remote sensing

MDPI

Article

Semi-Autonomous Learning Algorithm for Remote Image Object Detection Based on Aggregation Area Instance Refinement

Bei Cheng [1], Zhengzhou Li [1,2,*], Hui Li [1], Zhiquan Ding [3] and Tianqi Qin [3]

[1] College of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China; chengbei@cqu.edu.cn (B.C.); lijiahui@cqu.edu.cn (H.L.)
[2] Key Laboratory of Beam Control, Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu 610209, China
[3] Sichuan Institute of Aerospace Electronic Equipment, Chengdu 610100, China; 13350314996@163.com (Z.D.); tianqiqin2008@163.com (T.Q.)
* Correspondence: lizhengzhou@cqu.edu.cn; Tel.: +86-132-0601-5717

Abstract: Semi-autonomous learning for object detection has attracted more and more attention in recent years, which usually tends to find only one object instance with the highest score in each image. However, this strategy usually highlights the most representative part of the object instead of the whole object, which may lead to the loss of a lot of important information. To solve this problem, a novel end-to-end aggregate-guided semi-autonomous learning residual network is proposed to perform object detection. Firstly, a progressive modified residual network (MRN) is applied to the backbone network to make the detector more sensitive to the boundary features of the object. Then, an aggregate-based region-merging strategy (ARMS) is designed to select high-quality instances by selecting aggregation areas and merging these regions. The ARMS selects the aggregation areas that are highly related to the object through association coefficient, and then evaluates the aggregation areas through a similarity coefficient and fuses them to obtain high-quality object instance areas. Finally, a regression-locating branch is further developed to refine the location of the object, which can be optimized jointly with regional classification. Extensive experiments demonstrate that the proposed method is superior to state-of-the-art methods.

Keywords: semi-autonomous learning; aggregation area; residual network; object detection; remote sensing image

Citation: Cheng, B.; Li, Z.; Li, H.; Ding, Z.; Qin, T. Semi-Autonomous Learning Algorithm for Remote Image Object Detection Based on Aggregation Area Instance Refinement. *Remote Sens.* **2021**, *13*, 5065. https://doi.org/10.3390/rs13245065

Academic Editors: Mercedes E. Paoletti and Juan M. Haut

Received: 26 October 2021
Accepted: 6 December 2021
Published: 14 December 2021

1. Introduction

With the rapid development of modern remote sensing technology, remote sensing image processing has been widely used in various fields, such as object detection [1–3], road mapping [4–6], agricultural planning [7–9], semantic analysis [10–12] and urban planning [13–15]. Recently, thanks to the stronger feature representation ability of Convolutional Neural Networks (CNNs) [16,17] and the availability of rich datasets with case-level annotations, the object detection of remote sensing images has achieved breakthrough performance. However, collecting such accurate annotations is labor-intensive and time-consuming. In contrast, scene-level tags can be easily obtained. This paper mainly studies the application of weakly supervised scene-level tags to realize semi-autonomous learning of object detections in remote sensing images.

In the existing remote sensing image object detection methods, the data used for the training network usually have object-level markers, which include the specific number, location, size or direction of all objects in the scene. However, the label at the scene level only records the category type of the main objects in each scene, and does not contain the specific object information in the scene. In addition, scenes with the same label usually contain different numbers of objects with different positions, sizes and orientations. The lack of this object information brings great challenges to the learning process.

With the aid of multiple instance learning, researchers [16,18–27] have shown that deep networks trained with only image-level/scene-level tags can generate pseudo ground truth, thus effectively predicting object location information. Most semi-autonomous learning object detection methods are divided into two stages [22,28–30]. In the first stage, a series of candidate boxes are generated by the object proposal method [31,32], and in the second stage, the features obtained from each proposal are regarded as instance-level classification problems. The core of these methods is to treat each image as a bag of potential object instance, and then train the instance classifier under the constraint of multi-instance learning. Promising results have been reported in the aforementioned methods. In [23], Ren et al. proposed an unpacking algorithm which iteratively removes negative instances from positive bags; this algorithm can effectively reduce the ambiguity of positive bags. However, the method ignores the accuracy of positive bags and may result in less accurate positive bags. Bilen et al. [22] proposed an effective multi-instance learning method, namely, the weakly supervised deep detection network (WSDDN), which can operate at the image region level and perform region selection and classification at the same time. Nevertheless, this method lacks displayed labels to classify instances. Later, Tang et al. [28] proposed a new online instance classifier refinement (OICR) algorithm, which integrates the multiple instance learning and instance classifier optimization process into a single deep network, and at the same time, it proposes to optimize instance classifier online by using a spatial relationship. This method can learn more distinguished instance classifiers by assigning binary labels on display. However, this method makes the network pay more attention to part of objects and ignores the complete objects. Thereafter, a proposal cluster learning (PCL) is presented to [29], to realize a refined instance classifier by learning the generated proposal clusters. As the proposals in the same cluster are associated with the same object, this can make the network focus more on the whole object instead of the parts of objects. In [30], Feng et al. designed a dual contextual refinement strategy to shift the focus of the detection network from locally different parts to the whole object by using local and global context information.

Although these studies have achieved good results, the development of semi-autonomous learning in remote sensing images still faces two major challenges. The first challenge is that deep residual networks, such as ResNet [33] and DenseNet [34], have become the standard backbones of many computer vision tasks, while the advanced semi-autonomous learning methods for object detection still rely on common networks, such as VGG [35]. The fundamental problem is unstable for the header of semi-autonomous learning, and is particularly sensitive to model initialization. This may propagate uncertain and wrong gradients back to the backbone, thus deteriorating visual representation learning. Especially in remote sensing image, the background is complex, the features are chaotic and the object size is small. The large kernel convolution kernel and non-maximum down-sampling in traditional ResNet will lead to the loss of highly informative features in the original image [36].

The second challenge is that most methods of semi-autonomous learning tend to select only the candidate frames with the highest scores to train the corresponding detectors. However, the suggestions with the highest scores and their associated suggestions usually cover only a part of the object instead of the whole object, especially in the large-scale and chaotic remote sensing image. Therefore, the above methods may not be enough to extract the complete object features. In addition, remote sensing images usually contain multiple instances of the same kind. If the highest score and its associated suggestions are simply selected as pseudo GT, other instances of the same kind may be missed, resulting in suboptimal object detectors [29].

For the first challenge, we designed a modified residual network (MRN) to solve the problem of losing highly informative features in the original image. Specifically, this paper replaces large convolution kernel and non-maximum down-sampling in a traditional residual network with a small kernel convolution and maximum down-sampling. It could enhance the robustness of network information flow by extracting finer object boundary features from the original image and keeping the information of instances. In the process

of inputting remote sensing images into the deep neural network, continuous convolution and pooling will reduce the image size and increase the receptive field, which will lose some resolution and make some image details unable to be reconstructed. To solve this problem, the hole convolution is used in ResNet to increase the receptive field without the loss of feature information by the pooling operation. Each convolution output contains a wider range of feature information than an ordinary convolution, which is beneficial in obtaining the global information of the object features in remote sensing images.

In order to address the second challenge, this paper proposes an end-to-end aggregation-based region-merging strategy (ARMS), which combines similar regions in the cluster and deletes redundant regions in the cluster to select high-quality regions. In the process of network refinement, this strategy shifts the focus of network detection from part of the object to the whole object, and reduces the influence of background area on the network. More specifically, this paper does not directly select the suggestion with the highest score as pseudo-supervision, but selects other areas related to the suggestion with the highest score, and merges these related areas to form a complete object area, generating a new confidence level according to the merged form.

In this paper, MRN and ARMS are combined to form a progressive aggregation area instance refinement network. The MRN solve the problem of losing highly informative features in the original image. It extracts finer object boundary features in image processing, and preserves the information of instances. The ARMS is designed to select high-quality instances by selecting aggregation areas and merging these regions. ARMS selects the aggregation areas that are highly related to the object information through the association coefficient, and then evaluates the aggregation areas through the similarity coefficient and fuses them to obtain high-quality object instance areas. Moreover, a regression-locating branch is further developed to refine the location of the object, which can be optimized jointly with regional classification. Experiments on augmented NWPU VHR-10 and LEVIR datasets clearly demonstrate that the proposed method significantly outperforms previous state-of-the-art semi-autonomous learning object detection approaches on LEVIR and augmented NWPU datasets. In summary, the main contributions of this paper can be summarized as follows: (1) An end-to-end framework is proposed to realize semi-autonomous learning object detection based on image level annotations, which can jointly optimize the region classification and regression to improve the accuracy of object location. (2) The MRN is applied as the backbone to implement a lightweight network structure, which provides finer object boundaries and preserves the information of small instances, enhancing the robustness of information flow in the network. (3) A novel ARMS algorithm is proposed to utilize the cluster information of the region to obtain high-quality object features, and further suppress the interference of messy background information.

2. Materials and Methods

The overall architecture of the proposed framework is illustrated in Figure 1. The core goal of this method is to mine more accurate positive instances by making full use of aggregation areas under a weak supervision setting. For this reason, an instance refinement framework of the aggregation area based on a modified residual network is designed. Specifically, for each input image, the selective search method [32] is applied to generate about 2000 suggestions, MRN is used to extract the features of the image, and then the data enter the multiple instance learning (MIL) branch to generate instance-level supervision. Then, the AMSR further refines the instances, and finally uses the classification and regression branch to process the features at the same time to get the final result.

2.1. Modified Residual Network for Semi-Autonomous Learning

The residual network is a variant network proposed by Mike et al. in 2015, which has achieved excellent results in the field of image recognition and object detection [37,38]. The residual network has less computation and can solve the problems of network degradation and gradient dispersion. Deep residual networks, such as ResNet and DenseNet, have

become the standard backbone of many computer vision tasks [39–41], but the advanced semi-autonomous learning methods still rely on common networks, such as VGG. The fundamental problem is that the header of the semi-autonomous learning network is particularly sensitive to model initialization. This may propagate uncertain and wrong gradients back to the backbone and deteriorate visual representation learning. Training the semi-autonomous learning network with ResNet backbone will reduce the identification of the proposed features, and will be weak in localizing object instances. Discovered by [36], the proposed semi-autonomous learning algorithm in this paper takes MRN as the backbone network.

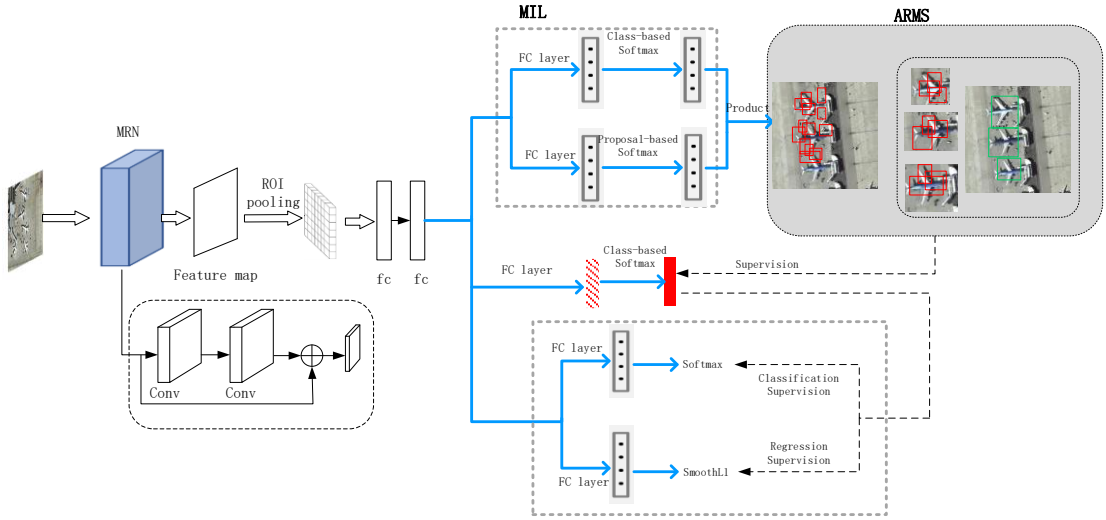

Figure 1. The overall architecture of the proposed framework.

In the traditional plain network, the data source of the network can only come from the previous layer; as shown in Figure 2a, the data flow down layer by layer. In the convolutional network, the data will flow into the next layer after down sampling, and data will be compressed in this process. In the residual structure, a design similar to the "short circuit" is introduced, in which the data output from the previous layer will directly skip over several layers and be introduced into the input part of the following data layer, as shown in the Figure 2b. Simply put, the uncompressed vector data of the previous layer and the further compressed data of the later layer will be used as the later input data together. More abundant dimension characteristic values are introduced into the network, so that the network can learn more abundant features. The residual convolution block is composed of several residual convolution submodules, as shown in Figure 2b. Assuming that the input of layer l is x_l, then the input of layer $l+1$ is

$$x_{l+1} = H\left(x_l, w_f\right) = x_l + F(x_l, w_l) \tag{1}$$

where $w_l = \{w_{l,k} | 1 \leq k \leq K\}$ is the parameter of layer l, and K represents the number of residual unit layers. Sequential cycle analogy is denoted as

$$x_{l+2} = x_{l+1} + F(x_{l+1}, w_{l+1}) = x_l + F(x_l, w_l) + F(x_{l+1}, w_{l+1}) \tag{2}$$

$$x_L = x_l + \sum_{i=1}^{L-1} H(x_i, w_i) \tag{3}$$

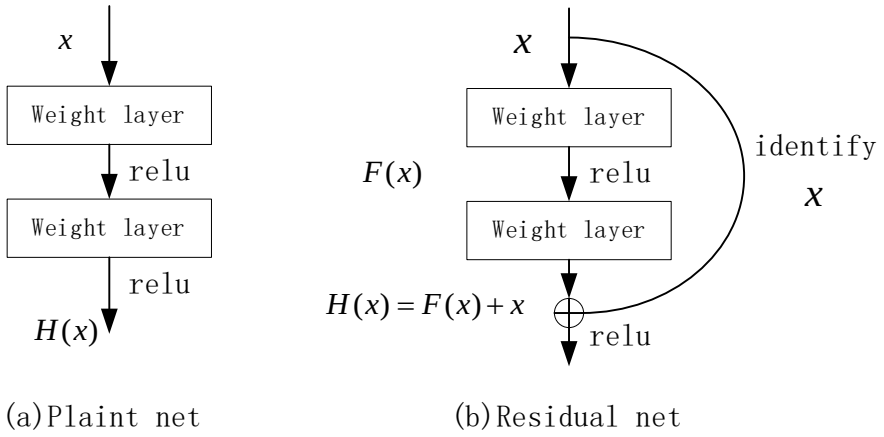

Figure 2. The data flow for plaint net and residual net.

In Formula (3), it can be seen that the feature x_L in layer L can be divided into two parts—the first part is shallow network representation x_l, and the second part is the mapping of residual function $\sum_{i=1}^{L-1} H(x_i, w_i)L$, which indicates that the model is a residual form in any unit. Different from the plain network, the feature in layer L is the product of a series of vectors, that is, $\prod_{i=0}^{L-1} W_i x_i$. In addition, a network often has the best network level; therefore, many deep networks have many redundant network layers. At this time, it is hoped that these redundant layers can complete identity mapping to ensure that the input and output through this identity layer are identical. As shown in Formula (3), in forward propagation, the input can be propagated directly from any lower level to higher level, which is equivalent to containing a natural identity mapping. This characteristic can solve the problem of network degradation to a certain extent. This also explains why the output in Figure 2b is $H(x) = F(x) + x$. In the residual network, for features with any depth, the result is the sum of all the previous residual modules, which increases the diversity of features. Similarly, the residual network also has good back propagation characteristics. Assuming that the loss is E, according to the chain derivation rule, we can get

$$\frac{\partial E}{\partial x_l} = \frac{\partial E}{\partial x_L} \cdot \frac{\partial x_L}{\partial x_l} = \frac{\partial E}{\partial x_L} \left(1 + \frac{\partial}{\partial x_l} \sum_{i=1}^{L-1} H(x_i, w_i) \right) \tag{4}$$

It can be seen from the Formula (4) that the gradient consists of two parts—one is the information flow without any weight, another one is the weighted information flow. The linear characteristics of the connection between the two parts ensure that the information can be transmitted back to the shallow layer. At the same time, the formula also shows that for a small batch-size, the possibility of gradient disappearance becomes smaller.

Although the residual network has powerful advantages for visual tasks, it cannot be directly applied to semi-autonomous learning. This is because semi-autonomous learning is particularly sensitive to model initialization. Firstly, the large kernel convolution weakens the information of the object boundary, which leads to the uncertainty around the object boundary. Secondly, non-maximum down-sampling may also damage the flow of information, which makes small instances difficult to be perceived. This is because non-maximum down-sampling may not retain the activation and gradient of the information flowing through the network under weak supervision. Inspired by [36], a MRN is proposed in this paper, and the small kernel convolution and max-pooling are used to improve the robustness of information flow, which makes the object boundary more detailed. Specifically, the original stem block is replaced by three conservative 3×3 convolutions, and the first and third convolutions are followed by the max-pooling layer. For down-sampling, max-pooling is applied to change the average-pooling operation. The general depth residual network usually uses five stages to extract the whole image feature map. In the process of inputting remote sensing images into the deep neural network, continuous convolution and pooling operations will be carried out on the images, and this

operate will lose some resolution and make some image details unable to be reconstructed. To solve this problem, this paper uses hole convolution to increase the receptive field when extracting the high-resolution weakly supervised feature so that each output of convolution contains a wider range of feature information than ordinary convolution, and it is beneficial to obtain the global information of object features. Specifically, the size of the space is fixed after the third stage, and the expansion convolution with a rate of 2 is added in the subsequent stage. Table 1 displays some parameters of MRN. Figure 3 visualizes the feature map of the VGG backbone network and MRN. The first column is the original image, the second column is the feature map extracted by VGG-16 backbone network and the third column is the feature map extracted by MRN; all these feature maps are come from one specific CNN layer. It can be seen that the MRN can obtain clearer boundary features, especially for the small object in the complex scene. For the scene with multiple objects, the MRN can obtain more complete global information.

Table 1. The design details of MRN.

Layer Name	Convolution Kernels	Rate of Dilated Convolutions
Conv1	3×3, conv 2×2, maxpool 3×3, conv 3×3, conv 2×2, maxpool	–
Conv2	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$	–
Conv3	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$	–
Conv4	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$	2
Conv5	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$	2
2×2, maxpool, stride2		

2.2. Multiple Instance Learning Branch (MIL Branch)

The image only contains the image-level label indicating whether the object category appears or not. In order to train a standard object detector, it is necessary to mine instance-level supervision, such as the bounding-box annotation. Therefore, this paper needs to introduce a MIL branch to initialize the pseudo ground truth (GT) annotations, such as [22,28]. The OICR network based on WSDDN is chosen in this paper because it is effective and can realize end-to-end training. WSDNN adopts a dual-stream network: classification stream x_{cls} and detecting data stream x_{det}. Instance-level prediction can be achieved by aggregating these two streams. The scores of region proposal can be defined as

$$x_{cr} = x_{cls} \odot x_{det} \tag{5}$$

$$x_{cls} = \frac{e^{x_{ij}^c}}{\sum_{k=1}^{C} e^{x_{ij}^c}}, \; x_{det} = \frac{e^{x_{ij}^d}}{\sum_{r=1}^{|R|} e^{x_{ir}^d}} \tag{6}$$

where x_{ij}^c comes from a matrix of data $\mathbf{x}^c \in \mathbb{R}^{C \times |R|}$ and x_{ij}^d comes from a matrix of scores $\mathbf{x}^d \in \mathbb{R}^{C \times |R|}$; $\odot$ means element-wise operation, C represents the number of image categories and R denotes the number of object proposals. Suppose the image label is denoted as $\mathbf{Y} = [y_1, y_2, \ldots, y_C]^T \in \mathbb{R}^{C \times 1}$, $y_C = 0$ indicates that there is no object in the image and

$y_C = 1$ indicates that one or more objects exist in the image. Then, the image-level class prediction score can be generated by summing up the scores over all proposals.

$$S_c = \sum_{r=1}^{|R|} x_{cr} \tag{7}$$

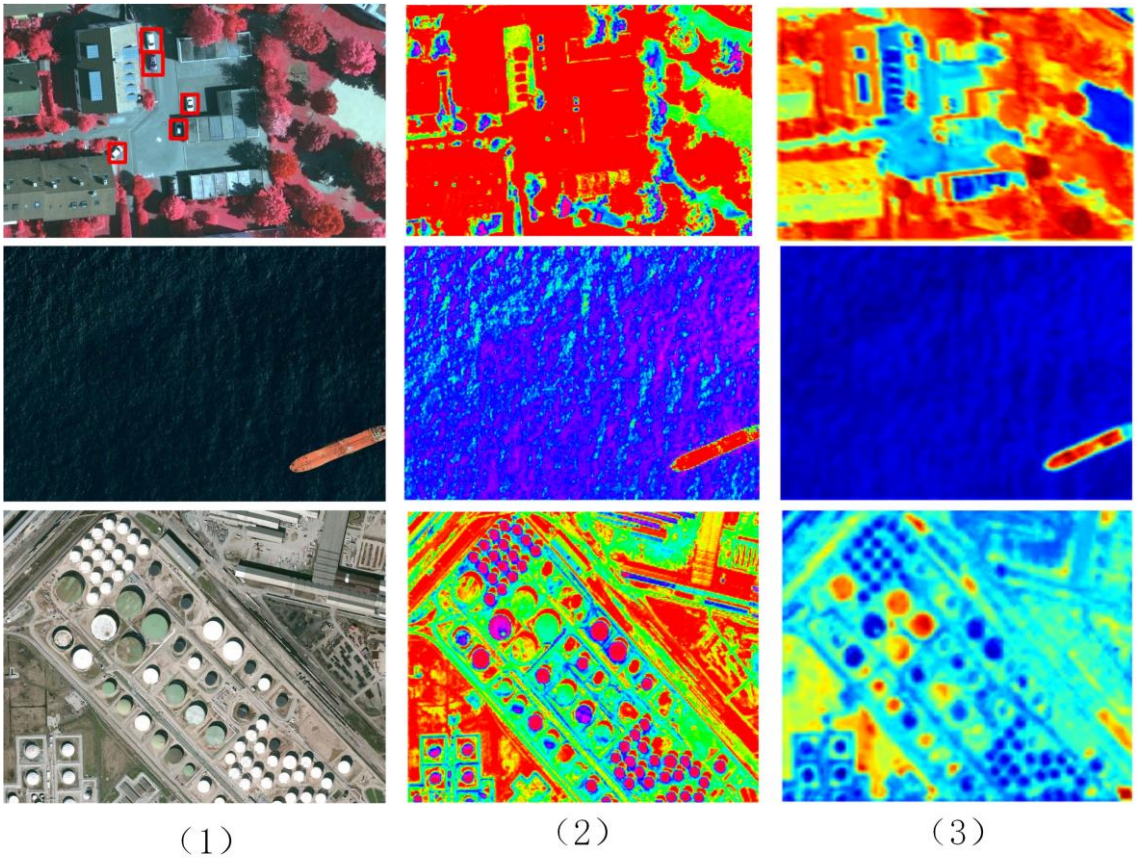

Figure 3. The feature map of VGG backbone network and MRN. (**1**) The original image. (**2**) The feature map extracted by VGG-16 backbone network. (**3**) The feature map extracted by MRN.

The two-stream network of multiple instance can be trained as:

$$loss_{wsddn} = -\sum_{c=1}^{C} \{y_C \log S_c + (1 - y_C) \log(1 - S_c)\} \tag{8}$$

A preliminary classification result about the object instance can be obtained by instantiating the branch, but the result at this time is a very rough initial result. In order to get a finer classification result, similar to [29], this paper designs a refinement branch to be added after the instantiation branch. Different from OICR, our refinement branch contains two sub-branches, namely, the classification branch and location branch. In each refinement branch, the proposed ARMS strategy is introduced to obtain the most representative object area by using the information of the aggregation area; further, to discover instances with more complete characteristics by merging aggregation areas.

2.3. Aggregate-Based Region Merging Strategy (ARMS)

This section will explain how to mine trusted instances to achieve semi-autonomous object detection when only image-level tags are available. Firstly, the aggregation area is generated to shift the focus of the detection network from local different areas to the whole

object area. Then, the aggregated local information is merged and the tag information is propagated to other possible instances.

2.3.1. Identification and Selection of Aggregation Regions

It is needed to select an aggregation center before selecting an aggregation area. As the top region can always detect a part of the object, the proposal with the higher score is selected as the clustering center, and then the aggregation region is obtained according to the degree of spatial overlap with the aggregation center after obtaining the proposal score. Although the proposal with the high score may not be the most complete, it can at least cover a part of the objects. The gathering area formed by taking it as the center will contain some objects or even the whole object, on the assumption that the R object proposals have boxes $B = (b_1, b_2, \ldots, b_R)$ and a score $S = (s_1, s_2, \ldots, s_R)$. Suppose the object category c exists in the image—that is $y_c = 1$. Firstly, a series of the highest-scores are selected by

$$r_0^c = \text{argmax}(s_r^c) \tag{9}$$

Then, this proposal is chosen as the nth center of the aggregation area, i.e., $R_0^c = (b_n, y_n, s_n) = (b_{r_0^c}, c, s_{r_0^c})$, where $b_{r_0^c}$ denotes the box of r_0^cth proposal region. To take advantage of aggregated information in the aggregation area, an association coefficient $\sigma(R_0^c, R_n^c)$ is introduced to describe the closeness between the center of the aggregation area and other proposals. The composition of the nth aggregation region corresponding to category c can be expressed as

$$R_n^c = \{R_0^c | \sigma(R_0^c, {R_i}^c) > \lambda\} \tag{10}$$

$$\sigma(R_0^c, R_i^c) = \frac{area(b_0^c \cap b_i^c)}{area(b_0^c)} \tag{11}$$

$$area = \left(x_{\max}^i - x_{\min}^0 + 1\right) * \left(y_{\max}^i - y_{\max}^0 + 1\right) \tag{12}$$

where $b = (x_{\min}, y_{\min}, x_{\max}, y_{\max})$ denotes the coordinates of proposals. $x_{\min}$ and $y_{\min}$ represent the coordinates of the upper left corner, $x_{\max}$ and $y_{\max}$ represent the coordinates in the lower right corner and λ is a threshold that is applied to select areas close to the central area to form the aggregation area.

2.3.2. Aggregation Region Merging

The selected aggregation area set R_n^c can provide supplementary information to the object, which is helpful to obtain complete object features. Although the area with the highest score is not the most complete object region, it can cover at least a part of the object. Then, taking this region as the clustering center and gradually merging the adjacent areas, a nearly complete region can be obtained. However, one of the main problems is that the number of these regions is not fixed and may contain useless background information. Merging arbitrary regions may lead to negative effects. In order to solve this problem, a graph is designed to evaluate the similarity of these aggregation areas, and then the aggregation areas with high similarities are merged by the greedy strategy, so as to obtain the object areas with more complete features. The concept of the similarity coefficient $SM_{mn}(f_{mn}, g_{mn})$ is introduced to evaluate the similarity between areas. The similarity coefficient is divided into two parts—feature similarity and geometric similarity. It is well known that regions with similar features will have close detection scores. At the same time, the regions with large overlapping areas are more closely related. Therefore, the label information of local instances through a similarity coefficient measurement can be propagated to the global instance. The similarity measure can be expressed as:

$$SM_{mn}(f_{mn}, g_{mn}) = \begin{cases} f_{mn} = |s_m / s_n - 1| \\ g_{mn} = \sigma(R_m, R_n) \end{cases} \tag{13}$$

where $R_m, R_n \in R_n^c$. f_{mn} and g_{mn} represent the feature similarity and geometric similarity between the region m and region n, respectively. The regions in the aggregation area R_n^c are sorted according to the score. A graph $G = (V, E)$ is established based on the similarity measure. The node V in the graph denotes the aggregation area. Edge $E = e(V_m, V_n)$ represents the connection between nodes and it is determined by the similarity between two nodes.

$$e(V_m, V_n) = \begin{cases} 1 \text{ if } f_{mn} < \tau_s \text{ and } g_{mn} > \tau_R \\ 0 \text{ otherwise} \end{cases} \tag{14}$$

where τ_s denotes the feature metric threshold and τ_R is the geometric metric threshold. Therefore, when the similarity between two nodes is high, they will be connected together. Then, the greedy merging strategy will be adopted to merge these two regions with high enough similarity to obtain a new region. Since regions with similar features have similar detection scores, redefining the merged regions can shift the focus of the detection network from partial object information to complete object information. The score and coordinates of the merged area are defined as:

$$s_A = s_m + s_n \tag{15}$$

$$A(x_{\min}, y_{\min}, y_{\min}, y_{\max}) = \begin{Bmatrix} x_{\min}=\max(x_{\min}^m, x_{\min}^n), y_{\min}=\max(y_{\min}^m, y_{\min}^n) \\ x_{\max}=\min(x_{\max}^m, x_{\max}^n), y_{\max}=\min(y_{\max}^m, y_{\max}^n) \end{Bmatrix} \tag{16}$$

As the aggregation regions are all closely related to the regions with the highest score, these regions contain at least a part of the features of the object, and further merging these parts of the features can obtain nearly complete features of the object. The ARMS mines more accurate instance information to a certain extent and obtains higher region scores, and provides more accurate monitoring information for subsequent object feature extraction. Figure 4 shows the process of aggregation region merging in ARMS. The Algorithm 1 summarizes the whole process of ARMS. In the first row, the relationship between region A and region C satisfies the description in Formula (14) according to the SM_{AC}, then region A and region C are merged. The same is done between region A and region B. In the last row, all regions do not meet the consolidation conditions in Formula (14), so there are no merge operations between regions.

Algorithm 1 Aggregate-Based Region Merging Strategy

Input: Image-level label $\mathbf{Y} = [y_1, y_2, \ldots, y_C]^T$; proposal boxes $B = (b_1, b_2, \ldots, b_R)$; proposal score matrix $S = (s_1, s_2, \ldots, s_R)$.
Output: merge region score s_A; bounding box $A(x_{\min}, y_{\min}, y_{\min}, y_{\max})$
1: for c = 1 to C do;
2: if y_c = 1 then;
3: Choose the highest-score proposal R_0^c by Equation (9);
4: Find aggregation region R_i^c by Equations (10)–(12);
5: Build a graph between aggregation areas;
6: Compute similarity between R_m and R_n by Equation (13);
7: Merge aggregation region by Equation (14);
8: Generates the instance-level score and coordinate by Equations (15) and (16);
9: End if;
10: End for.

2.3.3. Multi-Task Classification and Regression Branch

After executing the above process, the pseudo GT can be obtained, and the subsequent classification task can be carried out in a fully supervised way. Between WSDDN and OICR, the subsequent classification only includes one branch of classification. Inspired by [42], this paper uses a multi-task classification and regression branch to perform detection tasks at the same time, similarly to Faster-RCNN [43], in which the classification branch outputs the discrete probability distribution, which is calculated by a softmax on the $C + 1$ outputs of FC layer. The regression branch predicts a bounding-box regression of each

C object classes. Suppose the output of the classification branch and regression branch are $p^c \in \mathbb{R}^{(C+1)\times 1}$ and $b^c = \left(b_x^c, b_y^c, b_w^c, b_h^c\right)$, respectively. Suppose $loss_{det}$ is used for the classification and bounding-box regression. The loss function of the network is defined as

$$loss = loss_{wsddn} + loss_{det} \tag{17}$$

$$loss_{det} = loss_{cls}(y^c, p^c) + \lambda loss_{reg}(t^c, b^c) \tag{18}$$

where $loss_{cls}$ and $loss_{reg}$ represent classification loss and regression loss, respectively. λ is the balance parameter that controls the balance between these two losses. $loss_{cls}$ and $loss_{reg}$ can be calculated as:

$$loss_{cls}(y^c, p^c) = -\frac{1}{|R|}\sum_{r=1}^{|R|}\sum_{c}^{C+1} \omega_r y_r^c \log p_r^c \tag{19}$$

$$loss_{reg}(t^c, b^c) = \sum_{i\in(x,y,w,h)} smooth_{L_1}(t_i^c - b_i^c) \tag{20}$$

$$smooth_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if}|x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \tag{21}$$

where $t^c = \left(t_x^c, t_y^c, t_w^c, t_h^c\right)$ denote the pseudo GT bounding box and y^c represents the real label. ω_r denotes the loss weight and can be calculated following the weights calculation method in [28]. R is the number of region proposals.

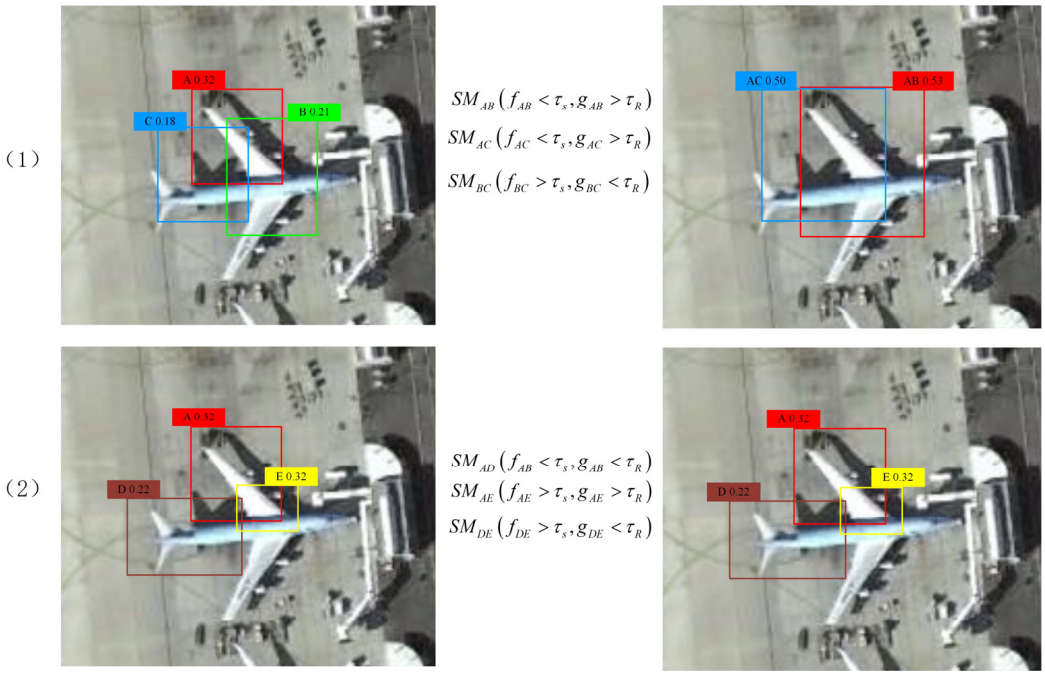

Figure 4. Illustration of the process of aggregation region merging in ARMS. (**1**) The relationship between region A and region C satisfies the description in Formula (14) according to the SM_{AC}, then region A and region C are merged. The same is done be-tween region A and region B. (**2**) All regions do not meet the consolidation conditions in Formula (14), so there are no merge operations between regions.

3. Results

3.1. Dataset and Evaluation Metrics

The proposed method is evaluated on the augmented NWPU VHR-10 [44] and LEVIR [45] datasets. The NWPU VHR consists of 650 images (400 × 400 pixels) from ten object categories. This dataset is composed of three groups: the train, test and valuation

set. In this experiment, 650 images were expanded to 1950 images by rotating the images 90 degrees, 180 degrees and 270 degrees, respectively. Half of the images are used for training data and half for testing. All of the 1950 images were scrambled, and the training data and test data were randomly selected from these images in proportion to ensure the balance and randomness of the training and test data. In the test, two standard metrics are used to evaluate the performance of the proposed method. Average precision (AP) is measured to evaluate the proposed model on the test set, and the correct position (CorLoc) is applied to evaluate the positioning accuracy of the proposed model.

3.2. Comparisons with State-of-the-Art Methods

In order to evaluate the robustness and effectiveness of the proposed method, several widely used semi-autonomous learning methods are taken for performance comparison. These semi-autonomous learning methods have made breakthrough achievements in the field of object detection, including CSC [46], WSDDN [22], OICR [28] and PCL [29]. CSC is a weak supervision method that designs a new change segmentation and classification module based on image-level semantic tags. WSDDN initiated the transformation of the weak supervision problem into an architecture of the implicitly learning object detector. OICR and PCL add new online instance classifier optimization branches on the basis of WSDDN. In addition, the results of the proposed method are compared with fully supervised object detection methods, such as Faster-RCNN [43], YOLOv4 [47], DCIFF-CNN [1], YOLOv4-CSP [48] and YOLOv5 (https://github.com/ultralytics/yolov5.git, accessed on 1 December 2021). Among them, Faster-RCNN and YOLOv4 are representative of CNN object detection methods, while DCIFF-CNN is a deep learning method for remote sensing images.

3.2.1. Detection Results on Augmented NWPU VHR-10 Dataset

Table 2 shows the AP detection results of the different methods for the augmented NWPU-VHR 10 dataset. It can be observed that the proposed method achieves the highest mean average precision (mAP) value, and the mAP value obtained by WSDDN is slightly lower. Compared with other methods, WSDDN lacks a multi-instance refinement branch, while OICR and PCL combine the multi-instance learning and multi-stage instance classifier. The refinement branch makes the ability of the instance classifier iteratively enhanced and improves the detection performance. The difference is that in the process of refinement, ARMS combines the regions with the highest scores and the aggregation regions closely related to it, so as to obtain more complete object features, while OICR simply selects a series of regions with the highest scores as corresponding categories. For objects of storage tanks and vehicles that are small in size and crowded, the proposed method obtains outstanding performance. This is because ARMS can distinguish multiple instances in dense areas more effectively while extracting the complete features of the object. At the same time, the added positioning branches can locate the object more accurately and eliminate the redundant bounding box. For objects of the ground-track-field and bridge, WSDDN gets better AP values. However, this method is unstable in detecting small objects. In addition, the method in this paper further narrows the gap between the semi-autonomous learning method and the fully supervised learning method. The dataset in the semi-autonomous learning method is based on the image level label, while the data in the full supervision method are based on the object level label, and the results of the two methods are quite different. The particularity of the dataset brings many challenges and difficulties to the detection of semi-autonomous learning methods. We strive to make the detection accuracy under the semi-autonomous learning condition close to that under a fully supervised condition. In particular, for the object categories such as the ship, storage tank, baseball diamond, ground-track-field, vehicles, etc., the proposed method has achieved equivalent or even better results than the fully supervised method. Therefore, the results in Table 2 show that the proposed method is robust and effective in multi-task remote sensing image object detection.

Table 2. AP detection results of different methods for augmented NWPU-VHR 10 dataset (Bold indicates the best result).

	Fully Supervised Methods					Semi-Autonomous Learning Methods			
Objects \ Methods	Faster-RCNN	YOLOv4	DCIFF-CNN	YOLOv4-CSP	YOLOv5	WSDDN	OICR	PCL	Ours
Airplane	89.41	94.60	89.65	97.61	96.81	12.90	22.54	30.28	**38.03**
Ship	76.73	67.58	74.82	67.01	79.87	50.62	70.00	70.00	**71.25**
Storage tank	57.20	83.78	77.62	94.15	92.31	31.43	58.00	56.00	**82.00**
Baseball diamond	88.74	94.30	89.84	92.92	95.76	90.91	94.07	94.92	**96.19**
Tennis court	64.39	74.09	79.66	86.23	92.43	5.96	1.23	2.45	**22.09**
Basketball court	69.37	57.93	89.87	90.09	89.50	39.20	29.49	16.10	**33.05**
Ground-track-field	89.67	91.42	90.41	92.63	95.00	98.05	**99.57**	99.11	99.13
Harbor	78.97	74.13	88.76	91.68	95.62	5.32	9.09	**20.45**	18.18
Bridge	72.75	56.42	78.99	67.50	80.67	1.21	3.61	3.35	**8.64**
Vehicle	57.80	69.06	76.14	82.42	85.01	9.30	19.23	20.41	**61.54**
mAP	74.50	76.33	83.58	86.22	90.30	34.49	40.68	41.31	**53.01**

Table 3 shows the performance of the augmented NWPU VHR-10 dataset measured by CorLoc. For the fully supervised method, the training data all have specific positioning labels. Since CorLoc is the evaluation criterion for positioning accuracy, it is only applicable to the semi-autonomous learning method with image-level labels. Compared with WSDDN, OICR and PCL, the proposed method gets 19.12%, 12.33% and 11.70% improvement, respectively. For the object of harbor, PCL gets a better CorLoc value. However, for the object of tennis court and basketball court, which are similar in shape and size, PCL gets a poor result. This is because the background is complex, and for objects with high similarity, choosing only the proposal with the highest score as a positive example will lead to the loss of a large amount of object information. Because ARMS combines the area with the highest score and the gathering area closely related to it, it can mine even more complete positive instance information, which will be of great improvement to the positioning results.

Table 3. CorLoc detection results of different methods for augmented NWPU-VHR 10 dataset (Bold indicates the best result).

Objects \ Methods	OICR	WSDDN	PCL	Ours
Airplane	22.54	12.9	30.28	**38.03**
Ship	70.00	50.62	70	**71.25**
Storage tank	58.00	31.43	56	**82**
Baseball diamond	94.07	90.91	94.92	**96.19**
Tennis court	1.23	5.96	2.45	**22.09**
Basketball court	29.49	39.2	16.1	**33.05**
Ground-track-field	**99.57**	98.05	99.11	99.13
Harbor	9.09	5.32	**20.45**	18.18
Bridge	3.61	1.21	3.35	**8.64**
Vehicle	19.23	9.3	20.41	**61.54**
mCorLoc	40.68	34.49	41.31	**53.01**

The precision-recall curves (PRCs) of several semi-autonomous learning methods are displayed in Figure 5. In general, the precision will decrease with the increase of recall. Therefore, the performance of a model cannot be comprehensively measured by only the precision and recall corresponding to a certain point, and it should be measured by the overall performance of the PRCs. It can be observed that the proposed method can achieve better performance than other methods in most object categories. For the

object of ground-track-field, all methods can achieve stable PRCs. This is because this object is large in size and has special features compared with other objects, which is of great help to object detection. For the object of tennis courts and basketball courts, which have similar characteristics and similar environments, the proposed method obviously has more outstanding precision and recall. It can not only accurately identify the object, but also obtain a high-precision detection rate. It is indicated that the proposed aggregation region instance refinement framework can effectively mine multi-task positive instances to accurately identify different objects. For the object of storage tank and vehicle, they are both object categories that are small in size. Moreover, there are a lot of instances in the image they appear, and the instances are relatively crowded. The proposed method can obtain more prominent PRCs for this kind of object. This benefits from the proposed ARMS, which utilizes the aggregation region to form a more complete object feature to solve the problem of missing object information in multiple instances. For the object of bridge, the proposed method obtains the suboptimal PRCs. This is due to the irregular shape of the bridge leading to some difficulties in object detection. However, the proposed method is still superior to most methods. Generally speaking, these results show that the proposed framework, based on aggregation region instance refinement, can mine richer positive instance features and generate more accurate object information by merging aggregation regions.

From the above results, it can be observed that the semi-autonomous learning method proposed in this paper can achieve better performance than other semi-autonomous learning methods in most object categories. WSDDN proposes an implicit learning object detector based on weak supervision, which can simultaneously select and classify regions at the level of image regions. However, compared with other methods, it is difficult and challenging to train classifiers because of the lack of multi-instance refinement branches. OICR proposes a novel online case classifier refinement algorithm, which integrates MIL and the instance classifier refinement procedure into a single depth network. However, OICR tends to select only the most confident candidate box to train the corresponding detector. However, the proposal with the highest score often covers only a part of the object rather than the whole object, which will affect the performance of the algorithm. PCL generates a proposed cluster, and learns the refined instance classifier through an iterative process. It spatially associates the proposals in the same cluster. This can prevent the network from paying too much attention to part of the object instead of the whole object. However, there are usually many similar examples in remote sensing images. Simply selecting the highest score and its related suggestions as pseudo GT may miss other similar examples, resulting in suboptimal object detectors. This paper proposes an end-to-end aggregation-based region merging strategy. By selecting other regions related to the proposal with the highest score and merging these related regions, a complete object region is formed, and a new confidence level is generated according to the merging form. In the iterative process, similar areas in the cluster are merged and redundant areas in the cluster are deleted to select high-quality areas. In the process of network refinement, this strategy shifts the focus of the detection network from part of the object to the whole object; at the same time, it reduces the influence of the background area on the network, and can generate its own cluster areas for many similar instances.

3.2.2. Detection Results on LEVIR Dataset

The LEVIR dataset contains three types of objects, namely, airplane, ship (including offshore and open sea) and oil port. Although it contains few kinds of objects, the data structure of this dataset is allocated reasonably. What is more, the image environment of the dataset is diverse and contains most types of ground features, such as a city, village, mountain and ocean. Table 4 shows the AP detection results of different methods for the LEVIR dataset. It can be observed that the proposed method is far superior to other semi-autonomous learning methods. Compared with other methods, the results obtained by CSC are slightly insufficient. This is because CSC does not add instance refinement

branch and lacks effective supervision information in the subsequent classification. As CSC is also one of the typical semi-autonomous learning methods, it is compared with the semi-autonomous learning method proposed in this paper despite its poor results. As shown in the table, the proposed method outperforms other methods by about 6% in the mAP metric. Especially for the airplane class, it is about 12% higher than that of PCL. In addition, the proposed method further narrows the gap between the semi-autonomous learning method and the full supervision method. The above results prove that the proposed method is robust to various background images.

Table 4. AP detection results of different methods for LEVIR dataset (Bold indicates the best result).

Methods \ Objects		Airplane	Ship	Oil Port	mAP
Fully supervised methods	Faster-RCNN	77.60	70.21	65.34	71.10
	YOLOv4	79.81	73.34	70.40	74.52
	DCIFF-CNN	80.62	77.34	70.73	76.23
	YOLOv4-CSP	93.68	86.79	87.79	89.42
	YOLOv5	95.63	88.62	88.90	91.05
Semi-autonomous learning methods	CSC	0.81	0.13	10.76	3.90
	WSDDN	16.02	0.54	16.10	10.88
	OICR	22.06	0.41	33.93	18.80
	PCL	39.92	13.65	39.92	29.93
	Ours	**52.41**	**14.57**	**40.06**	**35.68**

The comparisons of the CorLoc metric on the LEVIR dataset are reported in Table 5. It is clear that, on average, the proposed method outperforms the previous state-of-the-art methods by a large margin. The proposed method outperforms other methods by about 12% in the mCorLoc metric, especially for the airplane class. Because the dataset contains not only ships far from the coast but also ships close to the coast, with the coast being a complex existence for ships, the environment information of ships is extremely complex and changeable. As shown in Table 5, there is a big deviation between WSDDN and PCL in positioning ship objects. However, the proposed method can overcome these difficulties. This is because the proposed framework can obtain more complete object features in the process of mining more positive examples, which provides more accurate supervision information for the subsequent refinement process. This shows that the proposed method is robust in complex and changeable environments.

The PRCs of several semi-autonomous learning methods for the LEVIR dataset are displayed in Figure 6. It can be observed that the proposed method is superior to other methods. Especially for the objects that have complex background environments, such as the airplane, the proposed method obtains prominent and stable PRC. For the object of the ship, the PRC obtained by OICR is obviously a little worse; this is because the background of the ship is changeable, which include both the offshore scene and open sea scene. Even so, the proposed method still performs well and gets stable PRC. This is a benefit from the modified residual network proposed in this paper, which can extract highly informative features from the original image. Generally speaking, the proposed environment is robust in complex and changeable scenes.

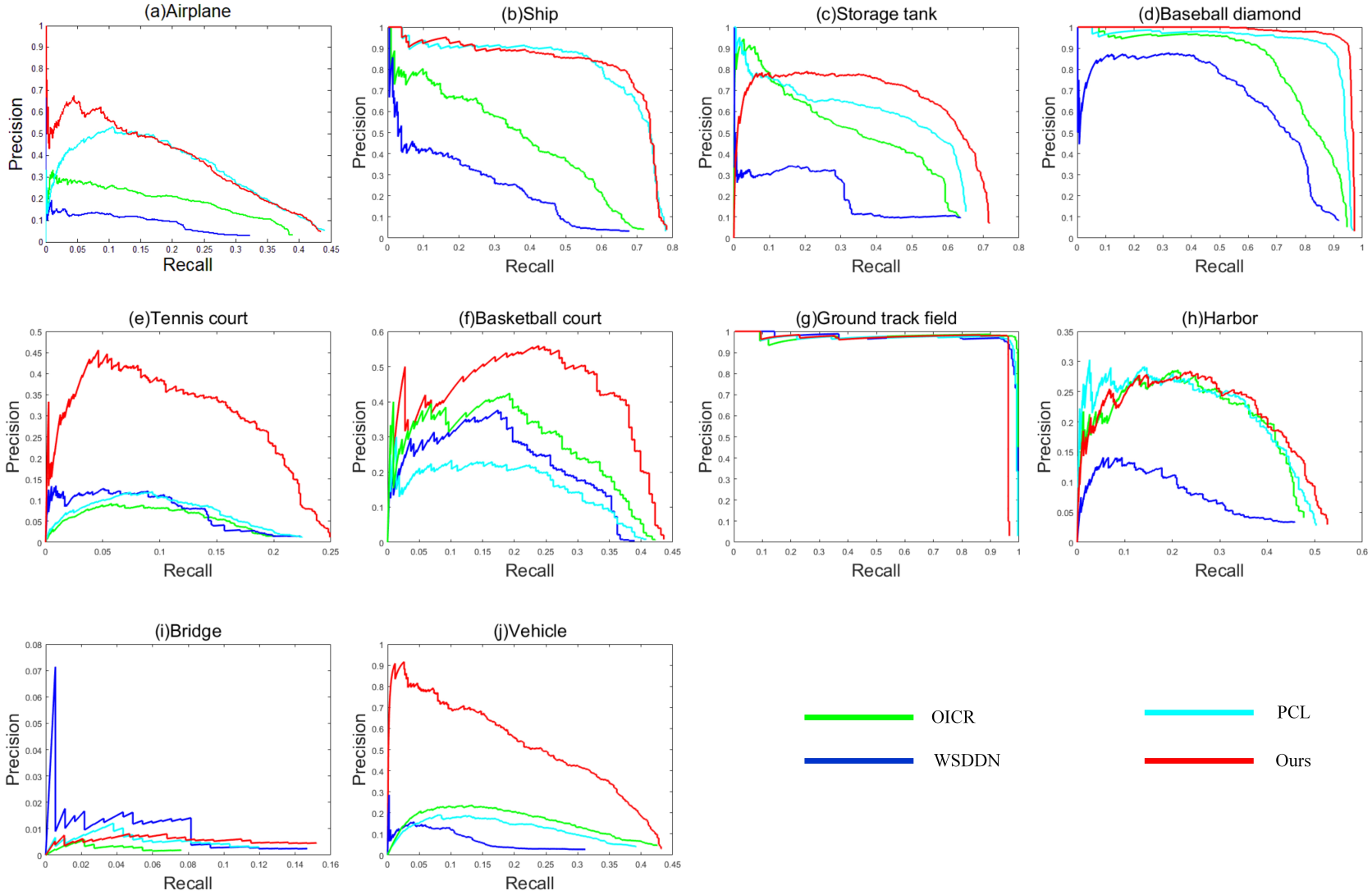

Figure 5. The PRCs of several semi-autonomous learning methods for augmented NWPU-VHR 10 dataset. (**a**) The PRC of airplane. (**b**) The PRC of ship. (**c**) The PRC of storage tank. (**d**) The PRC of baseball diamond. (**e**) The PRC of tennis court. (**f**) The PRC of basketball court. (**g**) The PRC of ground track field. (**h**) The PRC of harbor. (**i**) The PRC of bridge. (**j**) The PRC of vehicle.

Table 5. CorLoc detection results of different methods for LEVIR dataset (Bold indicates the best result).

Methods \ Objects	Airplane	Ship	Oil Port	mCorLoc
CSC	0.00	0.57	13.64	4.74
WSDDN	37.55	0.00	36.36	24.64
OICR	36.80	1.14	65.91	34.62
PCL	49.44	0.00	54.55	34.66
Ours	**72.12**	**24.00**	**75.00**	**57.04**

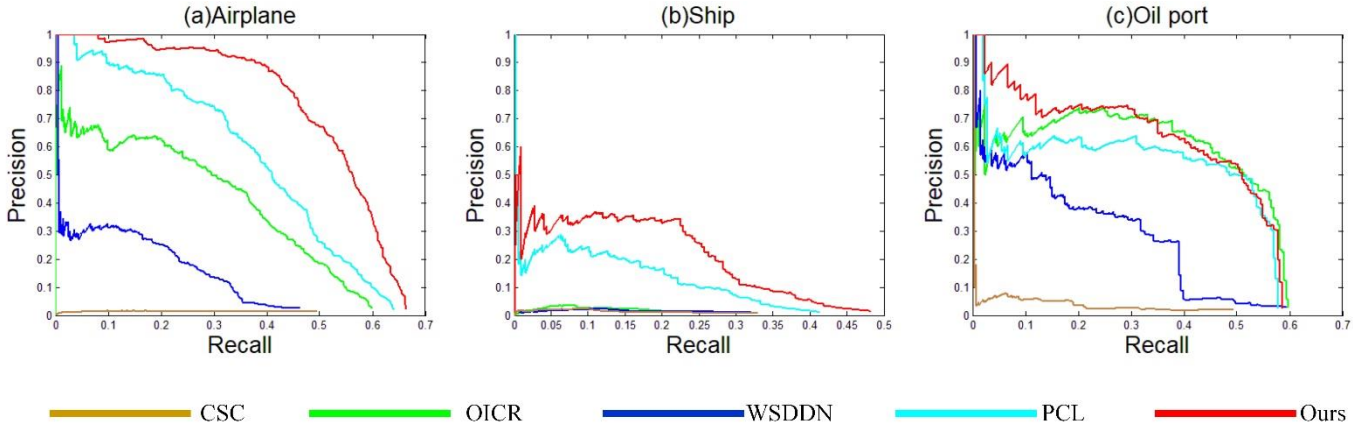

Figure 6. The PRCs of several semi-autonomous learning methods for LEVIR dataset. (**a**) The PRC of airplane. (**b**) The PRC of ship. (**c**) The PRC of oil port.

3.3. Qualitative Results

Some detection results on the augmented NWPU VHR-10 are displayed in Figure 7. It can be observed that for most objects, the proposed method can obtain accurate and tight bounding boxes. Figure 7(b(1),b(5)) show the detection result of the harbor. For this crowded object with large-scale change, the proposed method can still obtain a more accurate positioning frame. However, it can be observed in Figure 7(b(1)) that there are still some shortcomings. The detection ability in the irregular scene of the crowded harbor still needs to be improved. The detection of the storage tank is displayed in Figure 7(a(5),b(3)); the object is small in size and crowded in image. It can be observed that for this small object, the proposed method has a certain rate of missing inspection. This is because the proportion of the storage tank is especially small, which causes some difficulties to the detection. Figure 7(c(1),c(2)) show the detection results of vehicle, whose environment is complex. It can be observed from the figures that there is a certain error detection rate. However, for the ship, ground-track-field and other objects with obvious features or simple scenes, the proposed method can obtain excellent detection performance.

Figure 8 shows the detection results of some methods for LEVIR. The first row denotes the PCL method, and the second row is the method proposed in this paper. It can be observed from the figure that the proposed method is superior to PCL. For the airplane, PCL has some error detection results. For the ship, especially those near the coast, the results obtained by the two methods are not good enough, but in comparison, the error detection rate of the proposed method is less. This is because the coastal scene has certain influence on object detection. For the oil port object, it is obvious that the proposed method has a less missed detection rate. Generally speaking, these show that the proposed method is robust to complex and changeable environments.

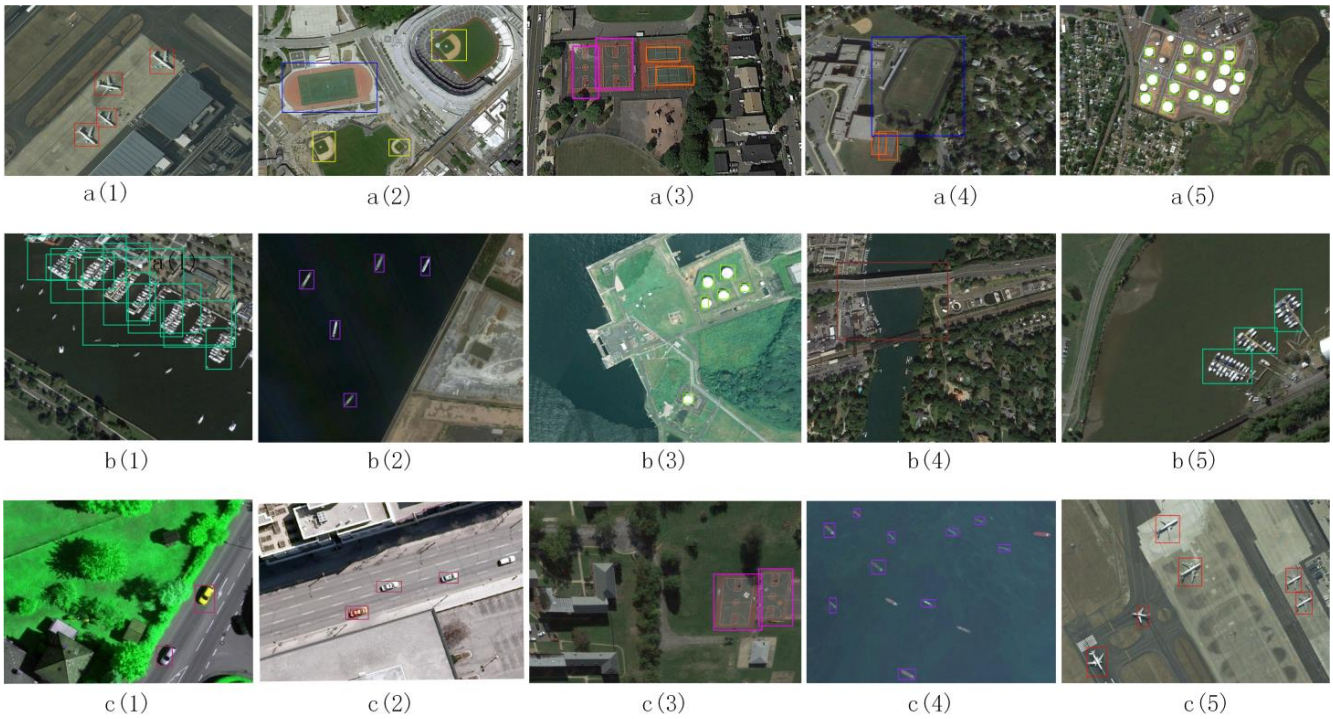

Figure 7. Some detection results on the augmented NWPU VHR-10. (**a**(**1**)) Detection results of airplane. (**a**(**2**)) Detection results of ground track field and baseball diamond. (**a**(**3**)) Detection results of basketball court and tennis court. (**a**(**4**)) Detection results of ground track field and tennis court. (**a**(**5**)) Detection results of storage tank. (**b**(**1**)) Detection results of harbor. (**b**(**2**)) Detection results of ship. (**b**(**3**)) Detection results of storage tank. (**b**(**4**)) Detection results of bridge. (**b**(**5**)) Detection results of harbor. (**c**(**1**)) Detection results of vehicle. (**c**(**2**)) Detection results of vehicle. (**c**(**3**)) Detection results of basketball court. (**c**(**4**)) Detection results of ship. (**c**(**5**)) Detection results of airplane.

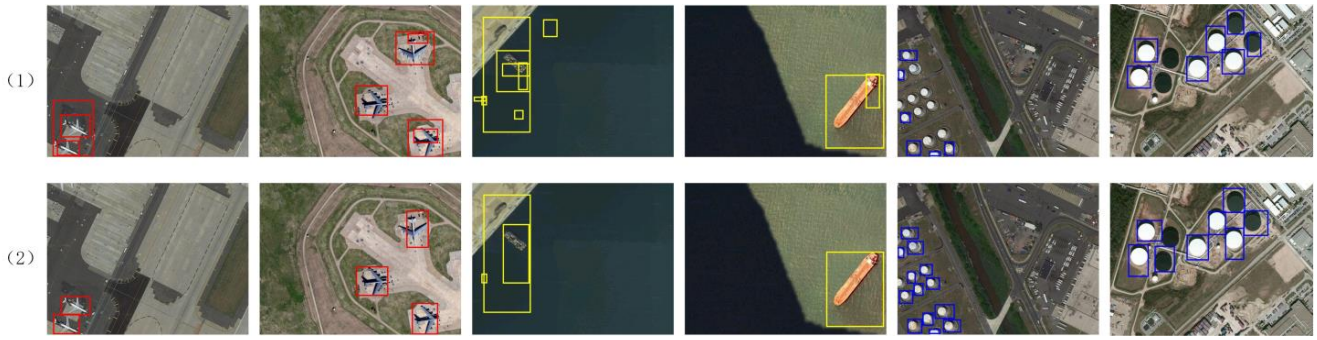

Figure 8. Some detection results on the augmented LEVIR. (**1**) Detection results of PCL method. (**2**) Detection results of our method.

The runtime comparisons between the proposed method and other semi-autonomous learning methods are shown in Table 6, where the runtime of the proposal generation is not considered. It can be observed that with the improvement of the method, the algorithm takes more and more time. The method of CSC is efficient due to the lack of MIL. Although the proposed method is more competitive than other methods, the proposed method takes almost the same testing time as PCL. This is because the proposed approach is an end-to-end network. Although a small number of additional test computations are required compared to WSDDN and OICR, the proposed method obtains much better detection results than other methods.

Table 6. Computation times for different methods in terms of LEVIR dataset.

Methods	CSC	WSDDN	OICR	PCL	Ours
Testing time (second/image)	0.55	1.64	1.69	2.12	2.14

4. Discussion

To evaluate the effectiveness of the proposed approach, ablation experiments are constructed to analyze the effects of the key components of ARMS, such as the modified residual network, ARMS and regression branches.

4.1. MRN

In order to verify the effectiveness of the improved residual network, experiments were carried out on different backbone networks, including VGG-16, RESNET-18 and the MRN, as shown in Figure 9. In this experiment, except for the different backbone networks, the other settings are the same. It can be observed that the performance of mAP decreased from 34.78% to 33.90% by directly replacing VGG-16 with resnet-18, especially for the objects with high similarity and indistinguishability, such as the tennis court and basketball court. For a vehicle with a small size and complex scene, the AP value drops from 23.23% to 7.02%. Therefore, it may be counterproductive to simply and rudely replace VGG-16 with resnet-18. On the contrary, it can be observed that the mAP value steadily increased from 34.78% to 38.72% after replacing VGG-16 with the modified residual network. For the small-sized object such as the harbor and vehicle, the AP value can be kept stable or even increased, which further confirms that the proposed method is effective and stable in the multi-task object detection under semi-autonomous learning.

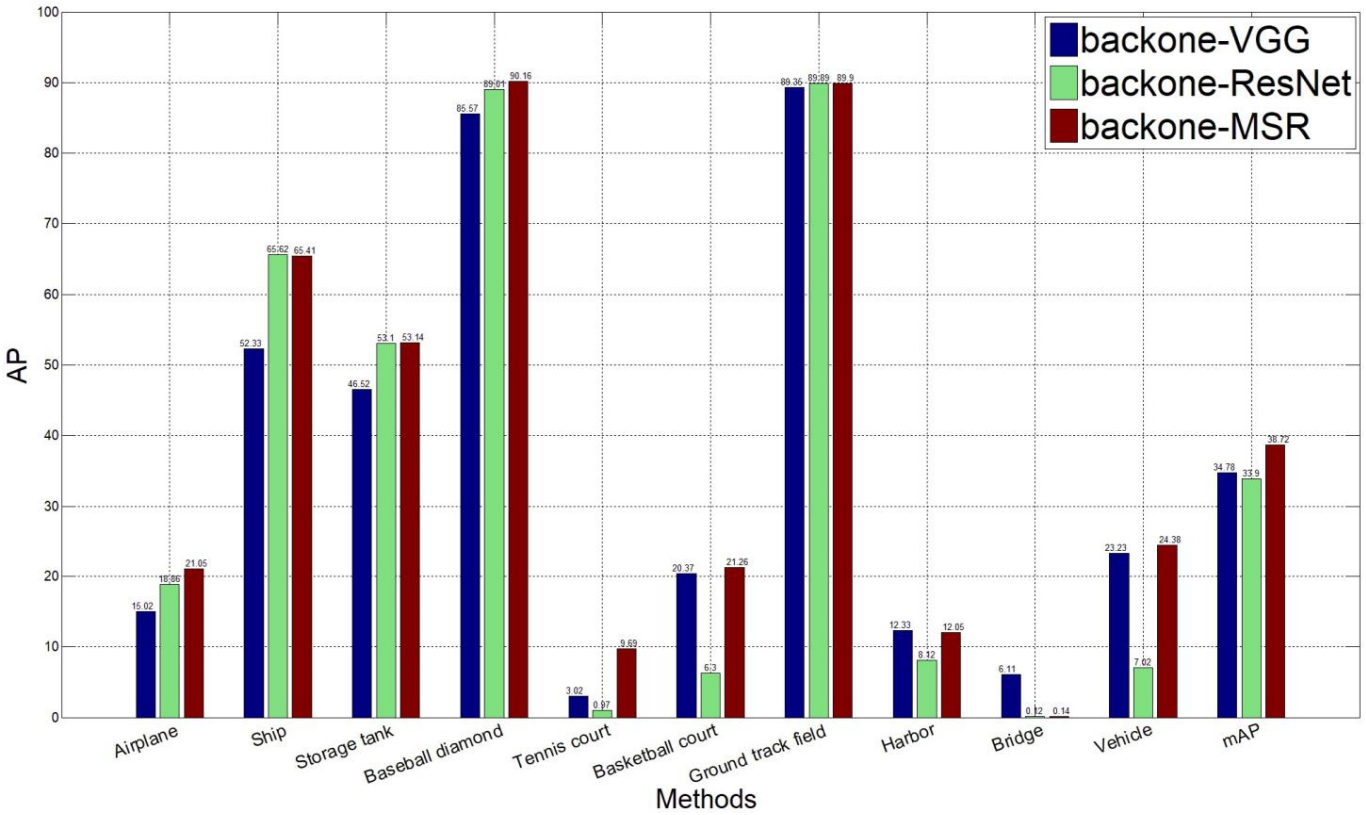

Figure 9. The experiment for different backbone networks.

4.2. ARMS

In this section, the proposed ARMS is compared to a training strategy similar to OICR, which only selected the candidates with the highest credibility to refine the next classifier. In this experiment, the way of instance refinement is different and other settings are the same. As shown in Figure 10, it can be observed that ARMS greatly improves the detection performance of the network, and the value of mAP increases from 31.9% to 38.72%. Interestingly, the proposed ARMS is more effective for small objects with particularly complex backgrounds and object classes with multiple instances in the image, such as the vehicle and airplane. Moreover, it can clearly distinguish the similar objects, such as tennis courts and basketball courts. Therefore, the proposed ARMS can mine more positive examples and classify more accurately. The main reason is that in the refinement process, OICR only chooses the suggestions with the highest scores as the supervision information in the next stage, which makes it easy to confuse the background with the positive instance, as well as with the different types of positive instances, while ignoring other similar goals in the scene. However, ARMS makes use of the suggestions with the highest scores and the aggregation areas closely related to them, which can fully mine the correct examples and classify them accurately.

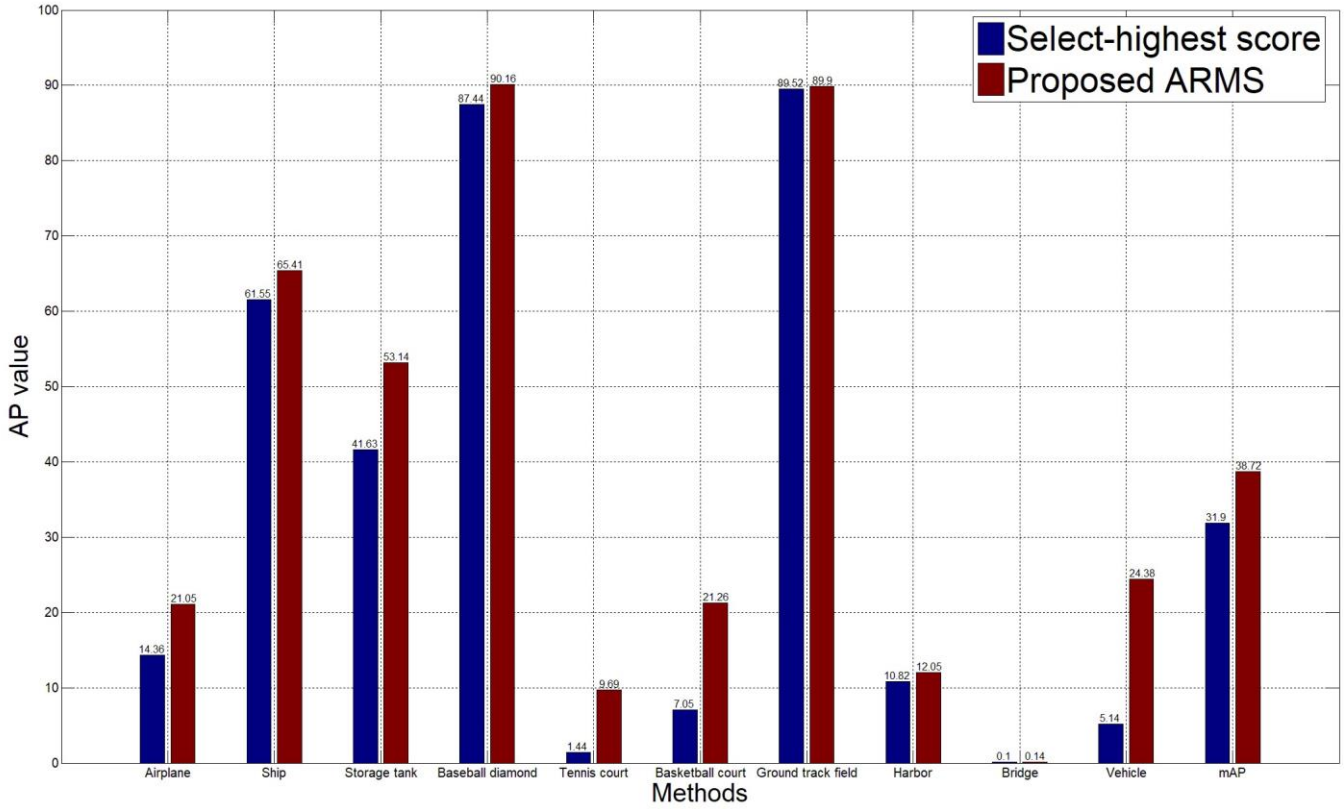

Figure 10. The ablation experiment for ARMS.

4.3. Regression Branches

Figure 11 shows the contribution of the regression branch to the whole framework. In this experiment, the way of regression is different and other settings are the same. It can be observed from the table that the performance is obviously improved after the regression branch is added, and the mAP is improved from 34.7% to 38.72%. Especially for small objects with multiple instances and close distance, such as oil tanks and airplanes, the positioning branch can accurately find its exact position. This is due to the combination of the regression branch and classification branch, which can effectively solve the over-fitting

problem caused by the weak detector when distinguishing parts, and at the same time, it can locate the object more accurately.

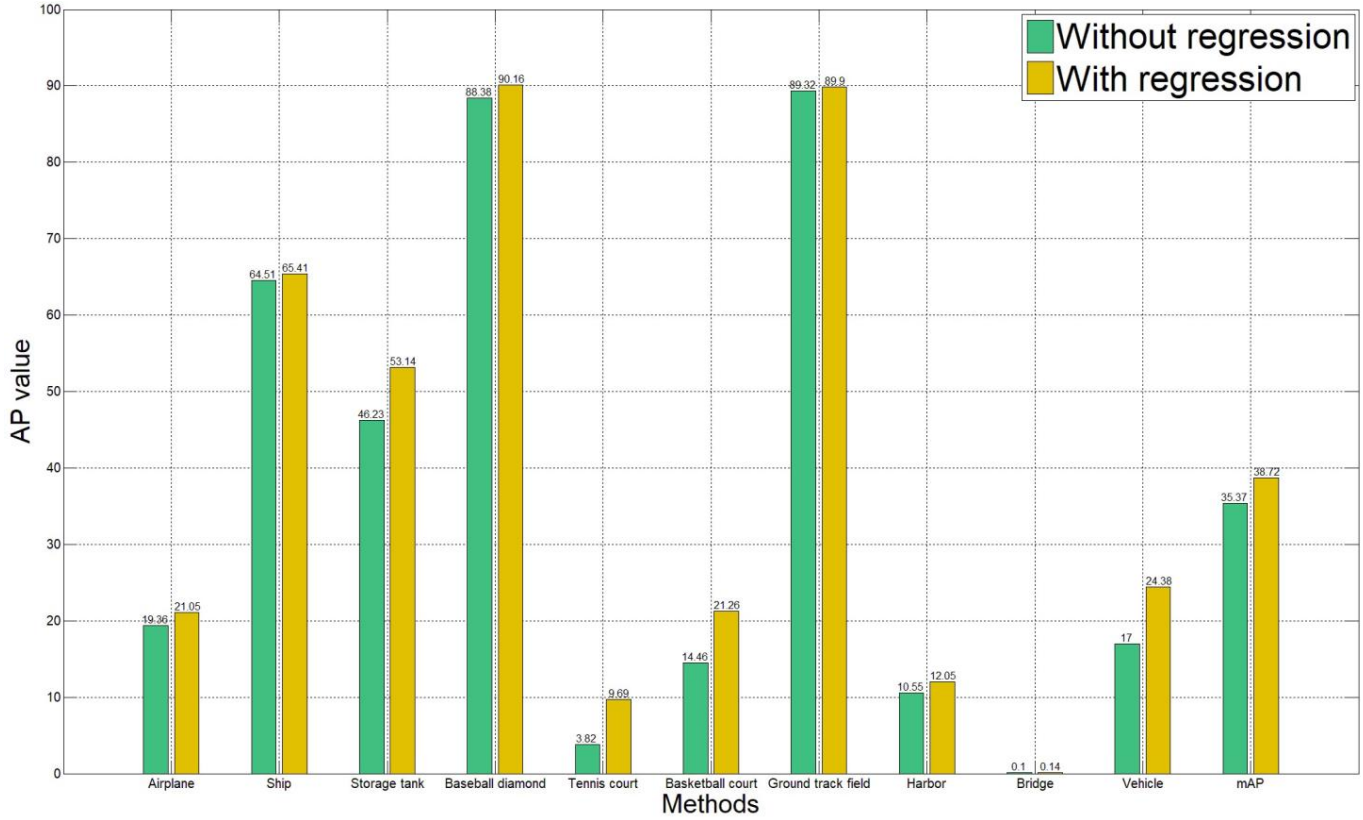

Figure 11. The contribution of regression branch to the whole framework.

5. Conclusions

A novel end-to-end aggregate-guided semi-autonomous learning residual network is presented in this paper to handle the object detection problem in remote sensing images. Specially, ARMS is designed to select high-quality regions by merging similar regions in the cluster and deleting redundant regions in the cluster. Moreover, a progressive residual network is applied to the backbone network to make the detector more sensitive to small objects. Meanwhile, a regression locating branch is further developed to refine the location of the object, which can be optimized jointly with regional classification. Extensive experiments demonstrate that the proposed method outperforms state-of-the-art object detection of the remote sensing image. For the NWPU dataset with a large number of object categories, the semi-autonomous learning methods WSDDN, OICR and PCL achieved 34.49%, 40.68% and 41.31% AP values, respectively, while the semi-autonomous learning method proposed in this paper achieved a 53.01% AP value. This indicates that the proposed method is obviously superior to other methods. For the LEVIR dataset with a relatively balanced number of objects, the semi-autonomous learning methods WSDDN, OICR and PCL have achieved 10.88%, 18.80% and 29.93% AP values, respectively, while the semi-autonomous learning method proposed in this paper has achieved a 35.68% AP value. It can be proved that the proposed method outperforms the state-of-the-art methods.

Author Contributions: B.C. and Z.L. conceived of the study; B.C. wrote the code, performed the analysis and wrote the article; Z.L. analyzed the results; H.L. collected the dataset; Z.D. and T.Q. revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the National Natural Science Foundation of China under Grant No. 61675036, 13th Five-year Plan Equipment Pre-research Fund under Grant

No. 6140415020312, and Chinese Academy of Sciences Key Laboratory of Beam Control Fund under Grant No. 2017LBC006.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Cheng, B.; Li, Z.; Xu, B.; Yao, X.; Ding, Z.; Qin, T. Structured Object-Level Relational Reasoning CNN-Based Target Detection Algorithm in a Remote Sensing Image. *Remote Sens.* **2021**, *13*, 281. [CrossRef]
2. 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]
3. Sun, X.; Wang, P.; Wang, C.; Liu, Y.; Fu, K. PBNet: Part-based convolutional neural network for complex composite object detection in remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *173*, 50–65. [CrossRef]
4. Lu, X.; Zhong, Y.; Zheng, Z.; Liu, Y.; Zhao, J.; Ma, A.; Yang, J. Multi-Scale and Multi-Task Deep Learning Framework for Automatic Road Extraction. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 9362–9377. [CrossRef]
5. Wei, Y.; Zhang, K.; Ji, S. Simultaneous Road Surface and Centerline Extraction from Large-Scale Remote Sensing Images Using CNN-Based Segmentation and Tracing. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 8919–8931. [CrossRef]
6. Li, X.; Wang, Y.; Zhang, L.; Liu, S.; Mei, J.; Li, Y. Topology-Enhanced Urban Road Extraction via a Geographic Feature-Enhanced Network. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 8819–8830. [CrossRef]
7. Tu, Y.; Lang, W.; Yu, L.; Li, Y.; Jiang, J.; Qin, Y.; Wu, J.; Chen, T.; Xu, B. Improved Mapping Results of 10 m Resolution Land Cover Classification in Guangdong, China Using Multisource Remote Sensing Data with Google Earth Engine. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 5384–5397. [CrossRef]
8. Mu, L.; Wang, L.; Wang, Y.; Chen, X.; Han, W. Urban Land Use and Land Cover Change Prediction via Self-Adaptive Cellular Based Deep Learning with Multisourced Data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 5233–5247. [CrossRef]
9. Georganos, S.; Grippa, T.; Vanhuysse, S.; Lennert, M.; Shimoni, M.; Wolff, E. Very High Resolution Object-Based Land Use–Land Cover Urban Classification Using Extreme Gradient Boosting. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 607–611. [CrossRef]
10. Zhou, P.; Han, J.; Cheng, G.; Zhang, B. Learning Compact and Discriminative Stacked Autoencoder for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 4823–4833. [CrossRef]
11. Yao, X.; Han, J.; Cheng, G.; Qian, X.; Guo, L. Semantic Annotation of High-Resolution Satellite Images via Weakly Supervised Learning. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 3660–3671. [CrossRef]
12. Han, J.; Cheng, G.; Li, Z.; Zhang, D. A Unified Metric Learning-Based Framework for Co-Saliency Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *28*, 2473–2483. [CrossRef]
13. Tuermer, S.; Kurz, F.; Reinartz, P.; Stilla, U. Airborne Vehicle Detection in Dense Urban Areas Using HoG Features and Disparity Maps. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2013**, *6*, 2327–2337. [CrossRef]
14. Zhong, P.; Wang, R. A Multiple Conditional Random Fields Ensemble Model for Urban Area Detection in Remote Sensing Optical Images. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 3978–3988. [CrossRef]
15. Dewi, C.; Chen, R.-C.; Liu, Y.-T.; Jiang, X.; Hartomo, K.D. Yolo V4 for Advanced Traffic Sign Recognition with Synthetic Training Data Generated by Various GAN. *IEEE Access* **2021**, *9*, 97228–97242. [CrossRef]
16. Zhang, X.; Feng, J.; Xiong, H.; Tian, Q. Zigzag Learning for Weakly Supervised Object Detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4262–4270.
17. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
18. Li, Y.; Chen, W.; Zhang, Y.; Tao, C.; Xiao, R.; Tan, Y. Accurate cloud detection in high-resolution remote sensing imagery by weakly supervised deep learning. *Remote Sens. Environ.* **2020**, *250*, 112045. [CrossRef]
19. Li, Y.; Zhang, Y.; Huang, X.; Yuille, A.L. Deep networks under scene-level supervision for multi-class geospatial object detection from remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2018**, *146*, 182–196. [CrossRef]
20. Jie, Z.; Wei, Y.; Jin, X.; Feng, J.; Liu, W. Deep Self-Taught Learning for Weakly Supervised Object Localization. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4294–4302.
21. Li, Y.; Shi, T.; Zhang, Y.; Chen, W.; Wang, Z.; Li, H. Learning deep semantic segmentation network under multiple weakly-supervised constraints for cross-domain remote sensing image semantic segmentation. *ISPRS J. Photogramm. Remote Sens.* **2021**, *175*, 20–33. [CrossRef]
22. Bilen, H.; Vedaldi, A. Weakly Supervised Deep Detection Networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
23. 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. [CrossRef]

24. Cinbis, R.G.; Verbeek, J.; Schmid, C. Weakly Supervised Object Localization with Multi-Fold Multiple Instance Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 189–203. [CrossRef]
25. Li, D.; Huang, J.-B.; Li, Y.; Wang, S.; Yang, M.-H. Weakly Supervised Object Localization with Progressive Domain Adaptation. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 3512–3520.
26. Song, H.O.; Lee, Y.J.; Jegelka, S.; Darrell, T. Weakly-supervised Discovery of Visual Pattern Configurations. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014.
27. Wang, C.; Ren, W.; Huang, K.; Tan, T. Weakly Supervised Object Localization with Latent Category Learning. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014.
28. 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.
29. Tang, P.; Wang, X.; Bai, S.; Shen, W.; Bai, X.; Liu, W.; Yuille, A. PCL: Proposal Cluster Learning for Weakly Supervised Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 176–191. [CrossRef]
30. 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]
31. 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.
32. Uijlings, J.R.R.; van de Sande, K.E.A.; Gevers, T.; Smeulders, A.W.M. Selective Search for Object Recognition. *Int. J. Comput. Vis.* **2013**, *104*, 154–171. [CrossRef]
33. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
34. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.
35. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
36. Shen, Y.; Ji, R.; Wang, Y.; Chen, Z.; Zheng, F.; Huang, F.; Wu, Y. Enabling Deep Residual Networks for Weakly Supervised Object Detection. In Proceedings of the Computer Vision—ECCV 2020, Glasgow, UK, 23–28 August 2020; pp. 118–136.
37. Song, G.; Song, K.; Yan, Y. EDRNet: Encoder–Decoder Residual Network for Salient Object Detection of Strip Steel Surface Defects. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 9709–9719. [CrossRef]
38. Chen, S.; Tan, X.; Wang, B.; Lu, H.; Hu, X.; Fu, Y. Reverse Attention Based Residual Network for Salient Object Detection. *IEEE Trans. Image Process.* **2020**, *29*, 3763–3776. [CrossRef]
39. Roy, S.K.; Manna, S.; Song, T.; Bruzzone, L. Attention-Based Adaptive Spectral–Spatial Kernel ResNet for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 7831–7843. [CrossRef]
40. Yu, X.; Kang, C.; Guttery, D.S.; Kadry, S.; Chen, Y.; Zhang, Y.D. ResNet-SCDA-50 for Breast Abnormality Classification. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2021**, *18*, 94–102. [CrossRef]
41. Zhu, H.; Sun, M.; Fu, H.; Du, N.; Zhang, J. Training a Seismogram Discriminator Based on ResNet. *IEEE Trans. Geosci. Remote. Sens.* **2020**, *59*, 7076–7085. [CrossRef]
42. Yang, K.; Li, D.-S.; Dou, Y. Towards Precise End-to-end Weakly Supervised Object Detection Network. In Proceedings of the International Conference on Computer Vision (ICCV2019), Seoul, Korea, 27 October–2 November 2019.
43. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
44. 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]
45. Zou, Z.; Shi, Z. Random Access Memories: A New Paradigm for Target Detection in High Resolution Aerial Remote Sensing Images. *IEEE Trans. Image Process.* **2018**, *27*, 1100–1111. [CrossRef]
46. Andermatt, P.; Timofte, R. A Weakly Supervised Convolutional Network for Change Segmentation and Classification. In Proceedings of the Asian Conference on Computer Vision, Kyoto, Japan, 30 November–4 December 2020.
47. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR2020), Seattle, WA, USA, 16–18 June 2020.
48. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. Scaled-YOLOv4: Scaling Cross Stage Partial Network. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR2020), Seattle, WA, USA, 16–18 June 2020.