



Article

Weakly Supervised Object Detection for Remote Sensing Images via Progressive Image-Level and Instance-Level Feature Refinement

Shangdong Zheng ¹, Zebin Wu ^{2,*}, Yang Xu ² and Zhihui Wei ²

¹ School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China; zhengsd@njupt.edu.cn

² School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China; xuyangth90@njust.edu.cn (Y.X.); gswei@njust.edu.cn (Z.W.)

* Correspondence: wuzb@njust.edu.cn

Abstract: Weakly supervised object detection (WSOD) aims to predict a set of bounding boxes and corresponding category labels for instances with only image-level supervisions. Compared with fully supervised object detection, WSOD in remote sensing images (RSIs) is much more challenging due to the vast foreground-related context regions. In this paper, we propose a progressive image-level and instance-level feature refinement network to address the problems of missing detection and part domination for WSOD in RSIs. Firstly, we propose a multi-label attention mining loss (MAML)-guided image-level feature refinement branch to effectively allocate the computational resources towards the most informative part of images. With the supervision of MAML, all latent instances in images are emphasized. However, image-level feature refinement further expands responsive gaps between the informative part and other sub-optimal informative ones, which results in exacerbating the problem of part domination. In order to alleviate the above-mentioned limitation, we further construct an instance-level feature refinement branch to re-balance the contributions of different adjacent candidate bounding boxes according to the detection task. An instance selection loss (ISL) is proposed to progressively boost the representation of salient regions by exploring supervision from the network itself. Finally, we integrate the image-level and instance-level feature refinement branches into a complete network and the proposed MAML and ISL functions are merged with class classification and box regression to optimize the whole WSOD network in an end-to-end training fashion. We conduct experiments on two popular WSOD datasets, NWPU VHR-10.v2 and DIOR. All the experimental results demonstrate that our method achieves a competitive performance compared with other state-of-the-art approaches.

Keywords: weakly supervised learning; remote sensing images; object detection; image-level feature refinement; instance-level feature refinement



Citation: Zheng, S.; Wu, Z.; Xu, Y.; Wei, Z. Weakly Supervised Object Detection for Remote Sensing Images via Progressive Image-Level and Instance-Level Feature Refinement.

Remote Sens. **2024**, *16*, 1203. <https://doi.org/10.3390/rs16071203>

Academic Editor: Pedro Melo-Pinto

Received: 17 January 2024

Revised: 14 March 2024

Accepted: 27 March 2024

Published: 29 March 2024



Copyright: © 2024 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

Convolutional neural networks (CNNs) have facilitated the progress of different computer visual tasks [1–6], including object detection [7–10] in remote sensing images (RSIs). However, accurate bounding box (BB)-level annotations [11–13] are difficult to obtain. Due to the lacking of professional knowledge, manually annotating image- and BB-level labels by staff can be labor intensive and imprecise. In this paper, we focus on weakly supervised learning (WSL) for object detection.

Most of the previous weakly supervised object detection (WSOD) methods [14–17] treated WSL as a multiple instance learning (MIL) problem. First, images are defined as bags and different region proposal methods [18,19] are utilized to generate the candidate regions. The weakly supervised deep detection network (WSDDN) was proposed in [14], which simultaneously performed region selection and category classification tasks. Tang et al. [15]

proposed a multiple online instance classifier refinement (OICR) to propagate category label inference for refining the object classifier. The OICR selected the top-scoring proposal and assigned the same labels for its spatially high-overlapped regions. Tang [20] also designed a proposal cluster learning (PCL) network to boost the performance of the base OICR. The PCL network divided each instance bag into various small new bags and reduced the equivocal supervisions for these bags.

According to [21–28], WSOD in RSIs still encounters two major challenges. First of all, most previous methods tend to focus on the most discriminative regions in an image (part domination). For example, airfoil and fuselage regions might have the greatest contributions to recognizing airplane instances. However, using an airfoil or fuselage part to represent the entire airplane instance is obviously problematic. This problem is further aggravated in RSIs due to the vast foreground-related context regions. Second, there are always multiple instances with the same category in RSIs. Most WSOD methods use the highest-scoring proposals as pseudo-supervision to optimize the detection model, which will result in omitting many valuable positive proposals (missing detection).

Due to the lack of BB annotations, WSOD methods have no proper guidance and can only rely on the ill-posed training operation. As an effective way to provide reliable priors, an attention mechanism (AM) is used as a component to allocate the computational resources towards the most informative part of images. Thus, all the useful information can be highlighted. However, as illustrated in the second row of Figure 1, image-level features refined by the attention mechanism denote ‘airfoil’ as a detection prediction for the class ‘airplane’, resulting in inferior performance of the WSOD (part domination). Meanwhile, the uncontrollable nature of the attention mechanism can further exacerbate the aforementioned problem. On the contrary, exploring the spatial diversification constraints is a valid method to alleviate the problem of part domination. As illustrated in the third row of Figure 1, instance-level features refined by the spatial diversification constraints have great potential for excavating the complete instance according to instance-associative learning.

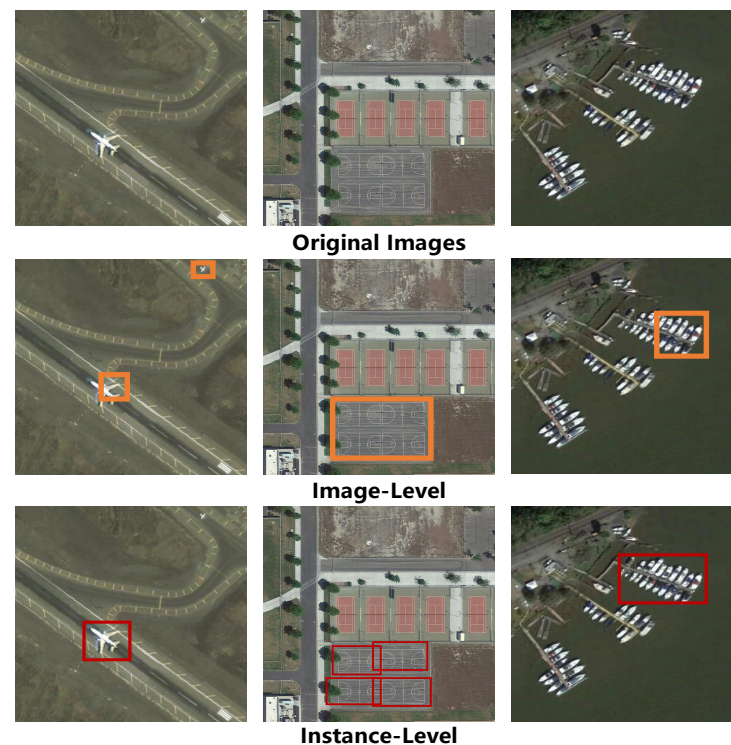


Figure 1. Visualizations of boosting deep features at image level and instance level. The orange bounding boxes demonstrate that refining features at image level could capture the most discriminative regions in one object (e.g., airfoil and fuselage). The red bounding boxes demonstrate that refining the instance-level features has great potential to perceive the full instance.

In this paper, we propose a progressive image-level and instance-level feature refinement network to perceive more latent positive proposals and excavate the complete instance for WSOD in RSIs. First, we propose an image-level feature refinement branch, which involves the attention mechanism to excavate all the informative regions in features. A novel multi-label attention mining loss (MAML) is constructed to train the attention mask as a category-specific descriptor. For the instance-level feature refinement branch, we propose a salient RoI feature selection strategy to pick out all crucial regions with its spatially high-overlapped proposals to model the position correlation among different instance-level features. An instance selection loss (ISL) is proposed to further boost the representation of the selected salient regions by exploring supervision from the network itself. Finally, we jointly optimize the region classification and box regression with our constructed MAML and ISL loss functions. Our contributions can be summarized as follows:

- We propose a progressive image-level and instance-level feature refinement network to perceive more latent positive proposals and explore the instance-associative spatial correlation among instance regions.
- A multi-label attention mining loss (MAML) and an instance selection loss (ISL) are constructed to boost the representation of image-level and instance-level features by exploring supervision from the network itself.
- We optimize the classification and regression tasks by the constructed MAML and ISL loss functions to boost the performance of WSOD in RSIs. The proposed method outperforms previous state-of-the-art approaches on the NWPU VHR-10.v2 and DIOR datasets, which demonstrates the effectiveness of boosting deep features at both image level and instance level for WSOD.

The organization of this paper is as follows: In Section 2, we introduce the studies related to our work. Then, in Section 3, we introduce the details of our method. In Section 4, we compare our approaches with other WSOD methods on the NWPU VHR-10.v2 and DIOR datasets. Finally, the conclusions on our proposed method are given in Section 5.

2. Related Work

2.1. Weakly Supervised Object Detection

Multiple instance learning (MIL) [29] is the most frequently used method in WSOD tasks. For example, ref. [14] proposed for the first time a WSDDN, a convolutional neural network-based WSOD method to perform region classification and selection tasks, simultaneously. With only image-level supervision, Tang et al. [15] combined MIL and the progressive refined instance classifier into a single detection network to construct an online instance classifier refinement (OICR) algorithm. The proposed OICR constructed multiple streams in a deep network, where each former branch is supervised by its latter one. Based on this, a host of follow-up studies [30,31] were developed.

Although WSOD has made significant progress in natural scenes, it is still a nontrivial task to directly apply these approaches in RSIs. According to [32–35], WSOD in RSIs still suffers from two major challenges: part domination and missing detection. Qian et al. [36] constructed a part-based topology distillation network (PTDNet) to perceive the combined instances via the extracted most informative parts of objects. Layer-wise relevance propagation (LRP) and point set representation (RepPoints) [37] were proposed to reduce the ambiguities in object recognition. The improved oriented loss function (IOLF) and pseudo-oriented bounding boxes were used as the supervisions to train the aforementioned modules. Qian et al. [38] proposed a pseudo-instance soft label (PSL) to evaluate whether each proposal covers a complete instance. Feng et al. [39] proposed a rotation-invariant aerial object detection network (RINet), which constructed a flexible multi-branch online detector refinement to perceive more rotated instances in RSIs. After coupling instance prediction with its different rotated variants, the rotation-consistent supervision was proposed to improve the performance of WSOD.

2.2. Feature Refinement For WSOD in RSIs

2.2.1. Image-Level Feature Refinement

In order to model complementary and discriminative visual patterns, a triple context-aware network (TCANet) [24] was proposed to capture the global visual context of image-level features. Moreover, the semantic discrepancy of the local context was leveraged by the constructed dual-local context residual (DLCR) module. Both the global and local image-level contexts were used to boost the performance of WSOD. Ma et al. [40] re-labeled an optical RSI dataset in a scribble manner and proposed a scribble embedding network (SEN) to extract discriminant regions in images. SEN unified correlation filtering technology into the proposed WSOD network, which successfully emphasizes the foreground features and suppresses the cluttered background. Shamsolmoali et al. [41] believed that most existing WSOD methods found it difficult to detect small objects in RSIs. The proposed multi-patch feature pyramid network (MPFP-Net) firstly divided different images into the class-affiliated subsets. An effective regularization strategy was constructed to perform a norm-preserving fusion transition. Therefore, image-level feature refinement has great potential for addressing the problem of missing detection in RSIs. However, re-balancing the computational resources towards all the informative parts of images inevitably exacerbates the problem of part domination. So, many researchers were devoted to modeling the interaction among instance-level features for perceiving the complete instances.

2.2.2. Instance-Level Feature Refinement

In [21], a multiple instance graph (MIG) learning framework was proposed with a spatial graph-based vote (SGV) mechanism to collect the highest-scoring proposals with its high spatial overlap regions as positive objects. Moreover, latent instances with the same category were excavated via the constructed appearance graph-based instance mining (AGIM) module. Xie et al. [42] introduced an attention erasing (AE) strategy into the WSOD model for dropping the most discriminative region. An IoU-balanced sampling component was constructed to excavate the complete instances, which further enhanced the performance of WSOD. Cheng et al. [23] proposed a self-guided proposal generation (SPG) module to excavate more reliable candidate proposals, which explicitly generated more high-quality instance-level features. Tan et al. [37] unified horizontal and oriented object detection tasks into a complete WSOD framework to detect different oriented instances in RSIs. Feng et al. [43] proposed an end-to-end progressive contextual instance refinement (PCIR) framework, which constructed a dual-contextual instance refinement (DCIR) strategy to re-balance the focus of the WSOD network from the local discriminative part to the complete instance. The complementary multiple instance detection network (CMIDN) and complementary feature learning (CEL) [44] module were proposed to excavate more latent objects. The contribution of each proposal was weighted in the final loss function.

3. The Proposed Method

The overview of our proposed method is illustrated in Figure 2. First, a VGG16 backbone network, which is pre-trained on a large-scale image-level classification dataset [11], is used to extract coarse image-level features F . Then, an image-level feature refinement branch is constructed to excavate all the informative regions in image. In order to guide the attention weights learning, we proposed a multi-label attention mining loss (MAML) function to enhance the localization ability of feature refinement with the supervision of image-level annotations. After feeding the enhanced image-level features F^{AM} into RoI Pooling layer, we obtain the original instance-level RoI features R . All the RoI features are fed into our instance-level feature refinement branch. Then, a set of salient RoIs R^{AB} can be obtained via the proposed salient RoI feature selection (SRFS) strategy. Because R^{AB} are selected by the category score vectors and GT image-level labels, these instance-level RoI features are more likely to contain the informative regions of certain categories. Therefore, we further construct an instance selection loss (ISL) function to provide the explicit indicator to guide the processing of instance-level feature selection. Finally, the enhanced instance-

level RoI features R^e are fed into a MIL branch to perform pseudo-location regression and proposal classification tasks. The algorithmic pipeline of our method is formulated in Algorithm 1.

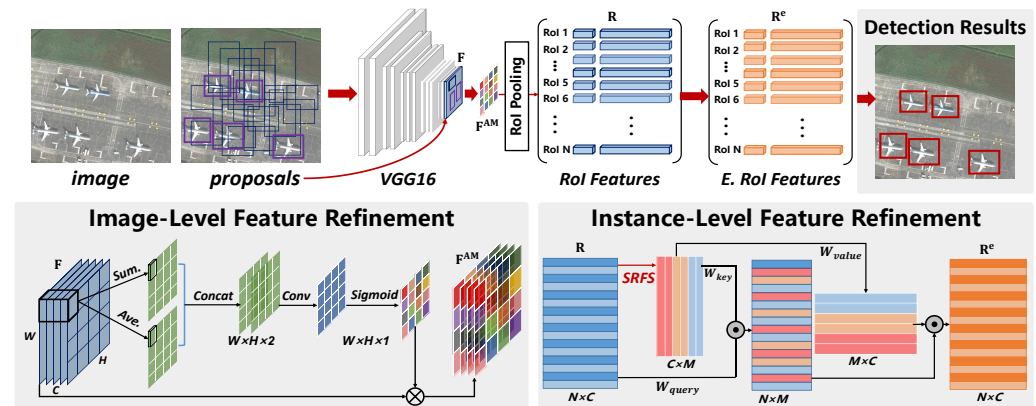


Figure 2. The overview of our proposed method. First, a pre-trained VGG16 backbone network is constructed to extract image features F . Then, an image-level feature refinement branch is constructed to progressively enhance the representation of backbone features. After feeding the enhanced image-level features F^{AM} into RoI Pooling layer, we obtain the original instance-level RoI features R . All the RoI features will be fed into our instance-level feature refinement branch. A set of salient RoIs and general RoIs can be obtained via the proposed salient RoI feature selection (SRFS). Finally, the enhanced (E.) instance-level RoI features R^e are used to attain the proposal classification results of our WSOD method. *Sum.*, *Ave.*, *Concat*, *Conv* and *FC* denote summation, average, concatenation, convolution and fully connection layers, respectively. W_{query} , W_{key} and W_{value} represent different linear layers, respectively.

Algorithm 1 Pipeline of image-level and instance-level feature refinement.

Require: Coarse backbone features F , image-level labels L , score threshold T_s , *IoU* threshold T_i , number of salient RoI T_n

Ensure: Enhanced instance-level features R^e , MAML loss function $\mathcal{L}_{MAML}$, ISL loss function $\mathcal{L}_{ISL}$

- 1: Generate attention map W by Equations (1) and (2)
 - 2: Calculate MAML loss $\mathcal{L}_{MAML}$ by Equations (4) and (5)
 - 3: Generate enhanced image-level features F^e by Equation (3)
 - 4: Generate instance-level features R by RoI Pooling layer
 - 5: Calculate proposal scores R^s by Equations (6)–(8)
 - 6: **if** image-level labels L contains category i **then**
 - 7: Sort scores R^s by confidence of category i
 - 8: **for** r in R **do**
 - 9: **if** sorted scores $r^s \geq T_s$ **then**
 - 10: Put r in R^A
 - 11: **for** r in $(R - R^A)$ **do**
 - 12: Calculate *IoU* between R^A and $(R - R^A)$
 - 13: **if** *IoU* $\geq T_i$ **then**
 - 14: Put r in R^B
 - 15: Calculate number of $(R^A$ and $R^B)$ by $M = A + B$
 - 16: **while** $M \leq T_n$ **do**
 - 17: **for** r in $(R - R^A - R^B)$ **do**
 - 18: Randomly put r in R^C
 - 19: $M = M + 1$
 - 20: Generate enhanced instance-level features R^e by Equations (9) and (10)
 - 21: Calculate ISL loss $\mathcal{L}_{ISL}$ by Equation (11)
-

3.1. Motivations

Due to the complex background information and the vast foreground-related context regions, it is easy for instances with small scales to be submerged by useless regions, resulting in the problem of missing detection. Thus, our proposed image-level feature refinement branch aims to excavate all the informative regions, which has great potential for perceiving all the target instances in images. However, taking the computational resources towards the most informative part of objects inevitably exacerbates the problem of part domination. For example, airfoil and fuselage regions might have the greatest contributions to detect airplane instances. These regions can be further emphasized by the constructed image-level feature refinement branch. However, using airfoil or fuselage to represent the entire airplane instance is obviously problematic. Therefore, we construct an instance-level feature refinement branch to explore the instance-associative spatial diversification constraints. As illustrated in the second and third rows of Figure 1, the image-level feature refinement branch successfully detects the small airplane while regarding the partial region of harbor as the final detection result. Conversely, our proposed instance-level feature refinement branch recognizes the complete harbor instance while missing the small airplane. The aforementioned analysis inclines us to unify the image-level and instance-level feature refinement branches into a complete network to boost the performance of WSOD in RSIs.

3.2. Image-Level Feature Refinement Branch

Given an intermediate feature map $F \in \mathbb{R}^{C \times W \times H}$ as input, the attention mechanism is adopted to determine a spatially normalized attention weight map to enhance the representation of F . Here, we follow the conventional pipeline to refine the image-level features.

Consider that the global max pooling (GMP) layer focuses on the area with the highest responsive value and reduces the impact of unrelated information, which aligns with the working principle of our proposed image-level feature refinement branch. Moreover, the channel-wise global average pooling (GAP) aggregates multiple representations of each pixel and encourages the network to distinguish instances with different categories. Therefore, we utilize a combination of GAP and GMP to generate the spatial context descriptor as

$$F^{AM} = [\frac{1}{c} \sum_{i=1}^c F_p^i, \max(F_p^i)_{i=1}^c], \quad (1)$$

where $F^{AM} \in \mathbb{R}^{2 \times W \times H}$. p and c denote the pixel position and channel dimension of F , respectively. $[\]$ represents the concatenation operation. These two features are fused to gather the contextual information of instances. We then use a convolutional layer C to boost F^{AM} as

$$W = \sigma(C(F^{AM})), \quad (2)$$

where $W \in \mathbb{R}^{1 \times W \times H}$ denotes the spatial attention mask and σ represents *Sigmoid* function. The final enhanced feature F^e can be formulated as

$$F^e = W \odot F, \quad (3)$$

where $\odot$ denotes the element-wise production.

Importantly, Attention mechanism is an effective way to re-balance the partialness of different regions in images which could perceive all the foreground instances. However, supervised by only image-level annotations, attention mechanism always confuses instances that are similar in appearances. To address the aforementioned limitation, we proposed a multi-label attention mining loss (MAML) function for excavating latent supervision from the network itself.

Sine we attain the attention mask W in Equation (2), we feed W into a convolutional layer C and the predicted vector $\hat{y}$ is determined by a GAP layer G . The aforementioned operations can be formulated as

$$\hat{y} = G(C(W)), \quad (4)$$

Finally, the predicted score vector $\hat{y}$ can be supervised by the MAML loss as

$$\mathcal{L}_{MAML} = - \sum_{c=1}^C (y_c \log \hat{y}_c + (1 - y_c) \log(1 - \hat{y}_c)), \quad (5)$$

where y_c denotes the ground truth of image-level classification annotations. Obviously, the constructed MAML supervision provides an explicit descriptor for detection network to estimate whether an image contains a certain category instance. The final attention-refined feature map F^e is fed into the subsequent network for instance-level feature refinement.

3.3. Instance-Level Feature Refinement Branch

Taking F^e with its candidate proposals into RoI Pooling layer [45], the fixed-size instance-level features $R \in \mathbb{R}^{N \times D}$ are normalized into the same scale, where N, D represent the number and the dimension of R , respectively. Commonly, previous WSOD methods sort these instance-level features by the predicted score vectors and use the highest scorer to optimize their detection networks. This process is problematic in WSOD because the candidate proposals contain a larger number of negative regions and a small number of positive proposals. Obviously, the positive proposals contribute more to detecting objects in RSIs. To alleviate aforementioned limitation, we propose a salient RoI feature selection (SRFS) strategy to pick out all crucial regions with its spatially high-overlapped proposals to model the position correlation among different instance-level features.

Given a set of instance-level features $R = [r_1, r_1, \dots, r_N]$, two fully connected layers are used to produce different vectors R^c and $R^d \in \mathbb{R}^{N \times C}$ where C denotes the number of categories. These two vectors are fused by two *Softmax* layers ρ as

$$[\rho(R^c)]_{ij} = \frac{e^{R_{ij}^c}}{\sum_{k=1}^C e^{R_{kj}^c}}, \quad (6)$$

$$[\rho(R^d)]_{ij} = \frac{e^{R_{ij}^d}}{\sum_{k=1}^N e^{R_{ik}^d}}, \quad (7)$$

where $\rho(R^c)$ and $\rho(R^d)$ denote classification and detection predictions. The proposal scores R^s are generated as

$$R^s = \rho(R^c) \odot \rho(R^d), \quad (8)$$

where $\odot$ denotes element-wise product.

3.3.1. Salient RoI Feature Selection

Since we attain the instance-level features $R \in \mathbb{R}^{N \times D}$ and the corresponding proposal scores $R^s \in \mathbb{R}^{N \times C}$, the top-ranking proposal with its highly spatially overlapped regions can roughly determine the localization of objects, because the aforementioned proposals cover the most discriminative regions in images. As illustrated in Figure 3, the image-level annotation is ‘airplane’. After transferring image-level label into the one-hot vector, we obtain the transferred annotation $y_c = [1, 0, 0, \dots, 0]$, where the index 0 of y_c denotes the corresponding category. Given the predicted scores R^s , we select the corresponding row (index 0) of the category score vectors and put RoI features with category score higher than a threshold T_s (0.5 in our experiments) into a subset of salient RoI features, named $R^A \in \mathbb{R}^{A \times D}$. Additionally, regions which are highly overlapped with R^A may belong to the same category. Therefore, we calculate the IoU between R^A and all other RoIs. When the IoU is larger than a threshold T_i (0.5 in our experiments), we put these RoI features into a subset of salient RoI features, named $R^B \in \mathbb{R}^{B \times D}$. Finally, we randomly pick the remaining set $(R - R^{AB})$ of RoI features to attain $R^C \in \mathbb{R}^{C \times D}$, until the number of the selected RoIs is equal to the defined threshold (300 in our model).

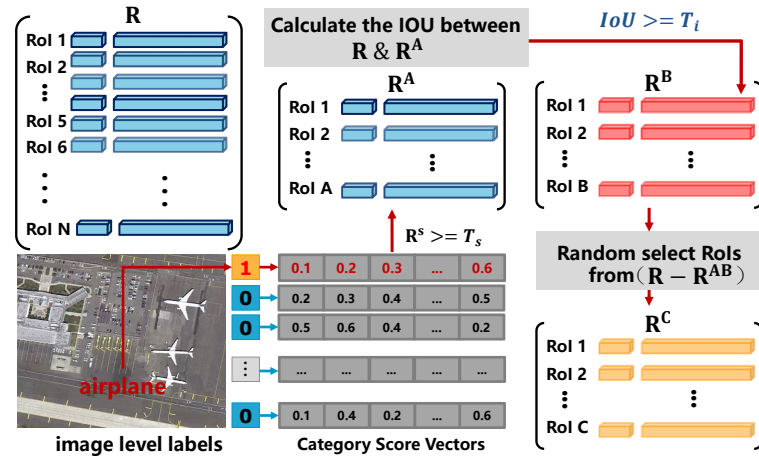


Figure 3. The overview of our salient RoI feature selection (SRFS). The collection of R^A includes the regions which are more likely to contain an informative image fragment of a certain class, which is attained by the predicted category score vectors. All the RoIs where the IoU is greater than 0.5 are put in set R^B . Finally, we randomly pick the remaining set of RoIs (R^C) to supplement the salient RoI feature set.

After obtaining the salient instance-level features $R^M \in \mathbb{R}^{M \times D}$ where $R^M = R^{ABC}$, we adopt self-attention mechanism [46] to model the correlation among these instance-level features.

$$w = \rho\left(\frac{W_{\text{query}}(R) \cdot W_{\text{key}}(R^M)^T}{\sqrt{d_{\text{key}}}}\right), \quad (9)$$

$$R^e = w \cdot W_{\text{value}}(R^M), \quad (10)$$

where M denotes the number of selected instance-level features; $w \in \mathbb{R}^{N \times M}$ represents the similarity weight; and W_{query} , W_{key} and W_{value} denote different *Linear* layers. The final enhanced instance-level features R^e are fed into the MIL branch for box regression and category classification.

3.3.2. Instance Selection Loss

We constructed an instance selection loss (ISL) function to progressively boost the representation of the selected RoI features. Firstly, all the selected salient proposals (except R^C) are more likely to contain the key region of one certain category. The rest should not trigger the network to recognize the object. Therefore, R^{AB} can be supervised by the loss [14]. The score of these proposals are generated by element-wise product $s^{AB} = \rho(W_4(R^{AB})) \cdot \rho(W_5(R^{AB}))$ where W_4 and W_5 denote different linear functions. The c -th class score can be obtained by summing up the score vectors over all proposals $y_c^{AB} = \sum_{k=1}^{AB} s_c^k$. Our selection loss is formulated as

$$\mathcal{L}_{ISL} = - \sum_{c=1}^C (y_c \log y_c^{AB} + (1 - y_c) \log(1 - y_c^{AB})), \quad (11)$$

where y_c is the ground truth of image-level classification annotations. With the supervision of $\mathcal{L}_{ISL}$, the network learns to model the inter-dependencies between the most salient instance-level RoI features and other proposals, which has great potential for perceiving the complete objects. We demonstrate the effectiveness of our instance-level feature refinement in Section 4.

3.4. MIL Branch

We use a pseudo-fully supervised classification and regression method [16] to refine the final results. Since the predicted instance labels can be obtained by the method in [14],

the standard bounding-box regression and category classification loss functions can be formulated as

$$\mathcal{L}_{DET} = \mathcal{L}_{CLS} + \lambda \mathcal{L}_{LOC}, \quad (12)$$

where $\mathcal{L}_{CLS}$ is classification loss and $\mathcal{L}_{LOC}$ is smooth L_1 loss. Moreover, our complete network is trained by optimizing the comprehensive loss function as

$$\mathcal{L} = \mathcal{L}_{MAML} + \mathcal{L}_{ISL} + \mathcal{L}_W + \mathcal{L}_R + \mathcal{L}_{DET}. \quad (13)$$

where $\mathcal{L}_W$ and $\mathcal{L}_R$ are the MLC loss of WSDDN [14] and refinement loss of OICR [15].

4. Experiments

4.1. Datasets and Evaluation Metrics

We evaluate our proposed method on the NWPU VHR-10.v2 <https://drive.google.com/file/d/15xd4TASVAC2irRf02GA4LqYFbH7QITR-/view?usp=sharing> (accessed on 20 March 2024) [22] and DIOR https://drive.google.com/drive/folders/1UdlgHk49iu6WpcJ5467iT-UqNPpx_CC (accessed on 20 March 2024) [47] datasets, which are the most frequently used benchmarks in RSI-based WSOD methods. In our experiments, we only adopt image-level annotations to train our model.

NWPU VHR-10.v2 [22] is modified from NWPU VHR-10 [48], which crops each image into 400×400 pixels. There are 1172 images in the NWPU VHR-10.v2 dataset. A total of 10 categories with 2775 instances are annotated in NWPU VHR-10.v2, namely, Airplane, Ship, Storage tank, Baseball diamond, Tennis court, Basketball court, Ground track field, Harbor, Bridge, and Vehicle. Following the standard protocols [21–23], our method uses 879 images for training and 293 images for testing.

DIOR is more challenging than NWPU VHR-10.v2 and contains 23,463 samples with a size of 800×800 pixels. A total of 20 categories with 192,518 instances are annotated in DIOR, namely, Airplane, Airport, Baseball field, Basketball court, Bridge, Chimney, Dam, Expressway service area, Expressway toll station, Golf field, Ground track field, Harbor, Overpass, Ship, Stadium, Storage tank, Tennis court, Train station, Vehicle, and Windmill. DIOR is divided into training, validation and testing sets. All the previous WSOD methods [14,15,20–24,37,38,43,49,50] use a combination of training and validation for training (11,725 images). The remaining 11,738 images in the testing set are used for testing. In order to facilitate fair comparisons, we also adopt the aforementioned settings.

The popular evaluation metrics for WSOD in RSIs are as follows: (1) We use the average precision (AP) and mean of AP (mAP) evaluation metrics to evaluate the performance of our method. (2) Correct localization (CorLoc) is another widely used WSOD evaluation metric to measure the accuracy of localization in the training set.

4.2. Implementation Details

In our experiments, only one RTX 1080ti GPU is used for training and testing. Following the standard settings in method [24], we use a pre-trained VGG16 network as the base feature extractor to extract image features. The selective search [18] tool is introduced to obtain roughly 2000 candidate proposals per image. For NWPU VHR-10.v2, the initial learning rate is set to 0.001 and a total of 30K iterations are used to train our model. The learning rate is decreased by 0.1 at iteration 20K. For DIOR, we train our model for 200K iterations with the learning rate decreased by 0.1 at iteration 100K. We set the batch size to 1 and the stochastic gradient descent (SGD) optimizer is adopted. The momentum and weight decay are set to 0.9 and 0.0005, respectively.

During training, we use five image scales {480, 576, 688, 864 and 1200} and adopt three rotation transformations {90°, 180° and 270°} as data augmentation strategies. During testing, we set the confidence threshold to 0.3 for performing non-maximum suppression (NMS).

4.3. Performance Evaluation

In order to evaluate the effectiveness of the proposed method, we compare our method with 12 WSOD approaches, including WSDDN [14], OICR [15], PCL [20], min-entropy latent model (MELM) [49], DCL [22], PCIR [43], MIG [21], TCANet [24], SPG [23], self-supervised adversarial and equivariant network (SAENet) [50], weakly supervised oriented detector (WSODet) [37] and PSL [38]. Most of the aforementioned approaches are published in the last two years. Following the methods in [37,38,50], we list the detection results reported in their published papers.

4.3.1. NWPU VHR-10.v2

As listed in Tables 1 and 2, WSDDN [14], OICR [15], PCL [20] and MELM [21] are the most frequently used baseline methods for WSOD, which achieve 35.1%, 34.5%, 39.4% and 42.3% mAP, and 35.3%, 40.0% 45.6% and 49.9% CorLoc, respectively. Although the detection performances of the aforementioned methods are inferior, they have laid solid foundations for the subsequent works. TCANet [24] crops the original image-level features into three sub-patches and models the interaction among these sub-features. However, re-balancing the computational resources towards the informative part of images inevitably exacerbates the problem of part domination, so that our proposed method achieves 6.4% mAP and 2.4% CorLoc gains over TCANet. Conversely, PCIR [43], MIG [21] and SPG [23] are all devoted to boosting the representation of instance-level features. However, due to the neglect of image-level feature refinement, it is difficult for the aforementioned three approaches to locate small instances. The best competitor, PSL [38], constructs the proposal quality score to excavate pseudo-bounding boxes of instances. Semantic segmentation is employed to perceive the complete object. However, it is difficult for the proposed pseudo-proposal generation strategy to cover objects with small scales. Through refining image-level features, our method achieves 1.4% mAP and 0.6% Corloc gains over PSL. Importantly, we also list the mAPs of many fully supervised object detection (FSOD) methods in Table 1. It can be clearly observed that the proposed method significantly narrows the accuracy gap between WSOD and FSOD. We also visualize many detection results in Figure 4 to analyze the effectiveness of our proposed method. The baseline method is biased towards bounding boxes which cover the most representative parts of the entire instances. Meanwhile, it is easier for objects with the same category in crowded scenes to be mis-detected by the baseline approach. As illustrated in the second row of Figure 4, our method can accurately cover most instances, which demonstrates that boosting deep features from the image level and instance level has great potential for addressing the problems of part domination and missing detection for WSOD in RSIs.

Table 1. Average precision (%) for different methods on the NWPU VHR-10.v2 testing set.

Methods	Airplane	Ship	Storage Tank	Baseball Diamond	Tennis Court	Basketball Court	Ground Track Field	Harbor	Bridge	Vehicle	mAP
Fully supervised:											
COPD [51]	62.3	69.4	64.5	82.1	34.1	35.3	84.2	56.3	16.4	44.3	54.9
Transferred CNN [52]	66.0	57.1	85.0	80.9	35.1	45.5	79.4	62.6	43.2	41.3	59.6
RICNN [53]	88.7	78.3	86.3	89.1	42.3	56.9	87.7	67.5	62.3	72.0	73.1
RCNN [54]	85.4	88.9	62.8	19.7	90.7	58.2	68.0	79.9	54.2	49.9	65.8
Fast RCNN [55]	90.9	90.6	89.3	47.3	100.0	85.9	84.9	88.2	80.3	69.8	82.7
Faster RCNN [45]	90.9	86.3	90.5	98.2	89.7	69.6	100.0	80.1	61.5	78.1	84.5
RICO [48]	99.7	90.8	90.6	92.9	90.3	80.1	90.8	80.3	68.5	87.1	87.1
Weakly supervised:											
WSDDN [14]	30.1	41.7	35.0	88.9	12.9	23.9	99.4	13.9	1.9	3.6	35.1
OICR [15]	13.7	67.4	57.2	55.2	13.6	39.7	92.8	0.2	1.8	3.7	34.5
PCL [20]	26.0	63.8	2.5	89.8	64.5	76.1	77.9	0.0	1.3	15.7	39.4
MELM [49]	80.9	69.3	10.5	90.2	12.8	20.1	99.2	17.1	14.2	8.7	42.3
DCL [22]	72.7	74.3	37.1	82.6	36.9	42.3	84.0	39.6	16.8	35.0	52.1
DPLG [28]	80.9	10.5	90.1	64.4	69.1	80.2	8.7	14.0	39.6	78.3	53.6
PCIR [43]	90.8	78.8	36.4	90.8	22.6	52.2	88.5	42.4	11.7	35.5	55.0
MIG [21]	88.7	71.6	75.2	94.2	37.6	47.7	100.0	27.3	8.3	9.1	56.0
TCANet [24]	89.4	78.2	78.4	90.8	35.3	50.4	90.9	42.4	4.1	28.3	58.8
SAENet [50]	82.9	74.5	50.2	96.7	55.7	72.9	100.0	36.5	6.3	31.9	60.7
WSODet [37]	95.3	75.6	81.9	98.0	20.9	56.7	100.0	29.8	5.1	48.1	61.3
SPG [23]	90.4	81.0	59.5	92.3	35.6	51.4	99.9	58.7	17.0	43.0	62.8
PSL [38]	87.6	80.1	57.3	94.0	36.4	80.4	100.0	56.9	9.8	35.6	63.8
Ours	90.8	81.6	56.6	91.7	51.9	69.5	100.0	53.4	16.3	40.5	65.2

Table 2. Correct localization (%) for different methods on the NWPU VHR-10.v2 trainval set.

Methods	Airplane	Ship	Storage Tank	Baseball Diamond	Tennis Court	Basketball Court	Ground Track Field	Harbor	Bridge	Vehicle	mAP
WSDDN [14]	22.3	36.8	40.0	92.5	18.0	24.2	99.3	14.8	1.7	2.9	35.3
OICR [15]	29.4	83.3	20.5	81.8	40.9	32.1	86.6	7.4	3.7	14.4	40.0
PCL [20]	11.8	50.0	12.8	98.7	84.5	77.4	90.7	0.0	9.3	15.6	45.6
MELM [49]	86.0	77.4	21.4	98.3	10.7	43.5	95.0	40.0	11.8	14.6	49.9
DPLG [28]	87.2	16.8	96.1	75.1	73.2	86.3	16.3	18.7	46.7	85.1	61.5
PCIR [43]	100.0	93.1	64.1	99.3	64.8	79.3	89.7	63.0	13.3	52.2	71.9
MIG [21]	97.8	90.3	87.2	98.7	54.9	64.2	100.0	74.1	13.0	21.6	70.2
TCANet [24]	96.9	91.8	95.1	88.7	66.9	62.8	96.0	54.2	19.6	55.6	72.8
SAENet [50]	97.1	91.7	87.8	98.7	40.9	81.1	100.0	70.4	14.8	52.2	73.5
WSODet [37]	96.7	93.2	82.5	99.5	40.5	58.0	100.0	67.7	9.7	73.3	72.4
SPG [23]	98.1	92.7	70.1	99.7	51.9	80.1	96.2	72.4	13.0	60.0	73.4
PSL [38]	94.4	86.6	68.5	97.8	69.8	87.5	100.0	68.6	16.0	56.6	74.6
Ours	98.6	94.7	76.4	83.9	61.3	82.4	100.0	78.1	19.5	57.2	75.2

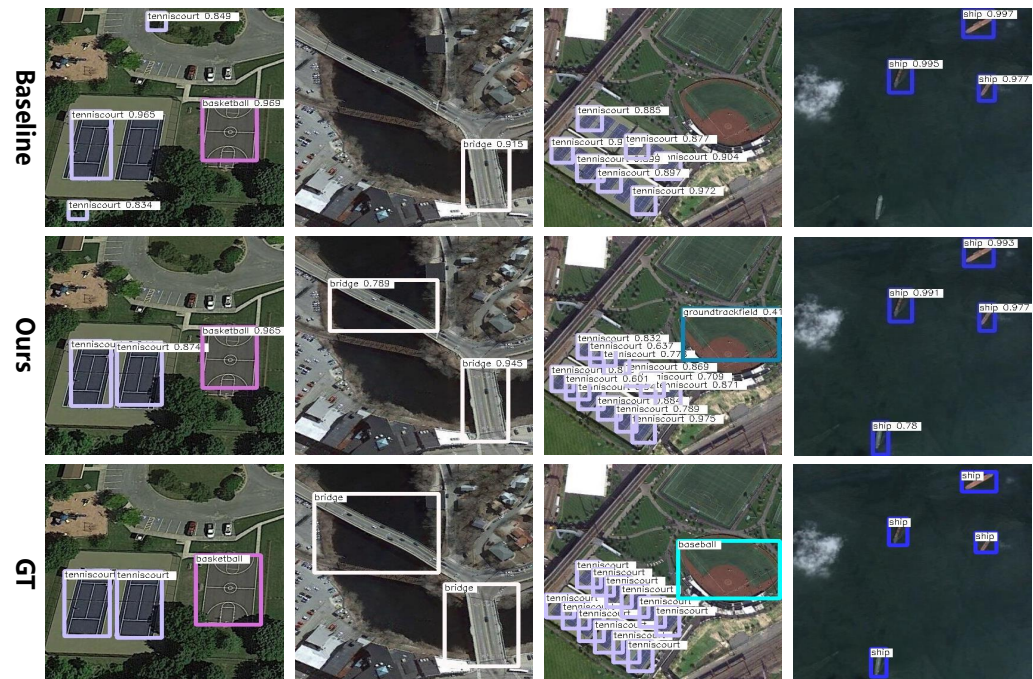


Figure 4. Visualizations of different methods on NWPU VHR-10.v2. The first row shows the detection results of baseline method and the second row shows the detection results of our proposed method. Taking the third row, visualizations of the ground true, as an indicator, we can clearly see that our method accurately detects the most instances.

4.3.2. DIOR

We list the detailed comparisons on the more challenging DIOR dataset in Tables 3 and 4. Consider that objects in DIOR are much smaller and more densely packed than in NWPU VHR-10.v2. Therefore, perceiving more latent positive proposals (our image-level feature refinement branch) and modeling the instance-associative spatial correlation among instances (our instance-level feature refinement branch) has great potential for alleviating the problems of missing detection and part domination. The evaluation criteria mAP and CorLoc also prove the effectiveness of our approach. Specifically, the proposed method outperforms the WSDDN [14], OICR [15], PCL [20], MELM [49], DCL [22], PCIR [43], MIG [21], TCANet [24], SPG [23], SAENet [50], WSODet [37] and PSL [38] by 15.8% (29.1% v.s. 13.3%), 12.6% (29.1% v.s. 16.5%), 10.9% (29.1% v.s. 18.2%), 10.4% (29.1% v.s. 18.7%), 8.9% (29.1% v.s. 20.2%), 4.2% (29.1% v.s. 24.9%), 4.0% (29.1% v.s. 25.1%), 3.3% (29.1% v.s. 25.8%), 3.3% (29.1% v.s. 25.8%), 2.0% (29.1% v.s. 27.1%), 1.8% (29.1% v.s. 27.3%) and 0.5% (29.1% v.s. 28.6%), which are notable margins in terms of mAP. Figure 5 illustrates some detection comparisons of many approaches on DIOR dataset. It can be clearly seen that the baseline method suffers from distinguishing instances with similar appearances. Conversely, our method captures more complete objects and reduces the instance ambiguity induced from extreme scales and poses, as well as similar appearances.

Table 3. Average precision (%) for different methods on the DIOR testing set.

Methods	Airplane	Airport	Baseball Field	Basketball Court	Bridge	Chimney	Dam	Expressway Service Area	Expressway Toll Station	Golf Field	mAP
WSDDN [14]	9.1	39.7	37.8	20.2	0.3	12.2	0.6	0.7	11.9	4.9	13.3
OICR [15]	8.7	28.3	44.1	18.2	1.3	20.2	0.1	0.7	29.9	13.8	16.5
PCL [20]	21.5	35.2	59.8	23.5	3.0	43.7	0.1	0.9	1.5	2.9	18.2
MELM [49]	28.1	3.2	62.5	28.7	0.1	62.5	0.2	13.1	28.4	15.2	18.7
DCL [22]	20.9	22.7	54.2	11.5	6.0	61.0	0.1	1.1	31.0	30.9	20.2
PCIR [43]	30.4	36.1	54.2	26.6	9.1	58.6	0.2	9.7	36.2	32.6	24.9
MIG [21]	22.2	52.6	62.8	25.8	8.5	67.4	0.7	8.9	28.7	57.3	25.1
TCANet [24]	25.1	30.8	62.9	40.0	4.1	67.8	8.1	23.8	29.9	22.3	25.8
SPG [23]	31.3	36.7	62.8	29.1	6.1	62.7	0.3	15.0	30.1	35.0	25.8
SAENet [50]	20.6	62.7	62.7	23.5	7.6	64.6	0.2	34.5	30.6	55.4	27.1
WSODet [37]	32.2	53.3	66.5	76.6	0.1	57.1	0.1	0.1	0.4	42.8	27.3
PSL [38]	29.1	49.8	70.9	41.4	7.2	45.5	0.2	35.4	36.8	60.8	28.6
Ours	32.9	70.5	63.2	45.7	0.2	69.7	0.2	12.4	39.4	56.4	29.1
Methods	Ground Track Field	Harbor	Overpass	Ship	Stadium	Storage Tank	Tennis Court	Train Station	Vehicle	Windmill	mAP
WSDDN [14]	42.4	4.7	1.1	0.7	63.0	4.0	6.1	0.5	4.6	1.1	13.3
OICR [15]	57.4	10.7	11.1	9.1	59.3	7.1	0.7	0.1	9.1	0.4	16.5
PCL [20]	56.4	16.8	11.1	9.1	57.6	9.1	2.5	0.1	4.6	4.6	18.2
MELM [49]	41.1	26.1	0.4	9.1	8.6	15.0	20.6	9.8	0.1	0.5	18.7
DCL [22]	56.5	5.1	2.7	9.1	63.7	9.1	10.4	0.0	7.3	0.8	20.2
PCIR [43]	58.5	8.6	21.6	12.1	64.3	9.1	13.6	0.3	9.1	7.5	24.9
MIG [21]	47.7	23.8	0.8	6.4	54.1	13.2	4.1	14.8	0.2	2.4	25.1
TCANet [24]	53.9	24.8	11.1	9.1	46.4	13.7	31.0	1.5	9.1	1.0	25.8
SPG [23]	48.0	27.1	12.0	10.0	60.0	15.1	21.0	9.9	3.2	0.1	25.8
SAENet [50]	52.7	17.6	6.9	9.1	51.6	15.4	1.7	14.4	1.4	9.2	27.1
WSODet [37]	66.6	0.1	1.9	2.0	52.6	22.4	68.8	0.2	1.2	0.3	27.3
PSL [38]	48.5	14.0	25.1	18.5	48.9	11.7	11.9	3.5	11.3	1.7	28.6
Ours	55.3	16.6	0.6	9.1	54.8	18.1	11.0	16.1	9.1	1.1	29.1

Table 4. Correct localization (%) for different methods on the DIOR trainval set.

Methods	Airplane	Airport	Baseball Field	Basketball Court	Bridge	Chimney	Dam	Expressway Service Area	Expressway Toll Station	Golf Field	mAP
WSDDN [14]	5.7	59.9	94.2	55.9	4.9	23.4	1.0	6.8	44.5	12.8	32.4
OICR [15]	16.0	51.5	94.8	55.8	3.6	23.9	0.0	4.8	56.7	22.4	34.8
PCL [20]	61.1	46.9	95.4	63.6	7.3	95.1	0.2	5.7	5.1	50.8	41.5
MELM [49]	77.0	28.9	92.7	63.0	13.0	90.1	0.2	17.0	37.9	44.6	43.3
PCIR [43]	81.6	51.3	96.2	73.5	5.0	94.7	15.9	32.8	46.0	48.6	48.4
MIG [21]	77.0	46.9	95.4	63.6	23.0	95.1	0.2	17.0	57.9	50.8	46.8
TCANet [24]	91.2	69.4	95.5	67.5	18.9	97.8	0.2	70.5	54.3	51.4	49.4
SPG [23]	80.5	32.0	98.7	65.0	15.2	96.1	22.5	17.0	46.1	51.0	48.3
SAENet [50]	91.2	69.4	95.5	67.5	18.9	97.8	0.2	70.5	54.3	51.4	49.4
WSODet [37]	95.2	81.0	95.7	88.0	5.9	94.1	1.4	1.1	3.7	92.1	49.5
Ours	88.3	69.1	98.8	69.4	19.9	97.8	0.3	24.7	56.2	54.4	52.3
Methods	Ground Track Field	Harbor	Overpass	Ship	Stadium	Storage Tank	Tennis Court	Train Station	Vehicle	Windmill	mAP
WSDDN [14]	89.9	5.5	10.0	23.0	98.5	79.6	15.1	3.5	11.6	3.2	32.4
OICR [15]	91.4	18.2	18.7	31.8	98.3	81.3	7.5	1.2	15.8	2.0	34.8
PCL [20]	89.4	42.1	19.8	37.9	97.9	80.7	13.8	0.2	10.5	6.9	41.5
MELM [49]	88.1	49.4	15.7	28.2	98.3	83.0	22.8	10.3	4.6	2.2	43.3
PCIR [43]	85.3	38.9	20.2	30.6	84.6	91.5	56.3	3.8	10.5	1.3	48.4
MIG [21]	89.4	42.1	19.8	37.9	97.9	80.7	13.8	10.3	10.5	6.9	46.8
TCANet [24]	88.3	48.0	2.3	33.6	14.1	83.4	65.6	19.9	16.4	2.9	49.4
SPG [23]	89.2	49.5	22.0	35.2	98.6	90.0	32.6	12.7	10.0	2.3	48.3
SAENet [50]	88.3	48.0	2.3	33.6	14.1	83.4	65.6	19.9	16.4	2.9	49.4
WSODet [37]	95.3	1.4	13.6	43.1	95.9	90.7	89.4	0.2	17.1	5.7	49.5
Ours	89.6	49.1	19.4	34.5	96.7	84.7	63.2	15.6	11.6	3.4	52.3



Figure 5. Visualizations of different methods on DIOR. The first row shows the detection results of baseline method and the second row shows the detection results of our proposed method. Taking the third row, visualizations of the ground true, as an indicator, we can clearly see that our method accurately detects the most instances.

Additionally, we further count the number of different categories in DIOR. As illustrated in Figure 6, most classes in the training and testing sets are approximately equal. The number ratio of five categories in the training and testing sets is approximately 1:2. The number ratio of two categories in the training and testing sets varies greatly. It can be clearly seen that the standard partitioning of the dataset indeed causes the unbalanced distribution of categories. The aforementioned statistical overview inspires our subsequent work. Based on the standard dataset partitioning principle, designing the category-balancing strategy is a potentially useful attempt to boost the performance of WSOD in the DIOR dataset.

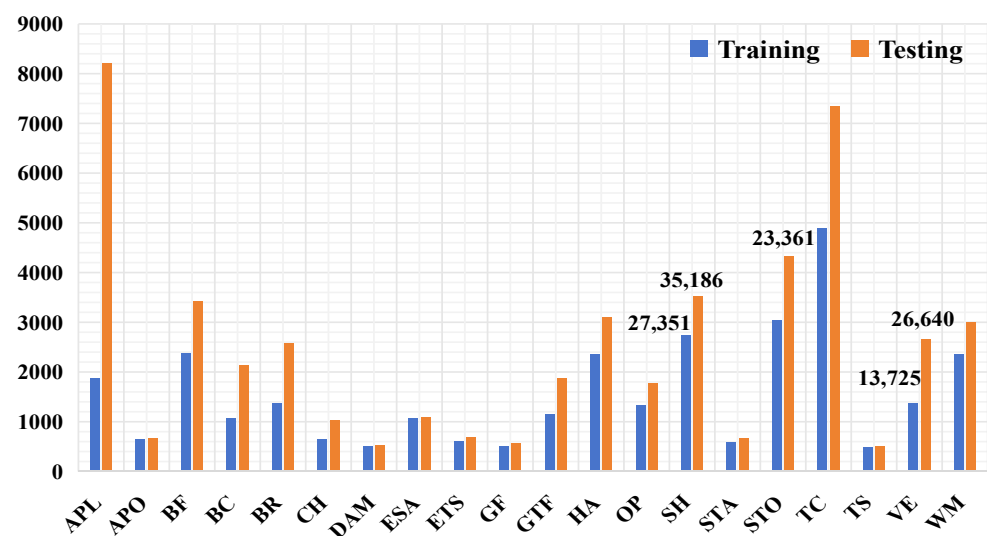


Figure 6. The numbers of different categories in training and testing sets.

4.4. Ablation Studies

We denote our proposed method with the image-level feature refinement branch and instance-level feature refinement branch as IM and IN, respectively. Moreover, we further identify our proposed image-level and instance-level feature refinement branches with the corresponding multi-label attention mining loss (MAML) and instance selection loss (ISL) as *w.* MAML, *w.o.* MAML, *w.* ISL and *w.o.* ISL, respectively. Importantly, we introduce rotation-invariant learning (RL) in MELM [49], which is defined as method A in Tables 5 and 6.

Table 5. Mean average precision (%) and correct localization (%) for different methods on NWPU VHR-10.v2.

Methods	IM		IN		mAP	CorLoc
	<i>w.o.</i> MAML	<i>w.</i> MAML	<i>w.o.</i> ISL	<i>w.</i> ISL		
A					56.3	70.4
B	✓				61.4	71.5
C		✓			62.3	72.6
D			✓		61.3	72.2
E				✓	62.0	73.1
F		✓		✓	65.2	75.2

Table 6. Mean average precision (%) and correct localization (%) for different methods on DIOR.

Methods	IM		IN		mAP	CorLoc
	<i>w.o.</i> MAML	<i>w.</i> MAML	<i>w.o.</i> ISL	<i>w.</i> ISL		
A					24.6	46.3
B	✓				25.5	49.5
C		✓			26.7	50.3
D			✓		26.4	50.6
E				✓	27.3	49.1
F		✓		✓	29.1	52.3

For the NWPU VHR-10.v2 dataset, the mAP and CorLoc of six methods with different configurations are compared in Table 5. Method A is the baseline network which achieves 56.3% mAP and 70.4% CorLoc, respectively. After integrating the proposed image-level feature refinement branch (without MAML) into the baseline method, method B achieves 61.4% mAP and 71.5% CorLoc, which is 5.1% and 1.1% higher than the baseline. However, with only image-level supervision, the attention mechanism tends to confuse instances with similar appearances. Method C is the combination of IM and MAML which achieves 0.9% mAP and 1.1% CorLoc gains compared to Method B. The aforementioned approaches focus on re-balancing the computational resources towards all the informative part of images so that most positive proposals can be retained. Method D is the combination of the instance-level feature refinement branch (without ISL) and the baseline method, which aims to explore the instance-associative spatial correlation among the instance regions to alleviate the problem of part domination. Specifically, method D achieves 61.3% mAP and 72.2% CorLoc, which is 5.0% and 1.8% higher than the baseline approach. After integrating ISL into the IN, method E achieves 0.7% mAP and 0.9% CorLoc gains compared to method D. Importantly, the best performance is still attained by refining deep features from the image level (with MAML) and instance level (with ISL), which achieves 65.2% mAP and 75.2% CorLoc, respectively.

The detailed comparisons on the DIOR dataset also validate the effectiveness of our method. Compared with the baseline network, IM and IN integrated into the baseline method achieve 0.9% mAP, 3.2% CorLoc and 1.8% mAP, 4.3% CorLoc gains, respectively. After integrating MAML and ISL into the aforementioned feature refinement branches, method C and method E attain 26.7% mAP and 50.3% CorLoc, and 27.3% mAP and 49.1% CorLoc, which are 2.1% mAP and 4.0% CorLoc, and 2.7% mAP and 2.8% CorLoc higher than the baseline method. Importantly, method E obtains the best mAP (29.1%) and CorLoc (52.3%) on the DIOR dataset, which demonstrates that the combination of image-level and instance-level feature refinement is effective. We further visualize many detection results in Figures 7 and 8. Compared with the baseline, the proposed image-level feature refinement branch successfully alleviates the problem of missing detection. However, allocating the computational resources towards the most informative part in

images inevitably exacerbates the problem of part domination. For example, airfoil and fuselage regions might have the greatest contributions to detect airplanes. However, using the airfoil or fuselage parts to represent the entire airplane is obviously problematic. In order to alleviate the aforementioned limitation, our instance-level feature refinement branch explores the instance-associative spatial diversification constraints, which has great potential for excavating the complete instances.



Figure 7. Visualization results of different methods. Compared with the baseline (first row), refining deep features at image level (second row) alleviates the problem of missing detection and reduces the ambiguities in object recognition.



Figure 8. Visualization results of different methods. Compared with the baseline (first row), refining deep features at instance level (second row) alleviates the problem of part domination and has great potential for perceiving the complete instances.

4.5. Extension: Performance on Natural Images

In order to verify the validity of our proposed method, we further evaluate the proposed model on PASCAL VOC2007 <http://host.robots.ox.ac.uk/pascal/VOC/voc2007/index.html> (accessed on 20 March 2024) [12], PASCAL VOC2012 <http://host.robots.ox.ac.uk/pascal/VOC/voc2012/index.html> (accessed on 20 March 2024) [12] and COCO <https://cocodataset.org/#download> (accessed on 20 March 2024) [13], which are the most popular datasets for natural image-based WSOD methods. Following the standard experimental settings [15,16], we only use image-level annotations for training.

We list the detailed comparisons on the PASCAL VOC2007 and 2012 datasets in Table 7. All the detection results are obtained from the original paper. Our proposed method achieves 57.5% mAP and 71.0% CorLoc, outperforming all the compared WSOD approaches. REG [16] uses an attention mechanism to boost the representation of image-level features but ignores modeling the correlation among different instances. Thus, our method outperforms REG by 8.9% mAP and 4.2% CorLoc, and 7.1% mAP and 1.8% CorLoc on the VOC2007 and 2012 datasets, respectively. Importantly, our proposed method outperforms the best two competitors, PSL [38] and Ren et al. [56], by a conspicuous margin. Compared with many other WSOD methods, our detection results on the VOC2007 and 2012 datasets demonstrate the effectiveness of refining deep features at both image level and instance level. Figure 9 also illustrates many detection results of many approaches on the VOC dataset. It can be clearly seen that our method has great potential for capturing more complete objects and reducing instance ambiguity induced from extreme scales, as well as similar appearances.



Figure 9. Visualizations of different methods on the VOC2007 dataset. The first row shows the detection results of baseline method and the second row shows the detection results of our proposed method. Taking the third row, visualizations of the ground true, as an indicator, we can clearly see that our method accurately detects the most instances.

Table 7. Mean average precision (%) and correct localization (%) for different WSOD methods on the VOC2007 and VOC2012 datasets.

Method	07-AP ₅₀	07-Cor	12-AP ₅₀	12-Cor
REG [16]	48.6	66.8	46.8	69.5
MPG-MIL [57]	50.4	67.1	46.9	67.4
PG + PS [17]	51.1	69.2	48.3	68.7
D-MIL [58]	53.5	68.7	49.6	70.1
Xu et al. [59]	53.9	69.8	52.8	73.3
CFB [60]	54.3	70.7	49.4	69.6
BGM + IBM [61]	54.3	72.1	50.5	71.9
DSW [62]	54.7	73.3	53.8	69.6
Ren et al. [56]	54.9	68.8	52.1	70.9
PSL [38]	56.3	70.3	—	—
Ours	57.5	71.0	53.9	71.3

Additionally, Table 8 shows the comparative experiments on the COCO dataset. The detection results are partially obtained from [17] and partially obtained by our reconstruction, REG* [16]. To facilitate a fair comparison, we adopt the same settings as [17]. Compared with other single-model-based methods, our method achieves 26.9% AP₅₀, which is 12.6% (26.9% v.s. 14.3%), 10.7% (26.9% v.s. 16.2%), 8.1% (26.9% v.s. 18.8%), 7.5% (26.9% v.s. 19.4%), 7.3% (26.9% v.s. 19.6%), 6.6% (26.9% v.s. 20.3%), 6.2% (26.9% v.s. 20.7%) and 1.1% (26.9% v.s. 25.8%) higher than OICR [15], ML-LocNet [63], MELM [49], PCL [20], REG [16], WS-JDS [64], PG + PS [17] and Ren et al. [56], respectively. For other evaluation metrics on the COCO dataset, the proposed method also achieves the best AP₇₅ (12.4%) and AP (12.9%), which is 1.9% and 0.5% higher than the best competitor, Ren et al. [56]. We also compare our method with other ensemble-learning-based approaches in Table 8. Following the methods in [14,15], we simply average the category scores predicted by our IM-integrated, IN-integrated and IM+IN-integrated models. Our method achieves 27.8% AP₅₀, 13.4% AP₇₅ and 14.7% AP, which attains 0.9% AP₅₀, 1.0% AP₇₅ and 1.8% AP gains compared to our single-model-based method.

Table 8. Mean average precision (%) and correct localization (%) for different single-model-based and ensemble-based methods on the COCO dataset.

Methods	Single			Ensemble		
	AP ₅₀	AP ₇₅	AP	AP ₅₀	AP ₇₅	AP
OICR [15]	14.3	6.1	7.7	16.1	8.1	8.4
ML-LocNet [63]	16.2	—	—	—	—	—
MELM [49]	18.8	7.6	7.8	20.7	9.1	9.7
PCL [20]	19.4	7.3	8.5	19.6	—	9.2
REG* [16]	19.6	8.1	9.3	21.2	9.7	10.8
WS-JDS [64]	20.3	—	10.5	—	—	—
PG + PS [17]	20.7	—	—	—	—	—
Ren et al. [56]	25.8	10.5	12.4	—	—	—
Ours	26.9	12.4	12.9	27.8	13.4	14.7

5. Conclusions

In this paper, we propose a progressive image-level and instance-level feature refinement network to boost the representation of deep features for weakly supervised object detection (WSOD) in RSIs. First, the image-level feature refinement branch is constructed to alleviate the problem of missing detection. With the supervision of the proposed multi-label attention mining loss (MAML), our method provides an explicit indicator to decrease the ambiguities in object recognition. Moreover, we further construct an instance-level feature refinement branch to model the region-associative spatial diversification for re-scoring the confidence of different adjacent bounding boxes. An instance selection loss (ISL) is proposed to excavate the salient RoI features for emphasizing the representation of instance-level features. The aforementioned feature refinement branches are unified into the complete WSOD network. The region classification and box regression tasks are optimized by our constructed MAML and ISL loss functions to boost the performance of WSOD in RSIs. The proposed method outperforms previous state-of-the-art approaches

on the NWPU VHR-10.v2 and DIOR datasets, which demonstrates the effectiveness of boosting deep features at both image level and instance level for WSOD.

Author Contributions: Conceptualization, S.Z. and Z.W. (Zebin Wu); methodology, Z.W. (Zebin Wu); software, Y.X.; validation, S.Z. and Z.W. (Zebin Wu); formal analysis, Z.W. (Zhihui Wei); investigation, S.Z. and Y.X.; resources, Z.W. (Zebin Wu); data curation, S.Z., Y.X. and Z.W. (Zebin Wu); writing—original draft preparation, S.Z.; writing—review and editing, S.Z. and Z.W. (Zhihui Wei); visualization, S.Z. and Y.X.; supervision, Z.W. (Zebin Wu) and Z.W. (Zhihui Wei); project administration, Z.W. (Zebin Wu); funding acquisition, Z.W. (Zebin Wu). All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded in part by the National Natural Science Foundation of China under Grant U23B2006; in part by the Jiangsu Provincial Innovation Support Program under Grant BZ2023046; in part by the Jiangsu Provincial Key Research and Development Program under Grant BE2022065-2; in part by the National Natural Science Foundation of China under Grant 62071233; and in part by the Jiangsu Provincial Natural Science Foundation of China under Grant BK20211570.

Data Availability Statement: The datasets in this study are available online from <https://drive.google.com/file/d/15xd4TASVAC2irRf02GA4LqYFbH7QITR-/view> (accessed on 20 March 2024) and https://drive.google.com/drive/folders/1UdIlgHk49iu6WpcJ5467iT-UqNPpx__CC (accessed on 20 March 2024).

Acknowledgments: The authors would like to thank the support from NVIDIA Corporation for providing the GeForce GTX 1080ti used in this research.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Zhang, T.; Zhuang, Y.; Chen, H.; Chen, L.; Wang, G.; Gao, P.; Dong, H. Object-Centric Masked Image Modeling-Based Self-Supervised Pretraining for Remote Sensing Object Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 5013–5025. [CrossRef]
2. Gao, L.; Li, J.; Zheng, K.; Jia, X. Enhanced Autoencoders With Attention-Embedded Degradation Learning for Unsupervised Hyperspectral Image Super-Resolution. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5509417. [CrossRef]
3. Gao, L.; Sun, X.; Sun, X.; Zhuang, L.; Du, Q.; Zhang, B. Hyperspectral anomaly detection based on chessboard topology. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5505016. [CrossRef]
4. Su, Y.; Gao, L.; Jiang, M.; Plaza, A.; Sun, X.; Zhang, B. NSCKL: Normalized Spectral Clustering With Kernel-Based Learning for Semisupervised Hyperspectral Image Classification. *IEEE Trans. Cybern.* **2023**, *53*, 6649–6662. [CrossRef] [PubMed]
5. Zhuang, L.; Ng, M.K.; Gao, L.; Michalski, J.; Wang, Z. Eigenimage2Eigenimage (E2E): A Self-Supervised Deep Learning Network for Hyperspectral Image Denoising. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, 1–15. [CrossRef]
6. Gao, L.; Wang, D.; Zhuang, L.; Sun, X.; Huang, M.; Plaza, A. BS3LNet: A new blind-spot self-supervised learning network for hyperspectral anomaly detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5504218. [CrossRef]
7. Small, C.; Sousa, D. Robust Cloud Suppression and Anomaly Detection in Time-Lapse Thermography. *Remote Sens.* **2024**, *16*, 255. [CrossRef]
8. Gui, S.; Song, S.; Qin, R.; Tang, Y. Remote Sensing Object Detection in the Deep Learning Era—A Review. *Remote Sens.* **2024**, *16*, 327. [CrossRef]
9. Feng, Y.; Han, B.; Wang, X.; Shen, J.; Guan, X.; Ding, H. Self-Supervised Transformers for Unsupervised SAR Complex Interference Detection Using Canny Edge Detector. *Remote Sens.* **2024**, *16*, 306. [CrossRef]
10. Zheng, S.; Wu, Z.; Xu, Y.; Wei, Z.; Plaza, A. Learning Orientation Information From Frequency-Domain for Oriented Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5628512. [CrossRef]
11. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. 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]
12. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
13. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September, 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755. [CrossRef]
14. Bilen, H.; Vedaldi, A. Weakly supervised deep detection networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2846–2854. [CrossRef]

15. Tang, P.; Wang, X.; Bai, X.; Liu, W. Multiple instance detection network with online instance classifier refinement. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July, 2017; pp. 2843–2851. [\[CrossRef\]](#)
16. Yang, K.; Li, D.; Dou, Y. Towards precise end-to-end weakly supervised object detection network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 7 October–2 November 2019; pp. 8372–8381. [\[CrossRef\]](#)
17. Cheng, G.; Yang, J.; Gao, D.; Guo, L.; Han, J. High-quality proposals for weakly supervised object detection. *IEEE Trans. Image Process.* **2020**, *29*, 5794–5804. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Uijlings, J.R.; Van De Sande, K.E.; Gevers, T.; Smeulders, A.W. Selective search for object recognition. *Int. J. Comput. Vis.* **2013**, *104*, 154–171. [\[CrossRef\]](#)
19. Zitnick, C.L.; Dollár, P. Edge boxes: Locating object proposals from edges. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September, 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 391–405. [\[CrossRef\]](#)
20. 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. [\[CrossRef\]](#) [\[PubMed\]](#)
21. 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. [\[CrossRef\]](#)
22. 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. [\[CrossRef\]](#)
23. 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. [\[CrossRef\]](#)
24. 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. [\[CrossRef\]](#)
25. Fasana, C.; Pasini, S.; Milani, F.; Fraternali, P. Weakly supervised object detection for remote sensing images: A survey. *Remote Sens.* **2022**, *14*, 5362. [\[CrossRef\]](#)
26. Choi, J.; Lee, S.J. Weakly Supervised Learning for Transmission Line Detection Using Unpaired Image-to-Image Translation. *Remote Sens.* **2022**, *14*, 3421. [\[CrossRef\]](#)
27. 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. [\[CrossRef\]](#)
28. 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. [\[CrossRef\]](#)
29. Foulds, J.; Frank, E. A review of multi-instance learning assumptions. *Knowl. Eng. Rev.* **2010**, *25*, 1–25. [\[CrossRef\]](#)
30. Huang, Z.; Zou, Y.; Kumar, B.; Huang, D. Comprehensive attention self-distillation for weakly-supervised object detection. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 16797–16807.
31. Lin, C.; Wang, S.; Xu, D.; Lu, Y.; Zhang, W. Object instance mining for weakly supervised object detection. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 11482–11489. [\[CrossRef\]](#)
32. 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. [\[CrossRef\]](#)
33. 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. [\[CrossRef\]](#)
34. Gungor, C.; Kovashka, A. Complementary Cues from Audio Help Combat Noise in Weakly-Supervised Object Detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023; pp. 2185–2194. [\[CrossRef\]](#)
35. Zhu, M.; Wan, S.; Jin, P.; Zhang, P. DFFNet: Dynamic Feature Fusion Network for Weakly Supervised Object Detection in Remote Sensing Images. In Proceedings of the 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan, 17–20 December 2022; pp. 1409–1414. [\[CrossRef\]](#)
36. Qian, W.; Yan, Z.; Zhu, Z.; Yin, W. Weakly Supervised Part-Based Method for Combined Object Detection in Remote Sensing Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 5024–5036. [\[CrossRef\]](#)
37. Tan, Z.; Jiang, Z.; Guo, C.; Zhang, H. WSODet: A Weakly Supervised Oriented Detector for Aerial Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5604012. [\[CrossRef\]](#)
38. Qian, X.; Huo, Y.; Cheng, G.; Gao, C.; Yao, X.; Wang, W. Mining High-Quality Pseudoinstance Soft Labels for Weakly Supervised Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5607615. [\[CrossRef\]](#)
39. Feng, X.; Yao, X.; Cheng, G.; Han, J. Weakly supervised rotation-invariant aerial object detection network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 14146–14155. [\[CrossRef\]](#)
40. Ma, L.; Luo, X.; Hong, H.; Zhang, Y.; Wang, L.; Wu, J. Scribble-attention hierarchical network for weakly supervised salient object detection in optical remote sensing images. *Appl. Intell.* **2023**, *53*, 12999–13017. [\[CrossRef\]](#)
41. 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. [\[CrossRef\]](#)

42. Xie, X.; Cheng, G.; Feng, X.; Yao, X.; Qian, X.; Han, J. Attention Erasing and Instance Sampling for Weakly Supervised Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *62*, 5600910. [\[CrossRef\]](#)
43. 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. [\[CrossRef\]](#)
44. Huo, Y.; Qian, X.; Li, C.; Wang, W. Multiple Instance Complementary Detection and Difficulty Evaluation for Weakly Supervised Object Detection in Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 6006505. [\[CrossRef\]](#)
45. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, Montreal, QC, Canada, 7–12 December 2015.
46. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
47. 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. [\[CrossRef\]](#)
48. Li, K.; Cheng, G.; Bu, S.; You, X. Rotation-Insensitive and Context-Augmented Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2337–2348. [\[CrossRef\]](#)
49. Wan, F.; Wei, P.; Jiao, J.; Han, Z.; Ye, Q. Min-entropy latent model for weakly supervised object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1297–1306. [\[CrossRef\]](#)
50. 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.* **2022**, *60*, 5610411. [\[CrossRef\]](#)
51. 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. [\[CrossRef\]](#)
52. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [\[CrossRef\]](#)
53. Cheng, G.; Zhou, P.; Han, J. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 7405–7415. [\[CrossRef\]](#)
54. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587. [\[CrossRef\]](#)
55. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448. [\[CrossRef\]](#)
56. 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. [\[CrossRef\]](#)
57. Ji, R.; Liu, Z.; Zhang, L.; Liu, J.; Zuo, X.; Wu, Y.; Zhao, C.; Wang, H.; Yang, L. Multi-peak Graph-based Multi-instance Learning for Weakly Supervised Object Detection. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2021**, *17*, 70. [\[CrossRef\]](#)
58. Gao, W.; Wan, F.; Yue, J.; Xu, S.; Ye, Q. Discrepant multiple instance learning for weakly supervised object detection. *Pattern Recognit.* **2022**, *122*, 108233. [\[CrossRef\]](#)
59. Xu, Y.; Zhou, C.; Yu, X.; Xiao, B.; Yang, Y. Pyramidal Multiple Instance Detection Network With Mask Guided Self-Correction for Weakly Supervised Object Detection. *IEEE Trans. Image Process.* **2021**, *30*, 3029–3040. [\[CrossRef\]](#)
60. Yin, Y.; Deng, J.; Zhou, W.; Li, H. Instance mining with class feature banks for weakly supervised object detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 3190–3198. [\[CrossRef\]](#)
61. Jia, Q.; Wei, S.; Ruan, T.; Zhao, Y.; Zhao, Y. Gradingnet: Towards providing reliable supervisions for weakly supervised object detection by grading the box candidates. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 1682–1690. [\[CrossRef\]](#)
62. Li, X.; Yi, S.; Zhang, R.; Fu, X.; Jiang, H.; Wang, C.; Liu, Z.; Gao, J.; Yu, J.; Yu, M.; et al. Dynamic sample weighting for weakly supervised object detection. *Image Vis. Comput.* **2022**, *122*, 104444. [\[CrossRef\]](#)
63. Zhang, X.; Yang, Y.; Feng, J. MI-locnet: Improving object localization with multi-view learning network. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 240–255. [\[CrossRef\]](#)
64. Shen, Y.; Ji, R.; Wang, Y.; Wu, Y.; Cao, L. Cyclic guidance for weakly supervised joint detection and segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 697–707. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.