A Similarity Constraint Divergent Activation Method for Weakly Supervised Object Detection in Remote Sensing Images

Mengmeng Zhu, Shouhong Wan(✉), Peiquan Jin, and Jian Xu

University of Science and Technology of China, No.96, JinZhai Road, Hefei 230026, People's Republic of China
{zhumeng,xxxujian}@mail.ustc.edu.cn, {wansh,jpq}@ustc.edu.cn

Abstract. With the development of remote sensing technology and object detection technology, many fully-supervised convolutional neural networks (CNN) methods based on object labeling information such as bounding box have achieved good results in remote sensing image object detection. However, due to the wide detection range of remote sensing images, diversity of objects, and the complexity of background, it is very difficult to manually label large-scale remote sensing images. Therefore, in recent years, more and more attention has been paid to the weakly supervision method using only image-level labels in object detection. Class activation mapping (CAM) method based on weakly supervision works well for object detection in natural scene images, but it has the problem when it is used in remote sensing images: a large number of small objects are lost. In this paper, we propose an object detection method for remote sensing image based on similarity constraint divergent activation (SCDA). The divergent activation (DA) module in SCDA improves the response intensity of the low response regions in the shallow layer feature map. According to the similarity between the objects, the similarity constraint module (SCM) is used to further improve the feature distribution and suppress background noise. By fusing DA and SCM, the missed rate of small objects can be reduced. Comprehensive experiments and comparisons with state-of-the-art methods on WSADD and DIOR data sets demonstrate the superiority of our proposed method.

Keywords: Object detection · Weakly supervised · Remote sensing image · Divergent activation · Similarity constraint

1 Introduction

With the rapid development of remote sensing technology, more and more high-quality remote sensing images with high spatial resolution and rich objects are emerging, which provide sufficient data and analysis conditions for the research of remote sensing image in various fields. And remote sensing image object detection has practical application scenarios and values in both civil and military fields.

Y. Peng et al. (Eds.): ICIG 2021, LNCS 12888, pp. 418–430, 2021.
https://doi.org/10.1007/978-3-030-87355-4_35

Here are many object detection algorithms proposed for remote sensing images, which are mainly divided into machine learning algorithm based on hand-crafted features and deep learning algorithm based on convolution neural network (CNN). Although hand-crafted features, such as Histogram of Oriented Gradient (HOG) [3], Scale-Invariant Feature Transform (SIFT) [5] and Bag of Words (BOW) [7] have achieved some results, their object detection effect is not good enough because they cannot express the high-level semantic information of the objects. CNN can not only describe the low-level features of the objects, but also express the high-level semantic information of the objects, so it has achieved good results in the field of natural scene target detection. Inspired by the deep learning technology in natural scenes, Cao et al. [1] And Yao et al. [13] introduced R-CNN and Faster-RCNN into remote sensing image object detection.

However, most of the existing deep learning methods that perform well require training data labeling the position information of objects. And if the location information of objects is labeled manually, the construction of large-scale image dataset is a huge workload and greatly increases the human cost of the dataset. As a result, Weakly Supervised Object Detection (WSOD) has received increasing attention. WSOD refers to the use of image-level labels in a given image to learn the location of objects, so it does not require expensive object bounding box labels for training.

It is a challenging task for object detection only using the image-level labels to learn a deep model. And some pioneer works have been proposed to achieve WSOD in natural scene. For example, a weakly supervised learning algorithm based on Class Activation Map (CAM) [16] is proposed to realize target detection. Next, adversarial erasing methods [2, 6, 10, 15] pursue learning full object extent by erasing the discriminative regions. The divergent activation method [9, 12, 14] designs multiple parallel branches or introduces attention modules to drive the network to locate the complete object region.

However, when these methods are applied to remote sensing images, there will be some problems with a large number of small objects, such as inaccurate positioning and object loss. Active Region Corrected (ARC) method [11] can effectively solve the problem of inaccurate positioning of objects by combining shallow feature positioning maps. And structure-preserving activation (SPA) method [8] extracts the structure preserving ability of features, so as to achieve accurate object location. But these methods cannot solve the problem of small object loss. This is because these missed objects may exist in the shallow layer of the network, but their responsiveness is insufficient, which results in the loss of features when they are transmitted backwards, and results in the loss of objects when they are last located.

Therefore, this paper introduces the divergent activation (DA) module and designed similarity constraint module (SCM) on the basis of ARC, and proposes an object detection method for remote sensing image based on similarity constraint divergent activation (SCDA). The idea is to enhance the response intensity of the low response region and the non-response region in the shallow layer network feature map by using DA module [12], and to enhance the intensity of similar features in the high response region and suppress background noise through SCM. It uses SCDA to improve the feature distribution in the shallow layer feature map and to solve the problem of large number of small object loss.

Our contributions are summarized as: 1) We propose a simple and effective SCDA method to improve the feature distribution in the shallow layer feature map to focus on more small object regions. 2) We experiment our method on two datasets, and compared with state-of-the-art methods based on CAM, precision and recall method have significant improvement.

2 Related Work

Weakly supervised object detection (WSOD) aims to detect the object with only image-level labels training data. Zhou et al. [16] proposed a weakly supervised learning algorithm based on Class Activation Map (CAM) to achieve object detection. They add a global average pooling layer and a full connection layer to the convolution neural network, and combine the weights of the final convolution layer features and the full connection layer to generate a CAM for the localization purpose. Wei et al. [10] proposed a method of adversarial erasing to locate a complete object. On this basis, Zhang et al. [15] propose an end-to-end network based on Adversarial Complementary Learning (ACoL), which uses two parallel classifiers with dynamic erasing and adversarial learning to discover complementary object regions more effectively. Xue et al. [12] proposed the Divergent Activation Network (DANet), which enlarges the difference of each dimension by increasing the dimension of the deep feature map, thereby increasing the response intensity of the unresponsive region and the weakly response region, and ultimately locating the complete object region. And structure-preserving activation (SPA) method [8] extracts the structure preserving ability of features by proposing a self-correlation map generating (SCG) module, so as to achieve accurate object location.

However, when these methods are applied to remote sensing images, there will be some problems with a large number of small objects, such as inaccurate positioning and object loss. We proposed an Active Region Corrected (ARC) [11] method, which can effectively solve the problem of inaccurate positioning of objects by combining shallow feature positioning maps, but it cannot solve the problem of object loss. This is because these missed objects may exist in the shallow layer of the network, but their responsiveness is insufficient, which results in the loss of features when they are transmitted backwards, and results in the loss of objects when they are last located.

3 Proposed Method

3.1 Framework

Because there are a large number of small objects in remote sensing images, it is not suitable to use a deeper network. We use ResNet34 as the basic network, and use a similarity constraint divergent activation (SCDA) framework on the basis of the ARC [11] network. As shown in Fig. 1, since the third-layer features of ResNet34 contain certain semantic information and also do not contain too much noise, the DA module and the similarity constraint module (SCM) are embedded in the third layer of ResNet34. First, through the DA module, the dimension of the network extracted shallow layer features is increased by 1×1 convolution. By activating the shallow layer feature map, the

DA module increases the response intensity of the low-response region, and activates the non-response region to increase the response intensity of the region, thereby increasing the number of response regions. Then, the output of the DA module is merged as the input of SCM. The SCM builds a similarity matrix to improve the response intensity of similar regions, suppress the response intensity of background noise regions, and improve feature distribution.

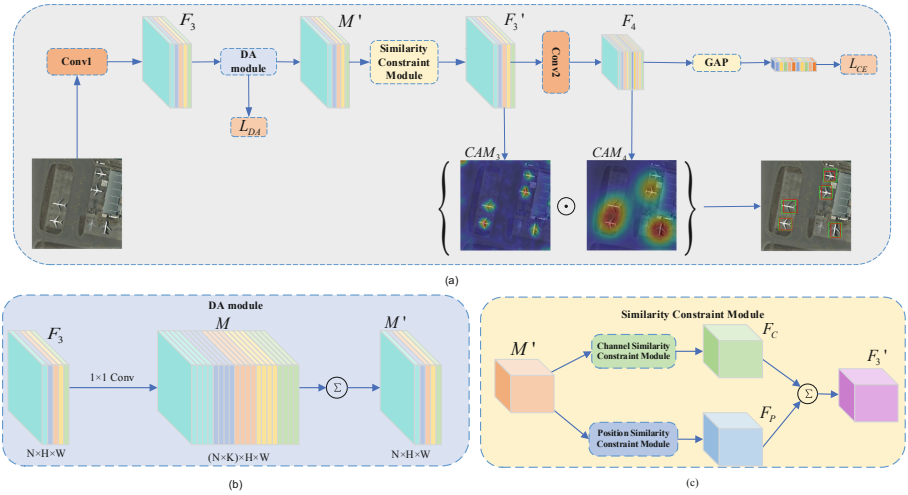

Fig. 1. The framework of the proposed SCDA approach. (a) We use ResNet34 as the basic network and use SCDA approach improve the feature distribution in the shallow feature map. (b) Illustration of the DA module. (c) Similarity constraint module completes feature fusion by summing the output of the channel similarity constraint module and position similarity constraint module.

In addition, we use divergent activation loss (L_{DA}) and classification loss (L_{CE}) to cooperate to drive the model to detect more objects in the training phase. The total loss of SCDA training is defined as:

$$L = L_{CE} + \lambda L_{DA} \tag{1}$$

where L_{CE} uses cross-entropy loss in two-class classification, and uses marge loss in multi-label classification. L_{DA} is the divergent activation loss in DA module, and we will introduce it in detail in next section. λ is a regularization factor to balance the two items.

3.2 Divergent Activation Module

The DA module increases the size of the shallow layer feature map to K times the original size, and expands the cosine distance between the corresponding K feature maps in each group to activate the regions and form more activation regions.

We denote the feature map output by the third layer network as $F_3 \in R^{N \times H \times W}$. As shown in Fig. 1, we use 1×1 convolution to increase the dimension of the feature map

F_3 from N × H × W to (N × K) × H × W. We denote the feature map obtained by 1 × 1 convolution as $M \in R^{(N\times K)\times H\times W}$. For clarity of description, we use M_n^k to represent each channel of the feature map M, where $k \in \{1, 2, \ldots, K\}$, $n \in \{1, 2, \ldots, N\}$.

In the initial stage, for each group of K feature maps M_n^k, they are equal, such as $M_n^1 = M_n^2 = \ldots = M_n^K$, where $n \in \{1, 2, \ldots, N\}$. In order to find more activation regions, we use the cosine distance between each feature map as a function loss constraint, so that during the network training, the activation regions in each feature map are different, which increases the response regions to a certain extent, and at the same time makes the response of the original weaker region stronger. We define the function loss constraint as:

$$L_{DA} = \frac{2}{N \times K \times (K-1)} \sum\nolimits_{1\le n\le N} \sum\nolimits_{1\le k_1\le k_2\le K} S\left(M_n^{k_1}, M_n^{k_2}\right) \tag{2}$$

where $S\left(M_n^{k_1}, M_n^{k_2}\right)$ is the cosine distance between the feature map $M_n^{k_1}$ and $M_n^{k_2}$, which is defined as:

$$S\left(M_n^{k_1}, M_n^{k_2}\right) = \frac{M_n^{k_1} \cdot M_n^{k_2}}{\|M_n^{k_1}\| \|M_n^{k_2}\|} \tag{3}$$

When the cosine distance between the feature maps $M_n^{k_1}$ and $M_n^{k_2}$ decreases, it means that the difference between the feature maps $M_n^{k_1}$ and $M_n^{k_2}$ becomes larger. By minimizing the loss function L_{DA}, the difference between the response regions of the K feature maps are enlarged, so as to achieve the purpose of increasing the intensity of the response region and increasing the number of response regions.

Finally, we sum each feature map M_n^k in the n-th group to obtain the fused feature map M_n' by sum module. The sum module is defined as:

$$M_n' = \frac{1}{K} \sum\nolimits_{k=1}^{K} M_n^k \tag{4}$$

where $n \in \{1, 2, \ldots, N\}$, M_n' corresponds to the n-th channel of the feature map F_3 output by the third layer network. The N feature maps M_n' are spliced to obtain the final output feature map $M' \in R^{N\times H\times W}$ of DA module.

3.3 Similarity Constraint Module

Through the DA module, more response regions are activated in the feature map output by the convolution neural network. However, when DA module is divergent activating, it mainly enlarges the cosine distance between the feature maps, without considering the feature of the object regions. Therefore, we use the similarity constraint module (SCM) to further modify the output of the DA module, suppress the response intensity of non-object regions, and improve the response intensity of object regions.

As shown in Fig. 1(c), the SCM is composed of the channel similarity constraint module (CSCM) and the position similarity constraint module (PSCM). The CSCM constructs a channel similarity matrix based on the feature map output by the DA module,

and measures the feature similarity of regions on each channel by using the dependency relationship between channel mappings. The PSCM measures the similarity of features in different locations by building a location similarity matrix. Through two different similarity constraints, the background noise in image is further removed and the response intensity of the object regions is improved.

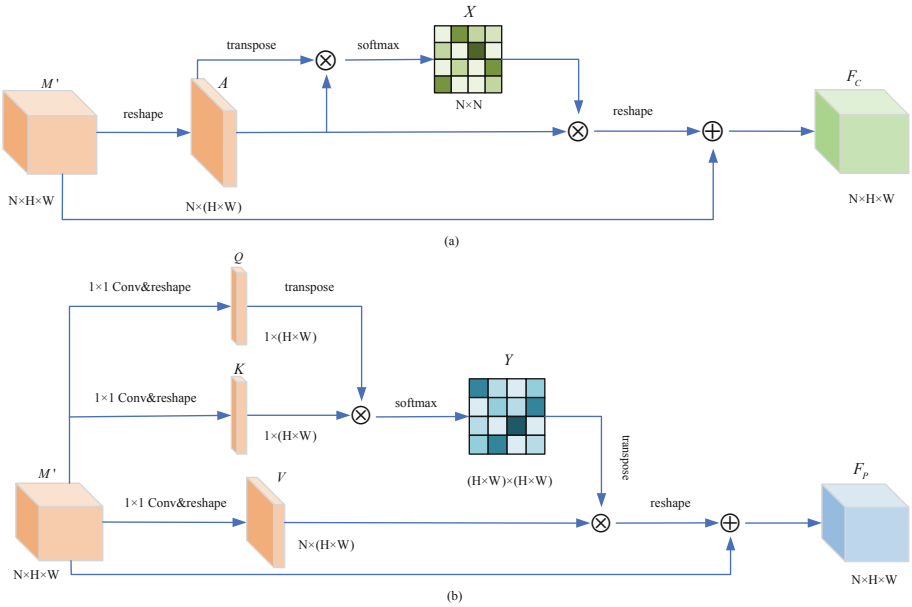

Fig. 2. Detail implementation of two different SCM. (a) CSCM aims to enhance features through the similarity between channels. (b) PSCM aims to enhance features through the similarity between positions.

Channel Similarity Constraint Module. We denote the feature map output by the DA module as $M' \in R^{N\times H\times W}$. The specific process of the channel similarity constraint module (CSCM) is shown in Fig. 2(a). First, we convert the dimension of feature map M to N × (H × W), and denote it as matrix A. Then we use matrix A to calculate the similarity between each channel, and construct a channel similarity matrix X to measure the degree of correlation between different channels. The calculation formula of the channel similarity matrix X is as:

$$X = softmax\left(AA^T\right) \tag{5}$$

where $X \in R^{N\times N}$. Through the channel similarity matrix X, we use the similarity between the feature map channels to update the feature map matrix A. Then multiply the result obtained by the scale parameter α, and convert its dimension to N × H × W. Finally, the result is added to the feature map M' to obtain the channel feature map F_C,

which is output by the CSCM. We define the calculation formula of the channel feature map F_C as:

$$F_C = reshape(\alpha XA) + M' \tag{6}$$

where $F_C \in R^{N \times H \times W}$.

Position Similarity Constraint Module. The specific process of the position similarity constraint module (PSCM) is shown in Fig. 2(b). First, we use three different 1×1 convolutions to perform a convolution operation on the feature map M', and reshape the dimension of the results to obtain three different feature matrices $Q, K \in R^{1 \times (H \times W)}$ and $V \in R^{N \times (H \times W)}$. Then, we use matrix Q and K to calculate the similarity between different positions, and construct a position similarity matrix Y to measure the degree of association of features at different positions. The calculation formula of the position similarity matrix Y is defined as:

$$Y = softmax\left(Q^T K\right) \tag{7}$$

where $Y \in R^{(H \times W) \times (H \times W)}$. Through the position similarity matrix Y, we update the feature map matrix V by using the correlation degree of the features at different positions of the feature map. Then multiply the result obtained by the scale parameter number β, and convert its dimension to N × H × W. Finally, the result is added to the feature map M' to obtain the position feature map F_P, which is output by the PSCM. We define the calculation formula of the position feature map F_P as:

$$F_P = reshape\left(\beta VY^T\right) + M' \tag{8}$$

where $F_P \in R^{N \times H \times W}$.

4 Experiments

4.1 Data Sets and Experimental Settings

Datasets. In order to evaluate the effectiveness of the weakly supervised object detection algorithm based on similarity constraint divergent activation (SCDA), we conducted verification experiments on the data sets WSADD and DIOR respectively. WSADD (Weakly Supervised Airplane Detection Dataset, WSADD) is an object detection dataset independently constructed by our laboratory. The data set has 700 remote sensing images in total, including 400 images containing the aircraft as the positive samples and the other 300 images, which are the background images mainly composed of the airstrip and apron, as the negative samples. In the training phase, we use 600 remote sensing images of 300 aircraft and 300 background images as the training set. In the test phase, 100 aircraft remote sensing images are used as the test set, which contains 308 aircraft. DIOR dataset [4] is a public remote sensing image target detection dataset, which contains 20 categories and 23463 remote sensing images. In the experiment, because there are too few dam

images in the DIOR dataset, we removed the dam category. Therefore, the Dior data set we used has only 19 categories and 23287 remote sensing images, including 11725 images in the training set and 11562 images in the test set. Since in the DIOR dataset, many large objects do not have the problem of the objects with a small proportion of the image or a large number. So, for large objects, such as expressway service area, we directly use deep positioning maps to locate these objects. While for small objects with problems, we use the proposed methods to locate such objects as airplane, baseball field, basketball court, chimney, ship, tennis court, vehicle, and so on.

Metrics. For WSADD dataset, there are only two categories, so we use precision rate and recall rate to evaluate the performance of the model. They are defined as:

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

where TP, FP, and FN represent the number of true positives, false positives, and false negatives respectively in the detection results where IoU > 0.5 is setting to evaluate the results as positive detection. If the recall rate is higher, it means that the correct object accounts for a larger proportion of all objects. And if the precision rate is higher, it means that the correct test results account for a larger proportion of all test results. For DIOR dataset, we choose mean average precision (mAP) as evaluation criteria where IoU > 0.3 is setting to evaluate the results as positive detection. The mAP is defined as:

$$AP = \sum_{k=1}^{N} \max_{\tilde{k} \geq k} P\left(\tilde{k}\right) \times (r(k) - r(k-1)) \tag{11}$$

$$mAP = \frac{1}{C} \sum_{i=1}^{C} AP_i \tag{12}$$

where N is the number of detection results, $P\left(\tilde{k}\right)$ is the precision rate of the top $\tilde{k}$ detection results, $r(k)$ is the recall rate of the top k detection results, and C is the number of categories of the dataset.

Experimental Environment. The experiments are conducted in Ubuntu 16.04 operation system. The CPU is Intel Core i7–7700. Memory is 32GB. The graphics card is GeForce GTX 1080 Ti. The program is coded by Python with the Pytorch deep learning framework.

4.2 Experimental Results and Comparisons

In order to verify the effectiveness of the method in this paper, it is compared with the four algorithms of CAM [16], DANet [12], ACoL [15] and SPA [8], which are outstanding in natural scenes, and with the ARC [11] algorithm which is suitable for remote sensing scenes.

The detection results of the data set WSADD are shown in Table 1. It can be seen from the table that the CAM, DANet, ACoL, and SPA methods that achieve good results in natural scenes can hardly complete the positioning effect. This shows that the method applicable in natural scenes is not suitable for the remote sensing scenes with a large number of small objects. The ARC method has a significant improvement in precision and recall compared with the method in natural scenes. However, it still cannot solve the problem of the loss of a large number of objects with a small proportion of images in remote sensing images. Compared with the ARC method, our method increases the precision rate from 0.65 to 0.80 and the recall rate from 0.81 to 0.91 after adding the DA module and the similarity constraint module (SCM). This shows that our method effectively alleviates the problem of the loss of a large number of objects with a small proportion of images in remote sensing images.

Table 1. The detection precision and recall of representative methods on WSADD test set

Method	TP	FP	FN	Precision	Recall
CAM [16]	33	167	275	0.17	0.11
DANet [12]	23	160	285	0.13	0.07
ACol [15]	23	153	285	0.13	0.07
SPA [8]	119	206	189	0.39	0.38
ARC [11]	251	137	27	0.65	0.81
Ours	280	70	28	**0.80**	**0.91**

Table 2. Object classes in the DIOR data set

C1	C2	C3	C4	C5	C6	C7
Airplane	Airport	Baseball field	Basketball count	Bridge	Chimney	Dam
C8	C9	C10	C11	C12	C13	C14
Expressway service area	Expressway toll station	Golf field	Ground track field	Harbor	Overpass	Ship
C15	C16	C17	C18	C19	C20	
Stadium	Storage tank	Tennis court	Train Station	Vehicle	Wind mil	

The object categories of the DIOR data set are shown in Table 2. And the detection results of the DIOR data set are shown in Table 3. Because our method is proposed for a large number of small objects in remote sensing images, it can be seen from the table that compared with ACR, the SCDA method improves more in small object categories, such as airplanes and tennis courts, but it basically does not improve in large object categories. Overall, compared to ARC, SCDA has a 1.22% improvement in mAP. And compared to SPA, SCDA has a 1.15% improvement in mAP.

Table 3. Detection average precision (%) of representative methods on the proposed DIOR test set. The entries with the best APs for each object category are bold-faced. We only show 9 categories in which the results are quite different.

Method	C1	C3	C4	C8	C16	C17	C19	C20	mAP
CAM [16]	2.94	17.92	18.20	**26.44**	1.19	7.05	1.74	9.78	18.78
DANet [12]	1.33	13.56	17.95	25.84	0.56	5.85	1.11	3.08	17.37
ACol [15]	0.15	2.38	0.00	7.82	0.63	2.27	0.17	0.27	5.89
SPA [8]	10.68	**28.28**	18.13	25.11	**7.36**	13.05	5.15	11.42	20.93
ARC [11]	13.52	19.04	18.28	**26.44**	5.39	17.48	5.68	14.41	20.86
Ours	**18.03**	24.06	**23.46**	24.82	7.08	**25.56**	**6.45**	**18.77**	**22.08**

4.3 Analysis of Components

In this section, the contributions of the two key components in our proposed method including DA module and Similarity Constraint Module (SCM) are further evaluated. To this end, we conducted a series of experiments on the WSADD dataset to verify the effectiveness of the DA module and SCM, and to determine the hyper-parameters K and λ when using the DA module.

Table 4. Effect of DA module and SCM.

Method	TP	FP	FN	Precision	Recall
baseline	251	137	57	0.65	0.81
baseline + CSCM	239	130	69	0.65	0.78
baseline + PSCM	243	126	65	0.66	0.79
baseline + CSCM + PSCM	245	110	63	0.69	0.80
baseline + DA	**284**	183	**24**	0.61	**0.92**
baseline + CSCM + PSCM + DA	280	**70**	28	**0.80**	0.91

Similarity Constraint Module. In order to verify whether the shallow layer feature maps that have not passed the DA module can improve the response strength of similar targets through the similarity constraint module (SCM), we remove the DA module from the basic network and only verify the effect of the SCM. The experimental results on the data set WSADD are shown in Table 4. From Table 4, it can be seen that both the CSCM and the PSCM can improve the precision rate, but the effect is not obvious enough. This shows that in the shallow layer feature map, the features are scattered and sparse. Only paying attention to the similarity of position or channel features between feature maps cannot improve the response strength of similar features. After combining the two similarities, the precision rate has increased significantly, indicating that the combination of

Table 5. Effect of the hyper-parameters K and λ

Method	TP	FP	FN	Precision	Recall
Ours without DA	245	110	63	0.69	0.80
$K = 4, \lambda = 0.01$	260	86	48	0.75	0.84
$K = 4, \lambda = 0.05$	272	82	36	0.77	0.88
$K = 4, \lambda = 0.1$	**280**	**70**	**28**	**0.80**	**0.91**
$K = 4, \lambda = 0.5$	273	88	35	0.76	0.89
$K = 8, \lambda = 0.1$	271	94	37	0.74	0.88
$K = 16, \lambda = 0.1$	279	98	31	0.73	0.90

the two similarities can better improve the feature distribution and increase the response strength of similar features.

DA Module. In the DA module, we need to explore whether different K values will affect network performance. Similarly, in the network training segment, the loss function L_{DA} is constrained by the hyper-parameter λ. So, we need to analyze the impact of different λ on network performance. We design a comparative experiment on the WSADD data set to discuss the role of two hyper-parameters in the DA module. The experimental results are shown in Table 5. From Table 5, we can see that when the hyper-parameters K and λ are small, the DA module activation strength is weaker, and it cannot effectively activate more regions and improve the response strength of the region. When the value of K and λ is too large, such as $K = 16$ and $\lambda = 0.5$, the DA module has a strong activation intensity, but the activation regions contain more noise, which causes the false accuracy and recall rate to drop to zero. When the values of K and λ are moderate, such as $K = 4$, $\lambda = 0.1$, the DA module activation intensity is moderate at this time, and the precision and recall rates are improved. This shows that when the DA module activation intensity is moderate, it can effectively activate more regions and improve the response intensity of the region without increasing too much background noise.

Through Table 4, we can also see that it is difficult to improve the performance of the model using only the DA module or SCM. Using only the DA module can detect more objects, but it will also cause more error-detection. And just using the SCM can reduce error-detection, but it will also cause some objects to be undetected. Our proposed SCDA model uses both the DA module and the SCM to detect more targets without increasing the error detection rate.

5 Conclusion

In this article, we consider that there are a large number of small objects in remote sensing images. In order to better identify small objects in remote sensing images based on WSOD, we propose a similarity constraint divergent activation (SCDA) method to

activate the response regions of small objects. SCDA uses the DA module to obtain more response regions in the shallow layer feature map, and then removes noise through the SCM, which can better detect the small objects in remote sensing images. Experiments on WSADD and DIOR have proved that this method can solve the problem of remote sensing image object loss to a certain extent when weakly supervised learning is transferred to remote sensing scenes.

References

1. Cao, Y., Niu, X., Dou, Y.: Region-based convolutional neural networks for object detection in very high resolution remote sensing images. In: 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), pp. 548–554. IEEE (2016)
2. Choe, J., Shim, H.: Attention-based dropout layer for weakly supervised object localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2219–2228 (2019)
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), vol. 1, pp. 886–893. IEEE (2005)
4. Li, K., Wan, G., Cheng, G., Meng, L., Han, J.: Object detection in optical remote sensing images: a survey and a new benchmark. ISPRS J. Photogramm. Remote. Sens. **159**, 296–307 (2020)
5. Lindeberg, T.: Scale invariant feature transform. Scholarpedia **7**(5), 10491 (2012)
6. Mai, J., Yang, M., Luo, W.: Erasing integrated learning: a simple yet effective approach for weakly supervised object localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8766–8775 (2020)
7. Nowak, E., Jurie, F., Triggs, B.: Sampling strategies for bag-of-features image classification. In: European Conference on Computer Vision, pp. 490–503. Springer, Cham (2006). https://doi.org/10.1007/11744085_38
8. Pan, X., et al.: Unveiling the potential of structure-preserving for weakly supervised object localization. arXiv preprint arXiv:2103.04523 (2021)
9. Singh, K.K., Lee, Y.J.: Hide-and-seek: forcing a network to be meticulous for weakly-supervised object and action localization. In: 2017 IEEE international conference on computer vision (ICCV), pp. 3544–3553. IEEE (2017)
10. Wei, Y., Feng, J., Liang, X., Cheng, M.M., Zhao, Y., Yan, S.: Object region mining with adversarial erasing: a simple classification to semantic segmentation approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1568–1576 (2017)
11. Xu, J., Wan, S., Jin, P., Tian, Q.: An active region corrected method for weakly supervised aircraft detection in remote sensing images. In: Eleventh International Conference on Digital Image Processing (ICDIP 2019), vol. 11179, p. 111792H. International Society for Optics and Photonics (2019)
12. Xue, H., Liu, C., Wan, F., Jiao, J., Ji, X., Ye, Q.: Danet: divergent activation for weakly supervised object localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6589–6598 (2019)
13. Yao, Y., Jiang, Z., Zhang, H., Zhao, D., Cai, B.: Ship detection in optical remote sensing images based on deep convolutional neural networks. J. Appl. Remote Sens. **11**(4), 042611 (2017)
14. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6023–6032 (2019)

15. Zhang, X., Wei, Y., Feng, J., Yang, Y., Huang, T.S.: Adversarial complementary learning for weakly supervised object localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1325–1334 (2018)
16. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929 (2016)