remote sensing

MDPI

Article

Dynamic Pseudo-Label Generation for Weakly Supervised Object Detection in Remote Sensing Images

Hui Wang [1,2,3,4], Hao Li [1,2], Wanli Qian [5], Wenhui Diao [1,2], Liangjin Zhao [1,2], Jinghua Zhang [1,2] and Daobing Zhang [1,2,*]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China; wanghui18@mails.ucas.edu.cn (H.W.); lihao@aircas.ac.cn (H.L.); diaowh@aircas.ac.cn (W.D.); zhaolj004896@aircas.ac.cn (L.Z.); zhangjh004940@aircas.ac.cn (J.Z.)
[2] Key Laboratory of Network Information System Technology, Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China
[3] University of Chinese Academy of Sciences, Beijing 100190, China
[4] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China
[5] Georgia Institute of Technology College of Computing, Atlanta, GA 30318, USA; michaelq@gatech.edu
* Correspondence: dbzhang@mail.ie.ac.cn; Tel.: +86-10-5888-7618

Abstract: In recent years, fully supervised object detection methods in remote sensing images with good performance have been developed. However, this approach requires a large number of instance-level annotated samples that are relatively expensive to acquire. Therefore, weakly supervised learning using only image-level annotations has attracted much attention. Most of the weakly supervised object detection methods are based on multi-instance learning methods, and their performance depends on the process of scoring the candidate region proposals during training. In this process, the use of only image-level labels for supervision usually cannot obtain optimal results due to the lack of location information of the object. To address the above problem, a dynamic sample pseudo-label generation framework is proposed to generate pseudo-labels for each proposal without additional annotations. First, we propose the pseudo-label generation algorithm (PLG) to generate the category labels of the proposal by using the localization information of the object. Specifically, we propose to use the pixel average of the object's localization map in the proposal as the proposal category confidence and calculate the pseudo-label by comparing the proposal category confidence with the preset threshold. In addition, an effective adaptive threshold selection strategy is designed to eliminate the effect of different category shape differences in computing sample pseudo-labels. Comparative experiments on the NWPU VHR-10 dataset demonstrate that our method can significantly improve the detection performance compared to existing methods.

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

Citation: 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. https://doi.org/10.3390/rs13081461

Academic Editor: Pedro Melo-Pinto

Received: 18 March 2021
Accepted: 6 April 2021
Published: 10 April 2021

1. Introduction

Object detection is an important task in the remote sensing image interpretation. With the application of deep learning in computer vision [1–4], an increasing number of object detection methods based on convolutional neural networks (CNNs) [5–11] have been proposed to achieve good performance. However, fully supervised object detection methods require a large number of samples with instance-level labels. For remote sensing images with many targets, obtaining instance-level annotation is laborious and time-consuming. Therefore, weakly supervised object detection methods that require only image-level labels have attracted increasing attention.

Most weakly supervised object detection methods [12–24] are based on multi-instance learning (MIL) [25]. For MIL, a set of packages is given, and each package is a collection of instances. MIL has the following constraints: (1) If a package is positive, at least one

instance of the package is positive. (2) If a package is negative, all instances in the package are negative. In WSOD, MIL treats each proposal as an instance and each image as a package of latent target instances. The proposal scores are summarized into image-level classification scores that can be supervised by category labels in all methods. Finally, the most contributing proposals will be selected as pseudo-instance-level labels used to optimize the object detector. With the development of CNN, Bilen and Vedaldi [16] combined convolutional neural networks with multi-instance learning to design a weakly supervised deep detection network (WSDDN). Then, a series of works [17–24,26–31] based on WSDDN were proposed to enhance the performance of weakly supervised object detection (WSOD). Tang et al. [17] proposed online instance classifier refinement (OICR) to improve performance by propagating inferred labels of instances. Then, Tang et al. [19] proposed proposal cluster learning (PCL) to optimize the process of online instance classifier refinement. These methods perform proposal score prediction under image-level supervision and use the predicted proposal scores to evaluate the final detection results.

Despite the promising results of the above studies, weakly supervised object detection is still widely considered to be an open problem. First, the main weakly supervised detection methods obtain the prediction scores of proposals by training an image-level classifier. In weakly supervised object detection, only the category labels of the image can be used to constrain the classifier, instead of having the spatial information of the proposal like the fully supervised detector. There is a large number of dense objects in remote sensing images, and without the constraint of ground truth information, as shown in Figure 1a, the detector may mistake neighboring instances as one object. Therefore, there is a potential ambiguity in the weakly supervised object detection method, which leads to the inferior performance of WSOD compared to fully supervised object detection.

Second, for lack of instance-level annotation, WSOD is affected by background noise in the learning process. Particularly in complex remote sensing images, many objects appear with individual-specific backgrounds. Such confusing backgrounds adversely affect the learning of the detector. For example, for aircraft that usually stay on the tarmac, as shown in Figure 1b, the detector tends to mistake the tarmac for an aircraft. The OICR approach mentioned above would use the proposal scores in the WSDDN as supervision for the refined classifier, which would exacerbate this problem to a certain extent.

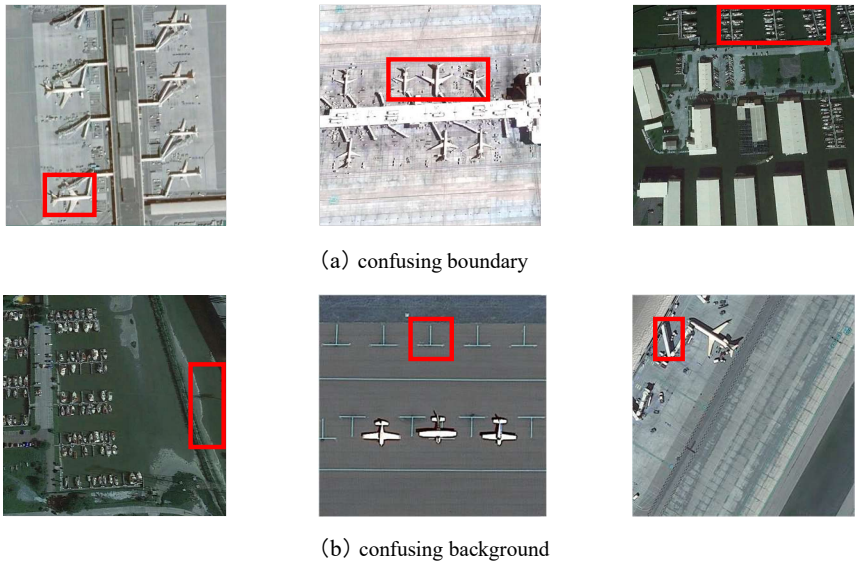

Figure 1. WSOD issues in remote sensing images : (**a**) confusing boundary: less tight candidates or fail to differentiate clustered instances (top); (**b**) confusing backgrounds: misjudge the background as object (bottom).

To address these challenges, we propose a novel pseudo-label generation (PLG) algorithm that combines the localization information of samples and image-level labels to generate instance-level pseudo-labels for each proposal to provide supervised information for training object detection networks. Specifically, a weakly supervised localization model is trained to obtain the localization map of the object, and negative samples are added to reduce the effect of the cluttered background. Next, during training, we map the region proposals to the localization maps generated by the pretrained localization model. Then, the proposal confidence is calculated based on the intersection between the proposal and the localization map, and the confidence score of the proposal is compared with a preset threshold to calculate the pseudo-label of the proposal. Finally, the instance-level pseudo-label is used as the supervision information for proposal classification in the weakly supervised object detection network.

In PLG, the proposal category confidence scores are measured based on the coverage of the proposal with the localization map. If the same threshold is used in the proposal pseudo-label calculation, the different geometric properties for categories will have different effects on the pseudo-label calculation. Therefore, it is important to choose the appropriate threshold value. We propose an effective adaptive threshold selection strategy to eliminate this effect. Specifically, we select the proposals with the highest prediction scores in all categories for each sample and then calculate the category confidence histogram distributions for all high-quality proposals, calculate the quantile of frequency histogram, and select the quantile as the new threshold. Finally, using the new threshold, the pseudo-label of the input image is calculated for the next iteration.

In summary, our main contributions and innovations are as follows.

(1) Based on image-level labels, a novel instance-level pseudo-label generation algorithm is designed in this paper for training the detection network. We propose to map region proposals into the localization map that is generated by a pretrained localization model. Then, confidence scores are calculated by computing the pixel average of the regional proposals in the localization map, and pseudo-labels are assigned by comparing confidence scores with the preset threshold.

(2) We design an adaptive threshold selection strategy that is used to continuously update pseudo-labels during the iteration process. First, we calculate the frequency histogram distribution of confidence scores for each category. Then, we propose to calculate the quantile on the frequency histogram and use the quantile as the new preset threshold to update the pseudo-labels for input image in the next iteration.

Experiments on the NWPU VHR-10 publicly available dataset shows that our weakly supervised method displays advanced performance. The remainder of the paper is structured as follows. Section 2 presents the framework of our method and describes its components in detail. Section 3 describes the experiments and results to analyze the impact of our method. Section 4 discusses the results of our method. Section 5 summarizes the paper.

2. Methods

We illustrate the overall framework of dynamic pseudo-label generation in Figure 2. The basic weakly supervised object detector extracts features on the backbone network, performs detection and classification branches by weighted MIL pooling, multiplies the outputs of the two branches, and accumulates them to obtain the image-level prediction scores. A proposal cluster learning (PCL [19]) strategy is also used to add the refined instance classifier to improve the performance. PLG generates a localization map for each image and assigns a pseudo-label and confidence score to each proposal based on the coverage of the proposal with the localization map. The pseudo-label is used to guide the training of classification branches. An adaptive threshold selection strategy selects the proposals with highest scores in the refined instance classifier as high quality proposal, and sets the quantile on the frequency histogram of high quality proposal confidence

scores as a new threshold value to update the pseudo-labels of the input images in the next iteration. Each part is described in detail below.

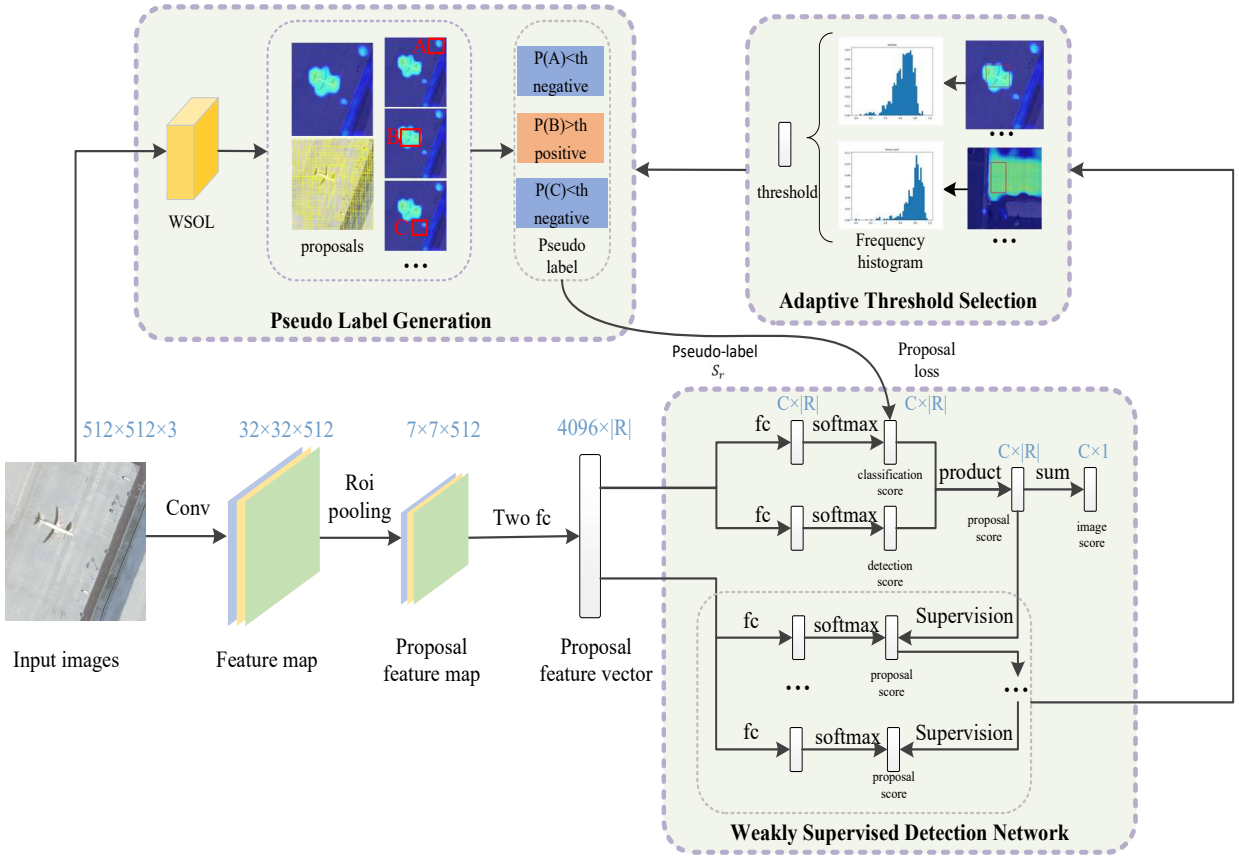

Figure 2. Pipeline of our weakly supervised learning strategy. The pseudo-label generation strategy generates category labels for each proposal during the training phase. The weakly supervised detector combines the proposal pseudo-labels and the image category labels for training. The adaptive threshold selection strategy is used to update the threshold for computing pseudo-labels. In "Weakly Supervised Detection Network", the ellipses represent the intermediate refinement classifier process.

2.1. Architecture of Weakly Supervised Object Detection Network

Figure 2 describes the overview architecture of the basic weakly supervised detection network. For each input image, the selective search method [32] is performed to produce approximately 2k proposals. Then, an ROI pooling [33] is used to obtain fixed-size convolutional feature maps. After two FC layers, the proposal features are divided into two branches: the classification branch and the detection branch. The proposal features are passed to the fully connected layer and softmax operation is performed to generate the classification score and detection score of each proposal. This is described by

$$[\sigma_{\text{cls}}(x^c)]_{i,j} = \frac{e^{x^c_{ij}}}{\sum_{k=1}^{C} e^{x^c_{kj}}} \tag{1}$$

$$\left[\sigma_{\text{det}}\left(x^d\right)\right]_{i,j} = \frac{e^{x^d_{ij}}}{\sum_{k=1}^{|R|} e^{x^d_{ik}}} \tag{2}$$

where $x^c \in \mathbb{R}^{C\times|R|}$ represents the proposal feature vector after the fully connected layer of the classification branch. $\sigma_{\text{cls}}(x^c)$ represents the output of the classification data stream, generated by performing softmax calculations on the classes. $x^d \in \mathbb{R}^{C\times|R|}$ represents the

proposal feature vector after the fully connected layer of the detection branch. $\sigma_{\text{det}}\left(x^d\right)$ represents the output of the localization data stream, generated by performing softmax operations on proposals. C represents the number of classes and $|R|$ represents the number of proposals.

The score of each proposal is obtained by multiplying the above two scores, $x_{cr} = \sigma_{\text{cls}}(x^c) \odot \sigma_{\text{det}}\left(x^d\right)$. Finally, the scores of each proposals are added up to obtain the image-level prediction scores: $\Phi_c = \sum_{r=1}^{|R|} x_{cr}^R$. By using the image level labels, the model can be trained with the composite loss function.

$$L_{MID} = L_{cls} + L_{proposal} \tag{3}$$

We use the pseudo-label generated for each proposal as supervision to learn the classification branches. Thus, the class branches can be trained by the cross-entropy loss function:

$$L_{\text{proposal}} = -\frac{1}{|R|} \sum_{r=1}^{|R|} \sum_{c=1}^{C+1} S_{rc} \log \sigma_{\text{cls}}(x^c) \tag{4}$$

where S_{rc} is the pseudo-label of each proposal, and $\sigma_{\text{cls}}(x^c)$ is the predicted output of the classification branch. S_{rc} is the proposal category pseudo-label generated by pseudo-label generation. When the score within a proposal is zero, the network tends to identify it as a negative sample. The multi-instance classifier is trained by minimizing cross-entropy loss functions using stochastic gradient descent, where y_c denotes the category label of the image:

$$L_{cls} = \sum_{c=1}^{C} \{y_c \log \Phi_c + (1 - y_c) \log(1 - \Phi_c)\} \tag{5}$$

Inspired by the work in [19], we adopt the PCL strategy to improve the performance. PCL adds a refined instance classifier to WSDDN. PCL has multiple output streams, treating WSDDN as the first data stream and the other streams as refined instance classifiers supervised by the previous stream. For each refined instance classifier, the proposal with the higher score is first used as the cluster center, and proposal clusters are generated based on the overlap with the cluster center. Next, the predicted scores of the previous streams at the cluster centers are used to compute the labels of the proposal clusters. Finally, each proposal cluster is considered as a package for training a refined instance classifier using a weighted softmax loss function. After the classifier is refined K times, the classifier tends to assign high scores to tight proposals.

During the training of WSOD, region proposals are obtained for every training image. Each proposal of the image is assigned a pseudo-label from the PLG algorithm to train classification branch. Then, the proposal prediction scores are obtained by multiplying the classification score with the detection score. Next, the scores of the proposals are summed to obtain the image-level prediction scores for training the basic multiple instance classifier. Additionally, the proposal prediction score is used as supervision of the first refined instance classifier and using the predicted scores of the preceding streams supervision information is calculated for the next output stream. Finally, the average output of all refine classifiers is chosen as the proposal's predicted score. Then, the adaptive threshold selection strategy is used to calculate a new threshold to update the pseudo-label of the input image for the next iteration. During testing, the average output of all refine classifiers is selected as the final predicted score.

2.2. Pseudo-Label Generation

Weakly supervised object detection can only utilize category labels at the image level. All of the proposals are assigned as positive packets when the image contains a positive instance in multi-instance learning. However, the presence of many negative instances in these proposals can affect the training of the weak supervision detector. Therefore,

we consider using the localization information generated by the weakly supervised localization model to generate pseudo-labels for each proposal. During the training stage, low-quality proposals are effectively suppressed, and the prediction scores of the proposals are generated more accurately.

Inspired by the work in [34], a global average pool (GAP) [35] layer is inserted in the classification module, followed by a classifier. The output features of the GAP are used as features of the fully connected layer. After the training is completed, the features output using the convolutional neural network are multiplied with the weights of the classifier to obtain a class activation map (e.g., Figure 3). We use the class activation map as the localization map of the samples. To obtain a cleaner localization map, negative samples are added in the training phase to give the model better discriminative power against the confusing background. The loss function used in training is a cross-entropy loss function containing C + 1 classes.

As shown in Figure 3, the localization map is first generated using a weakly supervised localization model. The proposal category confidence is designed to utilize the object localization information in the localization map. We consider $R_i, i \in (1, 2, 3, ...|R|)$ as a proposal from the given image, $H_c, c \in (1, 2, 3, ...C)$ represents the localization map for each category, and P_{ci} represents the confidence level of the proposal on a category, where C represents the number of classes and $|R|$ represents the number of proposals. The category confidence of the proposal is calculated by the pixel average of regional proposals in the localization map, and the confidence of the category for the ith proposal is expressed as follows:

$$P_{ci} = avg_{R_i}(H_c) \tag{6}$$

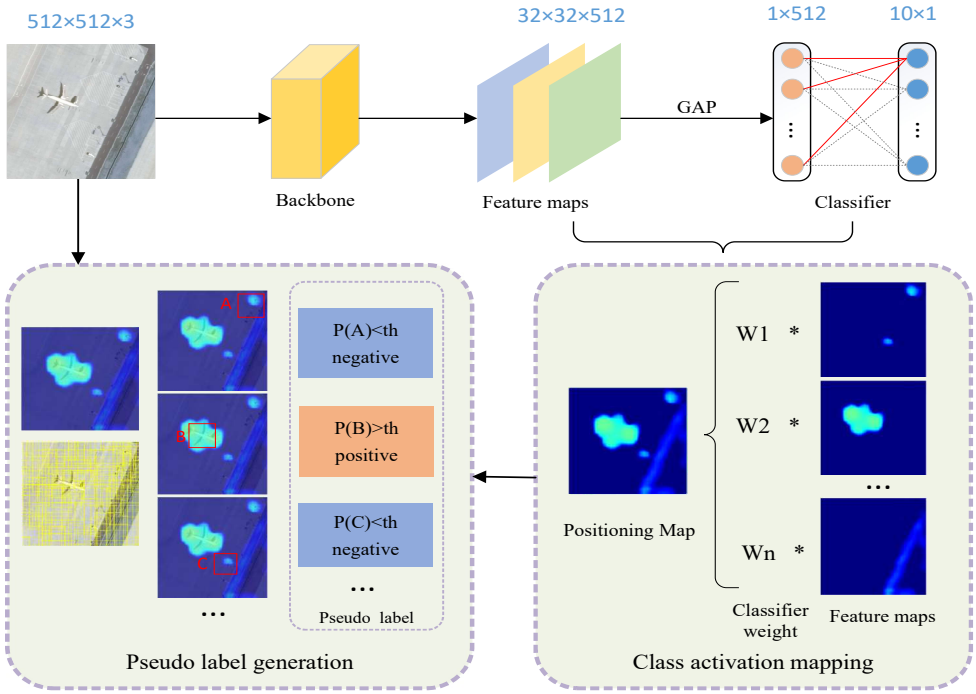

Figure 3. Illustration of the process of pseudo-label generation. For the classification model, the global average pooling layer is trained instead of the fully connected layer. Class activation mapping uses the feature map of the backbone network and the corresponding multiplication of the weights of the classifiers to obtain the localization map of the object. Pseudo-label generation calculates pseudo-labels and confidence scores based on the intersection between the proposal and the localization map.

Next, the pseudo-label of each proposal is calculated based on the proposal confidence scores. The proposal pseudo-label for the ith label is denoted using S_{ci}, where th is the threshold value.

$$S_{ci} = \begin{cases} 1, \text{ if } P_{ci} > th \\ 0, \text{ if } P_{ci} < th \end{cases} \quad (7)$$

The weakly supervised localization model used in our method does not show optimal performance. Works such as those in [36,37] have made many improvements on this basis. We insist that this stage can be further improved by using a more complex weakly supervised localization model. In this paper, our experiments show that using a simple localization model is enough to demonstrate the effectiveness of our method.

2.3. Adaptive Threshold Selection

We can use a weakly supervised localization model to generate a localization map of the target, and we can calculate the category confidence of each proposal by Equation (1). The proposal category scores are obtained by comparing them with a fixed threshold value. For the weakly supervised network, using the proposal scores as part of the supervised information suppresses the low-quality candidate frames.

In PLG, we designed the proposal category confidence scores is the pixel average of the object's localization map in the proposal. However, as shown in Figure 4, the proposal category confidence scores vary by categories. Figure 5 shows the frequency histograms of the airplane and tennis court confidence. If the pseudo-label is calculated by using the same threshold compared with the proposal confidence, it will have different effects on the detection results for different categories.

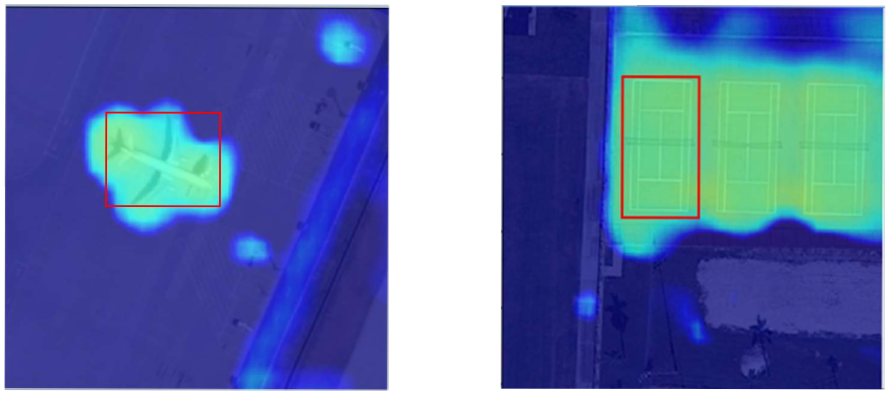

Figure 4. The location maps of plane (**left**) and tennis court (**right**).

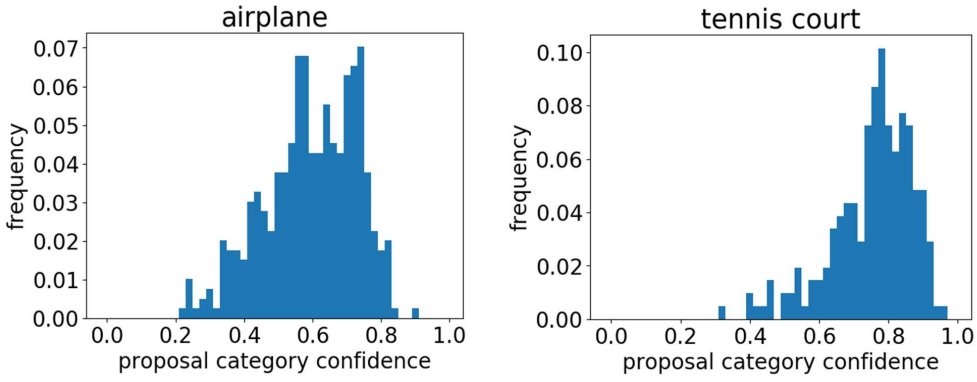

Figure 5. Frequency histogram of confidence for the airplane (**left**) and tennis court (**right**).

Therefore, we design an adaptive threshold strategy to select the threshold that is most adapted to each image category. First, the highest scoring proposals in all categories of each

sample are selected, and their category confidence scores are counted. Then, the frequency histograms of the confidence scores for each category are obtained. Finally, the quantile on the frequency histogram is calculated as the new threshold. Specifically, we use the average output of all refined classifiers as the final prediction score of the proposal and count the category confidence of the highest scoring proposals. The top-scoring proposal can be expressed as

$$r_i = arg\max_r x_{rc} \tag{8}$$

where r_i denotes the highest scoring proposal and x_{rc} denotes the final score of the proposal. Then, the proposal confidence of the high scoring proposal is taken as a sample and its frequency histogram is counted. The adaptive selection threshold is the quantile of the distribution of confidence scores for each category, and the probability of splitting is the hyperparameter $th^{'}$. The algorithm process is detailed in Algorithm 1.

In Algorithm 1, we use $x_{rk}, r \in (1, 2, 3, ...|R|), k \in (1, 2, ...K)$ as the proposal predicted score of the refine classifier, and x_r denotes the final predicted score. In the calculation of the frequency histogram of the confidence scores, using M denotes the numbers of groups, and $f_c, c \in (1, 2, ...C)$ denotes the relative frequency of sample confidence. $th^{'}$ is a hyperparameter. The new threshold is expressed using th_c.

Algorithm 1 Adaptive threshold selection.

Input:
training data set I**; proposals** r**; proposal category confidence** P**; image labels** $y = [y_1, ...y_C]^{\mathrm{T}}$**; refinement times** K**; the number of classes** C**; hyperparameters** $th^{'}$**;**
Output:
new threshold $th_c, c \in (1, 2, ...C)$

Input image I and its proposal r into the network to produce proposal predicted score x_{rk}
The final proposal score $x_r = \frac{1}{K}\sum_{k=1}^{K} x_{rk}$
for $c = 1$ to C **do**
  if $y_c = 1$ **then**
    Choose the top-scoring proposal r_i by Equation (8)
    Calculate the category confidence P_{ci} for the proposal r_i
  Divide confidence score distribution interval into M small intervals
  Count the frequency of confidence scores P_c falling into each interval
  Calculate the relative frequency f_c, acquire the confidence frequency histogram of each category.
  Set $m = 1$;
  while $P^{'} < th^{'}$ **do**
    Calculate the cumulative confidence $P^{'} = \sum_{j=1}^{m} f_{cj}$;
    $m = m + 1$;
  Calculate new threshold $th_c = \frac{m}{M}$
Update the threshold $th_c, c \in (1, 2, ...C)$.

At the beginning of training, proposal scores are calculated using a predefined threshold. In this paper, we use a PCL strategy where the predicted scores of the higher scoring proposals are propagated to the proposal boxes with their larger IOUs during the process of refine classification. Therefore, as the number of network iterations increases, the performance of the classifier continuously improves. Based on this, statistical information about the proposals with higher scores allows for the calculation of more reliable thresholds. The final experimental evaluation results confirm the effectiveness of this strategy.

3. Experiments and Result

In Section 3, the experimental setup including datasets, evaluation metrics, and hyperparameters used in training is described in detail. We conducted ablation experiments

to analyze the impact of the proposed method. Finally, comparisons with the existing advanced works are provided.

3.1. Implementation Details

3.1.1. Datasets and Evaluation Metrics

We conduct experiments on the proposed method on the publicly available NWPU VHR-10 dataset [38–40]. The dataset contains 3282 images (512 × 512 pixels) from object categories. The dataset is divided into three parts: 60% for training, 20% for validation, and 20% for test. For PLG, the negative samples are used in the training period of the weakly supervised localization model. In our experiments, two standard evaluation metrics are used to measure the performance of the proposed method. First, we evaluate our model by measuring the mAP on the test set. When the IOU between the ground truth and the bounding box is more than 0.5, the proposed method considers the bounding box as a positive test, which is the same as the PASCAL VOC standard. Second, the localization accuracy of our model is evaluated by using the correct position (CorLoc) [41]. CorLoc is the ratio of images containing at least one target, where the most confidently predicted box has an IOU greater than 0.5 with one of these targets. Furthermore, CorLoc is evaluated on the training set.

3.1.2. Train

We use VGG16 [1] pretrained on ImageNet [42] as the backbone network in which we replace the fifth max-pooling by RoI pooling. To enhance the features of small targets, we use a dilated convolutional layer instead of the fourth max-pooling layer and its subsequent convolutional layers. For initialization, the uninitialized layers are initialized by a Gaussian distribution with 0-mean and a standard deviation of 0.01. Prior to training, we use selective search [32] to produce approximately 2000 proposals for each image. For data augmentation, we horizontally mirror each image and rotate them by 180°. During training, the network performed 20k iterations. The initial learning rate is set to 0.001 for the first 15k and reduced it to 0.0001 for the last 5k iterations. The mini-batch size of the stochastic gradient descent optimizer is set to 4. Furthermore, we use the same five scales $\{480, 576, 688, 864, 1200\}$ as WSDDN. For the instance refinement classifier, we set the same refinement time K = 3 as PCl. In the pseudo-label generation algorithm, we use a threshold for calculating the proposal scores via proposal confidence with $th = 0.5$. The threshold value $th^{'}$ is set to 0.2 in the adaptive threshold selection strategy. The 0.3 IOU threshold in NMS [43] is set to calculate average precision (AP) and CorLoc.

3.2. Ablation Experiments

We performed ablation experiments in order to evaluate the effectiveness of our experimental approach and analyzed the effects of key components.

(1) Pseudo-label generation: To demonstrate the effectiveness of the pseudo-label generation algorithm, we used the training strategy of WSDDN + PCL as a baseline, using the localization information generated by the weakly supervised localization model as the supervision of the detection network. As shown in Table 1, the mAP improved from 46.7% to 50.9% and the performance of CorLoc improved from 52.0% to 58.1%, further confirming that the pseudo-label generation algorithm is effective for mining objects in a weakly supervised environment.

(2) Adaptive threshold selection strategy: to make the network adaptable to different classes, adaptive threshold ablation experiments are designed. As shown in Table 1, adaptive thresholding can improve the detection performance. The adaptive thresholding strategy mines higher quality instances as much as possible by statistically updating the threshold value with information of the detection frames with higher confidence. As shown in Table 1, the map improves from 50.9% to 53.6%, confirming the effectiveness of the proposed method. Figure 6 demonstrates the frequency confidence histogram of the proposed method for each class on the NWPU VHR-10 data set.

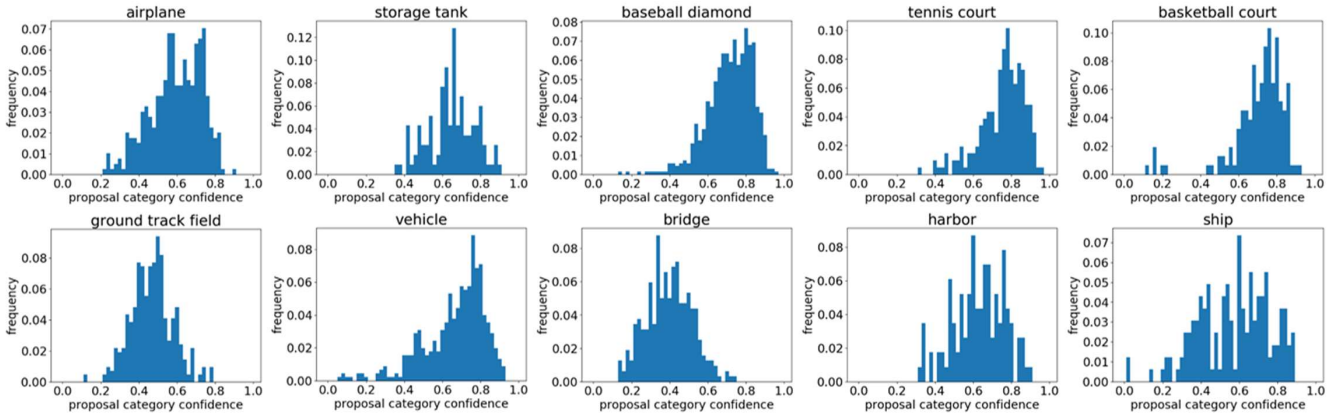

Figure 6. Frequency confidence histogram of the proposed method for each class on the NWPU VHR-10 data set.

Table 1. Result on the NWPU VHR-10 for Ablation experiments.

Method	mAP %	CorLoc %
baseline	46.7	52.0
baseline + PLG	50.9	58.1
baseline + PLG + Adaptive threshold	53.6	61.5

3.3. Comparative Experiment

We designed a series of experiments on NWPU VHR-10, and our method achieved advanced performance. Tables 2 and 3 show a summary of the experimental results of the various methods. We analyze the differences between the various methods and show the effectiveness of our approach.

WSDDN uses a two-branch structure to implement a multi-instance learning network but can easily fall into local optimization. OICR adds a refine classification module, but only the highest scoring proposal is selected as a positive sample, resulting in a large information loss. PCL improves on the OICR but relies on the detection results of WSDDN that can easily fall into the local optimization direction.

Table 2. mAP on the NWPU VHR-10 test set.

Method	Airplane	Storage Tank	Baseball Diamond	Tennis Court	Basketball	Ground Track Field	Vehicle	Bridge	Harbor	Ship	mAP
WSDDN [16]	0.008	0.016	0.297	0.175	0.414	0.482	0.005	0.007	0.01	0.034	0.145
OICR [17]	0.646	0.024	0.792	0.317	0.552	0.798	0.03	0.015	0.11	0.674	0.396
PCL [19]	0.740	0.049	0.901	0.504	0.682	0.791	0.009	0.018	0.251	0.744	0.467
ours	0.809	0.105	0.901	0.644	0.691	0.802	0.087	0.14	0.396	0.783	0.536

Table 3. CorLoc on the NWPU VHR-10 train set.

Method	Airplane	Storage Tank	Baseball Diamond	Tennis Court	Basketball	Ground Track Field	Vehicle	Bridge	Harbor	Ship	CorLoc
WSDDN [16]	0.005	0.029	0.374	0.139	0.500	0.757	0.005	0.008	0.001	0.034	0.185
OICR [17]	0.661	0.352	0.869	0.501	0.602	0.732	0.101	0.009	0.283	0.732	0.484
PCL [19]	0.862	0.035	0.937	0.516	0.738	0.886	0.049	0.012	0.283	0.877	0.520
ours	0.872	0.168	0.961	0.751	0.732	0.863	0.163	0.187	0.467	0.851	0.615

Our method achieves a 6.9% improvement in mAP values compared to the results of PCL examinations. This is mostly because of the following.

(1) Introduction of localization information as supervision in the detection network, use of a pretrained weakly supervised target localization model to generate the localization map of the target, and use of the localization information of the target as a constraint of the detection network. The low-quality proposals are suppressed, and the high-quality proposals are highlighted.

(2) Considering the complexity of the background of remote sensing images, in the process of pretraining the weakly supervised localization model, the background samples are added to participate in the training with the dataset to mitigate the influence of the background samples. Additionally, the weakly supervised labels are fully utilized in an effort to generate clean localization maps.

(3) Considering that the ratios of object area to detection frame vary by categories, the same threshold used in calculating the proposal score will have different effects on the detection structure with different categories. Therefore, the strategy of adaptive thresholds is proposed to adapt the network to different categories of objects.

Table 3 shows that our method improves from 52.0% to 61.5% on CorLoc compared to the results of PCL examinations. The main reason is that our proposed PLG algorithm uses the location information generated by the location model to calculate the instance pseudo-label. During the training of WSOD networks, we use the instance-level pseudo-label to mine as many instances as possible in the image, and the model performance is improved.

As observed in Figure 7, the accuracy is better for categories with similar distribution of fc curves, while for categories such as vehicles and bridges, the detection effect is not satisfactory. Figure 6 demonstrates the frequency confidence histogram of the proposed method for each category on the training set. It is observed from this that the confidence response intervals vary for different categories. This is shown in Figure 8 for a number of detection examples of the NWPU VHR-10 dataset. As observed from the figure, our method can provide accurate and tight bounding boxes for each object that appears in the image. However, for several categories such as bridges and oil storage tanks, our method may misdetect. This is because the coexistence of bridges and rivers causes the weakly supervision detector to misinterpret rivers as bridges. Furthermore, for multiple nonoverlapping storage tank clusters, the detector tends to identify them as a single target.

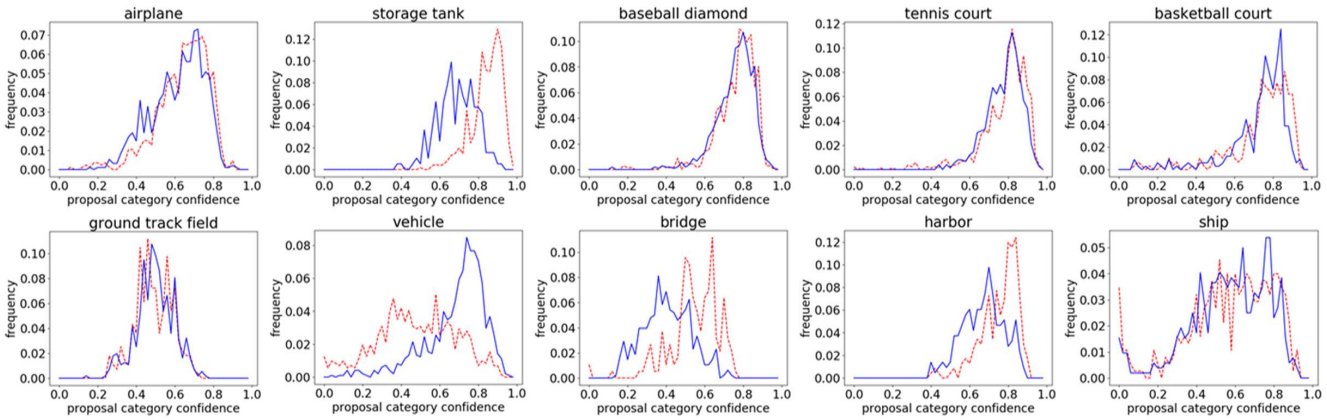

Figure 7. Frequency confidence curves of the proposed method (blue) and ground truth (red) for each class on the NWPU VHR-10 data set.

Figure 8. Example results on the NWPU VHR-10 test set for each class.

4. Discussion

The effectiveness of our proposed method and strategy is verified through careful analysis and comparison of multiple sets of experiments. Our proposed method makes full use of the existing weakly supervised information and extracts the localization information of the target as supervision in combination with the weakly supervised localization task.

From the experiments, it can be concluded that our proposed method can improve the performance of weakly supervised object detection in remote sensing images. Below, we list the main outcomes of this paper.

1. We propose a pseudo-label generation algorithm (PLG) to assign pseudo-labels to region proposals during the training phase of the model. Specifically, a weakly supervised localization model is first trained using image-level annotations. For any image, a localization map of the sample can be generated using the weakly supervised localization model. Then, during the final weakly supervised object detector, each proposal of the input image is assigned a label and a confidence score using the pregenerated localization map that is based on the intersection between the proposal and the generated localization of the PLG when the image is used as input. Then, "low-quality proposals", i.e., proposals with confidence scores below a given threshold, are considered as negative samples in the training phase.
2. We propose an adaptive threshold selection method. Considering the different object area to detection frame ratios in different categories, if the same threshold is used in calculating the proposal scores, it will have different effects on the detection structure with different categories. Specifically, the threshold T that is most adapted to each image category is selected by calculating the confidence score histogram for each category and analyzing which thresholds are associated with the proposals with the highest scores.

In addition, our proposed weakly supervised detection process uses a pretrained weakly supervised model to generate localization information. Therefore, the performance of the detector is influenced by the localization model. The localization strategy of CAM used in this paper is not the state-of-the-art performance method, and some studies have made many improvements on this basis. The detection effect should be improved if a better performing localization model is used.

Although the proposed method improves detection performance and works well on airplanes, tennis courts, and baseball fields, some problems still exist, and the detection results are less satisfactory in some classes. Bridges and vehicles are some examples of such targets. There are two main reasons for this.

1. For remote sensing images, many targets will appear along with a certain specific background. For example, in the detection of bridges, the coexistence of bridges and rivers causes the weakly supervision detector to misinterpret rivers as bridges. Although adding negative samples to the weakly supervised model for training can reduce the damage, it still cannot completely solve this problem.
2. The objects in remote sensing images are relatively dense. For categories such as oil storage tanks, multiple targets usually appear near, and a target appears less often. The lack of instance-level labeling makes it difficult for the detector to separate the adjacent object areas. In this case, the detector will tend to misjudge multiple targets as a single target, damaging the performance of the detector.

Because of the lack of instance-level annotation, the detector will not learn accurate information when the above situation occurs. During the training stage, the model will obtain suboptimal solutions. Therefore, compared with the fully supervised approach, the performance of the weakly supervision detector still needs to be improved, and these issues still need to be explored.

5. Conclusions

In this paper, we propose a novel weakly supervised object detection process that combines a weakly supervised localization method to process complex remote sensing images, detect objects in the images, and reduce false positive samples. The corresponding network structure is designed for this method. First, the corresponding network structure is designed that can be trained using the localization information of the samples to effectively suppress the low-quality samples from being misclassified as positive samples. Then, an adaptive threshold adjustment strategy is designed to calculate appropriate thresholds for different categories to improve the overall performance.

Detailed experiments show that our model obtains advanced performance on experimental datasets, particularly on targets such as aircraft, baseball fields, and tennis courts. Although the performance was improved, there were still some problems at that time. For example, the detection performance was not satisfactory on targets such as bridges, vehicles, and oil storage tanks, and this part of the problem still needs to be explored.

Author Contributions: Formal analysis, H.W.; Investigation, H.W.; Methodology, H.W.; Supervision, H.L.; Visualization, H.W.; Writing—original draft, H.W.; Writing—review & editing, H.W., H.L., W.Q., W.D., L.Z., J.Z., and D.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under Grants 41701508.

Acknowledgments: The authors would like to thank the anonymous reviewers for their very competent comments and helpful suggestions.

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

References

1. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
2. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9, doi: 10.1109/CVPR.2015.7298594.
3. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Seattle, WA, USA, 27–30 June 2016; pp. 2818–2826.
4. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 448–456.
5. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.; Berg, A.C. *SSD: Single Shot MultiBox Detector*; Springer: Cham, Switzerland, 2016.

6. Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; Fu, K. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 8232–8241.
7. Sun, X.; Liu, Y.; Yan, Z.; Wang, P.; Diao, W.; Fu, K. SRAF-Net: Shape Robust Anchor-Free Network for Garbage Dumps in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2020**, *99*, 1–15.
8. Wang, P.; Sun, X.; Diao, W.; Fu, K. FMSSD: Feature-Merged Single-Shot Detection for Multiscale Objects in Large-Scale Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 3377–3390.
9. Wei, H.; Zhang, Y.; Chang, Z.; Li, H.; Wang, H.; Sun, X. Oriented Objects as pairs of Middle Lines. *J. Photogramm. Remote Sens.* **2020**, *169*, 268–279.
10. Sun, X.; Shi, A.; Huang, H.; Mayer, H. BAS^{4}Net: Boundary-Aware Semi-Supervised Semantic Segmentation Network for Very High Resolution Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 5398–5413.
11. Ren, S.; He, K.; Girshick, R. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *39*, 1137–1149.
12. Yu, C.-N.J.; Joachims, T. Learning structural svms with latent variables. In Proceedings of the 26th International Conference on Machine Learning (ICML), Montreal, QC, Canada, 14–18 June 2009; pp. 1169–1176.
13. Song, H.O.; Lee, Y.J.; Jegelka, S.; Darrell, T. Weakly supervised discovery of visual pattern configurations. *Adv. Neural Inf. Process. Syst.* **2014**, *2014*, 1637–1645, .
14. Ren, W.; Huang, K.; Tao, D.; Tan, T. Weakly Supervised Large Scale Object Localization with Multiple Instance Learning and Bag Splitting. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 405–416.
15. Ye, Q.; Zhang, T.; Qiu, Q.; Zhang, B.; Chen, J.; Sapiro, G. Self-learning scene-specific pedestrian detectors using a pro-gressive latent model. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2057–2066.
16. Bilen, H.; Vedaldi, A. Weakly supervised deep detection networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2846–2854.
17. 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.
18. Gao, M.; Li, A.; Yu, R.; Morariu, V.I.; Davis, L.S. C-wsl: Count-guided weakly supervised localization. In Proceedings of the European Conference on ComputerVision (ECCV), Munich, Germany, 8–14 September 2018; pp. 152–168.
19. 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.
20. 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.
21. Wan, F.; Liu, C.; Ke, W.; Ji, X.; Jiao, J.; Ye, Q. C-mil: Continuation multiple instance learning for weakly supervised object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2199–2208.
22. 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.
23. 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.
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**, doi:10.1109/TGRS.2020.3030990.
25. Andrews, S.; Tsochantaridis, I.; Hofmann, T. Support vector machines for multiple-instance learning. *Adv. Neural Inf. Process. Syst.* **2002**, *2002*, 561–568.
26. Zhang, Y.; Bai, Y.; Ding, M.; Li, Y.; Ghanem, B. W2f: A weakly-supervised to fully-supervised framework for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 928–936.
27. Diba, A.; Sharma, V.; Pazandeh, A.; Pirsiavash, H.; Van Gool, L. Weakly supervised cascaded convolutional networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
28. Shen, Y.; Ji, R.; Zhang, S.; Zuo, W.; Wang, Y. Generative Adversarial Learning Towards Fast Weakly Supervised Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
29. Ge, W.; Yang, S.; Yu, Y. Multi-Evidence Filtering and Fusion for Multi-Label Classification, Object Detection and Semantic Segmentation Based on Weakly Supervised Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
30. Shen, Y.H.; Ji, R.R.; Wang, Y.; Wu, Y.J.; Cao, L. Cyclic guidance for weakly supervised joint detection and segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.

31. Li, X.; Kan, M.; Shan, S.; Chen, X. Weakly supervised object detection with segmentation collaboration. *arXiv* **2019**, arXiv:1904.00551.
32. Uijlings, J.; Sande, K.v.; Gevers, T.; Smeulders, A. Selective search for object recognition. *Int. Comput. Vis.* **2013**, *104*, 154–171.
33. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
34. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
35. Lin, M.; Chen, Q.; Yan, S. Network in network. In Proceedings of the Interna-Tional Conference on Learning Representations, Banff, AB, Canada, 14–16 April 2014.
36. Selvaraju, R.R.; Das, A.; Vedantam, R.; Cogswell, M.; Parikh, D.; Batra, D. Grad-CAM: Why did you say that? *arXiv* **2016**, arXiv:1611.07450.
37. Chattopadhay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 839–847, doi:10.1109/WACV.2018.00097
38. Cheng, G.; Han, J.; Zhou, P.; Lei, G. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *J. Photogramm. Remote Sens.* **2014**, *98*, 119–132.
39. Cheng, G.; Han, J. A survey on object detection in optical remote sensing images. *J. Photogramm. Remote Sens.* **2016**, *117*, 11–28.
40. 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.
41. Deselaers, T.; Alexe, B.; Ferrari, V. Weakly supervised localization and learning with generic knowledge. *Int. J. Comput. Vis.* **2012**, *100*, 275–293, .
42. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Fei, L. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252.
43. Hosang, J.; Benenson, R.; Schiele, B. Learning non-maximum suppression. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6469–6477.