remote sensing

MDPI

Review

Weakly Supervised Object Detection for Remote Sensing Images: A Survey

Corrado Fasana †, Samuele Pasini †, Federico Milani * and Piero Fraternali

Department of Electronics Information and Bioengineering, Politecnico di Milano, 20133 Milano, Italy
* Correspondence: federico.milani@polimi.it
† These authors contributed equally to this work.

Abstract: The rapid development of remote sensing technologies and the availability of many satellite and aerial sensors have boosted the collection of large volumes of high-resolution images, promoting progress in a wide range of applications. As a consequence, Object detection (OD) in aerial images has gained much interest in the last few years. However, the development of object detectors requires a massive amount of carefully labeled data. Since annotating datasets is very time-consuming and may require expert knowledge, a consistent number of weakly supervised object localization (WSOL) and detection (WSOD) methods have been developed. These approaches exploit only coarse-grained metadata, typically whole image labels, to train object detectors. However, many challenges remain open due to the missing location information in the training process of WSOD approaches and to the complexity of remote sensing images. Furthermore, methods studied for natural images may not be directly applicable to remote sensing images (RSI) and may require carefully designed adaptations. This work provides a comprehensive survey of the recent achievements of remote sensing weakly supervised object detection (RSWSOD). An analysis of the challenges related to RSWSOD is presented, the advanced techniques developed to improve WSOD are summarized, the available benchmarking datasets are described and a discussion of future directions of RSWSOD research is provided.

Keywords: weakly supervised object detection (WSOD); remote sensing; satellite images; aerial imagery; survey

check for updates

Citation: Fasana, C.; Pasini, S.; Milani, F.; Fraternali, P. Weakly Supervised Object Detection for Remote Sensing Images: A Survey. *Remote Sens.* **2022**, *14*, 5362. https://doi.org/10.3390/rs14215362

Academic Editors: Riccardo Roncella and Ricardo Torres

Received: 9 September 2022
Accepted: 20 October 2022
Published: 26 October 2022

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

cc BY

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

The availability of vast collections of aerial images has boosted the interest in methods for extracting information at various levels of detail, such as image classification [1], object detection [2], and instance segmentation [3]. In particular, Object Detection (OD) is one of the most challenging tasks of Computer Vision (CV) and has received significant attention over the last few years. Today's state-of-the-art object detectors can achieve outstanding performance under a fully supervised setting for natural images [2]. However, Fully Supervised Object Detection (FSOD) methods suffer from two major limitations:

- **Annotation effort**: the process of producing Bounding Box (BB) annotations, i.e., of delineating the object boundaries to provide the metadata necessary for the full supervision, is very time-consuming and non-trivial.
- **Domain generalizability**: most detectors have been developed to deal with natural images, i.e., images that portray one or more instances of common objects. However, other types of images (such as medical and aerial images) have more complex content, are less easy to collect and annotate, and may induce a drop in performance if domain-specific issues are not addressed (e.g., class imbalance, label noise, heterogeneous organs, and lesion appearance) [4,5].

Such drawbacks are critical when dealing with domain-specific applications such as earth observation and environmental monitoring, in which large collections of Remote Sensing Images (RSIs) acquired from satellites or other airborne sensors at different scales and resolutions are publicly available. Still, annotating such images is very time-consuming and challenging: the targets usually occupy a small portion of the whole image, which makes the annotation task particularly hard. Objects may be occluded, thus leading to less precise annotations. Regions and objects of interest can be highly domain-specific, e.g., in tasks such as illegal landfill detection [6], which requires expert knowledge for target identification and localization. As a consequence, general-purpose crowdsourcing platforms, which have been used to annotate popular natural image datasets such as ImageNet [7], cannot be exploited for RSI annotation.

Figures 1 and 2 illustrate the fundamental difference between RSIs and natural images. In RSIs:

- Objects normally occupy a small portion of the image, while in natural images few large objects are usually present.
- The background is complex and cluttered and multiple target objects coexist.
- Some target objects (e.g., ships and vehicles) can be extremely small and dense, while some other targets (e.g., ground track fields) can cover a large area.
- The target objects can have arbitrary orientations, whereas they often appear with horizontal orientation in natural images.
- The target objects are seen from an aerial viewpoint, whereas in natural images their profile is visible.
- The target objects may have high intra-class diversity (e.g., vehicles or aircraft of different shape, size, etc.) and inter-class similarity (e.g., a landfill vs. a quarry).

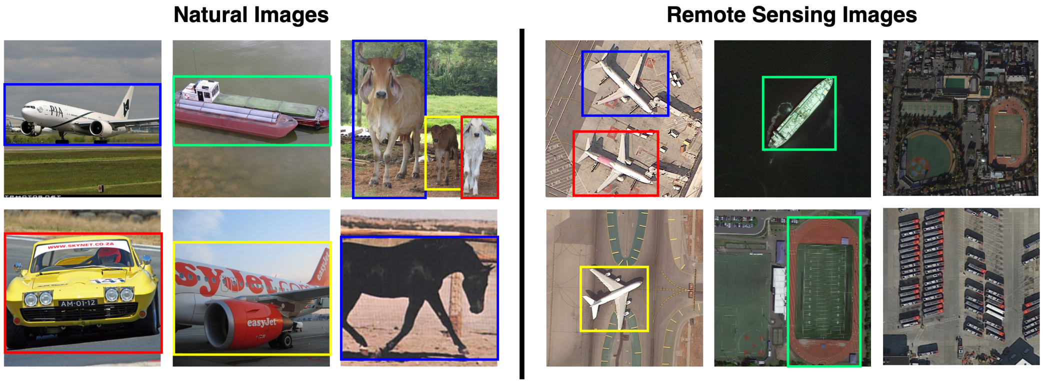

Figure 1. Examples of natural images from the PASCAL VOC dataset [8] and of RSIs from the DIOR [9] and DOTA [10] datasets. It is possible to observe that natural images usually contain few large objects while RSIs contain multi-scale, arbitrarily oriented objects with diverse spatial arrangements.

Unsurprisingly, models learned from natural images are hardly transferable to the remote sensing (RS) domain [9]. In addition, the difficulty of manually creating object bounding boxes hinders the production of effective FSOD methods relying only on fully annotated RSIs.

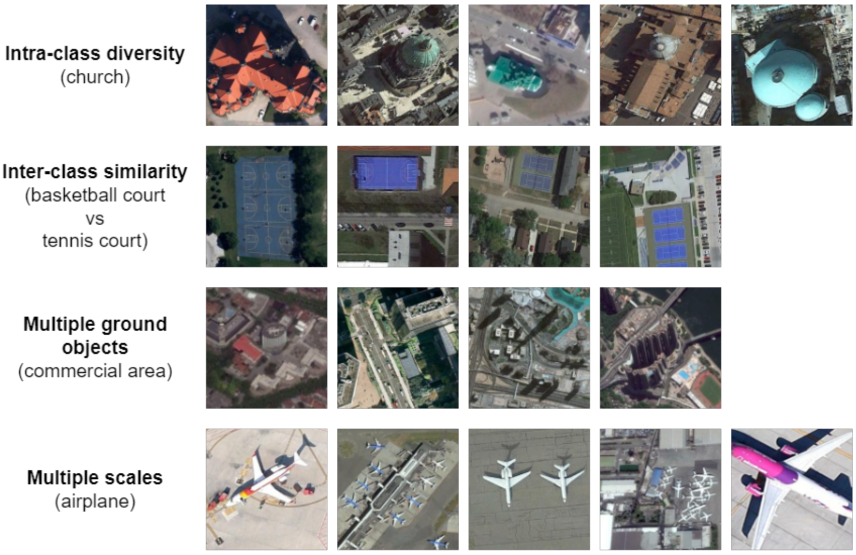

Figure 2. RSI challenges on images from the NWPU-RESISC45 dataset [11].

To account for the lack of fine-grain annotations such as object bounding boxes, several object detection methods have been developed that leverage only coarse-grain annotations (especially image-level labels indicating only the presence or absence of an object) [12]. This approach is called inexact Weak Supervision (WS) and introduces a new branch of OD called weakly supervised object detection (WSOD). WSOD has already been effectively applied to natural images [12–14] but is still an open problem in the RS domain [15–18] due to the described challenges. To the best of our knowledge, there is not yet a comprehensive review of the application of WSOD methods for RSIs.

1.1. Focus of the Survey

This survey reviews the recent advancements in WSOD techniques in the RS domain. The focus is twofold: accurately describing the available techniques and comparing their performance on RSI datasets.

In recent years, the RS domain has been gaining more interest and novel methods have been continuously proposed. In particular, WSOD offers many possibilities for applications that otherwise would not be possible due to the large number of manual annotations required, e.g., vehicle detection [19–21], marine animal detection [22], or defective insulator detection [23]. We introduce a timeline to precisely define the evolution of RSWSOD methods and describe the pros and cons and the challenges of each state-of-the-art method. Yue et al. [24] present a brief general overview of weakly supervised learning on RSIs. The authors describe the main concepts and divide the surveyed techniques into three main categories: inexact supervision, inaccurate supervision, and incomplete supervision. Based on their categorization, this survey focuses on the inexact supervision class, because fine-grained annotations are the most difficult type of metadata to obtain for RSIs.

Training and evaluating neural models on the right data is essential for any ML-based application. The dataset comparison proposes an analysis of the most popular RSI datasets and concentrates on their characteristics and per-class performance. Since many RSWSOD methods are use-case specific and cannot be directly compared with other works (e.g., tree species classification [25,26]), Section 3.2.4 briefly describes the main achievements obtained by use-case-specific methods.

1.2. Methodology

The research target comprises methods that leverage coarse-grained labels (e.g., image-level or point-based) to address the WSOD task in the RS domain. Only the proposals that provide an OD solution under inexact supervision are reported. This perimeter excludes those contributions that address OD under incomplete supervision, inaccurate supervision, and no supervision. For this reason, all methods that exploit self-supervised learning, active learning, unsupervised learning, noisy labels, or crowdsourcing-based approaches are excluded from the search.

The corpus of the relevant research has been identified by following the PRISMA procedure [27] for systematic reviews. Figure 3 illustrates the adopted workflow.

1. The search was conducted on the Scopus database since it has been demonstrated to support bibliographic analysis better than other repositories [28]. The search phrases were composed as follows:

```
<search> :- <task> AND <domain>
<task> :- weak supervision | inexact supervision |
          weakly supervised | weakly supervised learning |
          weakly supervised deep learning |
          weakly supervised object detection
<domain> :- remote sensing | remote sensing images |
            earth observations | aerial images |
            synthetic aperture radar images |
            satellite images | multispectral images |
            hyperspectral images
```

   The search results were filtered to retain only contributions in journals, conferences, and workshops.
2. The initial corpus, composed of 528 works, was reduced by removing duplicates: 196 works were kept. Next, we identified and eliminated the studies unrelated to RSWSOD by checking each contribution's title, keywords, and abstract. The reduced corpus contained 42 contributions.
3. In the remaining corpus, the full text of 3 articles was unavailable. Thus, the corpus was reduced to 39 contributions.
4. A final eligibility filter was applied and the full text of the remaining articles was read. In particular, 2 articles were removed because the task was not OD, 3 articles were removed because they were not related to inexact supervision, and 1 article was excluded because it was a draft of another already considered work. This final step yielded the 33 works considered in this survey.

1.3. Contributions

The contributions of this paper can be summarized as follows:

- A total of 33 RSWSOD methods are identified from an initial corpus of 528 papers resulting from a keyword search.
- The most suitable dimensions for analyzing the methods are identified and described (year, approach, annotation type, proposal generation method, addressed challenges, and use case). The techniques are described and compared based on these dimensions.
- A list of the most used datasets for RSWSOD is provided and the methods are compared based on their performance.
- A list of open issues in the RSWSOD field and possible future research directions are identified and discussed.

The rest of the paper is organized as follows: Section 2 describes the main RSWSOD challenges and describes state-of-the-art RSWSOD architectures; Section 3 presents the most common RS datasets and compares the performances of the surveyed methods; Section 4 highlights the open issues and discusses the relevant research directions; and Section 5 draws the conclusions.

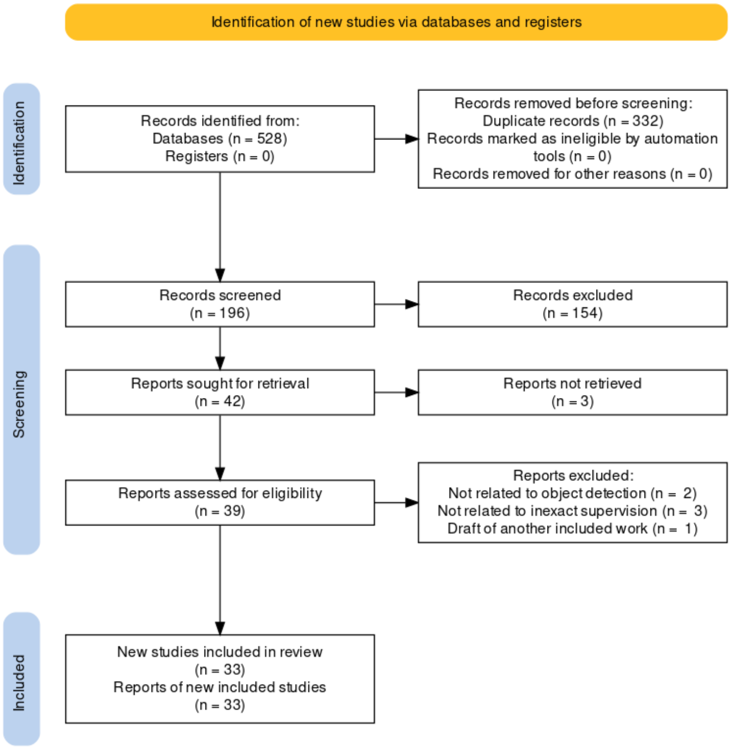
Identification of new studies via databases and registers
Identification
Records identified from:
Databases (n = 528)
Registers (n = 0)
Records removed before screening:
Duplicate records (n = 332)
Records marked as ineligible by automation tools (n = 0)
Records removed for other reasons (n = 0)
Screening
Records screened
(n = 196)
Records excluded
(n = 154)
Reports sought for retrieval
(n = 42)
Reports not retrieved
(n = 3)
Reports assessed for eligibility
(n = 39)
Reports excluded:
Not related to object detection (n = 2)
Not related to inexact supervision (n = 3)
Draft of another included work (n = 1)
Included
New studies included in review
(n = 33)
Reports of new included studies
(n = 33)

Figure 3. PRISMA flow diagram of the systematic review.

2. Remote Sensing Weakly Supervised Object Detection

Given an image, Remote Sensing Fully Supervised Object Detection (RSFSOD) aims to locate and classify objects based on BB annotations. Differently from RSFSOD, RSWSOD aims to precisely locate and classify object instances in RSIs using only image-level labels or other types of coarse-grained labels (e.g., points or scribbles) as ground truth (GT). Figure 4 presents an example. Due to the missing training information regarding the location of the objects, the performance gap between RSFSOD and RSWSOD is still large (≈30–40% mAP), even though several efforts have been made to improve the accuracy of RSWSOD techniques.

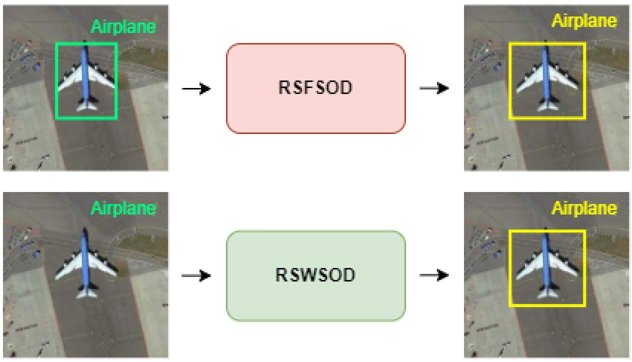

Figure 4. Visual explanation of RSFSOD and RSWSOD. In the top image, the green BB represents the ground-truth location information as given to the fully supervised detector. Such information is not present during the training of a weakly supervised detector, as shown in the bottom image.

This section presents the issues to be tackled in RSI OD, provides a taxonomy of the WSOD task, and categorizes and describes the 33 RSWSOD architectures under analysis. Table 1 reports a summary, indicating for each method: annotation type, approach category, proposal type, if it is use-case specific or generic, and the RSI challenges addressed. Proposal type indicates the technique used to generate areas of interest that are then processed to create bounding boxes.

Table 1. A summary of surveyed RSWSOD methods and of the main RSI challenges they address. The ✓symbol means that the work addresses the specific problem listed in the column. No check marks denote the works that study RSWSOD in general without focusing on specific problems.

Name	Year	Annotation Type	Approach	Proposals	Use Case-Specific	Partial Coverage	Multiple-Instance	Density	Speed	Generalizability
Zhang et al. [29]	2014	Image	TSI + TDL	Sb-SaS	Aircraft Detection					
Zhang et al. [30]	2014	Image	TSI + TDL	Sb-SaS	Generic					
Han et al. [31]	2014	Image	TSI + TDL	SW	Generic					
Cheng et al. [32]	2014	Image	TSI + TDL	SW	Generic					
Zhou et al. [33]	2015	Image	TSI + TDL	Sb-SaS	Generic					
Zhou et al. [34]	2016	Image	TSI + TDL	Sb-SaS	Generic					
LocNet [35]	2016	Image	TSI + TDL	RPN	Aircraft Detection				✓	
Cao et al. [36]	2017	Region	MIL	SW	Vehicle Detection					
MIRN [37]	2018	Region	MIL	n/d	Vehicle Detection		✓			
SLS [38]	2018	Scene	CAM	Heatmap	Generic				✓	✓
Du et al. [20]	2019	Image	TSI + TDL	CFAR	Generic					
WSA [39]	2019	Image	CAM	Heatmap	Aircraft Detection				✓	✓
Aygunes et al. [25]	2019	Image	MIL	SW	Tree-species classification					
FCC-Net [40]	2020	Image	MIL	SS	Generic					✓
DCL [41]	2020	Image	MIL	SS	Generic	✓				
PCIR [15]	2020	Image	MIL	SS	Generic	✓	✓			✓
AlexNet-WSL [42]	2020	Image	CAM	Heatmap	Aircraft Detection				✓	
Shi et al. [23]	2020	Coarse BB, Count	Other	RPN	Cap Missing Detection				✓	
TCANet [16]	2020	Image	MIL	SS	Generic	✓	✓	✓		
MPFP-Net [43]	2021	Image	MIL	Random	Generic	✓				✓
Aygunes et al. [26]	2021	Image	MIL	SW	Tree-species classification					
Sun et al. [44]	2021	HBB	Other	RPN	Generic				✓	✓
Wang et al. [45]	2021	Image	MIL + CAM	SS, Heatmap	Generic	✓				
Li et al. [46]	2021	Point	Other	SS	Generic	✓	✓			
MIGL [47]	2021	Image	MIL	SS	Generic	✓	✓			
Li et al. [48]	2021	Image, Count	Other	RPN	Terrain Feature Detection				✓	
SAENet [49]	2021	Image	MIL	SS	Generic	✓				
Berg et al. [22]	2022	Image	Other	Heatmap	Marine Animals Detection				✓	
Long et al. [50]	2022	Image	CAM	Heatmap	Generic				✓	✓
PistonNet [21]	2022	Image	Other	Heatmap	Ship Detection				✓	
SDA-RSOD [17]	2022	Image	CAM	Heatmap	Generic			✓	✓	✓
SPG + MELM [51]	2022	Image	MIL	RPN	Generic	✓			✓	
Qian et al. [18]	2022	Image	MIL	SS	Generic	✓				✓

2.1. Coarse-Grained Annotations

FSOD requires manual BB annotations, also referred as instance-level labels. Conversely, WSOD relies on coarse-grained annotations, i.e., all the types of labels considered less expensive to obtain than BBs.

The most common types of annotation used to perform RSWSOD are *image-level* labels, indicating the presence of at least one instance of a target object class. It is also possible to use other metadata, such as *region-level* annotations, suggesting the presence of at least one instance of an object in a portion of the image. A less popular coarse-grained annotation is represented by *scene-level* labels that record only the class of the most dominant object in the image.

Another weak RSWSOD annotation is the *count* of the number of class instances in an image. To alleviate the gap between fully supervised and weakly supervised approaches, *point* annotations have also been exploited. The idea behind these annotations is that point labels are far cheaper to obtain than BBs [46] and they significantly increase the model performance. Still, fully supervised performance has not been matched by any weakly supervised method.

2.2. Main Challenges

In general, WSOD presents three main challenges related to the use of coarse-grained annotations [12]:

- **Partial coverage problem**: This may arise from the fact that the object detection proposals computed by the WSOD method with the highest confidence score are those that surround the most discriminative part of an instance. If proposals are selected solely based on the highest score, the detector will learn to focus only on the most discriminative parts and not on the entire extent of an object (discriminative region problem). Another problem may derive from proposal generation methods such as Selective Search [52] and Edge Boxes [53], which output proposals that may not cover the entire targets well, reducing the performances of the detector (low-quality proposal problem).
- **Multiple-instance problem**: The model may have trouble trying to accurately distinguish multiple instances when there are several objects of the same class. This is due to the fact that most detectors [13,14] select only the highest-scoring proposal of each class and ignore other relevant instances.
- **Efficiency problem**: Current proposal generators (e.g., Selective Search [52] and Edge Boxes, [53]) largely used in WSOD are very time-consuming.

The characteristics of the RSIs discussed in Section 1 introduce additional challenges:

- **Density problem**: Images often contain dense groups of instances belonging to the same class. Models usually have difficulties in accurately detecting and distinguishing all the instances in such densely populated regions.
- **Generalization problem**: The high intra-class diversity in RSIs induces generalization problems mainly due to three factors:
  - **Multi-scale**: Objects may have varying sizes, and their representation strongly depends on the image resolution and ground sample distance (GSD).
  - **Orientation variation**: Instances present arbitrary orientations and may require the use of methods generating Oriented bounding boxes (OBB) instead of the classical horizontal bounding boxes (HBB).
  - **Spatial complexity**: In general, RSIs show varying degrees of complexity in the spatial arrangement of the objects.

Table 1 overviews the challenges addressed by each surveyed method. Three issues are studied by most methods: partial coverage, speed, and generalizability. Partial coverage and speed are mainly addressed by specific types of approaches: MIL-based methods deal with partial coverage, while CAM-based methods deal with speed. This is connected with the specific characteristics of the architectures (discussed in Section 2.3), e.g., MIL-based techniques require the use of external proposal generators that may not cover the object completely, thus harming the learning process. Instead, generalizability is a more common issue across all approaches. The multiple-instance and density problems are closely related and are very common issues in RSIs but are studied by very few methods. An example of these two challenges is shown in Figure 1, where a parking lot contains more than 60 instances of the same class.

2.3. Weakly Supervised Object Detection Approaches

As reported by Zou et al. [54], in the past two decades, the progress of OD has gone through two periods: "traditional object detection period (before 2014)" and "deep-learning-based object detection period (after 2014)". Being more specific branches of OD, both WSOD and RSWSOD have gone through the same historical phases. Figure 5 presents a timeline of FSOD, WSOD, and RSWSOD with important milestones (indicated by a green flag) for each task. More specifically, during the traditional object detection period, most WSOD approaches were based on the usage of support vector machines (SVMs), the MIL framework [55,56], and the exploitation of low-level and mid-level handcrafted features (e.g., SIFT [57] and HOG [58]). These methods obtained promising results on natural images but were difficult to apply to RSIs due to the previously discussed difficulties. With the advent of DL, OD architectures became more powerful and obtained successful results in many fields but required a large amount of annotated data. For this reason, many researchers shifted their focus to weakly supervised approaches. In Figure 5, it is interesting to note that most RSWSOD methods have been developed after specific WSOD milestones: CAM, WSDDN, and OICR.

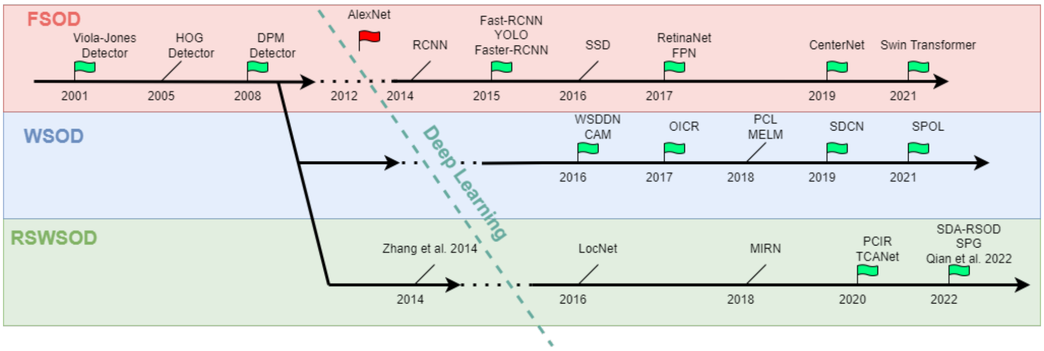

Figure 5. Timeline of the milestones in RSWSOD, with a comparison with FSOD [2,54] and WSOD [12] through the years. The flag symbol is used to represent milestones, while a simple line is used to denote other relevant methods. For the sake of clarity, not all methods have been reported [18,29].

Several approaches have been proposed to address the WSOD task. Four major categories can be identified depending on how the detector is trained:

- **TSI + TDL-based**: These approaches are based on a simple framework that consists of two stages: training set initialization (TSI) and target detector learning (TDL).
- **MIL-based**: These approaches are based on the Multiple Instance Learning (MIL) framework.
- **CAM-based**: These approaches are based on Class Activation Maps (CAMs), a well-known explainability technique.
- **Other DL-based**: Few methods reformulate the RSWSOD problem starting from the implicit results of other tasks, e.g., Anomaly Detection (AD).

This section provides a brief introduction to each category, describes the RSWSOD methods pertaining to them, and considers the addressed challenges.

2.3.1. TSI + TDL-Based

Before the advent of deep learning (DL), most object detectors were based on SVMs. The workflow behind these methods is to start by producing candidate proposals exploiting either a Sliding Window (SW) [31,32,36] or Saliency-based Self-adaptive Segmentation (Sb-SaS) [29,30,33,34] approach. SW generates proposals by sliding, on the entire image, multiple BBs with different scales while Sb-SaS produces saliency maps that measure the uniqueness of each pixel in the image and exploit a multi-threshold segmentation mechanism to produce BBs. Both methods try to deal with the variation in the target size and the resolution of the images. Each proposal is characterized using a set of low- and middle-level features derived from methods such as SIFT [57] and HOG [58]. The extracted features can be further manipulated to produce high-level ones. Then, sets of positive and negative candidates are chosen to initialize the training set. The training procedure is then composed of two steps: (1) the training of the detector and (2) the updating of the training set by modifying the positive and negative candidates. These steps are repeated until a stopping condition is met.

The first attempts to apply WSOD techniques to aerial images were performed by Zhang et al. [29,30] in 2014. The idea is to mine positive and negative samples to initialize the training set and then exploit an iterative training scheme to refine the detector and update the training set using a weakly supervised SVM. However, this method ignores some critical information, which could improve the detector performance, such as intra-class compactness and inter-class separability. For this reason, Han et al. [31] propose a probabilistic approach using the Bayesian rule [59] to jointly integrate saliency, intra-class compactness, and inter-class separability to better initialize the training set. This work also highlights the limitations of low- and mid-level feature extractors that are not powerful enough to effectively describe objects in RSIs due to the influence of the cluttered background. The authors propose the use of a Deep Boltzmann Machine (DBM) [60] to extract high-level features. All these methods focus on the problem of single-object detection. Cheng et al. [32] propose the Collection of Part Detectors (COPD) [19] composed of a set of weakly supervised SVM detectors to perform multi-object detection.

With the advent of convolutional neural networks (CNNs) [61], both WSOD and RSWSOD methods started to benefit from the powerful feature extraction capabilities of deep architectures. In 2015, Zhou et al. [33,34] proposed exploiting transfer learning on a CNN to extract high-level features to feed to an SVM-based detector. The authors further highlight the importance of the process used to select negative instances for training. Most previous methods select random negative samples, which may cause the deterioration or fluctuation of the performance during the iterative training procedure. The reason is that negative samples which are visually similar to positive samples tend to be easily misclassified. Thus, selecting ambiguous negative samples is fundamental to enhance the effectiveness and robustness of the classifier. The authors propose using negative bootstrapping instead of random selection for negative samples, building a more robust detector. This technique is still taken into consideration even in modern state-of-the-art methods.

Even though these methods improve over previous ones, they are still affected by two significant limitations. First, most of these techniques were proposed to address the task of single-object detection, but RSIs usually contain multiple instances and classes. Second, previous methods extracted proposals using either an SW or Sb-SaS approach, which are very time-consuming. In 2016, Zhang et al. [35] proposed the use of coupled CNNs that integrate a Region Proposal Network (RPN) [62] to perform aircraft detection more efficiently.

From Table 1 it can be seen that TSI + TDL-based methods do not focus on any specific RSI challenge. The first works aimed to propose working solutions for weakly supervised learning on RSIs while subsequent works concentrated on demonstrating the effectiveness of more powerful feature extractors.

Later on, researchers moved towards MIL-based and CAM-based methods thanks to the advancements in the DL field and the development of more powerful feature extractors and CV architectures.

2.3.2. MIL-Based

In MIL-based approaches, each image is treated as a collection of potential instances of the object to be found. Typically, MIL-based WSOD follows a three-step pipeline: proposal generation, feature extraction, and classification.

Proposal generation aims to extract a certain number of regions of interest, i.e., those areas that may contain object instances, from the image. This can be accomplished in several different ways, with the basic approach being Sliding Window. More advanced and efficient proposal generation methods have been proposed, such as Selective Search (SS) [52], which leverages the advantages of both exhaustive search and segmentation to generate initial proposals, or Edge Boxes (EB) [53], which uses object edges to generate proposals. These methods are built to have a high recall, so the generated candidates are very likely to contain an object instance. However, these methods are very time-consuming. To solve this issue, it is possible to either exploit CAM-based approaches in which there is no region proposal generation step or directly integrate the region proposal generation and feature extraction steps into the network using an RPN. The latter exploits CNNs and can extract more relevant features for the areas of interest and speed up the process. Despite their advantages, RPNs are not largely used in WSOD since traditional techniques have been proven to work well with natural images.

Feature extraction is needed to compute a feature vector for each candidate region extracted in the previous step. Features can be handcrafted or extracted by a CNN as in DL methods. **Classification** is the last step and performs WSOD by reformulating the problem as an MIL classification task. The MIL problem was first introduced in [55]. In image classification, each image is considered as a *bag*, containing a set of *feature vectors* to be classified (one for each region proposal). For the training step, each image (or bag) is assigned a positive or negative label based only on the image-level label, i.e., the presence or absence of a specific class. Thus, an image can be represented as a positive bag for one class, while a negative bag for another class not present inside such an image (Figure 6). The aim is to infer instance-level labels for the proposals inside each image.

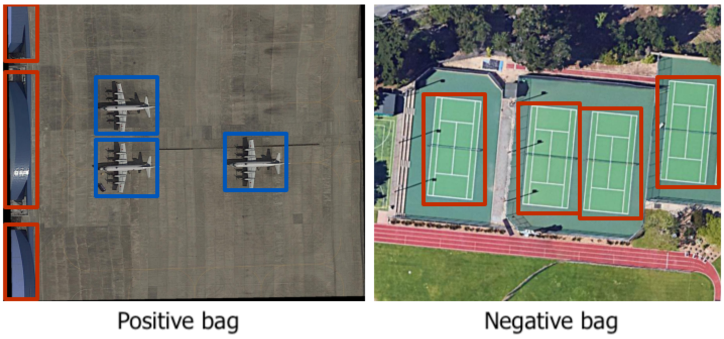

Figure 6. An example of positive/negative bags for the “airplane” class. Bounding boxes correspond to the proposals input to the network; blue indicates positive instances, while red indicates negative instances. MIL-based WSOD aims to differentiate between positive and negative instances based only on image-level labels.

In 2016, Bilen and Vedaldi proposed Weakly Supervised Deep Detection Network (WSDDN) [13], a MIL-based approach that can be considered a significant milestone for WSOD research. The idea behind WSDDN is a two-stream network that aims to perform both classification and localization. The classification branch computes the class score of each proposal, and the detection branch computes the contribution of each proposal to the image being classified as a specific class. These scores are then multiplied for each region and summed to obtain the final prediction score. Inspired by this work, one year later, Tang et al. proposed OICR [14], an improvement over WSDDN that tries to mitigate the discriminative region problem that characterizes the previous network by adding refinement branches. Many WSOD methods were developed based on OICR [63]. These methods focused on solving the limitations of the previous techniques, especially the discriminative region and the multiple-instance problems. The latter is essential because only the highest-scoring proposal was selected as a positive instance during training, ignoring that several instances may be present in the same image. Thus, a poor performance was obtained on multi-instance images. Shao et al. present an extensive review of WSOD methods on natural images [12].

The influence of WSDDN and OICR also affected the remote sensing community. However, the performance drop was severe. For this reason, many researchers focused on solving the RSWSOD problem by improving WSOD techniques and adding new modules that could overcome RSI challenges. Cao et al. [36] exploited MIL and density estimation to predict vehicles' locations starting from region-level labels. In 2018 (one year after OICR), Sheng et al. proposed MIRN [37], a MIL-based approach that tries to leverage the count information and an online labeling and refinement strategy, inspired by OICR, to perform vehicle detection, solving the multiple-instance problem.

Following the same direction, Feng et al. developed PCIR [15], an OICR-inspired method that tries to address the multiple-instance problem and the discriminative region problem. They exploit a context-based approach that diverts the focus of the detection network from the local distinct part to the whole object and further to other potential instances. PCIR alleviates the influence of negative samples induced by complex backgrounds by dynamically rejecting the negative training proposals. The same year, Feng et al. proposed another context-based method called TCANet [16], composed of two modules. The first module activates the features of the whole object by capturing the global visual scene context, alleviating the discriminative region problem. The second module captures the instance-level discriminative cues by leveraging the semantic discrepancy of the local context, thus distinguishing better adjacent instances and addressing the density problem. This network was further improved with the development of SAENet [49], which exploits an adversarial dropout–activation block to solve the discriminative region problem. The authors address the fact that most state-of-the-art methods ignore the consistency across different spatial transformations of the same image, causing them to be labeled differently.

In 2020, Yao et al. [41] observed that many current methods fail to provide high-quality initial samples, consequently deteriorating the detector performance. To solve this issue, the authors proposed considering the image difficulty while training using a dynamic curriculum learning strategy [64]. The reason behind the authors' work is that training the detector using curriculum learning, i.e., feeding training images with increasing difficulty that matches the current detection ability, improves the detector performance. This intuition was supported by the recent advances in WSOD [65,66]. Another interesting concept is considered by Chen et al. [40] for FCC-Net. They showed that training an RSWSOD and an RSFSOD network, alternatively, can improve the detector performance. In this case, the fully supervised ground truth is given by the refined BBs generated by the weakly supervised branch. This technique can also mitigate the multi-scale problem.

Wang et al. [47] proposed MIGL, an improved version of PCL [63], to find all possible instances based on the apparent similarity. This was achieved by exploiting clustering and a novel spatial graph voting strategy to identify high-quality objects, further alleviating the discriminative region problem. Other approaches have tackled the same problem. Shamsolmoali et al. [43] proposed a multi-patch feature pyramid network (MPFP-Net) trained using smooth loss functions based on the fact that the non-convexity of MIL is the major cause of the discriminative region problem. Using this network, the authors could also address the multi-scale problem.

Wang et al. [45] proposed an interesting MIL-based approach inspired by PCL to perform object detection. The key innovation is the proposal generation step. A novel pseudo-label generation (PLG) algorithm is developed combining Selective Search [52] with the information provided by a CAM-based weakly supervised localization model. This way, by intersecting the results of SS with those of the WSOL method, low-quality proposals can be effectively suppressed. This is a significant improvement because MIL-based approaches cannot produce high-quality detectors starting from low-quality proposals, as already observed in [41]. This is further pointed out by Cheng et al. [51] who propose SPG based on an RPN that exploits the objectness confidence score to generate high-quality proposals. The authors show that using the proposed RPN in place of standard techniques (e.g., Selective Search) can improve the performance of previous MIL-based methods such as OICR [14] and MELM [67]. The method is indicated as (SPG + MELM).

Recently, Qian et al. [18] considered image difficulty from a different perspective. The authors were the first to highlight that an imbalance between easy and hard samples causes the network to fail in detecting objects in the few available hard samples. To solve this challenge, they took advantage of the context information provided by a WS segmentation method [68] to evaluate the difficulty of each sample. They also exploited the idea of mining and regressing pseudo-GT BB to improve performance [69]. The authors also address the discriminative region problem.

Table 1 shows that MIL is the most widely used framework (≈42%) to solve RSWSOD. Most of the surveyed MIL methods are applied to generic scenarios, even though some specific applications have been studied (e.g., vehicle detection, tree species classification). Moreover, these approaches tend to focus on addressing the partial coverage, the multiple-instance, and the difficulty problems. Few works exist that address the density problem or the speed problem, which is a crucial point for real-time applications. Few MIL-based approaches do not address any of the described challenges. The reason is that either the work focuses on just reducing the effort of fully annotating datasets [36] or they are use-case-specific and thus solve more specific challenges [25,26]. Figure 7 presents a comparison of proposal generation techniques used in WSOD and RSWSOD. It can be noted that for both tasks, the most used method is Selective Search (≈55%), followed by Sliding Window (SW) and Edge Boxes (EB) which are employed by fewer works.

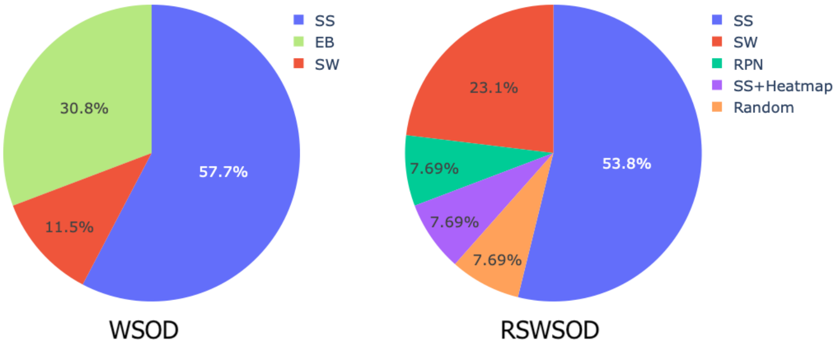

Figure 7. Type of proposals used in WSOD and RSWSOD MIL-based methods. SS stands for Selective Search, EB for Edge Boxes, SW for Sliding Window, and RPN for region proposal network. It is possible to notice that in both cases, Selective Search is the most used proposal generation technique. WSOD data are provided by [12].

2.3.3. CAM-Based

CAM-based approaches formulate the WSOD problem as a localizable feature map learning problem. The idea comes from the fact that every convolutional unit in the CNN is essentially an object detector and is able to locate the target object in the image [70]. For example, suppose the object appears in the upper left corner of the image; in that case, the upper left corner of a feature map after a convolutional layer will produce a greater response. These localization capabilities of CNNs have been further studied in other works such as [71,72].

Class activation maps [72] were introduced in 2016 as a weighted activation map (heatmap) showing the areas contributing the most to the classification. CAMs do not require any additional label or training and can be obtained from the last fully connected layer of a CNN. Bounding boxes can be produced by thresholding the CAM values. Figure 8 shows an example. After that, many different CAM variants and CAM-based methods were proposed for WSOD and especially for weakly supervised object localization (WSOL) [12].

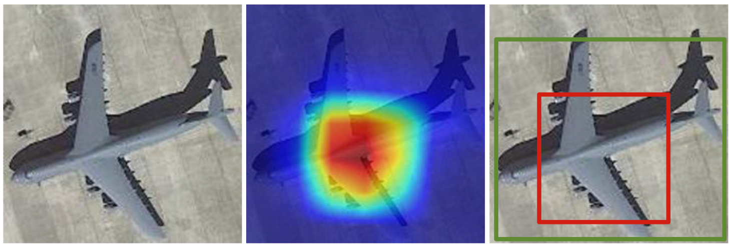

Figure 8. In the **middle**, an example of CAM for the "airplane" class obtained from the image on the **left**. On the **right**, the green BB is the ground truth, whereas the red BB is the one obtained by thresholding the CAM values.

In the remote sensing community, researchers have started to exploit CAM-based approaches for the task of aircraft detection. In 2018, Li et al. [38] proposed a Siamese network to overcome the fact that existing methods tend to take scenes as isolated ones and ignore the mutual cues between scene pairs when optimizing deep networks. Moreover, a multi-scale scene-sliding-voting strategy is implemented to produce the CAM and solve the multi-scale problem. The authors further propose different methods for thresholding the CAM and observe that detection results for each class have a strong dependence on the chosen thresholding method. Ji et al. [39] proposed a method to reduce the false detection rate that affects many aircraft detectors producing a more accurate attention map.

In 2020, Wu et al. [42] proposed AlexNet-WSL, an effective way of using CAMs for aircraft detection using a modified AlexNet CNN [61] as the backbone. In the following work [17], the authors state that it is difficult to adapt MIL to the complexity and diversity of image data in the real world and propose a CAM-based method (SDA-RSOD) based on two components: a divergent activation module and a similarity constraint module. These allow users to distinguish among adjacent instances and detect multiple instances, mitigating the density and multiple-instance problems. This is achieved by exploiting shallow CAMs extracted from the shallow layers of the network, and deep CAMs extracted from deep layers. The former can retain better location information while the latter allow a more accurate categorization. This work shows the importance of shallow features for RSIs [73].

A different approach has been proposed by Long et al. [50]. The authors define an RSWSOD method based on CAMs to perform object detection without extent. This means that objects are not identified using bounding boxes but instead by a single point (x, y). This type of localization has been rarely approached in the remote sensing community and further research needs to be pursued.

Table 1 shows that CAM-based methods are less used for RSWSOD (≈18%). This could be due to the fact that these methods work well when few big instances are present in the image. In fact, several CAM-based methods have been developed in the last few years, especially for the task of weakly supervised object localization (not part of the survey) [73–77]. It can also be observed that they are often used for tasks such as ship detection and airplane detection. Moreover, these methods can speed up the process of RSWSOD by avoiding the use of region proposal methods while only a couple of the analyzed works tried to address other RSI challenges.

2.3.4. Other DL-Based

In the literature, several studies present DL approaches that cannot be assigned to the TSI + TDL, MIL, or CAM categories.

A relevant characteristic of RSIs is that objects can have arbitrary orientations. However, in the majority of cases, horizontal bounding boxes (HBB) are used to enclose objects. This may cause the deterioration of the detector performance, in particular when the background is complex [23]. Oriented bounding boxes (OBB) have already been used under full supervision [78,79] and Sun et al. [44] proposed a method to address the task of weakly supervised oriented object detection in RSIs by exploiting HBB.

Aygunes et al. [25] addressed the task of weakly supervised fine-grained object recognition for tree species classification, which is more challenging than traditional RSWSOD given the very low inter-class variance. The problem is addressed by exploiting a modified WSDDN. The same authors, in a following work [26], address the issue under the presence of multiple sources. In this case, WSDDN is used to perform RSWSOD using multispectral images and LiDAR data, while the RGB images (assumed to have no location uncertainty) are exploited as a reference to aid data fusion, which is a critical step in multi-source scenarios.

Other works try to leverage the advancements of RSWSOD to perform more specific tasks or use different types of aerial images. For instance, Du et al. [20] proposed an RSWSOD method based on the TSI + TDL framework and image-level labels to detect objects in synthetic aperture radar (SAR) images. The training set is initially selected using the unsupervised latent Dirichlet allocation (LDA) [80] and iteratively updated by a linear SVM discriminator. Candidate proposals are generated using the log-normal-based constant false alarm rate (CFAR) and the target clustering methods [81]. Shi et al. [23] proposed a WS method for cap missing detection, exploiting OBB instead of HBB to reduce background interference. Li et al. [48] proposed a region proposal network for geospatial applications that considers Tobler's First Law of geography, stating that "Everything is related to everything else, but near things are more related than distant things". The idea is to convert the 2D object detection problem into a 1D temporal classification problem. The method is applied for terrain feature detection.

More recently, Yang et al. [21] addressed the task of ship detection, exploiting an image transformer [82] called PistonNet. PistonNet proposes the introduction of an artificial point in the feature map, whose aim is to bring the background values to 0 while keeping the important area to 1 to allow a clear separation between the two. The final bounding boxes can be obtained by applying standard thresholding methods to the activation map. Berg et al. [22] exploited an anomaly detection mechanism to detect marine animals from aerial images. The idea is to adapt and modify the patch distribution modeling method (PaDiM) [83], which is currently one of the state-of-the-art approaches used for visual industrial inspection. By training this model with empty ocean images, the model will then be able to detect animals as anomalies.

Even though impressive results have been obtained in recent years, there is still a big gap between RSFSOD and RSWSOD (≈30–40% mAP). To reduce this gap, Li et al. [46] proposed using point-based annotations, which are still far cheaper to obtain than BBs. The point annotations were used to guide the region proposal selection in a fully supervised detector, YOLOv5 [84]. The authors were able to reach a performance comparable to that of a fully supervised method.

In Table 1, it is possible to notice that many of the approaches analyzed in this section are use-case-specific. For this reason, they usually concentrate on solving challenges that are strictly related to the specific task or application domain.

3. Benchmarking

3.1. RSI Datasets

Several datasets have been proposed in the literature for evaluating detection performance on RSIs. Unfortunately, only a few have been released to provide a general benchmark, while many others are use-case-specific. Table 2 illustrates the main properties of the most common RSI datasets for the evaluation of RSWSOD techniques and reports how many times the datasets have been used. All datasets are annotated with bounding boxes, which means that they have been designed for FSOD. Still, they can be easily exploited for WSOD tasks by automatically labeling each training image with the class (or classes) of the manually annotated bounding boxes. The number of these positive training images is reported in Table 2. The use of fine-grained annotated datasets for RSWSOD offers the possibility to compare weakly and fully supervised methods on the same benchmark. Table 3 reports results of RSWSOD techniques while Tables 4–6 also report the upper bound of performance, defined by well-known fully supervised object detectors.

Table 2. Summary of the most common RSI datasets. For each dataset, we report the annotation type, number of images, number of BB annotations, number of positive training images (for WSOD tasks), number of classes, image characteristics, and number of times the dataset has been used for the evaluation of RSWSOD methods.

Name	Year	Annotation Type	# Images	# BB Annotations	# Positive Training Images	# Classes	Dimension (Pixels)	Spatial Resolution	Target Area (Pixels)	# Evaluated
ISPRS [85]	2010	BB	100	-	-	1 (Vehicle)	≈900 × 700	8–15 cm	1150∼11,976	3
Google Earth [29]	2013	BB	120	-	50	1 (Airplane)	≈1000 × 800	≈0.5 m	700∼25,488	5
Landsat-7 ETM+ [30]	2014	BB	180	-	-	1 (Airport)	400 × 400	30 m	1760∼15,570	3
NWPU-VHR-10 [19]	2014	BB, pixel	800	3896	150	10	533 × 597∼ 1728 × 1028	0.08–2 m	1122∼174,724	4
NWPU-VHR-10.v2 [86]	2017	BB, Pixel	1172	-	-	10	400 × 400	-	-	8
DOTA [10]	2018	Oriented BB	2806	188,282	1411	15	≈4000 × 4000	-	-	2
LEVIR [87]	2018	BB	21,952	11,028	400	3	600 × 800	0.2–1 m	10∼600	2
WSADD [42]	2020	BB	700	-	300	1 (Airplane)	768 × 768	0.3–2 m	-	2
DIOR [9]	2020	BB	23,463	192,472	5862	20	800 × 800	0.5–30 m	-	10

The dataset named **Google Earth** is a collection of 120 high-resolution images of airports collected using the homonym service. It was proposed by Zhang et al. [29] to demonstrate that their algorithm can deal with multi-size targets in large-scale RSIs with cluttered backgrounds. Zhang et al. [30] extended the airplane detection task to also include vehicles and airports and incorporated images from **ISPRS** and **Landsat-7 ETM+**. The ISPRS dataset provides vehicles with 100 high-resolution images provided by the German Association of Photogrammetry and Remote Sensing [85]. The Landsat-7 ETM+ dataset is acquired by the homonym sensor and includes 180 infrared RSIs of a variety of airports in China [30].

In 2014, Cheng et al. [19] proposed a dataset named **NWPU VHR-10** with ten classes, containing images from Google Earth and the German Association of Photogrammetry and Remote Sensing [85]. The classes are *Airplane*, *Ship*, *Storage Tank*, *Baseball Diamond*, *Tennis Court*, *Basketball Court*, *Ground Track Field*, *Harbor*, *Bridge* and *Vehicle*. A few years later, Li et al. [86] standardized NWPU VHR-10 proposing the **NWPU VHR-10.v2** dataset. In particular, 1172 images of 400 × 400 pixels were obtained by cropping the positive images of the NWPU VHR-10 dataset in which image sizes are different (from 533 × 597 to 1728 × 1028 pixels). The number of classes was instead left unmodified.

In 2018, Xia et al. [10] proposed **DOTA**, a dataset containing oriented BB annotations. DOTA has been presented as a large-scale benchmark dataset and an object detection challenge. Fifteen classes were annotated: *airplane*, *ship*, *storage tank*, *baseball diamond*, *tennis court*, *swimming pool*, *ground track field*, *harbor*, *bridge*, *large vehicle*, *small vehicle*, *helicopter*, *roundabout*, *soccer ball field*, and *basketball court*.

LEVIR was proposed by Zou et al. [87] and contains many high-resolution Google Earth images. LEVIR covers most types of ground features of the human living environment, e.g., city, country, mountain area, and ocean. There are three classes: *airplane*, *oil plot*, and *ship*. It is important to note that LEVIR is different from LEVIR-CD [88], a remote sensing building change detection dataset.

WSADD is an airplane detection dataset proposed by Wu et al. [42]. The images in this dataset include airports and nearby areas of different countries (mainly from China, the United States, the United Kingdom, France, Japan, and Singapore) taken from Google Earth.

In 2020, Li et al. [9] conducted a deep study on the existing RSI datasets and concluded by proposing **DIOR**. DIOR is one of the earth observation community's largest, most diverse, and publicly available object detection datasets. It is a particularly challenging dataset due to the variety of object sizes and imaging conditions such as weather conditions and seasons. The classes are: *airplane*, *airport*, *baseball field*, *basketball court*, *bridge*, *chimney*, *dam*, *expressway service area*, *expressway toll station*, *golf course*, *ground track field*, *harbor*, *overpass*, *ship*, *stadium*, *storage tank*, *tennis court*, *train station*, *vehicle*, and *wind mill*. It is important to note that DIOR has high inter-class similarity and intra-class diversity, and the number of object instances per class is not balanced.

3.2. Evaluation

In this section, a performance comparison between the surveyed methods is provided. The analysis is performed on the three most commonly used RSI datasets: Google Earth [29], DIOR [9] and NWPU VHR-10.v2 [86]. Table 3 reports the overall results in terms of average precision (AP) or mean average precision (mAP) and CorLoc [89], when available. It is possible to notice that only half of the analyzed methods evaluate their performance on these datasets. Indeed, many studies in the remote sensing field make use of custom or use-case-specific datasets and other metrics, e.g., precision–recall curve, overall accuracy, or F1 score. Thus, it is impossible to make a fair comparison between them. Table 7 summarizes the performance of such methods.

Table 3. Overall results of the methods on the main datasets. Only the methods with available results on DIOR, NWPU VHR-10.v2, or Google Earth are reported. The best performance for each data set is indicated in bold.

Name	Approach	Year	Google Earth	NWPU VHR-10.v2		DIOR	
			AP	mAP	CorLoc	mAP	CorLoc
Zhang et al. [29]	TSI + TDL	2014	54.18%	-	-	-	-
Han et al. [31]	TSI + TDL	2014	60.16%	-	-	-	-
Zhang et al. [30]	TSI + TDL	2014	66.42%	-	-	-	-
Zhou et al. [33]	TSI + TDL	2015	75.58%	-	-	-	-
Zhou et al. [34]	TSI + TDL	2016	**76.26**%	-	-	-	-
FCC-Net [40]	MIL	2020	-	-	18.30%	41.70%	-
DCL [41]	MIL	2020	-	52.11%	69.65%	20.19%	42.23%
PCIR [15]	MIL	2020	-	54.97%	71.87%	24.92%	46.12%
AlexNet-WSL [42]	CAM	2020	-	-	-	18.78%	-
TCANet [16]	MIL	2020	-	58.82%	72.76%	25.82%	48.41%
Wang et al. [45]	MIL + CAM	2021	-	53.60%	61.50%	-	-
SDA-RSOD [17]	CAM	2022	-	-	-	24.11%	-
MIGL [47]	MIL	2021	-	55.95%	70.16%	25.11%	46.80%
SAENet [49]	MIL	2021	-	60.72%	73.46%	27.10%	49.42%
SPG + MELM [51]	MIL	2022	-	**62.80%**	73.41%	25.77%	48.30%
Qian et al. [18]	MIL	2022	-	61.49%	**73.68%**	**27.52%**	**49.92%**

Looking at Table 3, it is possible to notice that CAM-based methods have been less evaluated on these three challenging datasets. This may happen because, as analyzed in Section 2.3.3, CAM-based approaches are used for specific tasks such as aircraft detection. Another factor influencing this trend is that CAMs work well when there are few large instances in the image [12]. For this reason, MIL-based approaches seem more effective than CAM-based ones in RSIs, where multiple instances are present. However, recently, this performance gap has been reduced by the work of Wu et al. [17] (SDA-RSOD).

It is important to note that the overall performance of the methods increases over the years independently of the dataset, demonstrating the effectiveness of the research in the remote sensing domain.

3.2.1. Google Earth Dataset

The methods developed before the advent of DL have been evaluated on the Google Earth dataset. These methods cannot be compared with DL-based approaches proposed after 2020 since the evaluation data are different (Table 3).

Table 4 shows a clear increasing trend in the performance over the years. This is not surprising given that most of the methods are based on the work by Zhang et al. [29,30] and try to leverage better feature extractors (e.g., DBM and CNN) or training set initialization techniques (negative bootstrapping).

Table 4. Results of the methods on the Google Earth dataset. Furthermore, well-known methods from the literature are reported for a fair comparison. The best-performing techniques are indicated in bold.

Name	Type	Year	AP
BOV [90]	FSOD	2010	52.75%
Han et al. [91]	FSOD	2014	54.21%
Zhang et al. [30]	FSOD	2014	**59.67%**
Zhang et al. [29]	RSWSOD	2014	54.18%
Zhang et al. [30]	RSWSOD	2014	66.42%
Han et al. [31]	RSWSOD	2014	60.16%
Zhou et al. [33]	RSWSOD	2015	75.58%
Zhou et al. [34]	RSWSOD	2016	**76.26%**

Regarding Zhang et al. [30], the reported result is the one obtained using the locality-constrained linear coding (LLC) [92] feature to represent each of the training examples since it provides the best performance. Moreover, in 2014, the authors showed that the proposed RSWSOD architecture could perform comparably to FSOD models and sometimes even outperform them. The clear improvements obtained by subsequent methods show that the usage of more powerful feature extractors such as CNN allows fully supervised approaches to be outperformed based on handcrafted features (BOV [93] and FDDL [91]). The results are consistent with the literature stating that handcrafted features are not powerful enough to accurately describe objects in RSIs [31,33,34].

3.2.2. DIOR Dataset

Table 5 presents the evaluation of various WSOD and RSWSOD methods on the DIOR dataset. The upper bound of performance is provided by two FSOD architectures. It can be noted that the performance of WSDNN [13] (a relevant milestone in WSOD) is improved by methods such as OICR [14] and PCL [63] which are based on its framework. The reason is that each of these methods tries to overcome the limitations of previous methods such as the discriminative region and the multiple-instance problems. However, these techniques are not capable of reaching satisfying performances when dealing with more challenging data such as RSIs. It can be seen that RSWSOD techniques significantly improve performances with respect to those developed for natural images. For instance, PCIR [15], which is an OICR-inspired method, can improve performance over OICR of 8.42% mAP, demonstrating the effectiveness of the developed specific network adaptations.

Considering two of the most recent works, TCANet [16] and its derivative work SAENet [49], an improvement of 1.28% can be seen. This can be because SAENet considers the consistency across different spatial transformations of the same image, which was ignored by previous approaches and could hurt the detector performances. Furthermore, Qian et al. [18] were able to obtain state-of-the-art performances for RSWSOD, taking into consideration the imbalance between easy and hard samples, which had not been previously considered.

Another interesting point concerns the region proposal generation step that is part of the MIL-based framework. As already pointed out in several works [41,45,46], high-quality proposals are needed to obtain high-quality detectors. However, most of the approaches make use of selective search [52] or similar methods to produce candidates (Table 1). It has been demonstrated that these proposals cannot cover the entire object well, severely hindering the performance of WSOD. This problem is effectively addressed by Cheng et al. [51] through a region proposal network. Using the RPN with the Min-Entropy Latent Model (MELM) [67], the authors can obtain an improvement of 7.11% over the basic MELM method [67], confirming the importance of high-quality proposals.

MIL-based approaches seem to be more effective than CAM-based methods, though the gap has been reduced by Wu et al. [17] (SDA-RSOD). The proposed CAM-based method has lower variance over the classes, while state-of-the-art MIL-based approaches tend to obtain good performances in some classes and terrible performances (almost 0% mAP) in others. For instance, the method of Qian et al. [18], the best-performing MIL-based approach on the DIOR dataset, shows an mAP of 27.52%, with the best precision on Baseball field (67.46%) and the worst precision on Dam (0.74%). Instead, SDA-RSOD shows an mAP of 24.11%, with the best precision on Golf field (61.04%) and the worst precision on Storage Tank (7.53%). The overall performance is slightly lower, but the results are far more balanced between classes.

Another thing that can be noticed is that most methods tend to perform quite well on classes such as Baseball field, Chimney, Ground track field, and Stadium. This is mainly because the objects of all these classes usually occupy a large part of the images and have a relatively low probability of co-occurrence with other classes [41]. Instead, the complex background and the fact that there can be coexisting objects may hurt the detector's discriminative power [16]. For instance, Bridge is considered a challenging class because it is almost always matched with a river in the background. The same holds for Dam as these objects usually coexist with reservoirs. A similar problem also arises for the Windmill class, whose shadow is usually detected instead of the object itself. Different imaging angles and illumination conditions can lead to the background features being more prominent than the object features [16].

In addition, it is interesting to note that there is not a method able to outperform all other techniques in (almost) every class, even when considering an approach and its immediate improvement (e.g., TCANet [16] and SAENet [49]). This shows that it is extremely difficult to improve the overall performance of RSWSOD without damaging the detection capability of some classes.

Differently from what happens in WSOD for natural images [13,14,63], almost every RSWSOD approach tackles the task from a brand new perspective without building on top of previous RSWSOD architectures. Previous work is explicitly used as a baseline and improved only when the authors are the same [16,49]. This may happen because the code is not publicly available.

Table 5 also shows that all the network adaptations proposed by RSWSOD techniques are still not enough to reach a performance comparable to the fully supervised counterparts. Qian et al. [18], the best-performing RSWOSD method, obtains 27.52% mAP, while Faster R-CNN can double the performance with 55.48% mAP.

Table 5. Results of the methods on the DIOR dataset. Two well-known fully supervised methods from the literature are reported for a fair comparison. The columns report the Average Precision (AP) for each category, the mean Average Precision (mAP), and the CorLoc for each method. The best-performing techniques for each class are indicated in bold.

Name	Type	Year	Airplane	Airport	Baseb. Field	Basketb. Court	Bridge	Chimney	Dam	Expr. Service Area	Expr. Toll Station	Golf Field	Ground Track Field	Harbor	Overp.	Ship	Stadium	Storage Tank	Tennis Court	Train Station	Vehicle	Windm.	mAP	CorLoc
Fast RCNN [94]	FSOD	2015	44.17%	**66.79%**	**66.96%**	60.49%	15.56%	**72.28%**	**51.95%**	**65.87%**	44.76%	**72.11%**	62.93%	**46.18%**	38.03%	32.13%	**70.98%**	35.04%	58.27%	37.91%	19.20%	38.10%	49.98%	-
Faster RCNN [62]	FSOD	2015	**50.28%**	62.60%	66.04%	**80.88%**	**28.80%**	68.17%	47.26%	58.51%	**48.06%**	60.44%	**67.00%**	43.86%	**46.87%**	**58.48%**	52.37%	**42.53%**	**79.52%**	**48.02%**	**34.77%**	**65.44%**	**55.48%**	-
WSDDN [13]	WSOD	2016	9.06%	39.68%	37.81%	20.16%	0.25%	12.18%	0.57%	0.65%	11.88%	4.90%	42.35%	4.66%	1.06%	0.70%	63.03%	3.95%	6.06%	0.51%	4.55%	1.14%	13.26%	32.40%
CAM [72]	WSOD	2016	2.66%	34.63%	16.87%	16.70%	10.59%	25.43%	17.66%	25.40%	25.87%	56.23%	17.68%	12.86%	25.32%	0.11%	9.27%	0.84%	6.64%	46.00%	1.48%	9.08%	18.07%	-
OICR [14]	WSOD	2017	8.70%	28.26%	44.05%	18.22%	1.30%	20.15%	0.09%	0.65%	29.89%	13.80%	57.39%	10.66%	11.06%	9.09%	59.29%	7.10%	0.68%	0.14%	9.09%	0.41%	16.50%	34.80%
PCL [63]	WSOD	2018	21.52%	35.19%	59.80%	23.49%	2.95%	43.71%	0.12%	0.90%	1.49%	2.88%	56.36%	16.76%	11.05%	9.09%	57.62%	9.09%	2.47%	0.12%	4.55%	4.55%	18.19%	41.52%
MELM [67]	WSOD	2018	28.14%	3.23%	62.51%	28.72%	0.06%	62.51%	0.21%	13.09%	28.39%	15.15%	41.05%	26.12%	0.43%	9.09%	8.58%	15.02%	20.57%	9.81%	0.04%	0.53%	18.66%	-
ACol [95]	WSOD	2018	0.15%	7.62%	2.38%	0.00%	0.00%	0.04%	6.10%	7.82%	0.78%	27.72%	13.18%	9.43%	20.56%	0.13%	0.00%	0.63%	2.27%	18.68%	0.17%	0.27%	5.89%	-
DaNet [75]	WSOD	2019	1.33%	33.41%	13.46%	17.95%	12.99%	21.60%	17.20%	25.84%	19.68%	53.98%	19.86%	12.63%	24.31%	0.43%	12.37%	0.56%	5.85%	**49.57%**	1.11%	3.08%	17.37%	-
MIST [69]	WSOD	2020	32.01%	39.87%	62.71%	28.97%	7.46%	12.87%	0.31%	5.14%	17.38%	51.02%	49.48%	5.36%	12.24%	**29.43%**	35.53%	**25.36%**	0.81%	4.59%	**22.22%**	0.80%	22.18%	43.57%
AlexNet-WSL [42]	RSWSOD	2020	2.94%	35.58%	17.92%	18.20%	12.10%	25.91%	18.71%	26.44%	25.46%	56.56%	19.24%	12.91%	**25.83%**	0.64%	10.39%	1.19%	7.05%	47.07%	1.74%	9.78%	18.78%	-
PCIR [15]	RSWSOD	2020	30.37%	36.06%	54.22%	26.60%	9.09%	58.59%	0.22%	9.65%	**36.18%**	32.59%	58.51%	8.60%	21.63%	12.09%	**64.28%**	9.09%	13.62%	0.30%	9.09%	7.52%	24.92%	46.12%
DCL [41]	RSWSOD	2020	20.89%	22.70%	54.12%	11.50%	6.03%	61.01%	0.09%	1.07%	31.01%	30.87%	56.45%	5.05%	2.65%	9.09%	63.65%	9.09%	10.36%	0.02%	7.27%	0.79%	20.19%	42.23%
FCC-Net [40]	RSWSOD	2020	20.10%	38.80%	52.00%	23.40%	1.80%	22.30%	0.20%	0.60%	28.70%	14.10%	56.00%	11.10%	10.90%	10.00%	57.50%	9.10%	3.60%	0.10%	5.90%	0.70%	18.30%	41.70%
TCANet [16]	RSWSOD	2020	25.13%	30.84%	62.92%	**40.00%**	4.13%	**67.78%**	8.07%	23.80%	29.89%	22.34%	53.85%	24.84%	11.06%	9.09%	46.40%	13.74%	**30.98%**	1.47%	9.09%	1.00%	25.82%	48.41%
MIGL [47]	RSWSOD	2021	22.20%	52.57%	62.76%	25.78%	8.47%	67.42%	0.66%	8.85%	28.71%	57.28%	47.73%	23.77%	0.77%	6.42%	54.13%	13.15%	4.12%	14.76%	0.23%	2.43%	25.11%	46.80%
SAENet [49]	RSWSOD	2021	20.57%	**62.41%**	62.65%	23.54%	7.59%	64.62%	0.20%	**34.52%**	30.62%	55.38%	52.70%	17.57%	6.85%	9.09%	51.59%	15.43%	1.69%	14.41%	1.41%	9.16%	27.10%	49.42%
SDA-RSOD [17]	RSWSOD	2022	19.51%	38.86%	26.40%	23.56%	**13.30%**	26.84%	**25.33%**	27.09%	27.17%	**61.04%**	20.89%	16.78%	25.57%	8.28%	10.34%	7.53%	26.52%	48.81%	9.28%	**19.16%**	24.11%	-
SPG + MELM [51]	RSWSOD	2022	31.32%	36.66%	62.79%	29.10%	6.08%	62.66%	0.31%	15.00%	30.10%	35.00%	48.02%	**27.11%**	12.00%	10.02%	60.04%	15.10%	21.00%	9.92%	3.15%	0.06%	25.77%	48.30%
Qian et al. [18]	RSWSOD	2022	**41.10%**	48.62%	**67.48%**	33.92%	4.32%	34.71%	0.74%	12.29%	24.33%	56.74%	**63.55%**	5.36%	23.11%	21.34%	57.44%	24.66%	0.85%	9.97%	18.34%	1.54%	**27.52%**	**49.92%**

A final comparison between the WSOD and RSWSOD architectures' performance on the DIOR dataset is reported in Figure 9. It is possible to notice a clear separation between the two categorizations, mainly because many RSWSOD works are based on a WSOD counterpart with the addition of specific adaptations. It can be noted that MIL-based and CAM-based WSOD methods obtain similar performances (≈18% mAP) except for MIST, while almost all RSWSOD methods have stable performances around ≈25–30% mAP. Only a few early RSWSOD techniques obtain negligible improvements over WSOD architectures. This figure confirms all the already discussed advantages brought by RSWSOD methods.

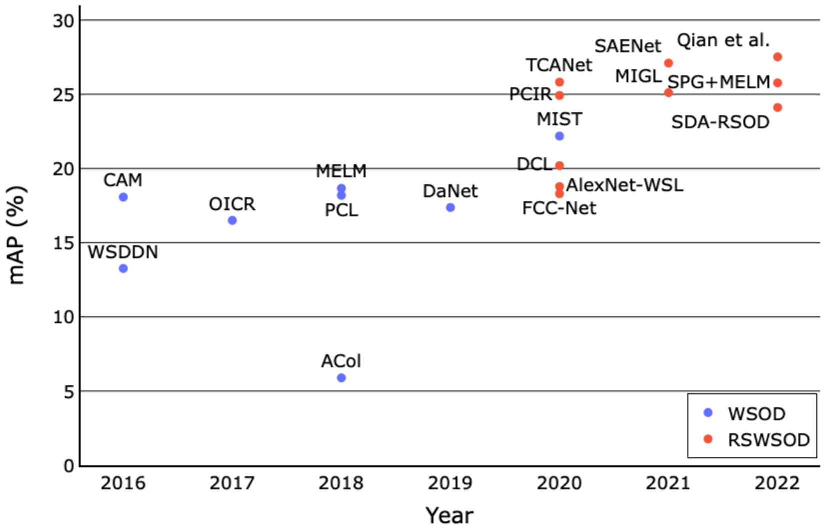

Figure 9. Mean Average Precision (mAP) trend of WSOD and RSWSOD methods on the DIOR dataset over the years. It is possible to notice that most RSWSOD methods are presented after 2020, and their performances are ≈5–10% better than the WSOD baselines [18].

3.2.3. NWPU VHR-10.v2 Dataset

As shown in Table 6, most of the considerations described for the DIOR dataset hold also for NWPU VHR-10.v2. The first thing that can be noticed is that performances are much higher on this dataset for all methods, both fully and weakly supervised (≈30–40% mAP more than DIOR presented in Table 5). This is because NWPU VHR-10.v2 is less challenging. In fact, besides having half of the classes of DIOR and very few images, the image diversity and variance are limited. As already outlined in Section 3.1, DIOR was built with the intent of having a large-scale dataset that could solve the limitations of other RSIs datasets, guaranteeing large size variations, high inter-class similarity, high intra-class diversity, and variations in terms of image condition, weather, and season.

From Table 6 the performance gap between WSOD and RSWSOD methods is evident. For instance, PCIR gains 20.45% mAP over OICR, and SPG boosts the performance of MELM by 20.51% thanks to the usage of an RPN. This again shows the importance of having high-quality proposals and specific network adaptations addressing the RSI challenges. Nonetheless, there is still a huge performance gap (≈25% mAP) between FSOD and RSWSOD methods, with Faster R-CNN obtaining 87.12% mAP and SPG + MELM 62.80% mAP. indicating the fact that the complexity of the dataset contributes only partially to the predominance of full supervision.

Furthermore, as shown in Figure 10, the performance of the various methods on each class is highly variant. Finally, there is also a considerable gap between the performance in some classes (e.g., Airplane, Baseball Diamond, and Ground Track Field) with respect to others (e.g., Bridge and Vehicle). Similarly to DIOR, this means that some classes are more challenging to identify than others or may be easily confused with the background (e.g., bridges may be confused with rivers).

3.2.4. Other Performances

The performances of the methods reported in this section are not comparable because they are use-case-specific and evaluated on custom datasets and with different metrics.

Looking at Table 7, it is possible to see that some methods obtain very high performances, while others slightly overcome 50% of accuracy or F1 score. A possible reason may reside in the difficulty of the task being solved. For instance, aircraft detection, for which the analyzed methods obtain impressive performance, is far easier than marine animal detection, for which the reported performance is much lower.

Moreover, the performances on the reported datasets are often far better than those reported on DIOR and NWPU VHR-10.v2, probably because these datasets contain a reduced amount of classes. Instead, DIOR and NWPU VHR-10.v2 were specifically constructed to build challenging benchmarks. These results further emphasize the difficulty induced by the high intra-class diversity and the high inter-class similarity, obviously mitigated when the number of classes in the dataset is lower.

Aygunes et al. [26] show a significant improvement in the tree species detection task using multiple sources from the previous work of the same authors [25].

Sun et al. [44] show the difficulty of the oriented object detection task. This can be observed by looking at the performance on the DOTA dataset with oriented BBs, which is less than half of the mAP obtained by MPFP-Net [43] using horizontal BBs. Still, the gap between Sun et al. [44] and a fully supervised method such as oriented R-CNN [96] is over 40% mAP.

Finally, another interesting observation concerns PistonNet [21]. In particular, it can be observed that the performance of this method on NWPU VHR-10 is 11.38% lower than MPFP-Net [43]. However, this result is promising since PistonNet was built to address the specific task of ship detection but shows interesting generalization capabilities when applied to a multi-class dataset.

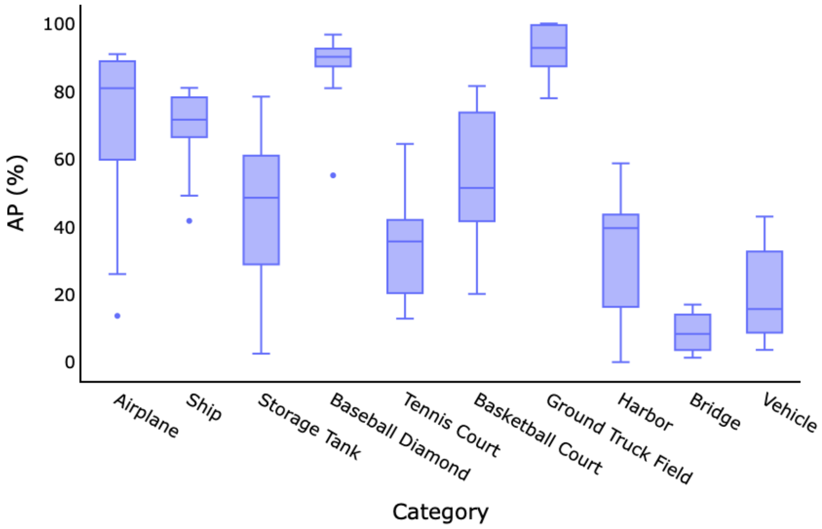

Figure 10. Box plot of average precision by class of all the WSOD/RSWOD methods on the NWPU VHR-10.v2 dataset. It is possible to notice that there is high inter-class performance variance based on the classes' difficulty (e.g., Ground Track Field and Bridge).

Table 6. Results of the methods on the NWPU VHR-10.v2 dataset. Three well-known fully supervised methods from the literature are reported for a fair comparison. The columns report the Average Precision (AP) for each category, the mean Average Precision (mAP), and the CorLoc for each method. The best-performing techniques for each class are indicated in bold.

Name	Type	Year	Airplane	Ship	Storage Tank	Basketball Court	Tennis Court	Basketball Court	Ground Truck Field	Harbor	Bridge	Vehicle	mAP	CorLoc
Fast RCNN [94]	FSOD	2015	90.91%	90.60%	89.29%	47.32%	**100.00%**	**85.85%**	84.86%	**88.22%**	**80.29%**	69.84%	82.71%	-
Faster RCNN [62]	FSOD	2015	90.90%	86.30%	90.53%	98.24%	89.72%	80.13%	**90.81%**	80.29%	68.53%	87.14%	84.52%	-
RICO [86]	FSOD	2017	**99.70%**	**90.80%**	**90.61%**	**92.91%**	90.29%	80.13%	**90.81%**	80.29%	68.53%	**87.21%**	**87.12%**	-
WSDDN [13]	WSOD	2016	30.08%	41.72%	34.98%	88.90%	12.86%	23.85%	99.43%	13.94%	1.92%	3.60%	35.12%	35.24%
OICR [14]	WSOD	2017	13.66%	67.35%	57.16%	55.16%	13.64%	39.66%	92.80%	0.23%	1.84%	3.73%	34.52%	40.01%
PCL [63]	WSOD	2018	26.00%	63.76%	2.50%	89.80%	**64.45%**	76.07%	77.94%	0.00%	1.30%	15.67%	39.41%	45.06%
MELM [67]	WSOD	2018	80.86%	69.30%	10.48%	90.17%	12.84%	20.14%	99.17%	17.10%	14.17%	8.68%	42.29%	49.87%
MIST [69]	WSOD	2020	69.68%	49.16%	48.55%	80.91%	27.08%	79.85%	91.34%	46.99%	8.29%	13.36%	51.52%	70.34%
PCIR [15]	RSWSOD	2020	**90.97%**	78.81%	36.40%	90.80%	22.64%	52.16%	88.51%	42.36%	11.74%	35.49%	54.97%	71.87%
DCL [41]	RSWSOD	2020	72.70%	74.25%	37.05%	82.64%	36.88%	42.27%	83.95%	39.57%	16.82%	35.00%	52.11%	69.65%
TCANet [16]	RSWSOD	2020	89.43%	78.18%	**78.42%**	90.80%	35.27%	50.36%	90.91%	42.44%	4.11%	28.30%	58.82%	72.76%
Wang et al. [45]	RSWSOD	2021	80.90%	78.30%	10.50%	90.10%	64.40%	69.10%	80.20%	39.60%	14.00%	8.70%	53.60%	61.50%
MIGL [47]	RSWSOD	2021	88.69%	71.61%	75.17%	94.19%	37.45%	47.68%	**100.00%**	27.27%	8.33%	9.06%	55.95%	70.16%
SAENet [49]	RSWSOD	2021	82.91%	74.47%	50.20%	**96.74%**	55.66%	72.94%	**100.00%**	36.46%	6.33%	31.89%	60.72%	73.46%
SPG + MELM [51]	RSWSOD	2022	90.42%	**81.00%**	59.53%	92.31%	35.64%	51.44%	99.92%	**58.71%**	**16.99%**	**42.99%**	**62.80%**	73.41%
Qian et al. [18]	RSWSOD	2022	81.64%	68.33%	65.31%	93.44%	36.43%	**81.54%**	98.67%	53.77%	9.86%	25.87%	61.49%	**73.68%**

Table 7. Results of surveyed methods evaluated on specific datasets. Works are sorted by year and grouped when using similar datasets. For each method, the used dataset, number of classes, and performance are reported. Performance metrics are: PR curve (no value), accuracy, F1 score, and mean average precision (mAP).

Name	Year	Dataset	Classes	Performance
Cheng et al. [32]	2014	Custom Google Earth	3	n/d (PR Curve)
Cao et al. [36]	2017	Custom Google Earth	1 (vehicles)	n/d (PR Curve)
MIRN [37]	2018	Custom Google Earth	1 (vehicles)	n/d (PR Curve)
LocNet [35]	2016	Tokyo Airport Sidney Airport Berlin Airport	1 (aircraft) 1 (aircraft) 1 (aircraft)	98.46% Acc. 89.13% Acc. 96.77% Acc.
WSA [39]	2019	Tokyo Airport Sidney Airport	1 (aircraft) 1 (aircraft)	96.92% Acc. 95.65% Acc.
SLS [38]	2018	NWPU VHR-9	9 (no vehicles)	11% mAP
Aygunes et al. [25]	2019	Custom 8-b MS WV-2	18 (trees) 40 (trees)	60.60% Acc. 42.50% Acc.
Aygunes et al. [26]	2021	Custom Seattle Trees, 8-b MS WV-2 Custom Seattle Trees, 8-b MS WV-2, LiDar DSM	40 (trees) 40 (trees)	51.70% Acc. 53.00% Acc.
MPFP-Net [43]	2021	NWPU VHR-10 LEVIR DOTA (HBB)	10 3 15	94.57% mAP 86.73% mAP 84.43% mAP
Sun et al. [44]	2021	DOTA (OBB)	15	38.6% mAP
Du et al. [20]	2019	miniSAR	1 (vehicles)	84.85% F1
Shi et al. [23]	2020	FIN GCAP	4 2	90% mAP 93% mAP
Li et al. [46]	2021	Custom NWPU VHR-10	4	92.40% mAP
Li et al. [48]	2021	Custom Mars craters	1	80.00% mAP
Berg et al. [22]	2022	Semmacape Custom Kelonia	8 2	53.00% F1 56.80% F1
Long et al. [50]	2022	Custom World Map, Google Maps	2	89% F1
PistonNet [21]	2022	GF1-LRSD NWPU VHR-10	1 10	81.25% mAP 83.19% mAP

4. Issues and Research Directions

The road map of the methods described in Section 2.3 shows that research in this field is gaining more and more interest. In recent years, there has been a promising improvement in the results obtained, meaning that taking into consideration specific RSI challenges is effective. However, the results show that the gap with respect to FS approaches is still relevant. The solutions proposed for these challenges only provide a partial improvement and the RSWSOD problem has yet to be solved efficiently.

A significant issue is related to the unavailability of the code (apart from a couple of works [17,22]) making it difficult for researchers to build novel methods upon existing architectures and to have a baseline for new contributions. This may be the reason why no surveyed methods used other RSWSOD works as a starting point. In WSOD for natural images, the authors responsible for WSDDN [13], OICR [14], and MELM [67] made their code publicly available, ensuring a faster advancement of the research.

Another important point is related to the comparability of the results, which can only be partially assessed, because of the massive usage of different datasets and metrics, making it difficult to evaluate and compare the performance of a method comprehensively.

The review of RSWSOD methods, the analysis of results, and the identification of issues allow the definition of a set of possible future research directions.

- **Coarser annotations**: This survey highlighted the fact that almost all RSWSOD approaches are based on image-level labels. Even though this type of annotation is the easiest to obtain, it does not provide any clue regarding the localization of the object.

Li et al. [46] recently showed that exploiting other types of labels, e.g., point-based, which are still cheaper than manual BBs, allows performances that are comparable to FS approaches to be obtained. Thus, more research in this direction should be carried out, considering a trade-off between annotation cost and overall model performance.

- **Interactive annotations**: given the difficulty in correctly detecting some classes, a viable option could be to learn a WS detector and then use human verification to check the correctness of the output BBs and refine them [97]. This could reduce the annotation time and produce high-quality annotations while decreasing the gap with fully supervised settings.
- **Hybrid architectures**: As reported in the survey, two major categories of approaches are currently leading the research in RSWSOD: MIL-based and CAM-based methods. MIL-based approaches usually provide better overall performances but the performance of each class is highly varied. On the other hand, CAM-based approaches are less widespread and effective than MIL-based approaches but they tend to be more stable in terms of performance over the classes. For this reason, it could be interesting to build hybrid approaches that exploit the advantages of both methods. A first approach was proposed by Wang et al. [45] and exploit CAMs to guide the selection of proposals that are then fed to a MIL-based detector. Still, there is a large room for improvement.
- **Transformer-based architectures**: Transformers were born to tackle natural language processing (NLP) problems, but their usage has gained much attention in the CV field due to their powerful capabilities. PistonNet [21] showed interesting results with the use of image transformers [82] and provided good generalization capabilities, despite being developed for the specific use-case of ship detection. This powerful family of architectures could be extended to general-purpose RSWSOD.
- **Better initial proposals**: As shown by Cheng et al. [51], proposal generation is a very critical step because the performance of an MIL-based method is strongly dependent on the quality of the initial proposal. For this reason, developing novel proposal generation methods is fundamental, especially when specific use-cases are addressed, and could boost the performance of WS approaches.
- **Transformation consistency and sample difficulty**: Recently, Feng et al. [49] brought to the attention of the remote sensing community that previous methods did not take into consideration the consistency across different spatial transformations of the same image, with different augmentations of the same image potentially being labeled differently. At the same time, Qian et al. [18] showed the importance of considering the samples' difficulty when training the detector. These factors should be carefully taken into consideration when developing future works.
- **Learn better representations of the data**: Self-supervised learning (SSL) [98] has recently gained much attention in the remote sensing field [99] since it allows better representations of the data to be learned. Instead of pre-training networks on huge datasets of natural images (e.g., ImageNet [7]), it could be interesting to combine self-supervised feature learning in RSIs and weak supervision. This could potentially help improve the performance of WS approaches. For instance, it could be especially useful for those classes that are easily misclassified, such as Bridges and Windmills.
- **Benchmark definition**: As highlighted in this survey, almost half of the analyzed methods rely on the use of custom datasets. However, this makes it extremely difficult to compare methods with each other. A step in this direction was achieved with the introduction of DIOR [9] and NWPU VHR-10.v2 [86]. However, this is still insufficient for all single-object detection methods. A possibility could be to assess the performance of these methods on single-object images extracted from these datasets. For instance, the performance of airplane detection on the images belonging to the Airplane class of DIOR could be assessed.

5. Conclusions

This survey presents a review of 33 state-of-the-art works for the task of remote sensing weakly supervised object detection. The main properties of these methods have been discussed and analyzed: annotation type, approach, and addressed challenges. The advantages and disadvantages of the surveyed works have been described to help the reader to obtain a clear and up-to-date view of the current literature in the field. The novel characteristics of the analyzed research are further emphasized.

A list of the most used datasets and an in-depth analysis of the performance are reported. Several issues emerge from the survey, explaining the large gap between weakly supervised and fully supervised techniques. For this reason, the most promising research directions have been highlighted to help improve future research in the remote sensing domain.

Author Contributions: Conceptualization, C.F., S.P., F.M. and P.F.; methodology, C.F., S.P., F.M. and P.F.; formal analysis, C.F., S.P., F.M. and P.F.; investigation, C.F., S.P., F.M. and P.F.; writing—original draft preparation, C.F., S.P., F.M. and P.F.; writing—review and editing, C.F., S.P., F.M. and P.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

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

Abbreviations

AD	Anomaly Detection
AP	Average Precision
BB	Bounding Box
CAM	Class Activation Map
CFAR	Constant False Alarm Rate
CNN	Convolutional Neural Network
COPD	Collection of Part Detector
CV	Computer Vision
DBM	Deep Boltzmann Machine
DL	Deep Learning
EB	Edge Boxes
FSOD	Fully Supervised Object Detection
GSD	Ground Sample Distance
GT	Ground Truth
HBB	Horizontal Bounding Boxes
mAP	Mean Average Precision
MIL	Multiple Instance Learning
OBB	Oriented Bounding Boxes
OD	Object Detection
PLG	Pseudo-label Generator
PR	Precision Recall
RPN	Region Proposal Network
RSFSOD	Remote Sensing Fully Supervised Object Detection
RSI	Remote Sensing Image
RSWSOD	Remote Sensing Weakly Supervised Object Detection
SAR	Synthetic Aperture Radar
Sb-SaS	Saliency-based Self-adaptive Segmentation
SOTA	State-of-the-art
SS	Selective Search
SSL	Self-supervised Learning

SVM	Support Vector Machine
SW	Sliding Window
TDL	Target Detector Learning
TSI	Training Set Initialization
WS	Weak Supervision
WSDDN	Weakly Supervised Deep Detection Network
WSOD	Weakly Supervised Object Detection
WSOL	Weakly Supervised Object Localization

References

1. He, Z. Deep Learning in Image Classification: A Survey Report. In Proceedings of the 2020 2nd International Conference on Information Technology and Computer Application (ITCA), Guangzhou, China, 18–20 December 2020; pp. 174–177. [CrossRef]
2. Zaidi, S.S.A.; Ansari, M.S.; Aslam, A.; Kanwal, N.; Asghar, M.; Lee, B. A survey of modern deep learning based object detection models. *Digit. Signal Process.* **2022**, *126*, 103514. [CrossRef]
3. Hafiz, A.M.; Bhat, G.M. A survey on instance segmentation: State of the art. *Int. J. Multimed. Inf. Retr.* **2020**, *9*, 171–189. [CrossRef]
4. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.A.; van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [CrossRef] [PubMed]
5. Aljabri, M.; AlGhamdi, M. A Review on the Use of Deep Learning for Medical Images Segmentation. *Neurocomputing* **2022**, *506*, 311–335. [CrossRef]
6. Torres, R.N.; Fraternali, P. Learning to identify illegal landfills through scene classification in aerial images. *Remote Sens.* **2021**, *13*, 4520. [CrossRef]
7. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Li, F.-F. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255. [CrossRef]
8. Everingham, M.; Eslami, S.; Van Gool, L.; Williams, C.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes Challenge: A Retrospective. *Int. J. Comput. Vis.* **2014**, *111*, 98–136. [CrossRef]
9. 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]
10. 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 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983.
11. Cheng, G.; Han, J.; Lu, X. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
12. Shao, F.; Chen, L.; Shao, J.; Ji, W.; Xiao, S.; Ye, L.; Zhuang, Y.; Xiao, J. Deep Learning for Weakly-Supervised Object Detection and Localization: A Survey. *Neurocomputing* **2022**, *496*, 192–207. [CrossRef]
13. 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. [CrossRef]
14. 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. [CrossRef]
15. 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]
16. 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]
17. Wu, Z.Z.; Xu, J.; Wang, Y.; Sun, F.; Tan, M.; Weise, T. Hierarchical fusion and divergent activation based weakly supervised learning for object detection from remote sensing images. *Inf. Fusion* **2022**, *80*, 23–43. [CrossRef]
18. Qian, X.; Huo, Y.; Cheng, G.; Yao, X.; Li, K.; Ren, H.; Wang, W. Incorporating the Completeness and Difficulty of Proposals Into Weakly Supervised Object Detection in Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 1902–1911. [CrossRef]
19. 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]
20. Du, L.; Dai, H.; Wang, Y.; Xie, W.; Wang, Z. Target discrimination based on weakly supervised learning for high-resolution SAR images in complex scenes. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 461–472. [CrossRef]
21. Yang, Y.; Pan, Z.; Hu, Y.; Ding, C. PistonNet: Object Separating From Background by Attention for Weakly Supervised Ship Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 5190–5202. [CrossRef]
22. Berg, P.; Santana Maia, D.; Pham, M.T.; Lefèvre, S. Weakly Supervised Detection of Marine Animals in High Resolution Aerial Images. *Remote Sens.* **2022**, *14*, 339. [CrossRef]
23. Shi, C.; Huang, Y. Cap-count guided weakly supervised insulator cap missing detection in aerial images. *IEEE Sens. J.* **2020**, *21*, 685–691. [CrossRef]

24. Yue, J.; Fang, L.; Ghamisi, P.; Xie, W.; Li, J.; Chanussot, J.; Plaza, A. Optical remote sensing image understanding with weak supervision: Concepts, methods, and perspectives. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 250–269. [CrossRef]
25. Aygüneş, B.; Aksoy, S.; Cinbiş, R.G. Weakly supervised deep convolutional networks for fine-grained object recognition in multispectral images. In Proceedings of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 1478–1481.
26. Aygunes, B.; Cinbis, R.G.; Aksoy, S. Weakly supervised instance attention for multisource fine-grained object recognition with an application to tree species classification. *ISPRS J. Photogramm. Remote Sens.* **2021**, *176*, 262–274. [CrossRef]
27. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Int. J. Surg.* **2021**, *88*, 105906. [CrossRef]
28. Falagas, M.E.; Pitsouni, E.I.; Malietzis, G.A.; Pappas, G. Comparison of PubMed, Scopus, web of science, and Google scholar: Strengths and weaknesses. *FASEB J.* **2008**, *22*, 338–342. [CrossRef] [PubMed]
29. Zhang, D.; Han, J.; Yu, D.; Han, J. Weakly supervised learning for airplane detection in remote sensing images. In *Proceedings of the Second International Conference on Communications, Signal Processing, and Systems*; Springer: Cham, Switzerland, 2014; pp. 155–163.
30. Zhang, D.; Han, J.; Cheng, G.; Liu, Z.; Bu, S.; Guo, L. Weakly supervised learning for target detection in remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2014**, *12*, 701–705. [CrossRef]
31. 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.* **2014**, *53*, 3325–3337. [CrossRef]
32. Cheng, G.; Han, J.; Zhou, P.; Guo, L. Scalable multi-class geospatial object detection in high-spatial-resolution remote sensing images. In Proceedings of the 2014 IEEE Geoscience and Remote Sensing Symposium, Quebec City, QC, Canada, 13–18 July 2014; pp. 2479–2482.
33. Zhou, P.; Zhang, D.; Cheng, G.; Han, J. Negative bootstrapping for weakly supervised target detection in remote sensing images. In Proceedings of the 2015 IEEE International Conference on Multimedia Big Data, Beijing, China, 20–22 April 2015; pp. 318–323.
34. Zhou, P.; Cheng, G.; Liu, Z.; Bu, S.; Hu, X. Weakly supervised target detection in remote sensing images based on transferred deep features and negative bootstrapping. *Multidimens. Syst. Signal Process.* **2016**, *27*, 925–944. [CrossRef]
35. 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]
36. Cao, L.; Luo, F.; Chen, L.; Sheng, Y.; Wang, H.; Wang, C.; Ji, R. Weakly supervised vehicle detection in satellite images via multi-instance discriminative learning. *Pattern Recognit.* **2017**, *64*, 417–424. [CrossRef]
37. Sheng, Y.; Cao, L.; Wang, C.; Li, J. Weakly Supervised Vehicle Detection in Satellite Images via Multiple Instance Ranking. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 2765–2770.
38. 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]
39. Ji, J.; Zhang, T.; Yang, Z.; Jiang, L.; Zhong, W.; Xiong, H. Aircraft detection from remote sensing image based on a weakly supervised attention model. In Proceedings of the IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 322–325.
40. 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]
41. 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.* **2020**, *59*, 675–685. [CrossRef]
42. 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]
43. Shamsolmoali, P.; Chanussot, J.; Zareapoor, M.; Zhou, H.; Yang, J. Multipatch feature pyramid network for weakly supervised object detection in optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5610113. [CrossRef]
44. Sun, Y.; Ran, J.; Yang, F.; Gao, C.; Kurozumi, T.; Kimata, H.; Ye, Z. Oriented Object Detection For Remote Sensing Images Based On Weakly Supervised Learning. In Proceedings of the 2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Shenzhen, China, 5–9 July 2021; pp. 1–6.
45. 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]
46. 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]
47. Wang, B.; Zhao, Y.; Li, X. Multiple instance graph learning for weakly supervised remote sensing object detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5613112. [CrossRef]
48. Li, W.; Hsu, C.Y.; Hu, M. Tobler's First Law in GeoAI: A spatially explicit deep learning model for terrain feature detection under weak supervision. *Ann. Am. Assoc. Geogr.* **2021**, *111*, 1887–1905. [CrossRef]
49. Feng, X.; Yao, X.; Cheng, G.; Han, J.; Han, J. Saenet: Self-supervised adversarial and equivariant network for weakly supervised object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5610411. [CrossRef]

50. Long, Y.; Zhai, X.; Wan, Q.; Tan, X. Object Localization in Weakly Labeled Remote Sensing Images Based on Deep Convolutional Features. *Remote Sens.* **2022**, *14*, 3230. [CrossRef]
51. Cheng, G.; Xie, X.; Chen, W.; Feng, X.; Yao, X.; Han, J. Self-guided Proposal Generation for Weakly Supervised Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5625311. [CrossRef]
52. 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]
53. Zitnick, C.L.; Dollár, P. Edge Boxes: Locating Object Proposals from Edges. In *Computer Vision—ECCV 2014, Proceedings of the 13th European Conference, Zurich, Switzerland, 6–12 September 2014*; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 391–405.
54. Zou, Z.; Shi, Z.; Guo, Y.; Ye, J. Object detection in 20 years: A survey. *arXiv* **2019**, arXiv:1905.05055.
55. Dietterich, T.G.; Lathrop, R.H.; Lozano-Pérez, T. Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.* **1997**, *89*, 31–71. [CrossRef]
56. Andrews, S.; Tsochantaridis, I.; Hofmann, T. Support vector machines for multiple-instance learning. *Adv. Neural Inf. Process. Syst.* **2002**, *15*, 561–568.
57. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
58. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.
59. Webb, G.I. Bayes' Rule. In *Encyclopedia of Machine Learning and Data Mining*; Sammut, C., Webb, G.I., Eds.; Springer: Boston, MA, USA, 2017; p. 99._21. [CrossRef]
60. Salakhutdinov, R.; Hinton, G. Deep Boltzmann Machines. In *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*; van Dyk, D., Welling, M., Eds.; PMLR: Clearwater Beach, FL, USA, 2009; Volume 5, pp. 448–455.
61. Krizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
62. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *arXiv* **2015**, arXiv:1506.01497.
63. 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.* **2018**, *42*,176–191. [CrossRef]
64. Bengio, Y.; Louradour, J.; Collobert, R.; Weston, J. Curriculum learning. In Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, QC, Canada, 14–18 June 2009; pp. 41–48.
65. Sangineto, E.; Nabi, M.; Culibrk, D.; Sebe, N. Self paced deep learning for weakly supervised object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 712–725. [CrossRef]
66. Zhang, D.; Han, J.; Zhao, L.; Meng, D. Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework. *Int. J. Comput. Vis.* **2019**, *127*, 363–380. [CrossRef]
67. Wan, F.; Wei, P.; Jiao, J.; Han, Z.; Ye, Q. Min-Entropy Latent Model for Weakly Supervised Object Detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1297–1306. [CrossRef]
68. Ahn, J.; Cho, S.; Kwak, S. Weakly Supervised Learning of Instance Segmentation with Inter-Pixel Relations. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 2204–2213. [CrossRef]
69. Ren, Z.; Yu, Z.; Yang, X.; Liu, M.Y.; Lee, Y.J.; Schwing, A.G.; Kautz, J. Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10598–10607.
70. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Object Detectors Emerge in Deep Scene CNNs. *arXiv* **2014**, arXiv:1412.6856.
71. Oquab, M.; Bottou, L.; Laptev, I.; Sivic, J. Is object localization for free? Weakly-supervised learning with convolutional neural networks. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 685–694. [CrossRef]
72. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.
73. Wei, J.; Wang, Q.; Li, Z.; Wang, S.; Zhou, S.K.; Cui, S. Shallow feature matters for weakly supervised object localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 5993–6001.
74. Yang, S.; Kim, Y.; Kim, Y.; Kim, C. Combinational Class Activation Maps for Weakly Supervised Object Localization. *arXiv* **2019**, arXiv:1910.05518.
75. Xue, H.; Liu, C.; Wan, F.; Jiao, J.; Ji, X.; Ye, Q. DANet: Divergent Activation for Weakly Supervised Object Localization. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 6588–6597. [CrossRef]

76. Mai, J.; Yang, M.; Luo, W. Erasing Integrated Learning: A Simple Yet Effective Approach for Weakly Supervised Object Localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
77. Wang, K.; Oramas, J.; Tuytelaars, T. MinMaxCAM: Improving object coverage for CAM-basedWeakly Supervised Object Localization. *arXiv* **2021**, arXiv:2104.14375.
78. Hou, L.; Lu, K.; Xue, J. Refined One-Stage Oriented Object Detection Method for Remote Sensing Images. *IEEE Trans. Image Process.* **2022**, *31*, 1545–1558. [CrossRef]
79. Dong, Z.; Wang, M.; Wang, Y.; Liu, Y.; Feng, Y.; Xu, W. Multi-Oriented Object Detection in High-Resolution Remote Sensing Imagery Based on Convolutional Neural Networks with Adaptive Object Orientation Features. *Remote Sens.* **2022**, *14*, 950. [CrossRef]
80. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
81. Gao, G.; Liu, L.; Zhao, L.; Shi, G.; Kuang, G. An adaptive and fast CFAR algorithm based on automatic censoring for target detection in high-resolution SAR images. *IEEE Trans. Geosci. Remote Sens.* **2008**, *47*, 1685–1697. [CrossRef]
82. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
83. Defard, T.; Setkov, A.; Loesch, A.; Audigier, R. Padim: A patch distribution modeling framework for anomaly detection and localization. In Proceedings of the International Conference on Pattern Recognition, Virtual, 10–15 January 2021; Springer: Cham, Switzerland, 2021; pp. 475–489.
84. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
85. Cramer, M. The DGPF-Test on Digital Airborne Camera Evaluation Overview and Test Design. *Photogramm.-Fernerkund.-Geoinf.* **2010**, *2010*, 73–82. [CrossRef]
86. Li, K.; Cheng, G.; Bu, S.; You, X. Rotation-insensitive and context-augmented object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2017**, *56*, 2337–2348. [CrossRef]
87. 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]
88. Chen, H.; Shi, Z. A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection. *Remote Sens.* **2020**, *12*, 1662. [CrossRef]
89. Deselaers, T.; Alexe, B.; Ferrari, V. Weakly supervised localization and learning with generic knowledge. *Int. J. Comput. Vis.* **2012**, *100*, 275–293. [CrossRef]
90. Xu, S.; Fang, T.; Li, D.; Wang, S. Object Classification of Aerial Images with Bag-of-Visual Words. *IEEE Geosci. Remote Sens. Lett.* **2010**, *7*, 366–370. [CrossRef]
91. Han, J.; Zhou, P.; Zhang, D.; Cheng, G.; Guo, L.; Liu, Z.; Bu, S.; Wu, J. Efficient, simultaneous detection of multi-class geospatial targets based on visual saliency modeling and discriminative learning of sparse coding. *ISPRS J. Photogramm. Remote Sens.* **2014**, *89*, 37–48. [CrossRef]
92. Wang, J.; Yang, J.; Yu, K.; Lv, F.; Huang, T.; Gong, Y. Locality-constrained linear coding for image classification. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 3360–3367.
93. Sun, H.; Sun, X.; Wang, H.; Li, Y.; Li, X. Automatic Target Detection in High-Resolution Remote Sensing Images Using Spatial Sparse Coding Bag-of-Words Model. *IEEE Geosci. Remote Sens. Lett.* **2012**, *9*, 109–113. [CrossRef]
94. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
95. Zhang, X.; Wei, Y.; Feng, J.; Yang, Y.; Huang, T.S. Adversarial complementary learning for weakly supervised object localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1325–1334.
96. Xie, X.; Cheng, G.; Wang, J.; Yao, X.; Han, J. Oriented R-CNN for Object Detection. *arXiv* **2021**, arXiv:2108.05699.
97. Papadopoulos, D.P.; Uijlings, J.R.; Keller, F.; Ferrari, V. We Don't Need No Bounding-Boxes: Training Object Class Detectors Using Only Human Verification. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; IEEE Computer Society: Los Alamitos, CA, USA, 2016; pp. 854–863. [CrossRef]
98. Albelwi, S. Survey on Self-Supervised Learning: Auxiliary Pretext Tasks and Contrastive Learning Methods in Imaging. *Entropy* **2022**, *24*, 551. [CrossRef]
99. Wang, Y.; Albrecht, C.M.; Braham, N.A.A.; Mou, L.; Zhu, X.X. Self-supervised Learning in Remote Sensing: A Review. *arXiv* **2022**, arXiv:2206.13188.