of objects. DarkNet, designed by YOLO, has been extensively researched. Other networks such as ZFNet [66] and MobileNet [67] have also achieved satisfactory effects.

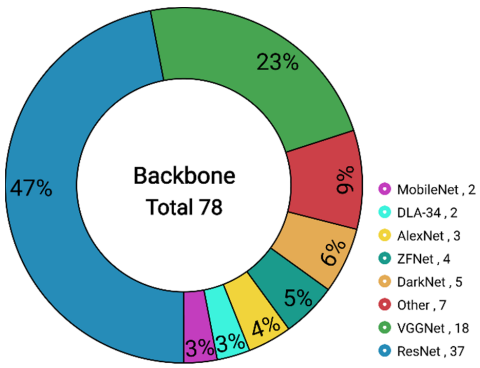

Figure 5. Statistical graph of feature extraction networks used in papers. The backbone and quantity are indicated in the lower right corner.

The complexity of RSIs may cause the features extracted from the backbone to be unfavorable in the final detection structure, affecting the subsequent results. To improve the accuracy, a variety of excellent feature processing strategies have been put forward, such as the attention mechanism, mining context information, multi-scale feature fusion, etc. These are discussed in Section 2.3.

2.2.3. Generating a Bounding Box

Detection tasks need to accurately locate the position of an object using a BBox, i.e., they must create high-quality BBoxes that match sufficiently well with the ground truth. Current methods for generating BBoxes include the traversal-based method, anchor-based method, and key point-based method.

Traversal Based Method: The sliding window traversal form using windows of different sizes sliding over the image by a fixed step to build a box was the first method used to generate the BBox; however, tiling such massive, redundant boxes significantly affects the model's efficiency. SS [28] is a selective traversal method which first generates a large number of candidate regions by the image segmentation method. It then calculates the degree of similarity between adjacent regions including the color, texture, size, and spatial overlap and merges the two candidates with the highest similarity. The above process is repeated until optimal BBoxes are generated. The SS algorithm increases the restrictions on the generated boxes so that most invalid and redundant boxes can be effectively eliminated, in contrast to the sliding window method. Edge Boxes [68] is a method to define BBoxes based on edge contours. The method first generates a map of possible edges of an object utilizing image processing. It then creates a segment of edge groups based on the points on the edge line in the map with a certain strategy, and calculates the similarity between groups and clusters to determine all the edges of an object and obtain boxes. Unlike SS, Edge Boxes can also provide an objective score for one box based on the number of enclosed contours, which is a more accurate method of generating BBoxes.

Anchor-Based Method: The anchor mechanism that appeared in Faster R-CNN [30] proposed by Ren et al. was proposed to further improve the quality of the BBox and to reduce the time required for that task. In contrast to the traversal method, this mechanism presets a series of anchor boxes of different sizes and aspect ratios on the final feature maps. Since this method operates on the feature map, no additional is required to create the boxes. Moreover, each point on the feature map sets anchor boxes and can cover almost all objects. The anchor mechanism has been widely applied in RSOD.

Key-points-Based Method: Anchor-free is another popular method. It utilizes key points in an innovative way to generate BBoxes. The main idea is to generate boxes by searching the key points of the object (corner points or center points). It then determines the BBox based on the predicted shape. CornerNet [47] is the cornerstone of key point detection. It determines the BBox by detecting the upper left and lower right corner points of objects in pairs, providing a new approach to object detection. To address the problem in CornerNet, i.e., when key points often fall outside the object, ExtremeNet [48] improved the network to generate BBox using four extreme points (top, bottom, left, right). At the same time, the center point was detected to combine these four points. Based on CornerNet, CenterNet [49] further restricted the generated boxes according to whether the region generated contains the center point. This model has also been widely applied in RSOD on account of its excellent performance. At present, the key points-based method and anchor-based method are developing in a mutually complementary manner.

This paper summarizes the methods of generating BBoxes described in 68 papers, as illustrated in Figure 6 and Table 1. It can be seen that the anchor method, as the most popular method at present, accounts for the majority of these papers. With this method, the hyperparameters need to be designed in advance. As such, inaccurate parameter design can produce poor network performance. The traversal method appears to be widespread in the early stages. However, due to the drawback of requiring inefficient input of BBoxes into the network multiple times for the same image, this method has been gradually phased out.

Nonetheless, the SS algorithm is often used for weakly supervised remote sensing detectors that do not have location labels. The key points-based method can eliminate the tediousness of creating boxes and generate BBoxes with suitable object scale and shape based directly on the key points by network prediction, thereby greatly improving the detection speed. It has therefore become an emerging research direction in RSOD. However, this method is more stringent in terms of the accuracy of key point predictions, and inaccurate predictions can easily cause missed detections and poor localization.

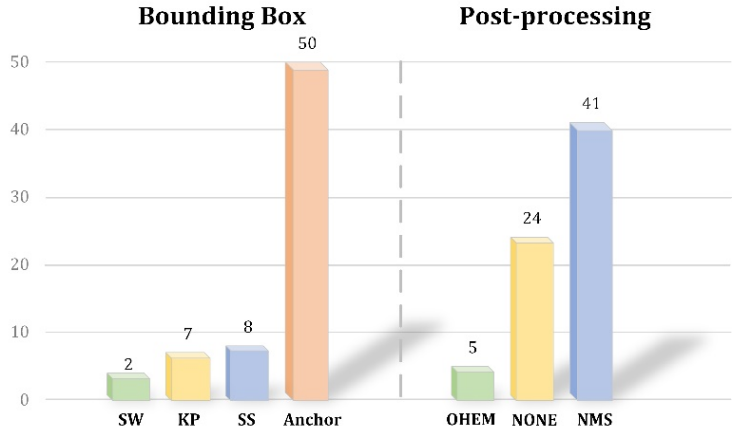

Figure 6. Statistical plot of the use of BBox and post-processing methods, where SW represents the sliding window method, KP represents the key point search method, and SS represents the selective search method.

Table 1. List of reviews of methods for generating BBoxes.

Methods		Characteristics	Representatives
Generate BBox	Based on Traversal method	• Produces many BBox • Low efficiency • Commonly used in weakly supervised detectors	Slide window Selective search Edge box
	Based on anchor	• Presets a series of BBox • High recall rate • Is heavily dependent on the super parameters settings	Anchor
	Based on key points	• Generates BBox from key points • No cumbersome BBox generation required • More dependent on the accuracy of key point detection	Anchor-free

2.2.4. Detection and Post-Processing

After obtaining the features and BBoxes, the network performs the final regression and classification tasks by predicting the head structure at the back end of the model. According to the difference between one- and two-stage algorithms, the network head detects the objects by different means. The two-stage process uses a fully connected network to link the detection head with the backbone and adopts the ROI Pooling layer [29] or ROI Align layer [31] to unify the feature map size to address the limitation of the fully connected network for the input features. One of the fully connected layers is used to predict the object category by the SoftMax layer, while the other outputs a four-dimensional vector corresponding to the center point coordinates of the BBox and the correction of the length and width information to more closely match the ground truth. On the other hand, the one-stage approach combines the classifier and regressor together and performs the convolution operation directly on the last layer of the feature map, outputting a multi-dimensional vector containing the prediction of the category and the correction of the position.

After obtaining the prediction result through the network head, post-processing, e.g., non-maximum suppression (NMS) [69] or online hard example mining (OHEM) [70], is performed to optimize the output.

NMS: The main role of the NMS algorithm [69] is to remove redundant detection results. All methods other than key points-based methods will flatten a large number of detection results in a single image in which the same object is being repeatedly detected. The purpose of this algorithm is to retain the best detection results for each object and to remove, as much as possible, redundant, low-quality, or background results. The method selects the box with the highest confidence score among all the detection results and calculates the intersection ratio of the remaining BBoxes to the detection results. Redundant results are eliminated by setting a fixed threshold.

OHEM: In general, the performance of the network in terms of detecting difficult objects reflects the power of the model, and the degree of difficulty encountered regarding object identification contributes to the network performance. OHEM [70] is designed to improve the ability to detect difficult objects by mining more complex results. Specifically, the method computes the detection difficulty of all outcomes reflected by the loss value; the larger the loss value, the greater the detection difficulty. Then, OHEM selects some of the difficult results as examples, and sets the loss of the remaining boxes to 0. Therefore, only the selected difficult examples are used for network training in the terminal phase. In RSIs, the background occupies most of the image, and a large number of background boxes are generated, which causes the network to focus on the background area. On the other hand, OHEM only selects a part of the background samples that contribute to the network for training, which effectively overcomes the problem of foreground background imbalance. Thus, OHEM improves the detection performance of RSIs.

As shown in Figure 6, most papers use NMS to deal with the results, but the threshold setting involves a trade-off, in that setting the threshold too high will not be effective in terms of removing redundant boxes, while setting it too low will remove useful boxes, resulting in a lower recall rate. Although OHEM can effectively improve the robustness of the network, this algorithm adds considerable computational burden, and as such, it has appeared in only a few papers. It is worth noting that post-processing is not necessary—some papers did not use post-processing and achieved suitable results.

2.2.5. Summary

In this subsection, we introduce the pipeline of RSOD based on DL, aiming to help readers to further understand object detection in RSIs. The pipeline of RSOD is the same as that of the natural domain, but it adds an image clipping step due to the larger size of RSIs compared with natural images. In addition, each step in RSOD is more detailed and stricter compared with those in the natural domain.

2.3. Improved Methods for Object Detection Based on Deep Learning

As mentioned above, because of the adaptation problem arising from source domain migration, the application of common object detection algorithms to RSIs does not produce satisfactory results; differences in appearance, shape, and features between remote sensing objects and natural domain objects are the main reasons for this. The features extracted from the source domain framework may contain useless and interfering information that is not suitable for the subsequent detection network. Additionally, not only feature extraction, but also the other detection parts of the general framework may not be appropriate for the detection environment of the remote sensing domain. This subsection introduces commonly applied improvement strategies and optimization methods related to these problems, as shown in Figure 7.

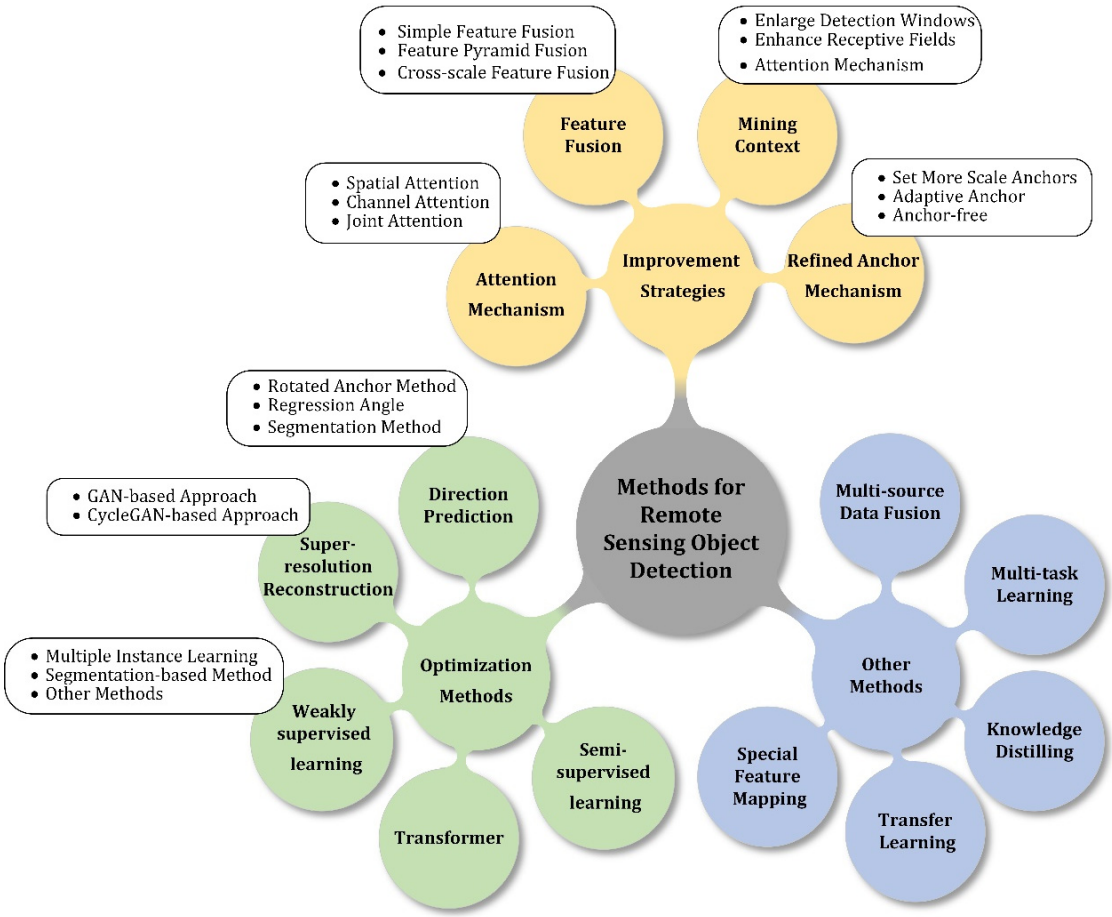

Figure 7. Prominent improvement strategies.

2.3.1. The Attention Mechanism-Based Method

RSIs have large and complex backgrounds in which objects are easily submerged, especially small objects, leading to low recall rates. Meanwhile, the extreme imbalance between the foreground and background in RSIs causes the network to pay more attention to the background, interfering with the operation of the detector. Drawing on the characteristics of visual perception, humans selectively focus on an object of interest and ignore most other information when observing complex scenarios. This phenomenon is known as the attention mechanism; it has been heavily researched and utilized in computer vision. The attention mechanism is regarded as a method of resource allocation, i.e., it redistributes initially evenly distributed resources according to the importance of the objects in the scenario [71]. In image analyses, the resources are assigned with different weights to emphasize diverse regions. Therefore, the attention mechanism is effective in terms of dealing with complex background problems in remote sensing.

This approach may be divided into the spatial attention mechanism, channel attention mechanism, and joint attention mechanism. The spatial attention mechanism captures pixel-to-pixel relationships at the image level to generate a mask map that emphasizes useful regions on the feature map by weight. This method differentiates between distinct image parts and focuses more on the object of interest to be detected. Currently, self-attention is frequently utilized to generate spatial attentional graphs; this process is shown in Figure 8a. Hua et al. [61] combined the attention feature maps generated by self-attention with a long short-term memory network to construct a deep feature pyramid. Wang et al. [72] embedded a self-attention module into the backbone to capture the correlation between the different regions and obtain discrimination features. Shi et al. [73] used the spatial attention mechanism to adaptively incorporate context information into feature maps to improve

recognition and localization accuracy. Chen et al. [74] proposed a cascade attention network (CA-CNN) composed of a patched self-attention module and a supervised spatial attention module to improve the feature representation of objects. Zhang et al. [75] designed an attention module for space and scale perception that directed the network's attention toward more informative features and suitable feature scales. Zhang et al. [76] built a center point detection network based on the spatial attention mechanism.

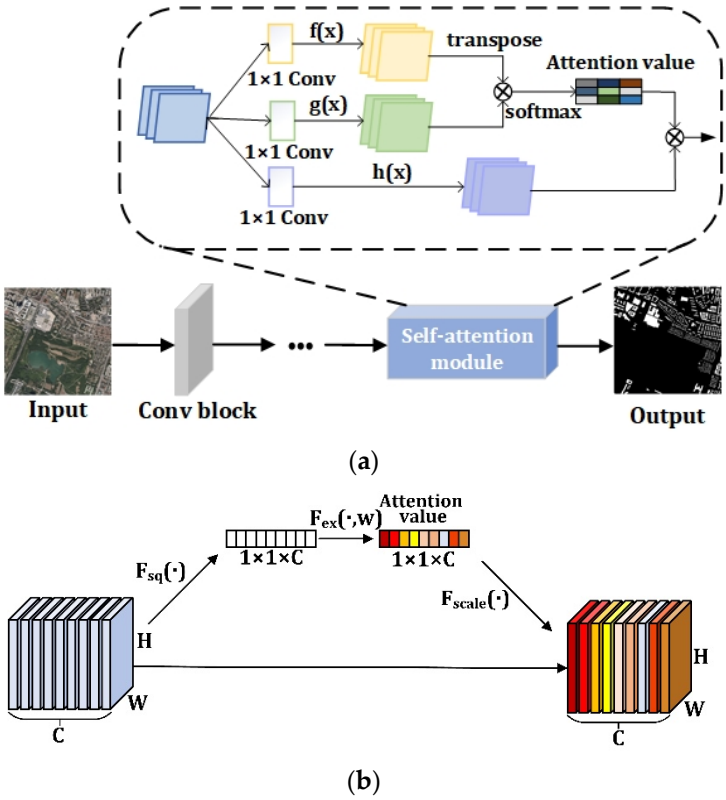

Figure 8. Schematic diagram of the attention mechanism. (**a**) The process of generating a spatial attention map by self-attention. (**b**) The process of generating a channel attention mechanism map by SENet. The colors of the attention value represent different weights.

The channel-level attention mechanism aims to obtain correlations between feature channels. This method processes information at the feature level to distinguish different channels and enhances the object feature channels. The method establishes dependencies between channels and strengthens the relationship between the features of objects. The most popular such mechanism is SENet [77] proposed by Hu, which obtains the global distribution of channel responses by a squeeze operation to grasp the relationship between features, and computes the weight of each feature channel by an excitation operation to emphasize the useful features and restrain the useless ones. SE-Net has pushed the channel attention mechanism into the spotlight, as shown in Figure 8b. Wu et al. [60] added the deformation convolution channel attention block (DCCAB), which highlights the features of objects and inhibits noise.

The joint attention mechanism employs both spatial and channel attention mechanisms, redistributing the originally uniformly assigned weights twice at the image and feature channel levels. This method not only determines the relationships among feature spatial locations but also captures the correlations of different features. Li et al. [78] modeled the spatial position dependence between global pixels to highlight object features and jointly explore a spatial attention network and a channel attention network to detect small objects surrounded by complex backgrounds. Chen et al. [79] used dilated convolution and global average pooling to obtain spatial and channel attention maps. Tian et al. [80] introduced

the attention mechanism to enhance the object features while reducing the influence of the background.

The attention mechanism adequately handles the problem of complex backgrounds in RSIs and alleviates the challenges of low contrast and lack of visual cues to a certain extent. The spatial attention mechanism aims to redistribute the process of reassigning weights to information on an image feature pixel level. The channel-level attention mechanism treats the feature map as a whole and instructs the network to devote more effort to the object features. The joint attention mechanism implements both operations at the same time. It is worth noting that the process of pixel-by-pixel computation increases the computational overhead slightly, while the network needs to learn the distribution of the image autonomously and assign weights both spatially and in terms of the channel, which increases the learning responsibility of the network.

2.3.2. The Multi-Scale Feature Fusion Based Method

Feature fusion combines various levels of features in the form of cascades or element sums to aggregate and enrich information. It has been widely accepted in RSOD. In order to extract effective features, the network is often designed with deep structures, and the resolution of the features at different levels leads to discrepancies in the amount of information contained and, in turn, expressed, especially for small objects. Meanwhile, as for the features of forward processing, the bottom features are generally extracted from the edges, gradients, and texture features of the object, which contain strong spatial location information and are more suitable for object localization. In contrast, the higher-level features are generally extracted with the discriminative part of the object, such as the wings and head of aircraft, which have strong semantic information and are more suitable for classification. Unlike the recognition task, which only requires features with strong semantic information to fully accomplish its goal, the detection task also has strict requirements for features with location information. The effective combination of features with different layers can adequately compensate for the lack of information in the single-layer map, and undoubtedly improve detections.

At present, general multi-scale feature fusion methods include simple feature fusion, feature pyramid fusion, and cross-scale feature fusion. Simple feature fusion combines the feature map of the top layer and adjacent layers, as shown in Figure 9a, fusing the multi-layer maps into the same size by a sampling strategy, and allowing features to consider adjacency information. After the layer-by-layer processing of the deep network, the information is gradually compressed, i.e., the amount of information contained in the features of the adjacent top layer becomes lower than that of the adjacent bottom layer. This strategy can compensate for the lack of information resulting from the feature transmission process. At the same time, since the features of adjacent layers have strong correlations and inheritance, fusing the features does not trigger the problem of network confusion caused by large differences in terms of information, and only a small number of additional parameters and calculations are required. [58,62,81–83].

Feature pyramid fusion is a popular fusion method in RSIs. Its name refers to its top-down fusion order, which evokes the shape of a pyramid. The method adopts layer-by-layer processing, which gradually passes down the information expressed in the upper layer, thus making each layer feature rich in semantic information. As shown in Figure 9b, the method adopts a lateral connection module to combine the information transmitted from the upper layer after scale amplification with the adjacent shallow information. This intermediate information is then to transmitted downward layer by layer until the bottom layer receives the fused information feedback, thereby completing the whole pyramid flow form of the transmission pipeline. The pyramidal fusion is perfectly suitable for multi-scale predictions in RSOD, since the feature information that needs to be processed and recognized by the multi-scale detection head is enhanced. The underlying features that contain all the information from the upper layer are the biggest beneficiaries, which also alleviates the small object problem. However, this fusion method complicates the structure

and introduces more complex computation than the first one; at the same time, it blindly fuses the features in each layer, some of which may contain noise and useless information, which is a drawback associated with the technique [12,73,75,84–86].

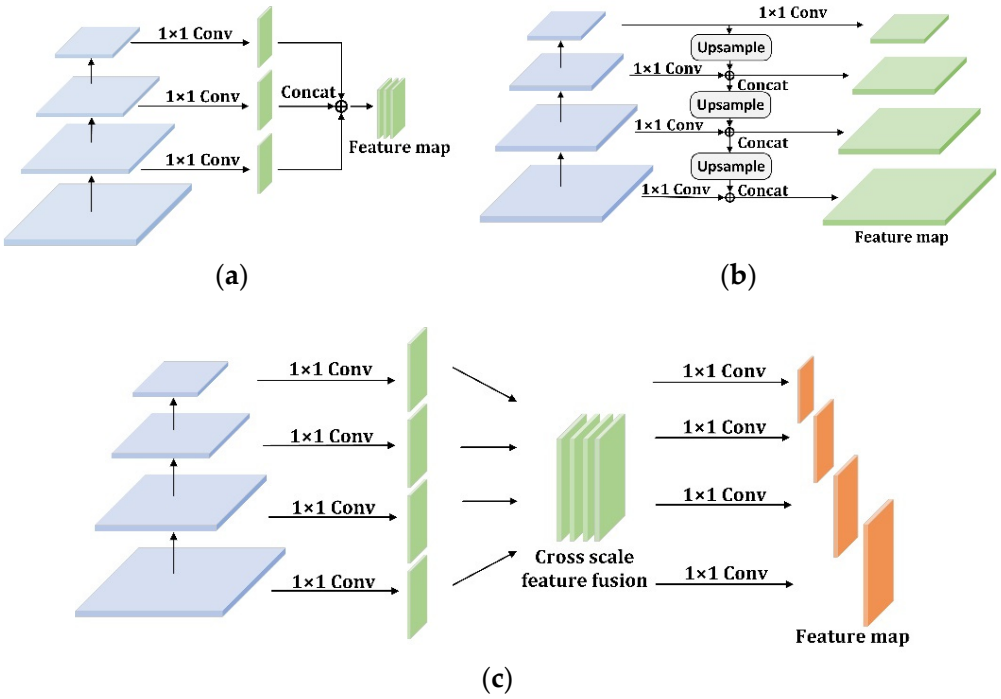

Figure 9. Schematic diagram of three multi-scale feature fusion methods. (**a**) Simple multi-layer feature fusion method. (**b**) Feature pyramid fusion method. (**c**) Cross scale feature fusion method.

Unlike the feature pyramid fusion method, the cross-scale feature fusion method fuses the features of all layers, as shown in Figure 9c. The fused features fully collect the information at all scales, which contains a large amount of global and local knowledge that is more beneficial for detecting objects at different scales [11,74,87,88]. The subsequent information separation process up- and down-samples the features to the original resolution of each layer in order to match the requirements of the subsequent detection head. However, this cross-scale fusion approach requires a larger magnification of the bottom and top layer features that deviate from the center, resulting in information compression and blurring problems that can severely compromise the original feature and cause serious information loss. Moreover, like pyramidal fusion, non-differentiated fusion imports useless information to each layer, which may negatively affect the performance of the network.

In addition to these three fusion methods, additional methods exist, such as those described in [89,90]. In general, multi-scale feature fusion has become an essential processing method for RSOD; it contributes greatly to the efficacy of the information represented by the features. The first approach of fusing neighboring layers only enriches the features at a single scale and is more suitable for single-scale detection; it does not make full use of the rich location information of the underlying layers, and single-scale detection cannot adequately adapt to remote sensing objects with complex scale variations. The latter two fusion approaches consider multiple layers of information and are more suitable for multi-scale predictions. Pyramidal fusion aims to pass down the rich semantic information from the top layer; however, the top layer does not receive any signal. Multiscale cross-fusion fuses all layer features in an undifferentiated way, aiming to enrich the information in each layer, but also increasing the computational complexity.

2.3.3. The Mining Context Information-Based Method

Contextual information generally refers to semantic information that contains strong symbiotic coupling between the object and the surrounding environment. With the help of such information, an object whose category originally could not be clearly distinguished according to its own characteristics can be effectively identified based on clues provided by the environment. For example, it is difficult to distinguish between a bridge and a road without using surrounding environment information such as the presence of lakes and towns, which are related to the object class and can reduce the ambiguity of categories. Thus, mining useful contextual information is valuable for distinguishing among similar objects. Contextual information can be divided into two categories: local and global. Local contextual information describes the correlation between an object and neighboring objects, or part of the environment in terms of color, texture, spatial distribution, and semantic representation. This can provide key semantic cues in weak feature responses embodied by pixels inside the object and enhance information representation from external factors. Global context information is closely related to the scene, which reflects the association between the object and all areas in the image. The vast scenes of RSIs contain a large number of spatial and semantic relationships. Spatial relationships can effectively assist in object localization, while semantic relationships reflect the strong symbiotic correlations between the scene and the internal object, which can be regarded as a type of a priori information that is useful for object identification.

At present, common methods for extracting contextual information can be roughly divided into three categories: expanding the BBox, enlarging the receptive field, and the attention mechanism. Expanding the window of the BBox is a simple method to increase the amount of contextual information. The detection network can obtain environmental information from outside the object by expanding the size of windows while extracting the object features. This environmental information as a supplement to the local contextual information can alleviate the problems of blurred appearance and poor structural information, effectively improving the network's ability to recognize objects. Li et al. [91] designed a network that simultaneously extracts features of $1\times$ and $1.5\times$ window sizes and fuses the features with a restricted Boltzmann machine (RBM). Gong et al. [92] designed a context mining layer in the network to adaptively generate context regions of appropriate size. Liu et al. [93] built a module to integrate global and local features, which produces horizontal minimum boundary rectangles of rotating boxes, thereby enlarging them.

The enlarging receptive field method is similar in principle to the first method, as it also makes the feature map cover a larger area of representations and achieve the purpose of learning contextual information. The receptive field is the result of the size of a region of information on the original input image, and how this is expressed by each pixel on the feature map. The scale of a feature's receptive field is based on how large that feature appears in a region of information in the original image—in other words, the size of the region from which the convolution kernel can extract information from the original image. Although high-level features contain less information, the receptive field of those features is much higher than that of the low-level features after multiple convolution operations. These high-level features provide strong semantic information generated by large receptive fields, and are more suitable for category identification. One effective way to increase the receptive field is to increase the size of the convolution kernel, although this has the shortcoming of introducing too many parameters into the computation and thus affecting the efficiency. The hole convolution module, specially designed to increase the receptive field, adds holes to the regular square kernel, allowing it to be scaled up without increasing the number of parameters so that more information can be extracted. Wang et al. [94] designed a basic receptive field module to extract context information by integrating feature maps obtained from three parallel dilated convolutions with different dilated rates. Liu et al. [86] proposed the receptive field module, which consists of three parallel branches, i.e., a convolution layer of different sizes, an asymmetric convolution layer, and a dilated convolution layer with different dilated rates. Wang et al. [95] designed a receptive field module with five branches,

in which four layers carry out dilated convolution to extract local context information while the other branches perform global pooling to extract global context information in order to obtain the discriminative features. Han et al. [58] inserted three serial dilated convolution layers into the residual module to extract context information. Yuan et al. [87] designed four parallel dilated convolution branches to increase the receptive field and help the network generate higher-resolution feature maps for local context information.

The attention mechanism calculates the correlation between neighboring pixel pairs in the image and highlights the dependency between the object and the scene. This can be supplemented as additional contextual information to enrich the features, and also guides the network to emphasize useful contextual information around the object. Thus, the attention mechanism is an effective strategy for contextual information mining. Shi et al. [73] used self-attention to generate a spatial attention map and adaptively included global context information in the feature map to improve object recognition accuracy. Li et al. [71] designed a cross-layer attention module to compute the global attention mapping and share that mapping with all locations. Zhang et al. [75] constructed a spatial scale-level attention module to instruct the network to focus on the regions with more context information at the correct scale.

Table A1 summarizes the characteristics of the three methods to enhance contextual information. The method of expanding the BBox is the most direct, but the window size should be set carefully; if it is too small, it cannot effectively to extract the contextual information, and if it is too large, noise will be introduced. Increasing the receptive field starts from the convolution operation, which increases the range of information areas that can be extracted by convolution to mine contextual information from around the object. However, there is a hole between the hole kernels, which may cause the loss of information due to the discontinuity between the features. Thus, it is necessary to add multiple branches to the hole convolution layer at the same time in order to solve this problem, which obviously expands the complexity of the network. Attention mechanisms focus more on connections between the objects and scenes, emphasizing useful contextual information and suppressing background noise.

2.3.4. The Refined Anchor Mechanism Based Method

The anchor mechanism is a strategy of pre-defined BBoxes designed to ensure detection recall. Its initial application object is the natural object, but there are large differences in terms of the scale, appearance, and orientation between remote sensing objects and natural objects. The direct application of the original anchor mechanism by remote sensing detectors generates maladaptive problems arising from source domain migration leading to performance degradation. Anchor improvement strategies have been designed to deal with complex objects and solve the limitation of detector performance resulting from the presence of unsuitable anchors in the original settings.

In general, the anchor mechanism sets multiple discrete-scale base boxes at each pixel point of the image to ensure that the recall closely matches the ground truth. The authors of [96] pointed out that the anchor can only regress the box in a limited range, and the object beyond the bound regression is easily ignored by networks. As such, the anchor scale discontinuity problem needs to be addressed. Presetting more scale anchors at each location to roughly cover complex scale variations of remote sensing objects can effectively alleviate scale variability issues, and is an effortless operation. Dong et al. [97] calculated the scale range of all kinds of objects in datasets and designed more suitable prior scale parameters. However, this method only has a favorable effect on the statistical dataset, and the data must be re-counted when the dataset is replaced. Han et al. [58] chose to generate anchor boxes with all scales on the feature map of each level; however, this method completely abandoned the idea of "divide and rule", i.e., that large, medium, and small objects need to be detected simultaneously on a single scale, which may make the detector insensitive to scale. Wang et al. [96] proposed a full-scale detection network with a scale-invariant regression layer that contains 14 detection heads in the regression layer,

allowing the discrete BBox to cover full-scale objects in the regression process. However, setting up so many heads inevitably increases the time cost.

Setting massive anchor boxes on each image pixel point has become a commonly accepted practice. However, this undifferentiated tiling strategy requires significant computing resources and does not achieve suitable results for special remote sensing objects. For this reason, various adaptive anchor strategies have been proposed. Yu et al. [98] designed a sub-network for an orientation guided anchor mechanism in which the anchor is generated only at the location where the object may exist, and the network only generates a few high-quality anchor boxes. Hou et al. [12] found that all categories of objects have their own aspect ratios, which can be regarded as prior information. The network generates BBoxes by setting base rectangle boxes with aspect ratio information without generating a large number of redundant anchors. Tian et al. [80] used the attention mechanism to highlight the object area and generate anchors accordingly. This intelligent method eliminates the need to set super-parameters and better matches the scene and object.

The setting of the anchor hyperparameters including scale, aspect ratio, and orientation requires a general comprehension of the distribution of data; however, different datasets have varying data characteristics. For example, there are almost no objects with an area less than 900 in the NWPU VHR-10 dataset, while in the DIOR dataset, such objects are commonplace. As such the problem of hyperparameter settings results in the low generalization of the anchor strategy. As the twin of the anchor mechanism, the anchor-free approach treats object location as the key point detection mechanism and assigns the problem of setting information, such as the size and direction of the object, to the network, completely avoiding the headache of setting the priority information. Meanwhile, the anchor-free approach also minimizes the problem of inefficient anchors. For backgrounds without key points, the network does not allocate resources to the generation of useless boxes. Thus, the anchor-free approach has the advantages of both high efficiency and high generalization. Wang et al. [99] used CenterNet with three parallel layers to predict the heat map, the compensation for the center, and the aspect of an object. Shi et al. [71] designed a central perception module, also based on CenterNet, to gradually guide the network to focus on the central regions. They also proposed a feature selection module to select the most suitable features for the scale feature layer to detect the objects. Huang et al. [100] generated BBoxes by predicting the four vertices of an object. They also introduced the M-sigmoid function to solve the instability problem introduced by large-scale regression. Cui et al. [101] designed a new anchor-free remote sensing ship detection model named SKNet, which constructs an orthogonal pool to highlight the features of the central point and its surroundings. On this basis, it then predicts the morphology of the central point. Shi et al. [102] proposed a remote sensing vehicle detection framework based on multi-task learning, which enables the network to simultaneously learn the vehicle center, direction, scale, and compensation. Liu et al. [103] constructed a module to highlight the boundary and central area of an object by using the dual attention mechanism. The aspect ratio constraint term was added to the angle regression to emphasize the effect of the aspect ratio for different objects. The network achieved real-time detection speed with guaranteed accuracy.

Table A1 summarizes the three anchor-based improvement strategies. Setting multi-scale anchors can solve the problem of remote sensing objects at different scales to a certain extent, but it is not enough to detect the super-scale objects which are sometimes contained in RSIs. Meanwhile, the alternate methods adopted in the aforementioned papers have certain limitations. The adaptive anchor strategy effectively solves the inefficient anchor tiling problem, and the network has the ability to apply more prediction work to the object, thereby reducing the number useless calculations. The anchor-free method is similar to the adaptive anchor mechanism, which eliminates the hyperparameter design as well as the redundancy calculation problem of anchor-based methods. However, the accuracy of key point prediction will directly determine the performance, and the detection model may not be able to efficiently deal with dense distributions of objects.

2.3.5. Direction Prediction-Based Method

Remote sensing objects often present a haphazard distribution of directions caused by overhead imaging, which makes their accurate detection a challenging task. CNN-based detection networks do not have rotational invariance, and therefore, are insensitive to object orientation information, resulting in unsatisfactory processing for rotating objects. The angle of remote sensing objects is a significant piece of information in back-end applications, such as military strikes; thus, accurate orientation predictions are an important part of the detection task. In this respect, Cheng et al. [104] added a rotation-invariant layer to AlexNet and enhanced the performance of the layer through data augmentation to make the detector more robust with regard to rotating objects. Shi et al. [73] designed geometric transformation modules to generate multi-angle images through random rotation and random flipping transformation, allowing the detector to learn the rotation features effectively. Huang et al. [105] added a deformable convolution layer to the network to learn the rotation-invariant features.

The methods mentioned above only enhance the network's ability to detect irregularly oriented objects; however, the orientation information of the objects is not reflected in the axis-aligned horizontal bounding boxes (HBB) results. HBB cannot accurately locate tilted objects, which will contain a large number of background or surrounding objects. Additionally, HBB are unfriendly to densely distributed objects, and a large overlap between boxes can easily cause the boxes to be filtered out by the post-processing procedure, resulting in missed detections. Thus, rectangular boxes with directions are more suitable for locating objects with irregular orientations. Directional BBox detection strategies may broadly be divided into three categories: rotating anchors, regression angle terms, and segmentation methods. The rotating anchor matches the directional objects by presetting multiple directional anchors at each pixel location with a fixed angular interval. This method does not need to change the existing network structure, and only needs to set the angular information regarding the anchor hyperparameters to generate the rotating box; a schematic diagram of this process is shown in Figure 10a. Ma et al. [106] designed a rotating regional network for the first time to generate anchor boxes with directions. The idea of the rotating anchor has been widely adopted. Fu et al. [84] also adopted a three-angle anchor to locate rotating objects and designed a new evaluation index to strictly limit the orientation object. Liu et al. [93] adopted a rotating anchor with 12 angles and proposed the length of the diagonal for the BBox to replace the regression of angle loss, achieving good results. Concerning the problem resulting from the small angle deviation of the object, which significantly contributes to the variation in IOU, Bao et al. [107] designed ArIOU to be more robust for evaluations of small angles. Xiao et al. [108] proposed an anchor selection method for an adaptive allocation anchor and designed the anchor with six angles at each position to locate directional objects.

The regression angle method predicts the direction of objects directly through the network function, treating the angle information independently of the four-boundary position information of the BBox for regression processing. Additionally, it adds the loss function as a constraint. The network generally predicts angles between 0° and 180°, finally combining them with the HBB to determine an object's rotation. Compared with the first method, this approach does not require the setting of hyperparameters, but rather, only uses an angle channel in addition to the original prediction layer. Yang et al. [85] constructed a ship heading detection model which uses convolution to directly predict the ship heading direction. Hua et al. [61] used a 1×1 convolution kernel to predict direction and designed a new angle loss function for constraint. The authors of [75,109,110] set the horizontal and direction prediction head at the back end of a detector that generates the horizontal HBB and orientational bounding boxes (OBB) at the same time. To address the problem of the boundary mutation of rotating boxes, Chen et al. [74] designed the OBB selection strategy, in which three parameters with the same shape are defined and the parameter with the least loss is selected.

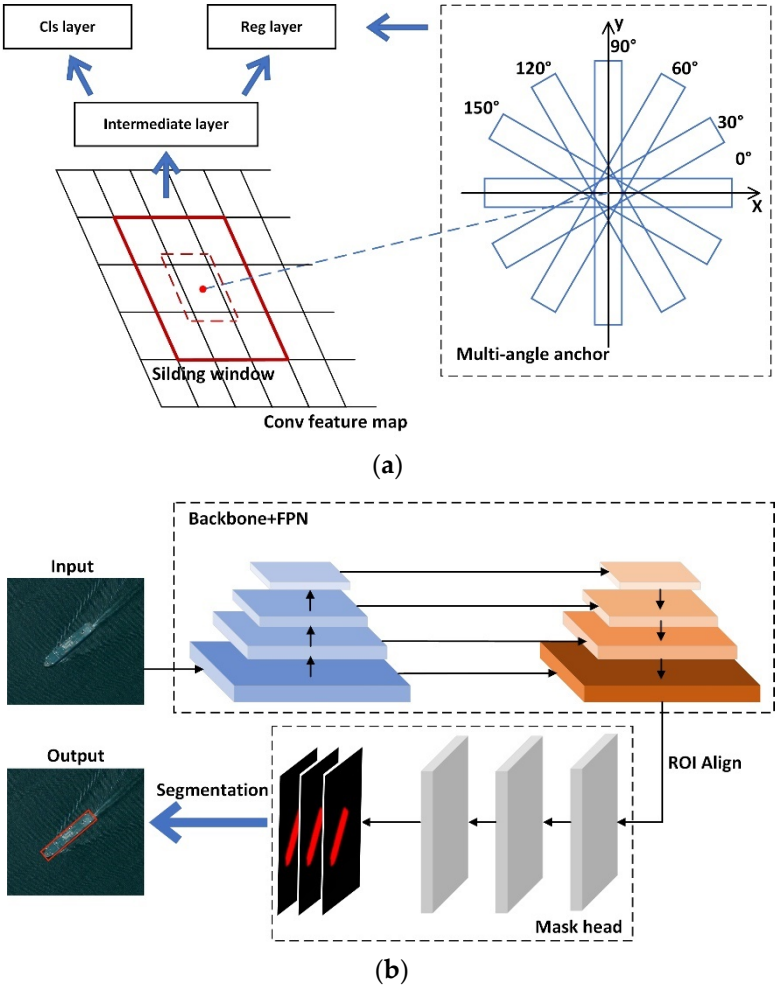

Figure 10. Schematic diagram of two methods for predicting object direction. (**a**) Multi-angle rotated anchors. (**b**) Direction prediction based on segmentation.

The segmentation angle approach draws on the idea of segmentation by generating a mask map of the object of interest. The mask map fully highlights the location information of the object, while the network segments the object area on the map to obtain the object size, shape, angle, and other information, as shown in Figure 10b. Li et al. [78] applied the improved Mask RCNN to RSOD. The network generated a mask by multi-task learning and calculated the minimum rectangular region in the mask to obtain the boxes. This represented a new way to solve the problem of arbitrary object direction in remote sensing.

2.3.6. The Super-Resolution Based Method

The processing of low-resolution images (LR) to obtain high-resolution images (HR) without changing the imaging equipment is called super-resolution reconstruction. It can effectively improve the resolution of an image, increasing the number of pixels, expanding the size, and enriching detail for LR. The authors of [111] proved that HR effectively improves the performance of RSOD. Image super-resolution reconstruction techniques show great potential as strategies with which to supplement additional information in order to improve object discrimination. On the one hand, super-resolution reconstruction technology can effectively deal with the small object problem, which is caused by a lack of sufficient semantic information for network identification resulting from the low pixel occupancy of an object of interest. The technology can reconstruct the discriminative information of a small object which is lacking feature expression; this is the key basis for network identification. On the other hand, super-resolution reconstruction technology

can solve the problem of image quality degradation due to environmental factors. Environmental conditions such as light, clouds, and weather can affect the image quality; the originally clear object may appear distorted in terms of its color, appearance, and clarity, which increases the difficulty of detection. Super-resolution reconstruction technology can improve the resolution of images, thereby reducing the recognition interference of object blur in the network. Meanwhile, this multitasking process can increase the application capacity of each technology and promote their simultaneous development.

Most super-resolution networks based on DL are Generative Adversarial Networks (GANs) [112], which comprise a generator and a discriminator. The generator produces near-real super-resolution images to cheat the discriminator. The discriminator then needs to determine whether the image is real or fake. The two compete against each other in the training process to obtain realistic super-resolution images, as shown in Figure 11. A super-resolution reconstruction network oriented toward the demands of a back-end detection network continuously reconstructs the HR, while the detection network recognizes the reconstructed images outputted by the front-end network and sends feedback signals to guide the super-resolution network. Front- and back-end networks can encourage each other to improve their capabilities during the training process. Mostofa et al. [113] designed a joint super-resolution remote sensing vehicle detection network which used the multi-scale MsGAN structure to output 2× and 4× super-resolution images (SR). They selected YOLOv3 [35] as the detection network to detect objects and designed a joint loss function. Bai et al. [114] chose to improve the resolution of the ROI and used a discriminator to simultaneously distinguish the authenticity of images, predict image categories, and perform boundary regression. Rabbi et al. [115] proposed an edge enhancement GAN named SERGAN, in which an image edge enhancement module was added to highlight the edge of the objects in SR for high-precision detection.

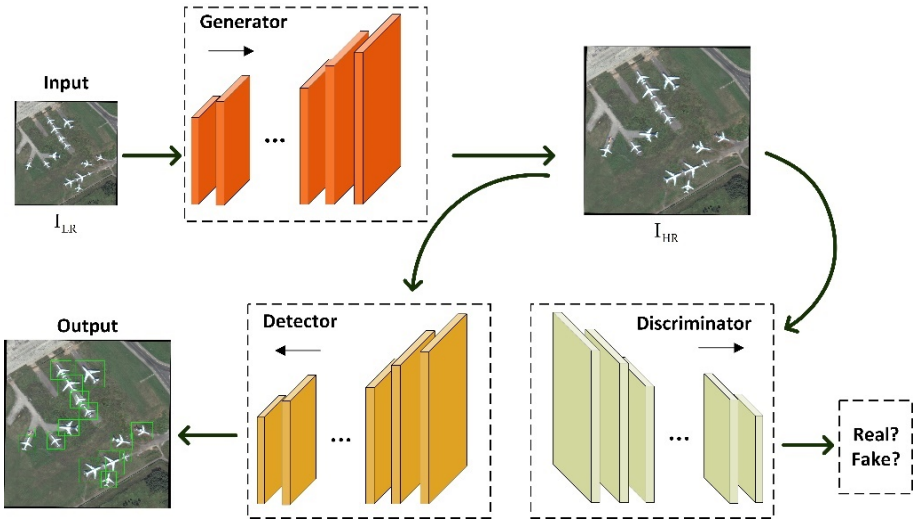

Figure 11. Schematic diagram of super-resolution reconstructed object detection based on the GAN network. I_{LR}—the input low-resolution RSI, I_{HR}—the input high-resolution RSI. The detector implements the localization and classification task from I_{HR} and the green box represents the BBox for locating the object.

The aforementioned work suffered from the limitation that training super-resolution networks require high- and low-resolution image pairs, and obtaining real image pairs requires imaging sensors with different imaging heights, which is difficult to achieve in practice. In order to reduce costs, the use of image processing to generate LR that are close enough to real images by downsampling is an appealing option. CycleGAN [116] was applied to the field of super-resolution reconstruction, integrating the generation of high- and low-resolution image pairs into the network. The network can be viewed as

a combination of two GAN networks containing two generators and two discriminators, where the former generate high- and low-resolution images respectively, and the latter determine the authenticity of the corresponding high- and low-resolution images, with the whole network forming a closed-loop-like structure. In addition, the network needs an additional dataset to guide the reconstruction effect, which could be image pairs from other domains. The authors of [117,118] adopted the improved CycleGAN to enhance the detection effect for the remote sensing of small objects. Ji et al. [119] designed the joint loss function with the idea of multi-task learning based on CycleGAN. This allowed detection loss to be propagated back to the super-resolution network during training to improve the performance of the two models. Gao et al. [120] proposed a new CycleGAN to guide the detection network, in which detection guidance branches were added after the network to improve the quality of object regions in RSIs. Liu et al. [121] synthesized remote sensing data by rendering a 3D CAD model and used CycleGAN for style transfer. Meanwhile, a multi-scale attention module was designed and embedded in CycleGAN to enhance the detailed information.

The strategy of combining super-resolution reconstruction with detection is a novel was to effectively improve the accuracy of detecting objects. The method opens up other possibilities, i.e., reconstruction under the current status quo, where the network structure, loss function, and feature processing in object detection are gradually maturing. Meanwhile, reconstructing RSIs with different object state distributions can enrich the diversity of data, solve the imbalance problem of extreme samples in the dataset, and alleviate the overfitting phenomenon of the network caused by a single scene distribution of the object.

2.3.7. The Transformer-Based Method

The Transformer [122], which was first proposed for use in sequence transduction tasks, adopts a multi-encoder-decode structure based on self-attention, and has become the dominant model in natural language processing. The encoder structure mainly consists of multi-head self-attention layers and feedforward neural network layers. The former is concerned the feature representation of subspaces at different positions, while the latter filters and collects multiple groups of Q, K, and V space features and feeds them into the decoder. To avoid the danger of gradient disappearance, the residual structure is introduced into the encoder. The decoder has an additional masked multi-head self-attention layer to prevent information at subsequent locations from interfering with input predictions. The huge and complex characteristics of the Transformer model increase the consumption of training data, but multiple encoders and decoders can be processed in parallel to fully exploit computational resources. This model can establish global relations between word vectors, even those that are particularly distant, and produce a surprising performance. Thus, the Transformer-based methods have great potential for development.

At present, the Transformer is extended to various computer vision fields, DETR [123] first introduced the Transformer to object detection, breaking the inherent rule of an original pre-defined anchor used to tile images. DETR gives the location prediction to the decoder structure in Transformer. The decoder inputs are Object Query sets, which are expressed as the coordinate vectors of BBox. The number of vectors in a Query set is calibrated to the number of BBox, which is much larger than the number of objects that actually exists in the image to ensure the recall rate. The encoder is responsible for extracting object features. Due to the existence of multiple self-attention layers in this structure, the encoder can extract fine features with rich global context information, which provides the conditions for the detection work. At the same time, this model adopts the Hungarian algorithm to ensure a one-to-one matching relationship between ground truths and prediction. DETR completely avoids troublesome anchor generation and time-consuming NMS post-processing, simplifying the original detection process and generating a competitive performance. However, the computational burden of the huge model must be solved urgently.

In RSOD, Zheng et al. [124] embedded the Transformer into a lightweight FPN and designed a self-transformer, a grounding transformer, and a rendering transformer to enhance the semantic information of the feature maps by connecting features at different levels. Li et al. [125] proposed a 2D position encoding by adding position information to the embedding in the encoder to alleviate the permutation invariance problem, and the proposed multi-head deformable self-attention layer allowed features to converge in an adaptive field. Zhang et al. [126] added a parallel Transformer branch with the backbone to improve the CNN's ability to capture global features. This branch replaced the original multi-head self-attentive layer with the SRGAN network to reduce the number of parameters and obtain similar results. Xu et al. [127] designed a Swin Transformer network that can effectively obtain the local perception of objects to alleviate the struggling performance of small objects. Zhu et al. [128] embedded the encoder structure into YOLOv5 to replace some convolutional layers, which can better capture global information and rich contextual information. The encoder structure was also inserted into the prediction heads to reduce the expensive computational and memory costs. Ma et al. [129] first extended the Transformer to remote sensing directional object detection, adding a directional dimension to the position prediction head. The author innovatively divided the self-attention structure of the Transformer into deeply separable convolution operations to extract lightweight features, which dramatically reduced the original computational burden.

At present, DETR has proved the applicability of the Transformer in object detection tasks and has received consistent acceptance in the natural domain. However, in the remote sensing field, the application of this technology is only the beginning stage. The complexity of RSIs makes the implementation of the Transformer frustrating, requiring a variety of refinements to accomplish satisfactory results. DETR also points out that the model produces a poor performance when facing small objects, and the problem becomes more prominent when it comes to remote sensing, where the situation can dramatically deteriorate the performance.

2.3.8. Non-Strongly Supervised Learning-Based Method

The previously mentioned works are processed in the fully supervised form with precise boundaries and category labels, but the labeling of detection tasks requires the tedious marking of the object's boundary locations, which is a heavy workload for remote sensing objects that are often densely distributed, so weakly supervised learning object detection methods have been developed to reduce the difficulty of labeling tasks. Ref. [130] summarizes weakly supervised learning into three categories: incompletely supervised, inexactly supervised, and inaccurately supervised. The first case represents semi-supervised learning in which the training set contains only a small part of samples annotated with precise labels giving the network reference. The model makes judgments based on the learning ability of the annotated samples to mine the internal laws of a large part of unlabeled data. Inaccurate supervision corresponds to weakly supervised learning in which the training data all provide only coarse-level labels, and the network needs to make judgments about unknown tasks. Inaccurate supervision means that the labeled data may not always be authentic, and the model needs to discriminate the labels in this case. This is subsection reviewed the RSOD for semi-supervised learning and weakly supervised learning. The rarely mentioned inaccurate supervision is not discussed here.

Semi-supervised Learning Based Method: The performance of the current DL-based detector relies heavily on a large-scale, high-quality labeled dataset, but collecting such a dataset is undoubtedly difficult and expensive. Semi-supervised learning can mitigate the model's requirement for labels and thus reduce the number of labeled samples, which can effectively solve the problem of labeling complex samples of RSIs. Zhong et al. [131] proposed an online parameter updating model using active learning. By network prediction, high-scoring BBox is selected for active learning and label up-dating, and the network training is further carried out to achieve real-time updating. Wu et al. [60] constructed a semi-supervised pseudo-label generation module that adopted curriculum learning. By

gradually increasing the threshold of the pseudo-label to train the detector from easy to difficult, the model's dependence on labels was reduced. Refs. [132,133] applied few-shot learning to RSOD by learning meta-knowledge from abundant known categories of data to learn unknown categories of samples.

Given that most algorithms in RSOD pursue supreme performance, the model that learns only a small amount of labeled data has a certain gap in performance compared to the fully supervised detector. However, the advantages of the model in terms of data make it worthy of more attention as a small branch.

Weakly supervised Learning-Based Method: As mentioned above, the precise annotation of the object's boundary information is a time-consuming and labor-intensive task, especially in remote sensing. The continuous development of remote sensing satellites and other technologies has reduced the difficulty of acquiring data, resulting in the current situation that the data volume of high-resolution RSIs is constantly rising. As shown in Figure 12c, annotating irregular objects and objects under small and dense areas in RSIs is bound to consume much time, not to mention the existence of object occlusion. Weakly supervised learning only needs to provide more readily available coarse-grained labels that are image-level labels for object categories as shown in Figure 12d. Thus, the training network with weakly supervised learning alleviates the difficulty of annotating complex labels for remote sensing.

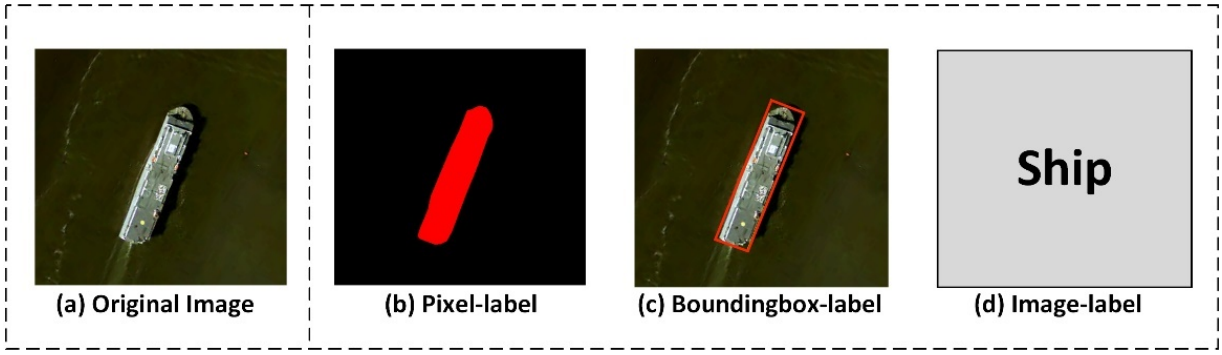

Figure 12. Schematic diagram of three different labels.

Current weakly supervised learning-based object detection methods for RSIs can be roughly divided into three categories: multiple instances learning-based methods, segmentation-based methods, and other special methods. Multiple instance learning (MIL) is the first strategy for weakly supervised object detection, which treats the image as a package, the object in the image as an instance, and the package as a collection of unlabeled instances. The network will accept a set of labeled packages to train and predict the unlabeled instances in the package. The method has two basic principles: (1) if there is at least one foreground instance in the image, the package is positive, and (2) if all the instances in the image are background, the package is negative. The first DL-based weakly supervised object detection (WSOD) method was the WSDDN [134], which initially combined MIL with object detection. Subsequently, PCL [135,136] adopted a pseudo-label mechanism for the supervised regression, which has been continuously developed by scholars [135–140] and gradually extended to remote sensing. To address the drawback of WSOD that often locates the object part, Feng et al. [141] calculated the correlation between pixels in feature extraction to help to discover the whole instance, while internal and external context modules were designed to extract contextual information and obtain more accurate detection results. Feng et al. [142] designed a dual-context instance refinement strategy to merge some candidate boxes and increase the confidence score of the box containing the whole object. Yao et al. [143] introduced the idea of dynamic curriculum learning to RSOD by training the detector from easy to difficult. Wang et al. [14] used OICR as the base framework and de-signed a pseudo-label generation mechanism for the

classification branch in WSDDN so that classification can be trained in a supervised form. Chen et al. [144] used asynchronous iterative training alternating between strongly and weakly supervised detectors to achieve the detection.

The segmentation-based method generates the object activation map and detects the object by extracting the region on the map. Li et al. [145] used pixel-level labels of scenes to learn the activation weight of specific categories and designed a multi-scale scene voting strategy to calculate the activation map of category-specific objects. Wu et al. [146] generated heatmaps to locate objects through reverse weighting based on AlexNet.

The third category of WSOD methods discussed here includes other special methods. Zhang et al. [147] developed a weakly supervised learning framework based on coupled CNN to automatically mine and enhance the training datasets from original images and continuously update the weak detector iteratively. Han et al. [148] used a deep Boltzmann machine to infer spatial and structural information encoded in low-level and middle-level features, and proposed a weak supervisor based on the Bayesian framework. Li et al. [149] proposed weak labels of object centroids and trained the network by using the generated pseudo-label.

Although weakly supervised learning has been widely developed, the performance of weakly supervised detectors without accurate boundary box annotations is far behind that of fully supervised detection models, and the gap in performance makes it difficult to apply in practice. Moreover, the method does not have a boundary regression process, leading to the problems of inaccurate object localization in the detection results, such as locating part of the object area, locating multiple instances of the same category, and locating the background. Therefore, the effective solution to the problem of sloppy positioning in the absence of location information deserves more in-depth consideration.

2.3.9. Other Methods

This subsection summarizes the general improvement strategies in addition to the above-mentioned methods, including special feature mapping, transfer learning, knowledge distillation, multi-task learning, and multi-source data fusion.

Special feature mapping methods map feature information onto different spaces and reprocess the information on subspaces by different means. Zheng et al. [150] designed a hyper-scale module to assign the convolution layer into sub-layer groups. Li et al. [151] mapped top-level features to three subspaces through different convolution kernels and obtained receptive fields of different sizes in each sub-space. Deng et al. [57] designed a multi-scale region proposal network that added three con-volution layers of different sizes to obtain the features of different receptive fields.

Transfer learning refers to the strategy of applying information learned in a domain or task to solve the corresponding problem in another related domain. The commonality of knowledge between the source domain and the transfer domain leads to a facilitation effect when the network trained in the source domain is transferred to the target domain. The common knowledge makes transfer learning an effective tool to deal with the challenges in remote sensing. The transfer model needs to find the connection between the original knowledge and the new knowledge to make the model produce a stable performance on the target domain, and how to find the similarity between the knowledge reasonably is the core problem of transfer learning. At the same time, transferring and fine-tuning the model according to the data distribution in the target domain will largely reduce the time of training the model from scratch. In addition, transferring similar knowledge will reduce the demand for samples in the target domain. To address the problem of a few small objects, Dong et al. [152] transferred the trained detector to remote sensing and realized automatic labeling for small objects. Li et al. [153] used CNN to transfer the model to solve the over-fitting problem caused by insufficient remote sensing data. Zhong et al. [154] used the pre-trained network to accelerate the training process of the model.

Similar to transfer learning, the goal of the knowledge distillation [155] strategy is to distill the knowledge acquired by complex models (teacher networks) to lightweight models (student networks) so that lightweight models have similar capabilities. The main purpose is to achieve model compression by adopting this knowledge transfer method. The prediction results of the teacher network, as soft labels, along with truth results, as hard labels, are sent to the student network as constraints to realize the delivery of model knowledge. In RSOD, Li et al. [156] used the knowledge distillation strategy to design distillation soft label loss in order to impart the capabilities of the larger teacher network ResNet-50 to the smaller student network ResNet-18 and achieve a light weight model. Liu et al. [157] enlarged the input image size of the teacher network and maintained the input image size of the student network to obtain cross-scale features. Meanwhile, positive-level L2 loss was adopted to constrain the difference between the features of the two networks. Zhang et al. [158] designed a dynamic knowledge distillation framework to solve the negative problems caused by blind inheritance of knowledge, and proposed a train-status-aware loss to enable the student network to dynamically focus on hard case objects, such as small-scale and extreme aspect ratio instances. Chen et al. [159] adopted knowledge distillation to prevent the impact of the introduction of new classes to the model on the prediction ability of old classes.

Multi-task learning applies the supervision of multiple related tasks to the same loss function so that the model learns various information concurrently. The method helps the model to explore the complementary information between subtasks, as it is difficult to obtain cues from a single task. Lei et al. [160] designed a reconstruction network that makes the network learn the reconstructed binary map that is used as the label to make the network focus on the object region. Chen [74] constructed a constraint attention network supervised by a binary segmentation map to guide the network to filter the background while retaining context information. Refs. [78,110,161] all adopt the means of multi-task learning to carry out detection and segmentation tasks simultaneously.

The multi-source data fusion strategy can effectively obtain the missing information for a single RSI from other external data. Due to the current imaging hardware performance limitations, the sensor can only take advantage in one of the three spectral, temporal, and spatial resolutions to obtain high-resolution effects; the remaining two in a single sensor can only present poor performance, which causes the lack of information. The fusion of RSIs obtained by multiple sensors in the same scene can adequately make up for this lack of information and effectively achieve the role of interconnection between information. Wu et al. [55] proposed an unsupervised multi-source active fine-tuning remote sensing vehicle detection framework. By integrating DSM images with RGB images, potential objects can be segmented through height information for automatic vehicle labeling. By sending RSIs and infrared images in the same scene to the detection network for feature extraction, Chachlakis et al. [59] obtained enhanced multi-source features to improve efficiency. However, multiple source images of the same scene are hard to acquire, and image alignment is also challenging, which seriously hinders the implementation of this method.

The various methods described above are summarized in Table A1. There are other unusual improvement methods, but we do not discuss these individually here.

3. Results

In recent years, many available datasets with reliable evaluation metrics have been released in RSOD. This section briefly introduces common public datasets and standard performance metrics, and compares the performance of statistical algorithms.

3.1. Benchmark Datasets

The common large-scale RSOD datasets are as follows:

1. NWPU VHR-10 [162] is a very high-resolution dataset with 800 images and 3651 instances for optical RSI object detection and contains ten categories of objects, where the categories are: airplane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, and vehicle.
2. DOTA [163] is a fifteen categories of RSOD dataset containing 2806 optical RSIs and a total of 188,282 instances. The dataset is labeled by experts with horizontal annotation and rotating annotation. The categories of objects are as follows: plane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, large vehicle, small vehicle, helicopter, roundabout, soccer ball field, and basketball court.
3. DIOR [16] contains 23,463 optical RSIs and 192,472 instances in total, with a spatial resolution of 0.5 to 30 m. Images present different states such as environment, weather, season, illumination with 800 × 800 in size. The twenty categories of objects are as follows: airplane, airport, base-ball field, basketball court, bridge, chimney, dam, expressway service area, expressway toll station, harbor, golf course, ground track field, overpass, ship, stadium, storage tank, tennis court, train station, vehicle, and windmill.
4. UCAS-AOD [164] contains vehicle data as well as aircraft data, selected from the Google Earth aerial image dataset. The vehicle data contains 310 images with 2819 vehicle instances. The aircraft data contains 600 images, including 3210 aircraft instances, with an image size of approximately 1000 × 1000.
5. HRSC2016 [165] is a dataset created specifically for ship detection. The dataset contains 1070 images with a total of 2917 ship instances, ranging from 300 × 300 to 1500 × 900 in size.
6. RSOD [166,167] is made by Wuhan University, which collected 976 RSIs from Google Earth and Tianditu with 6950 instances in total, including four categories: oil tank, aircraft, overpass, and playground.
7. LEVIR [168] consists of 3791 high-resolution RSIs from Google Earth, with the size of 800 × 600 and a spatial resolution of 0.2–1 m. There are three categories of objects in the dataset: aircraft, ships, and oil tanks.
8. HRSSD [81] is a category-balanced RSI dataset, the images are cropped from Google Earth and BaiDu Map. The dataset contains 26,722 images, totaling 13 categories of objects, which are: airplane, baseball diamond, basketball court, bridge, cross-road, ground track field, harbor, parking lot, ship, storage tank, T junction, tennis court, vehicle.
9. AI-TOD [169] is a challenging dataset specially designed for remote sensing tiny object detection. 28,036 images with 700,621 instances are collected in the dataset, and the average size of the objects is only 12.8 pixels, which is much smaller than other datasets. Images are mainly collected from multiple datasets with the 800 × 800 pixels, including eight categories of objects: airplane, bridge, storage-tank, ship, swimming-pool, vehicle, person, and wind-mill.
10. VEDAI [170] is a dataset for remote sensing vehicle detection, which contains a total of 1210 aerial images with 1024 × 1024 resolution. The nine categories of objects included are: plane, boat, camping car, car, pick-up, tractor, truck, van, and the other category, which contain five categories of vehicles with different appearance. The scale of each category varies widely and presents different orientations.

Table 2 lists the parameters of the above datasets to provide an intuitive comparison. In addition to the public datasets mentioned above, some scholars also created their datasets to meet their tasks, such as Refs. [62,97,98].

Table 2. Comparison of common optical RSI datasets, where HBB represents the horizontal bounding box, OBB represents the oriented bounding box.

Dataset	Quantity	Category	Size	Instance	Resolution	Label
NWPU VHR-10 [162]	800	10	350 × 350–1200 × 1200	3651	0.5–2 m, 0.08 m	HBB
DOTA [163]	2806	15	800 × 800–4000 × 4000	188,282	0.1–1 m	HBB, OBB
DIOR [16]	23,463	20	800 × 800	192,472	0.5–30 m	HBB
UCAS-AOD [164]	910	2	1000 × 1000	6029	-	HBB, OBB
HRSC2016 [165]	1070	1	300 × 300–1500 × 900	2917	0.4–2 m	HBB, OBB
RSOD [166,167]	976	4	800 × 1000	6950	0.3–3 m	HBB
LEVIR [168]	3791	3	800 × 600	11,028	0.2–1 m	HBB
HRSSD [81]	26,722	13	-	55,740	0.15–1.2 m	HBB
AI-TOD [169]	28,036	8	800 × 800	700,621	-	HBB
VEDAI [170]	1210	9	1024 × 1024	-	0.125 m	OBB

3.2. Performance Metrics

The universally accepted standard evaluation indicators are recall (R), precise (P), average precise (AP), PRC, mean average precise (mAP), and frame per second (FPS). The first five measure the detection accuracy, while the latter measures the detection speed. The intersection of union (IOU) measures the ratio of the area of intersection between proposals and ground truths to the area of union and usually judges whether the object is detected, defined as follows:

$$\mathrm{IOU} = \frac{\mathrm{area}(\mathrm{Proposal} \cap \mathrm{Ground\ Truth})}{\mathrm{area}(\mathrm{Proposal} \cup \mathrm{Ground\ Truth})} \tag{1}$$

According to different results, the definitions are as follows: the number of true positives (TP), false negatives (FN), false positives (FP), and true negatives (TN). The definitions of recall rate and accuracy rate are given as follows:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{2}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

where TP + FN represents the number of BBoxes and TP + FP represents the total number of objects to be detected. Therefore, P evaluates by the number of correctly detected objects in the total object, while R evaluates by how many objects are detected. AP takes into account both accuracy and recall, as defined below:

$$\mathrm{AP} = \int_0^1 \mathrm{P}(\mathrm{R})\mathrm{dR} \tag{4}$$

which represents the average precise of the recall rate between 0 and 1. PRC is a curve drawn according to the detection of maximum P at each R, and the area under the curve is equal to AP. The mAP represents the average precise of all categories, expressed as follows:

$$\mathrm{mAP} = \frac{1}{\mathrm{C}} \sum_{\mathrm{i}=1}^{\mathrm{C}} \mathrm{AP} = \frac{1}{\mathrm{C}} \sum_{\mathrm{i}=1}^{\mathrm{C}} \int_0^1 \mathrm{P_i}(\mathrm{R_i})\mathrm{dR_i} \tag{5}$$

where C represents the number of object categories. mAP comprehensively considers the detection accuracy of objects. Therefore, the higher the mAP value is, the more accurate the detector.

In addition to accuracy evaluation, detection speed is also another criterion for comprehensive evaluation of detection performance. The general evaluation of detection speed is FPS, that is, the number of images that can be detected within a second. Meanwhile, the speed can also be measured by the time used to calculate an image.

3.3. Performance Comparison

In this subsection, we compare the DL-based models of RSOD with three challenging common datasets, DOTA, NWPU VHR-10, and DIOR. Table A2 shows the performance comparison of each model in the three datasets, and mAP is used as the only indicator to measure the accuracy. The items listed in the comparison table also include the backbone network, the main strategies, the types of BBoxes, and the division of the dataset.

In Table A2, the excellent performances of the algorithms on the dataset reflect the great potential shown by DL techniques in RSOD, and illustrates that the implementation of DL techniques has contributed to the progress of detection effectiveness. In terms of datasets, the DOTA and DIOR datasets contain more categories and instances so that detection becomes more difficult. The clear division between training and test data in datasets makes performance comparisons fairer. The NWPU VHR-10 dataset contains relatively few images, making the average performance of the algorithms higher than the first two. But at the same time, this dataset requires autonomous division, and control variables cannot be guaranteed in data selection, leading to some errors in performance comparison. In terms of performance, various excellent algorithms have contributed to the evolution of detection efforts in part. However, the fact that the growth rate of performance is slowing down as the development pipeline of DL lengthens also reflects the trend that current detection algorithms are maturing and may have encountered an invisible bottleneck. Thus, how to continue to resolve possible flaws in the model from other aspects to further improve detection should be considered in current and future research. In terms of the backbone network, except for the early and individual papers adopting a shallow network of AlexNet, most of the papers chose VGGNet and ResNet with deeper layers and stronger extraction ability, which also indicates that the feature extraction ability of the backbone is particularly important. The deeper network that extracts strong feature semantic information is more conducive to the analysis work. However, the pursuit of the deeper network did not earn satisfactory feedback in performance, which also means that the method of extending network depth in RSOD may have inadequate parameter training or a gradient update problem. In terms of improvement strategies, the various strategies designed according to the complex objects, complex backgrounds, and complex sample annotation problems of RSIs have effectively alleviated the embarrassing situation of unsuitable performance faced by the general model. Although the superiority of various strategies cannot be reflected through the comparison of performance, it is certain that each strategy must play a certain role in promoting the detection effect, which is also one of the key factors in the rapid development of RSOD. The attention mechanism captures associations between different regions, enhances beneficial information and compresses interfering signals, which fully mitigates the challenges of complex backgrounds. The attention-based models also exhibit a superior in performance. Multi-scale feature fusion can effectively combine various types of information and reduce the loss between information transmission. In dealing with the remote sensing of complex objects, the method provides an additional reference for prediction work and is a strategy that all models must consider, thereby validating its effectiveness. Mining context information is another strategy that can be used to effectively address the challenges of complex objects. This approach can mine valuable external information to reduce the ambiguity of identifying difficult objects, which is also reflected in the model performance. The refined Anchor mechanism improves the defects of the original anchor from various aspects and is more suitable for the complex situation of remote sensing. Direction prediction improves the functionality of detection and increases the practicability of the model in remote sensing situations. Meanwhile, the lower performance of the rotation detector compared to the horizontal detector indicates

that it is more difficult to locate an object by applying the rotation box. The model needs to predict the object boundary in addition to the angle, which is not an easy task. The super-resolution reconstruction strategy is effective for dealing with complex small objects with regard to remote sensing. The strategy can expand the object scale and obtain more detailed information, which is very important for the identification and localization of small objects. The Transformer model can be interpreted as performing multiple attention operations and adequately establishing the interactions between the objects. This strategy delivers satisfactory results in terms of its performance. In addition, the incompletely supervised detector with the mission to alleviate the label annotation problem has a gap in performance with the strongly supervised model, where incomplete data are the main reason for the poor performance. The current performance also introduces more possibilities to this method. The model with incompletely supervised learning has great development prospects worthy of further development.

4. Discussion

Although RSOD algorithms based on DL techniques have made satisfactory progress and are gradually maturing, there are still shortcomings that deserve further discussion. In this section, we analyze and summarize the problems that exist in RSOD at present, and propose solutions and possible future research trends so that the reader can better grasp the current circumstances.

- **Improve network structures:** At present, the slowing improvement rate of remote sensing detector performance indicates that existing methods have reached their limitations, making it difficult to achieve a breakthrough. Thus, the question of how to further improve the technology is the key problem that needs to be solved. The underlying network structure, as the model's foundation, is likely the key to overcoming the problem. A state-of-the-art network structure designed specifically for RSIs will serve complex objects more effectively; this is certainly a worthwhile research direction.
- **Improve light weight models:** In order to extract features with rich information representation, networks are mostly designed with extremely deep structures, requiring the optimization of huge numbers of parameters. This increases the model's demand for data while increasing the burden on computing facilities. Current low arithmetic portable embeddable devices cannot implement such weighty models. The question of how to reduce the parameter scale of existing models in order to improve their practicality is particularly significant. Light weight models involve the participation of various aspects such as network structure and optimization methods.
- **Improve weakly supervised learning:** Defects in performance restrict the application scope of weakly supervised learning, and consequently, this direction is seldom explored. The advantages of labeling also broaden development prospects, and the further use of detection capabilities is a topic that is worthy of in-depth study. In addition, weakly supervised rotation detectors have not been developed due to the absence of boundary information, and the HBB used in current models do not accurately locate remote sensing objects with complex directional distributions. Thus, weakly supervised learning for rotation detection is advancing.
- **Improve the direction prediction strategy:** Direction is one of the essential manifestations of object position information, and a variety of direction recognition systems have been established for accurate object orientation. However, most such models set the direction in the range of 0–180°, which does not take orientation into account. For instance, a model defining the bow and stern of a ship does not provide a discriminant. Object orientation detection is of great significance for practical applications and deserves further attention. Meanwhile, there is no standard for determining how the position of a rotating object shall be correctly detected. Current IOU evaluation criteria have restrictions due to the drawback that slight deviations in the angle between the two directional boxes will lead to a drastic decrease in the IOU, which hinders the mea-

surement of the IOU for the directional boxes. Therefore, the metric of angle needs to be carefully examined to reasonably assess the rotation of objects. Moreover, direction detection models struggle with objects that have no obvious directional information, such as storage tanks, which is another a matter that is worthy of discussion.

- **Improve super-resolution detection:** For weak object detection without sufficient structural knowledge, super-resolution reconstruction technology can effectively expand the object scale and provide additional details to boost its identification effect. As such, this approach which has attracted widespread research interest. However, the effective combination of the two tasks turns out to be challenging due to the question of how to guide the super-resolution network to purposefully enhance details that are not adequately expressed in the object itself, instead of enhancing some irrelevant information in a general way. In addition, the double parameters produced by the joint network restrict the actual speed, and optimization algorithms need to be tailored to further reduce the time cost. The implementation of specific reconstruction strategies to enhance the super-resolution effect on the object and weaken the network's focus on the background is worth investigating.
- **Improve small object issues:** Small object detection has always been a priority in RSOD. Small objects—whose pixel occupancy is small and features are difficult to extract, making them prone to being obscured during the forward propagation—are commonplace in RSIs. Although researchers are currently studying this phenomenon and proposing various solutions, these tend to only alleviate the issue. Small object issues become problematic in the following three respects: **(1) Sample imbalance problem:** small objects account for a low proportion of remote sensing data. After statistics, objects with pixels less than 16×16 only account for 10% of the DIOR dataset. As such, these few small objects do not get enough attention from the system, resulting in missed detections. **(2) Loss imbalance problem:** during network training, the contribution of small objects to losses is much smaller than that of others, which is mainly due to the minor regression distance, thus yielding negligible loss. This loss imbalance phenomenon also leads to poor results for small objects. **(3) Matching imbalance problem:** positive and negative samples have been determined by the IOU threshold selection method. Slight deviations among small objects during label assignment results in large IOU variations, which limits small objects to produce only a small number of matching BBoxes, thereby reducing the chances of small objects being selected and increasing the number of missed matches. Therefore, it would be useful to boost performance regarding small objects. Small objects, such as ships in ports and aircraft in airports, tend to be densely distributed in RSIs, but their detection remains an arduous task. Indeed, there is no definitive way to solve the problem of dense distribution. In the future, densely packed small objects must be taken into account to achieve accurate localization and identification.

5. Conclusions

In recent years, RSI analysis technology has experienced major progress thanks to the advent of artificial intelligence. DL has driven the continuous development of object detection technology in the direction of intelligence. In this work, we first reviewed the successful combination of DL and detection techniques, such as one- and two-stage families. Then, we summarized in detail the mainstream detection ideologies and divided them into pre-processing, feature extraction and processing, BBox generation, detection and post-processing steps. Numerous improvement strategies for complex problems in relation to remote sensing are presented in detail and divided into a taxonomy comprising attention mechanisms, multi-scale feature fusion, mining contextual information, refined anchor strategies, direction prediction strategies, super-resolution reconstruction techniques, Transformer-based methods, semi-supervised learning, and weakly supervised learning, among other methods. Benchmark datasets, performance metrics, and performance comparisons of representative models are also discussed. Finally, considering the

obstacles that exist in current detection technologies, we provide research directions that could be explored in the future.

Currently, DL is a mainstream technology in the field of object detection. However, there are still some limitations. Firstly, various models perform well on large-scale general datasets; however, the question of how to further improve the accuracy of detection is a major challenge. Secondly, RSOD based on DL has strict requirements regarding training data. This severely affects the model performance. The question of how to reduce this dependence on data and improve the generalization ability of the models is a major challenge. Finally, small objects always pose a challenge in remote sensing, and existing small object detection methods are not ideal. Thus, research on small object detection needs to continue.

DL has become a key technology in RSOD and has achieved satisfactory results in various applications. However, numerous shortcomings remain, requiring researchers to make continuous efforts to achieve the intellectual development of RSOD.

Author Contributions: Z.L. wrote the manuscript; Y.W. gave professional guidance and edited; N.Z. and Y.Z. gave advice and edited; D.X. and G.B. gave advice; Z.Z. and Y.G. revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available in the article.

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

Appendix A

Table A1. Summary of the improvement strategies, including an attention mechanism, multi-scale feature fusion, mining context information, refined anchor, weakly supervised learning, and other methods.

Method		Characteristics	Ref.
Attention mechanism	Spatial attention	a. Captures spatial location relationships b. Assigns different weights to different locations c. Highlights the object region and suppresses background region	[61,72–76]
	Channel attention	a. Captures feature channel relationships b. Assigns different weights to different channels c. Highlights object features and suppresses others features	[60,77]
	Joint attention	a. Captures the relationship between the spatial location and feature channel b. Comprehensively enhances object information	[78–80]
Multi-scale feature fusion	Simple feature fusion	a. Simply fuses the features of adjacent layers b. Enhances the ability to detect a single scale c. Easy operation	[58,62,81–83]
	Feature pyramid fusion	a. Enhances the feature information of the shallow layer b. Deals with multi-scale changes of objects c. Improves the ability to detect small objects	[12,73,75,84–86]
	Cross-scale feature fusion	a. Fuses all layer features together b. Enhances the feature information of each layer c. Requires a large number of calculations	[11,74,87,88]

Table A1. *Cont.*

Method		Characteristics	Ref.
Mining context information	Enlarge detection window	a. Enlarges the ROI region b. Increases the surrounding information c. Adds more local features	[91–93]
	Enhance receptive field	a. Expands convolution region area b. Increases the surrounding features c. Extracts more information	[58,86,87,94,95]
	Attention mechanism	a. Captures the relationship between the objects and their surroundings b. Highlights the object and surrounding information	[71,73,75]
Refined anchor	Set more scale anchors	a. Presets more anchors b. Full coverage of remote sensing objects c. Many super-parameters and low detection speed	[58,96,97]
	Adaptive anchor mechanism	a. Adaptively generates a few high-quality anchors b. No superparameter set c. More intelligent and fast detection speed	[12,80,98]
	Anchor-free	a. Uses the key points detection method b. Avoids the shortcomings of the anchor mechanism c. More efficient	[71,99–103]
Weakly supervised learning	Multiple instance learning	a. Judges the positive and negative aspects of the package according to the instances b. Coordinates cannot be accurately regressed c. Detection accuracy is far less than with a strong supervision detector	[14,141–144]
	Based on segmentation	a. Segments the activation map to locate the object b. Depends strongly on an activation map c. Detection position may be inaccurate	[145,146]
	Other methods	a. Novel methods b. Advantages in some specific aspects	[147–149]
Other methods	Special feature mapping	a. Maps features to other spaces b. Processes information from different spaces c. Enhances the ability of information mining	[57,150,151]
	Transfer Learning	a. Reduces the training time of the model b. Reduces model dependence on data c. Requires data similarity between source domain and transfer domain	[152–154]
	Knowledge Distilling	a. Achieves knowledge transfer between models b. Compresses the model parameters c. Reduces the inference time of the model	[156–159]
	Multi-task learning	a. Learns multiple tasks at the same time b. Learns complementary information from different tasks c. Saves computing time for multiple tasks	[74,78,110,160,161]
	Multi-source data fusion	a. Requires images from different sources but the same scene b. Supplies information from other source images c. Compensates for limited performance with a single data source	[55,59]

Appendix B

Table A2. The performance of each algorithm with the DOTA, NWPU VHR-10, and DOIR datasets is tabulated in the table, where WSL stands for weakly supervised learning and SSL for semi-supervised learning. The data are divided into a training set, a validation set, and a test set.

Dataset: DOTA						
Models	**Years**	**Backbones**	**Methods**	**Proposals**	**Dataset Set**	**mAP (%)**
			Strongly Supervised			
Ref. [79]	2019	VGG-16	Attention mechanism	HBB	50%, 16%, 34%	49.16
Ref. [85] [1]	2018	ResNet-101	Feature fusion	OBB	50%, 16%, 34%	81.25
CAD-Net [75]	2019	ResNet-101	Context information	OBB	50%, 16%, 34%	69.9
SE-SSD [94]	2019	VGG-16	Context information	HBB	50%, 16%, 34%	70.8
FSoD-Net [96]	2021	MSE-Net	Improved anchor	HBB	50%, 16%, 34%	75.33
SARA [12]	2021	ResNet-50	Improved anchor	OBB	50%, 16%, 34%	79.91
LO-Det [100]	2021	MobileNetv2	Improved anchor	OBB	50%, 16%, 34%	66.17
SKNet [101] [1]	2021	Hourglass-104	Improved anchor	OBB	75%, 25%	83.9
CBDA-Net [103]	2021	DLA-34	Improved anchor	OBB	50%, 16%, 34%	75.74
Rs-Det [105]	2019	ResNet-50	Direction prediction	HBB	50%, 16%, 34%	65.33
FFA [84]	2020	ResNet-101	Direction prediction	OBB	50%, 16%, 34%	75.7
F3-Net [109]	2020	ResNet-50	Direction prediction	OBB HBB	50%, 16%, 34%	76.02 76.48
AMFFA-Net [74]	2021	ResNet-101	Direction prediction	OBB HBB	50%, 16%, 34%	76.27 78.06
A2S-Det [108]	2021	ResNet-50 ResNet-101	Direction prediction	OBB	50%, 16%, 34%	70.42 70.64
HyNet [150]	2020	ResNet-50	Feature mapping	HBB	50%, 16%, 34%	62.01
Ref. [110]	2019	ResNet-101	Multi-task learning	OBB HBB	50%, 16%, 34%	67.96 69.88
RADet [78]	2020	ResNet-101	Multi-task learning	OBB	50%, 16%, 34%	69.09
ADT-Det [124]	2021	ResNet-50	Transformer	OBB	50%, 16%, 34%	79.95
O2DETR [129]	2021	ResNet-50	Transformer	OBB	50%, 16%, 34%	79.66
Dataset: NWPU VHR-10						
			Strongly Supervised			
RICNN [104]	2016	AlexNet	Direction prediction	HBB	20%, 20%, 60%	72.63
HRCNN [81]	2019	AlexNet	Feature fusion	HBB	20%, 20%, 60%	73.54
RECNN [160]	2020	VGG-16	Multi-task learning	HBB	20%, 20%, 60%	79.2
PSBNet [154]	2018	ResNet-101	Transfer learning	HBB	20%, 20%, 60%	82.0
Sig-NMS [152] [1]	2019	VGG-16	Transfer learning	HBB	20%, 20%, 60%	82.9
TRD [125]	2022	ResNet-50	Transformer	HBB	20%, 20%, 60%	87.9
Ref. [110]	2019	ResNet-101	Direction prediction	HBB	20%, 20%, 60%	89.07
F3-Net [109]	2020	ResNet-50	Direction prediction	HBB	20%, 20%, 60%	91.89
CA-CNN [92]	2019	VGG16	Context information	HBB	40%, 10%, 50%	90.97
MSNet [58]	2020	DarkNet53	Attention mechanism	HBB	40%, 60%	95.4
HyNet [150]	2020	ResNet-50	Feature mapping	HBB	40%, 60%	99.17
RADet [78]	2020	ResNet-101	Direction prediction	HBB	60%, 20%, 20%	90.24
FMSSD [95]	2019	VGG-16	Context information	HBB	60%, 20%, 20%	90.40
CANet [88]	2020	ResNet-101	Context information	HBB	60%, 20%, 20%	92.2
YOLOv3-Att [73]	2020	DarkNet-53	Attention mechanism	HBB	60%, 20%, 20%	94.49
DCL-Net [86]	2020	ResNet-101	Feature fusion	HBB	60%, 20%, 20%	94.55
Ref. [57]	2018	VGG-16	Feature mapping	HBB	60%, 40%	94.87
Ref. [91]	2017	ZFNet	Context information	HBB	75%, 25%	87.12
CAD-Net [75]	2019	ResNet-101	Context information	HBB	75%, 25%	91.5
LFPNet [90]	2021	ResNet-101	Feature fusion	HBB	75%, 25%	93.23
CANet [71]	2021	RestNet-101	Attention mechanism	HBB	75%, 25%	93.33
Ref. [79]	2019	VGG-16	Attention mechanism	HBB	80%, 20%	85.08

Table A2. *Cont.*

Models	Years	Backbones	Methods	Proposals	Dataset Set	mAP (%)
Dataset: NWPU VHR-10						
			Non-strongly Supervised			
Ref. [143]	2020	VGG-16	WSL	HBB	58%, 17%, 25%	20.19
Ref. [14]	2021	VGG-16	WSL	HBB	60%, 20%, 20%	53.6
PCIR [142]	2020	VGG-16	WSL	HBB	58%, 17%, 25%	54.97
TCANet [141]	2020	VGG-16	WSL	HBB	75%, 25%	58.82
FSODM [133]	2021	DarkNet-53	SSL	HBB	-	65.0
Ref. [149]	2021	CSPDarkNet-53	WSL	HBB	70%, 10%, 20%	92.4
Dataset: DIOR						
			Strongly Supervised			
LO-Det [100]	2021	MobileNetv2	Improved anchor	HBB	50%, 50%	65.85
TRD [125]	2022	ResNet-50	Transformer	HBB	50%, 50%	66.8
Ref. [11]	2020	ResNet-101	Attention mechanism	HBB	50%, 50%	68.0
MFPNet [87]	2021	VGG-16	Context information	HBB	50%, 50%	71.2
FRPNet [72]	2020	ResNet-101	Attention mechanism	HBB	50%, 50%	71.8
FSoD-Net [96]	2021	MSE-Net	Improved anchor	HBB	50%, 50%	71.8
Ref. [80]	2020	ResNet-50	Improved anchor	HBB	50%, 50%	73.6
CANet [88]	2020	ResNet-101	Context information	HBB	50%, 50%	74.3
			Non-strongly Supervised			
FCC-Net [144]	2020	ResNet-50	WSL	HBB	50%, 50%	18.1
PCIR [142]	2020	VGG-16	WSL	HBB	50%, 50%	24.92
TCANet [141]	2020	VGG-16	WSL	HBB	50%, 50%	25.82
prototype-CNN [132]	2021	ResNet-101	SSL	HBB	50%, 50%	32.6
FSODM [133]	2021	DarkNet-53	SSL	HBB	other	36.0
Ref. [143]	2020	VGG-16	WSL	HBB	50%, 50%	52.11

[1] means that the model used only partial data.

References

1. Lim, J.-S.; Astrid, M.; Yoon, H.-J.; Lee, S.-I. Small object detection using context and attention. In Proceedings of the 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Jeju Island, Korea, 13–16 April 2021; pp. 181–186.
2. Zhang, J.; Zhang, L.; Liu, T.; Wang, Y. YOLSO: You Only Look Small Object. *J. Vis. Commun. Image Represent.* **2021**, *81*, 103348. [CrossRef]
3. Van der Meer, F. Remote-sensing image analysis and geostatistics. *Int. J. Remote Sens.* **2012**, *33*, 5644–5676. [CrossRef]
4. Van der Meer, F.D.; van der Werff, H.M.A.; van Ruitenbeek, F.J.A.; Hecker, C.A.; Bakker, W.H.; Noomen, M.F.; van der Meijde, M.; Carranza, E.J.M.; Smeth, J.B.D.; Woldai, T. Multi- and hyperspectral geologic remote sensing: A review. *Int. J. Appl. Earth Obs. Geoinf.* **2012**, *14*, 112–128. [CrossRef]
5. ElMikaty, M.; Stathaki, T. Detection of Cars in High-Resolution Aerial Images of Complex Urban Environments. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 5913–5924. [CrossRef]
6. Rawat, W.; Wang, Z. Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Comput.* **2017**, *29*, 2352–2449. [CrossRef]
7. Liu, L.; Wang, Z.; Qiu, T.; Chen, Q.; Lu, Y.; Suen, C.Y. Document image classification: Progress over two decades. *Neurocomputing* **2021**, *453*, 223–240. [CrossRef]
8. Yu, D.; Ji, S. A New Spatial-Oriented Object Detection Framework for Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4407416. [CrossRef]
9. 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.* **2021**, *59*, 6121–6137. [CrossRef]
10. Ma, W.; Li, N.; Zhu, H.; Jiao, L.; Tang, X.; Guo, Y.; Hou, B. Feature Split–Merge–Enhancement Network for Remote Sensing Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5616217. [CrossRef]
11. Cheng, G.; Si, Y.; Hong, H.; Yao, X.; Guo, L. Cross-Scale Feature Fusion for Object Detection in Optical Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 431–435. [CrossRef]

12. Hou, J.-B.; Zhu, X.; Yin, X.-C. Self-Adaptive Aspect Ratio Anchor for Oriented Object Detection in Remote Sensing Images. *Remote Sens.* **2021**, *13*, 1318. [CrossRef]
13. 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.* **2022**, *60*, 5614914. [CrossRef]
14. Wang, H.; Li, H.; Qian, W.; Diao, W.; Zhao, L.; Zhang, J.; Zhang, D. Dynamic Pseudo-Label Generation for Weakly Supervised Object Detection in Remote Sensing Images. *Remote Sens.* **2021**, *13*, 1461. [CrossRef]
15. Cheng, G.; Han, J. A survey on object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28. [CrossRef]
16. 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]
17. Alganci, U.; Soydas, M.; Sertel, E. Comparative Research on Deep Learning Approaches for Airplane Detection from Very High-Resolution Satellite Images. *Remote Sens.* **2020**, *12*, 458. [CrossRef]
18. Zheng, Z.; Lei, L.; Sun, H.; Kuang, G. A Review of Remote Sensing Image Object Detection Algorithms Based on Deep Learning. In Proceedings of the 2020 IEEE 5th International Conference on Image, Vision and Computing (ICIVC), Beijing, China, 10–12 July 2020; pp. 34–43.
19. Kim, T.; Park, S.R.; Kim, M.G.; Jeong, S.; Kim, K.O.J.P.E.; Sensing, R. Tracking Road Centerlines from High Resolution Remote Sensing Images by Least Squares Correlation Matching. *Photogramm. Eng. Remote Sens.* **2004**, *70*, 1417–1422. [CrossRef]
20. Chaudhuri, D.; Kushwaha, N.K.; Samal, A. Semi-Automated Road Detection From High Resolution Satellite Images by Directional Morphological Enhancement and Segmentation Techniques. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2012**, *5*, 1538–1544. [CrossRef]
21. Akcay, H.G.; Aksoy, S. Building detection using directional spatial constraints. In Proceedings of the Geoscience & Remote Sensing Symposium, Honolulu, HI, USA, 25–30 July 2010.
22. Ok, A.O. Automated detection of buildings from single VHR multispectral images using shadow information and graph cuts. *ISPRS J. Photogramm. Remote Sens.* **2013**, *86*, 21–40. [CrossRef]
23. Huang, Y.; Wu, Z.; Wang, L.; Tan, T. Feature Coding in Image Classification: A Comprehensive Study. *IEEE Trans. Softw. Eng.* **2013**, *36*, 493–506.
24. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision & Pattern Recognition, San Diego, CA, USA, 20–25 June 2005.
25. Fei-Fei, L.; Perona, P. A Bayesian hierarchical model for learning natural scene categories. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR05), San Diego, CA, USA, 20–25 June 2005.
26. Krizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012.
27. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
28. Uijlings, J.R.R.; van de Sande, K.E.A.; Gevers, T.; Smeulders, A.W.M. Selective Search for Object Recognition. *Int. J. Comput. Vis.* **2013**, *104*, 154–171. [CrossRef]
29. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Washington, DC, USA, 7–13 December 2015; pp. 1440–1448.
30. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
31. He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 386–397. [CrossRef] [PubMed]
32. Lin, T.-Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
33. 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.
34. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
35. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
36. Bochkovskiy, A.; Wang, C.Y.; Liao, H. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
37. 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 ECCV 2016, Amsterdam, The Netherlands, 8–16 October 2016; Springer: Cham, Switzerland, 2016.
38. Yi, J.; Wu, P.; Metaxas, D.N. ASSD: Attentive single shot multibox detector. *Comput. Vis. Image Underst.* **2019**, *189*, 102827. [CrossRef]
39. Shen, Z.; Liu, Z.; Li, J.; Jiang, Y.-G.; Chen, Y.; Xue, X. DSOD: Learning Deeply Supervised Object Detectors from Scratch. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1937–1945.

40. Fu, C.Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. DSSD: Deconvolutional Single Shot Detector. *arXiv* **2017**, arXiv:1701.06659.
41. Li, Z.; Zhou, F. FSSD: Feature fusion single shot multibox detector. *arXiv*, 0096.
42. Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; Li, S.Z. Single-Shot Refinement Neural Network for Object Detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4203–4212.
43. Zhang, S.; Wen, L.; Lei, Z.; Li, S.Z. RefineDet++: Single-Shot Refinement Neural Network for Object Detection. *IEEE Trans Circuits Syst Video Technol* **2021**, *31*, 674–687. [CrossRef]
44. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef]
45. Jeong, J.; Park, H.; Kwak, N. Enhancement of SSD by concatenating feature maps for object detection. In Proceedings of the British Machine Vision Conference 2017, London, UK, 4–7 September 2017.
46. Liu, S.; Huang, D.; Wang, Y. Receptive Field Block Net for Accurate and Fast Object Detection. In Proceedings of the Computer Vision—ECCV 2018, 15th European Conference, Munich, Germany, 8–14 September 2018; Lecture Notes in Computer Science. Springer: Cham, Switzerland, 2018; pp. 404–419.
47. Law, H.; Deng, J. CornerNet: Detecting Objects as Paired Keypoints. In Proceedings of the Computer Vision—ECCV 2018, 15th European Conference, Munich, Germany, 8–14 September 2018; Lecture Notes in Computer Science. Springer Verlag: Berlin, Germany, 2018; pp. 765–781.
48. 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, 16–20 June 2019; pp. 850–859.
49. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. CenterNet: Keypoint Triplets for Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 6568–6577.
50. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 9626–9635.
51. Simonyan, K.; Zisserman, A.J.a.p.a. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
52. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
53. Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; Yang, Y. Random Erasing Data Augmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 34.
54. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
55. Wu, X.; Li, W.; Hong, D.; Tian, J.; Tao, R.; Du, Q. Vehicle detection of multi-source remote sensing data using active fine-tuning network. *ISPRS J. Photogramm. Remote Sens.* **2020**, *167*, 39–53. [CrossRef]
56. Van Etten, A. You only look twice: Rapid multi-scale object detection in satellite imagery. *arXiv* **2018**, arXiv:1805.09512.
57. Deng, Z.; Sun, H.; Zhou, S.; Zhao, J.; Lei, L.; Zou, H. Multi-scale object detection in remote sensing imagery with convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 3–22. [CrossRef]
58. Han, W.; Kuerban, A.; Yang, Y.; Huang, Z.; Liu, B.; Gao, J. Multi-Vision Network for Accurate and Real-Time Small Object Detection in Optical Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 6001205. [CrossRef]
59. Sharma, M.; Dhanaraj, M.; Karnam, S.; Chachlakis, D.G.; Ptucha, R.; Markopoulos, P.P.; Saber, E. YOLOrs: Object Detection in Multimodal Remote Sensing Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 1497–1508. [CrossRef]
60. Wu, Y.; Zhao, W.; Zhang, R.; Jiang, F. AMR-Net: Arbitrary-Oriented Ship Detection Using Attention Module, Multi-Scale Feature Fusion and Rotation Pseudo-Label. *IEEE Access* **2021**, *9*, 68208–68222. [CrossRef]
61. Hua, X.; Wang, X.; Rui, T.; Zhang, H.; Wang, D. A fast self-attention cascaded network for object detection in large scene remote sensing images. *Appl. Soft Comput.* **2020**, *94*, 106495. [CrossRef]
62. Pang, J.; Li, C.; Shi, J.; Xu, Z.; Feng, H. R^2-CNN: Fast Tiny Object Detection in Large-Scale Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5512–5524. [CrossRef]
63. Wang, Y.; Dong, Z.; Zhu, Y. Multiscale Block Fusion Object Detection Method for Large-Scale High-Resolution Remote Sensing Imagery. *IEEE Access* **2019**, *7*, 99530–99539. [CrossRef]
64. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
65. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.
66. Zeiler, M.D.; Fergus, R. Visualizing and Understanding Convolutional Networks. In *European Conference on Computer Vision, Proceedings of the 13th European Conference, Zurich, Switzerland, 6–12 September 2014*; Springer: Cham, Switzerland, 2014; pp. 818–833.
67. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.

68. Zitnick, C.L.; Dollár, P. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision, Proceedings of the 13th European Conference, Zurich, Switzerland, 6–12 September 2014*; Springer: Cham, Switzerland, 2014; pp. 391–405.
69. Neubeck, A.; Van Gool, L. Efficient non-maximum suppression. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR06), Hong Kong, China, 20–24 August 2006; pp. 850–855.
70. Shrivastava, A.; Gupta, A.; Girshick, R. Training region-based object detectors with online hard example mining. In Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 761–769.
71. Shi, L.; Kuang, L.; Xu, X.; Pan, B.; Shi, Z. CANet: Centerness-Aware Network for Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5603613. [CrossRef]
72. Wang, J.; Wang, Y.; Wu, Y.; Zhang, K.; Wang, Q. FRPNet: A Feature-Reflowing Pyramid Network for Object Detection of Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 8004405. [CrossRef]
73. Shi, G.; Zhang, J.; Liu, J.; Zhang, C.; Zhou, C.; Yang, S. Global Context-Augmented Objection Detection in VHR Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 10604–10617. [CrossRef]
74. Chen, L.; Liu, C.; Chang, F.; Li, S.; Nie, Z. Adaptive multi-level feature fusion and attention-based network for arbitrary-oriented object detection in remote sensing imagery. *Neurocomputing* **2021**, *451*, 67–80. [CrossRef]
75. 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]
76. Zhang, S.; Mu, X.; Kou, G.; Zhao, J. Object Detection Based on Efficient Multiscale Auto-Inference in Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 1650–1654. [CrossRef]
77. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef]
78. 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]
79. Chen, J.; Wan, L.; Zhu, J.; Xu, G.; Deng, M. Multi-Scale Spatial and Channel-wise Attention for Improving Object Detection in Remote Sensing Imagery. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 681–685. [CrossRef]
80. Tian, Z.; Zhan, R.; Hu, J.; Wang, W.; He, Z.; Zhuang, Z. Generating Anchor Boxes Based on Attention Mechanism for Object Detection in Remote Sensing Images. *Remote Sens.* **2020**, *12*, 2416. [CrossRef]
81. Zhang, Y.; Yuan, Y.; Feng, Y.; Lu, X. Hierarchical and Robust Convolutional Neural Network for Very High-Resolution Remote Sensing Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5535–5548. [CrossRef]
82. Deng, Z.; Lin, L.; Hao, S.; Zou, H.; Zhao, J. An enhanced deep convolutional neural network for densely packed objects detection in remote sensing images. In Proceedings of the 2017 International Workshop on Remote Sensing with Intelligent Processing (RSIP), Shanghai, China, 18–21 May 2017.
83. Ding, P.; Zhang, Y.; Deng, W.-J.; Jia, P.; Kuijper, A. A light and faster regional convolutional neural network for object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2018**, *141*, 208–218. [CrossRef]
84. 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]
85. 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]
86. Liu, E.; Zheng, Y.; Pan, B.; Xu, X.; Shi, Z. DCL-Net: Augmenting the Capability of Classification and Localization for Remote Sensing Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 7933–7944. [CrossRef]
87. Yuan, Z.; Liu, Z.; Zhu, C.; Qi, J.; Zhao, D. Object Detection in Remote Sensing Images via Multi-Feature Pyramid Network with Receptive Field Block. *Remote Sens.* **2021**, *13*, 862. [CrossRef]
88. Li, Y.; Huang, Q.; Pei, X.; Chen, Y.; Jiao, L.; Shang, R. Cross-Layer Attention Network for Small Object Detection in Remote Sensing Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 2148–2161. [CrossRef]
89. Li, C.; Cong, R.; Guo, C.; Li, H.; Zhang, C.; Zheng, F.; Zhao, Y. A parallel down-up fusion network for salient object detection in optical remote sensing images. *Neurocomputing* **2020**, *415*, 411–420. [CrossRef]
90. Zhang, W.; Jiao, L.; Li, Y.; Huang, Z.; Wang, H. Laplacian Feature Pyramid Network for Object Detection in VHR Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5604114. [CrossRef]
91. Li, K.; Cheng, G.; Bu, S.; You, X. Rotation-Insensitive and Context-Augmented Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2337–2348. [CrossRef]
92. Gong, Y.; Xiao, Z.; Tan, X.; Sui, H.; Xu, C.; Duan, H.; Li, D. Context-Aware Convolutional Neural Network for Object Detection in VHR Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 34–44. [CrossRef]
93. Liu, Q.; Xiang, X.; Yang, Z.; Hu, Y.; Hong, Y. Arbitrary Direction Ship Detection in Remote-Sensing Images Based on Multitask Learning and Multiregion Feature Fusion. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 1553–1564. [CrossRef]
94. Wang, G.; Zhuang, Y.; Wang, Z.; Chen, H.; Shi, H.; Chen, L. Spatial Enhanced-SSD For Multiclass Object Detection in Remote Sensing Images. In Proceedings of the IGARSS 2019–2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 318–321.
95. Wang, P.; Sun, X.; Diao, W.; Fu, K. FMSSD: Feature-Merged Single-Shot Detection for Multiscale Objects in Large-Scale Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 3377–3390. [CrossRef]

96. Wang, G.; Zhuang, Y.; Chen, H.; Liu, X.; Zhang, T.; Li, L.; Dong, S.; Sang, Q. FSoD-Net: Full-Scale Object Detection From Optical Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5602918. [CrossRef]
97. Dong, Z.; Wang, M.; Wang, Y.; Zhu, Y.; Zhang, Z. Object Detection in High Resolution Remote Sensing Imagery Based on Convolutional Neural Networks With Suitable Object Scale Features. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 2104–2114. [CrossRef]
98. Yu, Y.; Guan, H.; Li, D.; Gu, T.; Tang, E.; Li, A. Orientation guided anchoring for geospatial object detection from remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2020**, *160*, 67–82. [CrossRef]
99. Wang, P.; Niu, Y.; Xiong, R.; Ma, F.; Zhang, C. DGANet: Dynamic Gradient Adjustment Anchor-Free Object Detection in Optical Remote Sensing Images. *Remote Sens.* **2021**, *13*, 1642. [CrossRef]
100. Huang, Z.; Li, W.; Xia, X.-G.; Wang, H.; Jie, F.; Tao, R. LO-Det: Lightweight Oriented Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5603515. [CrossRef]
101. Cui, Z.; Leng, J.; Liu, Y.; Zhang, T.; Quan, P.; Zhao, W. SKNet: Detecting Rotated Ships as Keypoints in Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 8826–8840. [CrossRef]
102. Shi, F.; Zhang, T.; Zhang, T. Orientation-Aware Vehicle Detection in Aerial Images via an Anchor-Free Object Detection Approach. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 5221–5233. [CrossRef]
103. Liu, S.; Zhang, L.; Lu, H.; He, Y. Center-Boundary Dual Attention for Oriented Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5603914. [CrossRef]
104. 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]
105. Huang, H.; Huo, C.; Wei, F.; Pan, C. Rotation and Scale-Invariant Object Detector for High Resolution Optical Remote Sensing Images. In Proceedings of the IGARSS 2019–2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 1386–1389.
106. Ma, J.; Shao, W.; Ye, H.; Wang, L.; Wang, H.; Zheng, Y.; Xue, X. Arbitrary-Oriented Scene Text Detection via Rotation Proposals. *IEEE Trans Multimed.* **2018**, *20*, 3111–3122. [CrossRef]
107. Bao, S.; Zhong, X.; Zhu, R.; Zhang, X.; Li, Z.; Li, M. Single Shot Anchor Refinement Network for Oriented Object Detection in Optical Remote Sensing Imagery. *IEEE Access* **2019**, *7*, 87150–87161. [CrossRef]
108. Xiao, Z.; Wang, K.; Wan, Q.; Tan, X.; Xu, C.; Xia, F. A2S-Det: Efficiency Anchor Matching in Aerial Image Oriented Object Detection. *Remote Sens.* **2020**, *13*, 73. [CrossRef]
109. Ye, X.; Xiong, F.; Lu, J.; Zhou, J.; Qian, Y. $\mathcal{F}^3$-Net: Feature Fusion and Filtration Network for Object Detection in Optical Remote Sensing Images. *Remote Sens.* **2020**, *12*, 4027. [CrossRef]
110. 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]
111. Shermeyer, J.; Van Etten, A. The effects of super-resolution on object detection performance in satellite imagery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019.
112. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*; NeurIPS: Vancouver, BC, Canada, 2014; Volume 27.
113. Mostofa, M.; Ferdous, S.N.; Riggan, B.S.; Nasrabadi, N.M. Joint-SRVDNet: Joint Super Resolution and Vehicle Detection Network. *IEEE Access* **2020**, *8*, 82306–82319. [CrossRef]
114. Bai, Y.; Zhang, Y.; Ding, M.; Ghanem, B. SOD-MTGAN: Small Object Detection via Multi-Task Generative Adversarial Network. In Proceedings of the Computer Vision—ECCV 2018, 15th European Conference, Munich, Germany, 8–14 September 2018; Lecture Notes in Computer Science. Springer: Cham, Switzerland, 2018; pp. 210–226.
115. Rabbi, J.; Ray, N.; Schubert, M.; Chowdhury, S.; Chao, D. Small-Object Detection in Remote Sensing Images with End-to-End Edge-Enhanced GAN and Object Detector Network. *Remote Sens.* **2020**, *12*, 1432. [CrossRef]
116. Zhu, J.-Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
117. Courtrai, L.; Pham, M.-T.; Lefèvre, S. Small Object Detection in Remote Sensing Images Based on Super-Resolution with Auxiliary Generative Adversarial Networks. *Remote Sens.* **2020**, *12*, 3152. [CrossRef]
118. Bashir, S.M.A.; Wang, Y. Small Object Detection in Remote Sensing Images with Residual Feature Aggregation-Based Super-Resolution and Object Detector Network. *Remote Sens.* **2021**, *13*, 1854. [CrossRef]
119. Ji, H.; Gao, Z.; Mei, T.; Ramesh, B. Vehicle Detection in Remote Sensing Images Leveraging on Simultaneous Super-Resolution. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 676–680. [CrossRef]
120. Gao, P.; Tian, T.; Li, L.; Ma, J.; Tian, J. DE-CycleGAN: An Object Enhancement Network for Weak Vehicle Detection in Satellite Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 3403–3414. [CrossRef]
121. Liu, W.; Luo, B.; Liu, J. Synthetic Data Augmentation Using Multiscale Attention CycleGAN for Aircraft Detection in Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 4009205. [CrossRef]
122. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30.

123. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *European Conference on Computer Vision, Proceedings of the Comput Vis ECCV 2020, 16th European Conference, Glasgow, UK, 23–28 August 2020*; Springer: Cham, Switzerland, 2020; pp. 213–229.
124. Zheng, Y.; Sun, P.; Zhou, Z.; Xu, W.; Ren, Q. ADT-Det: Adaptive Dynamic Refined Single-Stage Transformer Detector for Arbitrary-Oriented Object Detection in Satellite Optical Imagery. *Remote Sens.* **2021**, *13*, 2623. [CrossRef]
125. Li, Q.; Chen, Y.; Zeng, Y. Transformer with Transfer CNN for Remote-Sensing-Image Object Detection. *Remote Sens.* **2022**, *14*, 984. [CrossRef]
126. Zhang, Y.; Liu, X.; Wa, S.; Chen, S.; Ma, Q. GANsformer: A Detection Network for Aerial Images with High Performance Combining Convolutional Network and Transformer. *Remote Sens.* **2022**, *14*, 923. [CrossRef]
127. Xu, X.; Feng, Z.; Cao, C.; Li, M.; Wu, J.; Wu, Z.; Shang, Y.; Ye, S. An Improved Swin Transformer-Based Model for Remote Sensing Object Detection and Instance Segmentation. *Remote Sens.* **2021**, *13*, 4779. [CrossRef]
128. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 2778–2788.
129. Ma, T.; Mao, M.; Zheng, H.; Gao, P.; Wang, X.; Han, S.; Ding, E.; Zhang, B.; Doermann, D. Oriented object detection with transformer. *arXiv* **2021**, arXiv:2106.03146.
130. Zhou, Z.-H. A brief introduction to weakly supervised learning. *Natl. Sci. Rev.* **2018**, *5*, 44–53. [CrossRef]
131. Zhong, Y.; Zheng, Z.; Ma, A.; Lu, X.; Zhang, L. COLOR: Cycling, Offline Learning, and Online Representation Framework for Airport and Airplane Detection Using GF-2 Satellite Images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 8438–8449. [CrossRef]
132. 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]
133. Li, X.; Deng, J.; Fang, Y. Few-Shot Object Detection on Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5601614. [CrossRef]
134. Bilen, H.; Vedaldi, A. Weakly Supervised Deep Detection Networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2846–2854.
135. Tang, P.; Wang, X.; Bai, X.; Liu, W. Multiple Instance Detection Network with Online Instance Classifier Refinement. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3059–3067.
136. Tang, P.; Wang, X.; Bai, S.; Shen, W.; Bai, X.; Liu, W.; Yuille, A. PCL: Proposal Cluster Learning for Weakly Supervised Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 176–191. [CrossRef] [PubMed]
137. Wan, F.; Liu, C.; Ke, W.; Ji, X.; Jiao, J.; Ye, Q. C-mil: Continuation multiple instance learning for weakly supervised object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2199–2208.
138. Kantorov, V.; Oquab, M.; Cho, M.; Laptev, I. ContextLocNet: Context-Aware Deep Network Models for Weakly Supervised Localization. In *European Conference on Computer Vision, Proceedings of the 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Springer: Cham, Switzerland, 2016; pp. 350–365.
139. Zhang, M.; Zeng, B. A progressive learning framework based on single-instance annotation for weakly supervised object detection. *Comput. Vis. Image Underst.* **2020**, *193*, 102903. [CrossRef]
140. Yi, S.; Ma, H.; Li, X.; Wang, Y. WSODPB: Weakly supervised object detection with PCSNet and box regression module. *Neurocomputing* **2020**, *418*, 232–240. [CrossRef]
141. Feng, X.; Han, J.; Yao, X.; Cheng, G. TCANet: Triple Context-Aware Network for Weakly Supervised Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 6946–6955. [CrossRef]
142. Feng, X.; Han, J.; Yao, X.; Cheng, G. Progressive Contextual Instance Refinement for Weakly Supervised Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 8002–8012. [CrossRef]
143. Yao, X.; Feng, X.; Han, J.; Cheng, G.; Guo, L. Automatic Weakly Supervised Object Detection From High Spatial Resolution Remote Sensing Images via Dynamic Curriculum Learning. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 675–685. [CrossRef]
144. Chen, S.; Shao, D.; Shu, X.; Zhang, C.; Wang, J. FCC-Net: A Full-Coverage Collaborative Network for Weakly Supervised Remote Sensing Object Detection. *Electronics* **2020**, *9*, 1356. [CrossRef]
145. Li, Y.; Zhang, Y.; Huang, X.; Yuille, A.L. Deep networks under scene-level supervision for multi-class geospatial object detection from remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2018**, *146*, 182–196. [CrossRef]
146. Wu, Z.-Z.; Weise, T.; Wang, Y.; Wang, Y. Convolutional neural network based weakly supervised learning for aircraft detection from remote sensing image. *IEEE Access* **2020**, *8*, 158097–158106. [CrossRef]
147. Zhang, F.; Du, B.; Zhang, L.; Xu, M. Weakly Supervised Learning Based on Coupled Convolutional Neural Networks for Aircraft Detection. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 5553–5563. [CrossRef]
148. Han, J.; Zhang, D.; Cheng, G.; Guo, L.; Ren, J. Object Detection in Optical Remote Sensing Images Based on Weakly Supervised Learning and High-Level Feature Learning. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 3325–3337. [CrossRef]
149. Li, Y.; He, B.; Melgani, F.; Long, T. Point-Based Weakly Supervised Learning for Object Detection in High Spatial Resolution Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 5361–5371. [CrossRef]

150. Zheng, Z.; Zhong, Y.; Ma, A.; Han, X.; Zhao, J.; Liu, Y.; Zhang, L. HyNet: Hyper-scale object detection network framework for multiple spatial resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2020**, *166*, 1–14. [CrossRef]
151. Li, Q.; Mou, L.; Liu, Q.; Wang, Y.; Zhu, X.X. HSF-Net: Multiscale Deep Feature Embedding for Ship Detection in Optical Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 7147–7161. [CrossRef]
152. Dong, R.; Xu, D.; Zhao, J.; Jiao, L.; An, J. Sig-NMS-Based Faster R-CNN Combining Transfer Learning for Small Target Detection in VHR Optical Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 8534–8545. [CrossRef]
153. Li, S.; Xu, Y.; Zhu, M.; Ma, S.; Tang, H. Remote Sensing Airport Detection Based on End-to-End Deep Transferable Convolutional Neural Networks. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1640–1644. [CrossRef]
154. Zhong, Y.; Han, X.; Zhang, L. Multi-class geospatial object detection based on a position-sensitive balancing framework for high spatial resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2018**, *138*, 281–294. [CrossRef]
155. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
156. Li, Y.; Mao, H.; Liu, R.; Pei, X.; Jiao, L.; Shang, R. A Lightweight Keypoint-Based Oriented Object Detection of Remote Sensing Images. *Remote Sens.* **2021**, *13*, 2459. [CrossRef]
157. Liu, B.-Y.; Chen, H.-X.; Huang, Z.; Liu, X.; Yang, Y.-Z. ZoomInNet: A Novel Small Object Detector in Drone Images with Cross-Scale Knowledge Distillation. *Remote Sens.* **2021**, *13*, 1198. [CrossRef]
158. Zhang, Y.; Yan, Z.; Sun, X.; Diao, W.; Fu, K.; Wang, L. Learning Efficient and Accurate Detectors with Dynamic Knowledge Distillation in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5613819. [CrossRef]
159. Chen, J.; Wang, S.; Chen, L.; Cai, H.; Qian, Y. Incremental Detection of Remote Sensing Objects with Feature Pyramid and Knowledge Distillation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5600413. [CrossRef]
160. Lei, J.; Luo, X.; Fang, L.; Wang, M.; Gu, Y. Region-enhanced convolutional neural network for object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 5693–5702. [CrossRef]
161. Wu, Q.; Feng, D.; Cao, C.; Zeng, X.; Feng, Z.; Wu, J.; Huang, Z. Improved Mask R-CNN for Aircraft Detection in Remote Sensing Images. *Sensors* **2021**, *21*, 2618. [CrossRef]
162. 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]
163. 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, 18–22 June 2018; pp. 3974–3983.
164. Zhu, H.; Chen, X.; Dai, W.; Fu, K.; Ye, Q.; Jiao, J. Orientation robust object detection in aerial images using deep convolutional neural network. In Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP), Quebec City, QC, Canada, 27–30 September 2015; pp. 3735–3739.
165. Liu, Z.; Wang, H.; Weng, L.; Yang, Y. Ship Rotated Bounding Box Space for Ship Extraction From High-Resolution Optical Satellite Images With Complex Backgrounds. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 1074–1078. [CrossRef]
166. 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]
167. 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]
168. Zou, Z.; Shi, Z. Random Access Memories: A New Paradigm for Target Detection in High Resolution Aerial Remote Sensing Images. *IEEE Trans. Image Process.* **2018**, *27*, 1100–1111. [CrossRef] [PubMed]
169. Wang, J.; Yang, W.; Guo, H.; Zhang, R.; Xia, G.-S. Tiny Object Detection in Aerial Images. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 3791–3798.
170. Razakarivony, S.; Jurie, F. Vehicle detection in aerial imagery: A small target detection benchmark. *J. Vis. Commun. Image Represent.* **2016**, *34*, 187–203. [CrossRef]