SAENet: Self-Supervised Adversarial and Equivariant Network for Weakly Supervised Object Detection in Remote Sensing Images

Xiaoxu Feng, Xiwen Yao, *Member, IEEE*, Gong Cheng, *Member, IEEE*, Jungong Han, and Junwei Han, *Senior Member, IEEE*

***Abstract*— Weakly supervised object detection (WSOD) in remote sensing images (RSIs) remains a challenge when learning a subtle object detection model with only image-level annotations. Most works tend to optimize the detection model via exploiting the most contributed region, thereby to be dominated by the most discriminative part of an object. Meanwhile, these methods ignore the consistency across different spatial transformations of the same image and always label them with different classes, which introduces potential ambiguities. To tackle these challenges, we propose a unique self-supervised adversarial and equivariant network (SAENet) and aim at learning complementary and consistent visual patterns for WSOD in RSIs. To this end, an adversarial dropout–activation block is first designed to facilitate the entire object detector via adaptively hiding the discriminative parts and highlighting the instance-related regions. Besides, we further introduce a flexible self-supervised transformation equivariance mechanism on each potential instance from multiple spatial transformations to obtain spatially consistent self-supervisions. Accordingly, the obtained supervisions can be leveraged to pursue a more robust and spatially consistent object detector. Comprehensive experiments on the challenging LEarning, VIsion and Remote sensing Laboratory (LEVIR), NorthWestern Polytechnical University (NWPU) VHR-10.v2, and detection in optical RSIs (DIOR) datasets validate that SAENet outperforms the previous state-of-the-art works and achieves 46.2%, 60.7%, and 27.1% mAP, respectively.**

***Index Terms*— Multiple instance learning (MIL), remote sensing images (RSIs), self-supervised learning, weakly supervised object detection (WSOD).**

I. Introduction

THE development of object detection in remote sensing images (RSIs) has facilitated many practical applications [1]–[11] and attracted much attention in recent years. Driven by the advent of the era of convolutional neural networks (CNNs) [12], [13], object detection [14]–[18] in both natural scene images and RSIs has achieved impressive progress by leveraging the availability of abundant datasets with subtle manually labeled annotations [19]–[21]. Nevertheless, manually labeling such subtle annotations for each instance from the enormous number of RSIs is laborious, time-consuming, or even impractical. To reduce the heavy labeling effort, weakly supervised object detection (WSOD) has been extensively studied in recent years.

Of late, with the emergence of several constructive works [22], [23], a series of advanced methods [6], [22]–[31] first leveraged multiple instance learning (MIL) constraints to transform the WSOD problem into a multilabel classification problem and treated redundant object proposals as inputs. Next, the most contributing proposal was treated as the pseudo instance-level label to learn the more discriminative object detector. Based on it, some researches aim to leverage curriculum learning [32], better initialization models [23], [27], knowledge distillation [24], [33], and regularization [34], [35] to further improve the performance of WSOD. Our work also follows the aforementioned two-stage strategy to train an end-to-end MIL network.

Despite the remarkable progress of WSOD methods in natural scene images, there are still two major obstacles which need to be resolved immediately. First, previous WSOD methods inclined to over-fit on the most discriminative object parts rather than the entire object, leading to a part domination problem. It is the dominant reason for the large performance gap between WSOD and fully supervised methods. Obviously, these methods will get worse when it comes to the large-scale cluttered background of RSIs so that these methods cannot be directly leveraged to address the WSOD problem in RSIs.

It is worth noting that the existing approaches also ignored the consistency across data augmentation of the same images. However, some common data augmentation methods, such as flipped image and scaled image, lead to significant inconsistency and introduce potential ambiguities under weakly supervised settings, as the same instance across different spatial transformations may be labeled with different categories. It is another obstacle that seriously impedes the performance of

Manuscript received April 18, 2021; revised August 2, 2021; accepted August 12, 2021. Date of publication August 27, 2021; date of current version January 31, 2022. This work was supported in part by the National Science Foundation of China under Grant 62071388, Grant 61701415, Grant 61772425, and Grant 61773315; in part by the Fundamental Research Funds for the Central Universities under Grant 3102019ZDHKY05; in part by China Postdoctoral Science Foundation under Grant 2018T111094 and Grant 2017M620468; in part by the Postdoctoral Science Foundation of Shaanxi Province under Grant 2017BSHYDZZ36; and in part by the National Key Research and Development Program of China under Grant 2017YFB0502900. *(Corresponding author: Xiwen Yao.)*

Xiaoxu Feng, Xiwen Yao, Gong Cheng, and Junwei Han are with the School of Automation, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: yaoxiwen517@gmail.com).

Jungong Han is with the Department of Computer Science, Aberystwyth University, Aberystwyth SY23 3FL, U.K. (e-mail: juh22@aber.ac.uk).

Digital Object Identifier 10.1109/TGRS.2021.3105575

WSOD in RSIs. Note that the inconsistency is not an issue for the full supervised paradigm, as the instance-level label can naturally encourage consistency.

To tackle the first obstacle, an adversarial dropout–activation (ADA) block is designed to encourage the detection model to activate the entire object rather than focus on the most discriminative parts. Specifically, a parametric spatial dropout block is first introduced to adversarially maximize the detection objective through adaptively hiding the most discriminative region. Next, an activation block is further introduced to active instance-related features via capturing cross-channel interaction. The cooperation of spatial dropout block and activation block formulates an adversarial mechanism which can effectively highlight the entire object features. Accordingly, the high-quality instances can be effectively mined to train more robust object detectors.

To tackle the second obstacle, a novel and flexible self-supervised transformation equivariance (SSTE) mechanism is further constructed to enforce the same instance with different spatial transformations to communicate with each other so that more consistent and robust features can be learned. More specifically, the same image with different transformations is fed into the detection framework at the same time. Equivariance regularization is applied on the positive instance from various transformed images to obtain spatially consistent self-supervisions. Next, the obtained supervisions are applied to learn the spatially consistent object detector via enforcing the same instances with multiple spatial transformations to be mapped close to each other. Benefitted from it, more robust and exclusive features can be captured no matter what transformations exist between instances, thereby further enhancing the performance of WSOD in RSIs.

Cooperating ADA with SSTE formulates a flexible end-to-end self-supervised adversarial and equivariant network (SAENet). ADA is designed to force the detector model to pursue the entire instance via an adversarial mechanism, and SSTE focuses on capturing more consistent information and alleviating the potential ambiguities introduced by spatial transformations. Extensive experiments, including both quantitative and qualitative results on the challenging public LEarning, VIsion and Remote sensing Laboratory (LEVIR) [19], NorthWestern Polytechnical University (NWPU) VHR-10.v2 [20], and detection in optical RSIs (DIOR) [21] datasets, clearly demonstrate the meliority of our method. Our key contributions can be summarized as follows.

1) We introduce a unique learnable ADA block via an adversarial mechanism, which can effectively activate the high-quality entire object.
2) A novel and flexible SSTE mechanism is developed to tackle the potential ambiguities introduced by different spatial transformations.
3) Comprehensive quantitative and qualitative results clearly demonstrate that our SAENet significantly boosts the performance of state-of-the-art results.

II. Related Work

Of late, almost all WSOD works formulate the object detection task as an image classification task with MIL. Meanwhile, self-supervised learning also is introduced to the MIL for better performance. In this section, some related works about MIL and self-supervised learning for remote sensing analysis will be reviewed in detail.

A. Multiple Instance Learning

MIL, as a classical weakly supervised learning strategy, has been widely applied to address the WSOD problem. In MIL, the object proposals generated by object proposal method [36] for each training image are divided into different "bags." Then the high-scored proposals from each positive bag are iteratively selected to learn the corresponding detectors. Under this strategy, a lot of advanced WSOD approaches have been proposed and achieved remarkable progress. For instance, Tang *et al.* [23] designed a remarkable online instance classifier refinement (OICR) framework which aims to find high-quality instances through propagating image-level labels to spatially overlapped regions. The work [27] introduced a novel cluster learning strategy to pursue high-quality instances. Wan *et al.* [28] proposed a min-entropy latent model (MELM). However, the above works cannot be directly used to address the WSOD problem in RSIs owing to the large-scale and cluttered background in RSIs.

To this end, a novel instance mining strategy [29] was proposed to address the WSOD problem in RSIs by iteratively mining positive instances from the negative data and then refining the corresponding object detector. This is the first attempt to address the object detection problem in RSIs under weakly supervised paradigm. Based on it, Zhou *et al.* [37] successfully boosted the detection performance by integrating a negative bootstrapping scheme into iterative learning. More recently, Yao *et al.* [32] introduced a unique dynamic curriculum learning (DCL) strategy to learn a more robust detector in a learning sequence from easy to difficult. To pursue high-quality instances, the work [38] constructed an end-to-end progressive contextual instance refinement framework and introduced a proposal self-pruning algorithm. More currently, triple context-aware network (TCANet) [35] designed a triple context-aware framework to address the densely packed objects in RSIs.

B. Self-Supervised Learning

Self-supervised learning approaches aim to learn a more robust representation by designing pretext tasks (e.g., image colorization [39], spatial transformation [40]) without the necessity of additional manual annotations. There are many studies [41], [42] that have applied self-supervised learning to address the object detection problem and achieved good performance. For instance, the work [41] aimed to generate more comparable features in a self-supervised learning manner via taking different scale downsampled images as input. Wang *et al.* [42] proposed a principled self-supervised sample mining approach. It aims to find reliable proposals via feeding proposals into different labeled images and then evaluating their difference under different image contexts. More recently, Wang *et al.* [43] proposed an impressive self-supervised equivariant attention mechanism for weakly

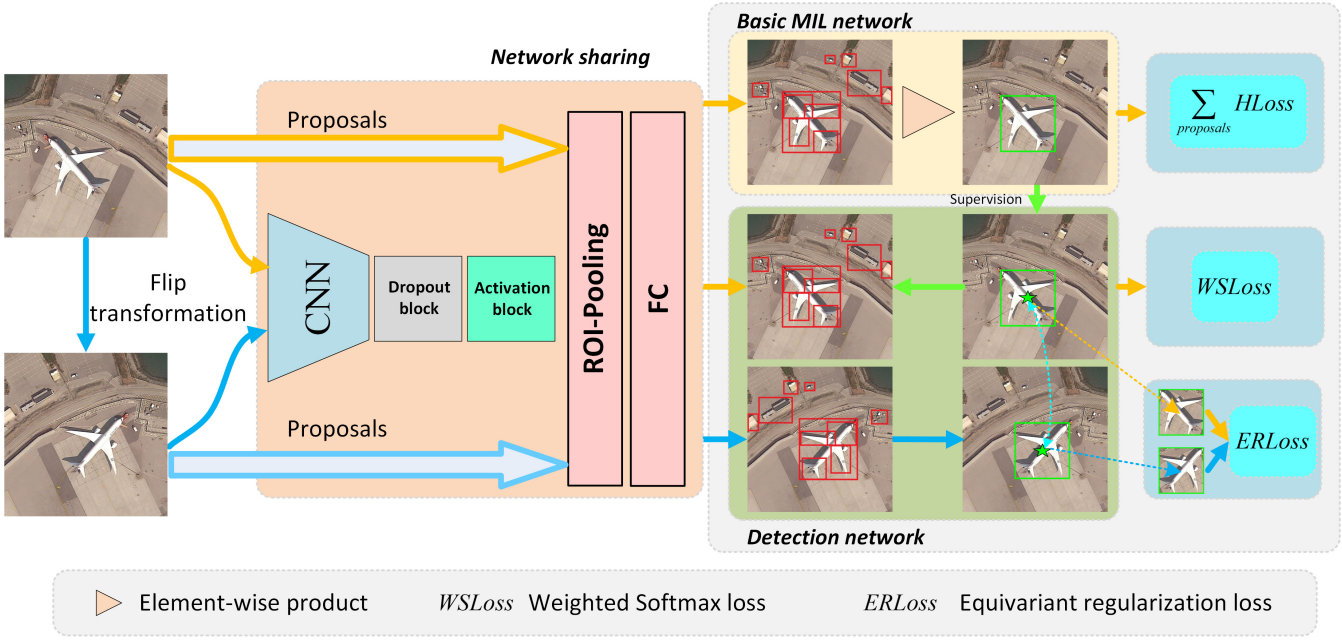


Fig. 1. Illustration of SAENet. To address the obstacles of part domination and potentially ambiguities introduced by multiple spatial transformations in RSIs, we first encourage the detection model to pursue entire object via adaptively hiding the most discriminative region and activating the remaining instance-related region. Then the same instance with different transformations is enforced to communicate with each other, thereby alleviating the challenge of ambiguities.

supervised semantic segmentation problems via constructing consistency regularization on various transformed images.

Our work is inspired by Wang *et al.* [43]. Compared with it, in this article, we first develop a new ADA model to pursue the full object extent. Besides, we design a flexible self-supervised mechanism that is applied on the positive instance predicted by the detection model from various transformed images rather than capturing context appearance information. There is no doubt that our method is the first attempt to address the obstacle of transformed variation in RSIs under weakly supervised settings.

III. Proposed Method

A. Overview

The inconsistency between inexact supervision and learning instance-level object detector introduces erratic object localization. It misleads the detection model to converge to the most discriminative part and even cluttered background in RSIs. Meanwhile, the potential ambiguities introduced by spatial transformations further increase the difficulty of the task and impair the discriminability of the detector. To address these challenges, we introduce a unique SAENet illustrated in Fig. 1. It aims at alleviating the part domination problem and learning a transformed invariant object detector via capturing the complementary and consistent patterns. Specifically, the ADA block is first developed to capture the complementary patterns through suppressing the most discriminative region and activating the remaining instance-related regions so that the entire object features can be mined to train a more robust object detector. Next, the SAENet drives the same instance with different spatial transformations to communicate with each other and enforces them to be mapped close to each other. Accordingly, more consistent and exclusive features can be learned to facilitate the more robust object detectors.

B. ADA Block

Due to the absence of instance-level labels, the current WSOD approaches are inclined to find the most discriminative part of an object. A natural solution to this challenge is a crude strategy, that is, suppress the most discriminative regions. However, it has some shortcomings: 1) the existing dropout operation is less effective for contiguous regions and 2) simply dropping the most discriminative regions will hurt the feature representation to some extent and may cause cluttered backgrounds to be considered as an object. Hence, the proposed ADA block aims to capture the complementary patterns via adaptively hiding the discriminative regions and simultaneously activating the remaining instance-related regions under an adversarial paradigm.

Given an input image $\mathcal{I} = \{(I_i, Y_i, \mathcal{B}_i)\}$, where $Y_i = [y_1, \ldots, y_c, \ldots, y_C] \in \{-1, 1\}$ is the image-level label to declare which object category is present or absent in an image and $\mathcal{B}_i$ denotes its region proposals. First, the feature maps $\Psi(W, \mathcal{I}) \in \mathbb{R}^{C\times H\times W}$ with C channels and $H \times W$ resolutions are generated by feeding the input image $\mathcal{I}$ into the backbone convolutional layers. Next, a probability map $p(\theta, \mathcal{I}) \in \mathbb{R}^{H\times W}$ is first generated by Ghiasi *et al.* [34] to control which contiguous regions will be dropped, and it obeys Bernoulli distribution. Then, $p(\theta, \mathcal{I})$ is fed into a spatial Gumbel-Softmax to transform it into a hard mask $m(\mathcal{I}) \in \{0, 1\}^{H\times W}$. For each $m(\mathcal{I}_{(x,y)}) = 0$, a spatial square mask with the center being $m(\mathcal{I}_{(x,y)})$ and with the size of drop_size is leveraged to set each feature channel value of $\Psi(W, \mathcal{I}) \in \mathbb{R}^{C\times H\times W}$ in the square to be zero. Accordingly, the

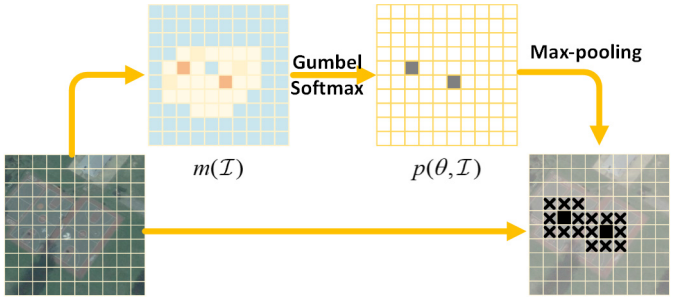


Fig. 2. Illustration of parametric spatial dropout block. The most discriminative region will be zeroed out, thereby enforcing the detection model to adversarially maximize the detection objective.

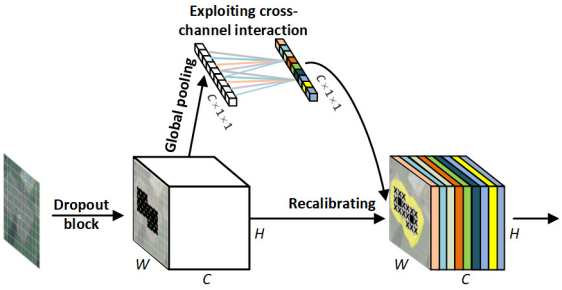


Fig. 3. Illustration of activation block. The interactions between each channel and its k neighbors are exploited to activate the instance-related features.

detection model will become more robust, and the challenge of part domination will be alleviated to some extent. The process of dropout block is illustrated in Fig. 2.

To avoid the problem of cluttered background domination introduced by dropping the discriminative regions, as shown in Fig. 3, we further develop an activation block to highlight the instance-related regions so that the object detectors are driven to look at the entire object. Given the dropped feature $\Psi_d(W,\mathcal{I}) \in \mathbb{R}^{C\times H\times W}$, the instance-related features are adaptively highlighted via learning a channel attention, which is given by

$$w = \sigma\left(W_1 \Psi_d(W,\mathcal{I})\right) \tag{1}$$

where W_1 denotes the $C \times C$ parameter matrix. Existing work [44] shows that avoiding dimensionality reduction can obtain more effective channel attention and using appropriate cross-channel interaction can achieve better performance while significantly decreasing the model complexity. Thus, we obtain the weights of the dropped feature $\Psi_d(W,\mathcal{I}) \in \mathbb{R}^{C\times H\times W}$ via exploiting the interactions between each channel and its k neighbors, which is formulated as

$$w_i = \sigma\left(\sum_{j=1}^{k} w_1^{ij} \Psi_d^{ij}\right), \quad \Psi_d^{ij} \in \Psi_d^{ik} \tag{2}$$

where $\Psi_d^{ik}(W,\mathcal{I})$ denotes the group of k neighbor channels of $\Psi_d^{i}(W,\mathcal{I})$.

To further decrease the model complexity, the corresponding interaction channels are made to share the same parameters. It can be readily done by the convolution layers with $1 \times k$ filter. Thence, the interaction channel attention module can be denoted as

$$w = \sigma\left(\mathrm{Conv}_k(\Psi_d)\right) \tag{3}$$

where Conv_k indicates the convolution layers with $1 \times k$ filter. In our experiment, k is set to 3. Accordingly, the captured local cross-channel interaction can be leveraged to activate the instance-related features, as in

$$\Psi_{\mathrm{ad}}(W,\mathcal{I}) = w \cdot \Psi_d(W,\mathcal{I}) \tag{4}$$

where $\cdot$ indicates the channel-wise multiplication.

The cooperation of the dropout block and the activation block formulates an adversarial mechanism where the dropout block aims at hiding the most discriminative regions, and the activation model endeavors to highlight the remaining instance-related regions so that the object detector can be driven to pursue the entire object.

C. SSTE Mechanism

Flipped image and scaled image are common data augmentations, which have been widely used in fully supervised object detection, and the instance-level labels should also have the same transformation. It brings in an extra latent equivariant constraint for the model training. Unfortunately, due to the absence of instance-level labels, it cannot be directly applied to train the detection model under weakly supervised settings. What is more, for the same input images, spatial transformations also introduce feature changes that do not correlate with the feature distribution between classes. Accordingly, the WSOD methods always label the same instance after different spatial transformations with different categories, which significantly hurts the discriminative of the detector. To this end, we propose a weakly supervised SSTE mechanism, that is, the label of instances predicted by the WSOD model should not be changed with spatial transformations.

Let $\mathcal{F}_{\mathcal{H}} = \mathrm{ROI}(\Psi_{\mathrm{ad}}, \mathcal{H}), \mathcal{H} \in \mathcal{B}_i$ denote the proposal features obtained using the region of intersect (ROI) pooling layer. As shown in Fig. 1, all proposal features are branched into two paralleled modules, termed MIL module and detection model, to generate the classification scores $x_c \in \mathbb{R}^{|\mathcal{H}|\times C}$ and detection scores $x_d \in \mathbb{R}^{|\mathcal{H}|\times(C+1)}$. $|\mathcal{H}|$ denotes the number of the proposal $\mathcal{B}_i \cdot \{C+1\}$ indicates C different object categories and one background. The corresponding image and its proposals after flipped or scaled transformation are also fed into the detection model to generate its detection scores $x_{Td} \in \mathbb{R}^{|\mathcal{H}|\times(C+1)}$. For each class c appearing in the image ($y_c = 1$), we first obtain the most confident region by selecting the top scoring proposal from the MIL module. Next, its neighbor proposals with enough spatial overlap and lowly spatial overlaps are labeled as positive instances and background, respectively. On one hand, the labeled instances are defined as pseudo instance-level labels to learn the detection model. On the other hand, we also introduce an SSTE mechanism to facilitate the consistent object detector learning.

By leveraging the aforementioned latent equivariant constraint, we can obtain extra self-supervision and draw the following conclusion, that is, the labeled instance remains

the same label as before and after flipping or scaling transformations. Inspired by it, we introduce a simple but effective equivariance regularization, which leverages the obtained self-supervisions to enforce the same instances with multiple spatial transformations to be mapped close to each other. Accordingly, more abundant and exclusive features can be captured to learn the more robust and spatially consistent object detectors. The concrete form of equivariance regularization is as follows:

$$\mathcal{R}_{\text{er}} = \left\| D(\mathcal{H}_p) - D(\text{Tran}(\mathcal{H}_p)) \right\|_1 \tag{5}$$

where $D(\cdot)$ denotes the detection module, Tran$(\cdot)$ indicates the spatial transformation, e.g., rescaling and flip, and $\mathcal{H}_p$ represents the labeled positive instances. It is worth noting that spatial transformations are operated before the feedforward of the network and they share the same network.

D. Loss for Model Learning

In our WSOD model, only image-level labels can be applied to declare which object category is present or absent in an image. We use the MIL model in [35] to diagnose the localization of instances and adopt the hinge loss function for MIL model learning. The prediction of image class can be generated by summation over all classification scores of proposals $\Phi^c = \sum_{r=1}^{|\mathcal{H}|} x_c$. The MIL loss is defined for C foreground object category, which is given by

$$\text{Loss}_{\text{MIL}} = \frac{1}{C} \sum_{c=1}^{C} \max(0, 1 - Y_i \cdot \Phi^c). \tag{6}$$

Next, the top scoring proposals and its neighbor proposals are preliminarily selected as pseudo instance-level labels for detection model learning. According to additional supervisions provided by equivariance regularization, we also apply the hinge loss function to encourage the detection model to capture more robust features as in

$$\mathcal{R}_{\text{er}} = \frac{1}{N} \sum_{n=1}^{N} \max\left(0, \left\| D(\mathcal{H}_{p_n}) - D(\text{rot}(\mathcal{H}_{p_n})) \right\|_1\right) \tag{7}$$

where N denotes the number of positive instances. Next, the obtained instances $\mathcal{P}$ are treated as the pseudo instance-level annotations. The weighted softmax loss function is used for detection model learning, which is denoted as

$$\text{Loss}_{\text{Det}} = -\frac{1}{|\mathcal{H}|} \left(\sum_{r \in \mathcal{H}} w_{\text{cd}} \mathcal{P} \log x_{dr} \right) \tag{8}$$

where w_{cd} denotes the loss weight. In our experiment, the top scoring proposal is selected as w_{cd} for each foreground object category.

Collaborating aforementioned losses formulates the final loss of our WSOD framework, which is denoted as

$$\mathcal{L} = \text{Loss}_{\text{MIL}} + \mathcal{R}_{\text{er}} + \text{Loss}_{\text{Det}}. \tag{9}$$

The MIL loss Loss_{MIL} is applied to roughly diagnose the localization of instances and equivariance regularization $\mathcal{R}_{\text{er}}$ forces the detector to capture feature changes introduced by different transformations so that the potentially ambiguities can be relieved to some extent. The detection model Loss_{Det} is leveraged to facilitate precise object localization.

IV. Experiments

A. Datasets and Evaluation Metrics

The LEVIR dataset includes 22k images covering three object categories with the size of 800×600, which comes from high-resolution Google Earth images. It is a challenging dataset and includes large-scale cluttered backgrounds which cover different human living environments, (i.e., city, country, mountain area, and ocean). In experiments, we keep the same training–testing settings (70% for training and 30% for testing) with comparison methods.

The proposed method is further evaluated on the commonly used NWPU VHR-10.v2 dataset with ten object categories. It includes 1172 images with the size of 400×400 and is divided into a train set with 679 images, a validation set with 200 images, and a test set with 293 images. Following the common experimental protocol for object detection, both the training set and the validation set are treated as the training split and the test set is used as the testing split.

We also evaluate the proposed approach on the more challenging DIOR dataset which includes 23 463 images with a total of 192 472 instances. It covers 20 object categories and is evenly divided into the training set with 11 725 images and testing set with 11 738 images.

As with all WSOD methods, the WSOD performance on both datasets are evaluated in terms of correct localization (CorLoc) and average precision (AP). The intersection over union (IoU) threshold with 50% is also treated as a criterion for both metrics. Here, CorLoc is used to evaluate the localization accuracy on the training set. AP is used to measure the accuracy of object detection on the testing set.

B. Implementation Details

In our experiments, we select the Dual-local context residual network in [35] as our baseline network where VGG16 [45] is adopted as the backbone. All settings of the detection model (i.e., learning rate with 0.001, mini-batch with 2, weight decay with 0.005, and momentumwith 0.9) are kept identical to [35], [38] for a fair comparison. For training, the detection model runs 30k iterations with 10k stepsize, 30k iterations with 20k stepsize, and 200k iterations with 100k stepsize for the NWPU VHR-10.v2, the LEVIR, and the DIOR datasets, respectively.

Similar to [35], [38], Selective Search [36] is adopted as a region proposal method to generate about 2000 proposals per image. We also augment both train and test sets by horizontal flipping them with five image scales {480, 576, 688, 864, 1200}. In dropout block, θ and drop_size are set as 0.3 and 3, respectively. For testing, an non-maximum suppression (NMS) of 0.3 is leveraged to remove duplicated bounding boxes.

C. Ablation Experiments

Our ablation experiments are carried out on the DIOR dataset to analyze the proposed method, including the ADA block and equivariance regularization.

TABLE I
RESULTS ON THE DIOR DATASET FOR EACH KEY COMPONENT

Architecture	Baseline	√	√	√	√	√
	Dropout block		√	√		√
	Activation block			√		√
Strategy	Equivariance regularization				√	√
mAP(%)		22.26	24.39	25.11	25.32	27.10
CorLoc(%)		45.12	46.36	47.44	46.91	49.42

1) Contribution of ADA: To disclose the contribution of the ADA block and each block in it, we elaborately conduct two group experiments to simply evaluate the dropout block and the ADA block. The ablation results are shown in Table I. Using the single dropout block and the ADA block significantly brings improvement with 2.13% and 2.85% mAP and 1.24% and 2.32% CorLoc, respectively. Based on it, we can draw the following conclusions.

1) Dropout block can alleviate the problem of part domination in WSOD to some extent.
2) Collaborating the dropout block with the activation block can adaptively pursue instance-related regions and meanwhile mitigate the influence of the cluttered background introduced by dropping the discriminative regions.

2) Contribution of Equivariance Regularization: We further integrate equivariance regularization into our network to investigate its contribution. As presented in Table I, it brings about large improvements (mAP from 22.26% to 25.32% and CorLoc from 45.12% to 46.91%). Besides, joining the ADA block and equivariance regularization also improves the detection performance by 1.78% mAP and 2.51% CorLoc. This fully demonstrates the effectiveness of equivariance regularization. It is mainly because equivariance regularization can effectively drive the detection model to capture the changes in feature caused by transformations. Meanwhile, it also enforces the features before and after transformations to be mapped close to each other. Accordingly, the potential ambiguities introduced by the different transformations can be alleviated to some extent.

D. Comparisons With State-of-the-Arts

Table II first provides our detection performance for each class and comparisons with advanced WSOD methods on the LEVIR dataset. It can be observed that the proposed SAENet achieves 46.2% mAP, which significantly boosts the baseline work by a large margin (46.2% versus 27.1%) and consistently improves the performance for each class. Compared with the existing WSOD methods in RSIs, we outperform the second best TCANet method 7.9% mAP and achieve a new state of the art, which fully demonstrates the superiority of the proposed method.

With the proposed approach, we further carry out object detection on the NWPU VHR-10.v2 dataset and indicate the quantitative comparisons with the existing state-of-the-art methods in Tables III and IV. As reported in

TABLE II
PERFORMANCE COMPARISONS IN TERMS OF AP (%) AND MAP (%) AMONG DIFFERENT METHODS ON THE LEVIR TEST SET

Method	plane	ship	oilpot	mAP
TINY-RAM [19]	50.5	19.3	39.6	37.8
MEDIUM-RAM [19]	76.0	50.0	48.1	58.0
LARGE-RAM [19]	71.7	60.8	43.0	58.5
Faster RCNN [46]	87.6	81.6	71.9	80.4
OICR[23]	21.1	0.3	48.7	23.4
PCL[27]	31.3	2.0	43.4	25.6
MELM[28]	45.0	1.8	44.2	30.3
TCANet [35]	51.2	13.2	50.6	38.3
Baseline [35]	31.5	11.6	38.2	27.1
Ours	**59.7**	**25.3**	**53.7**	**46.2**

Tables III and IV, among the existing weakly supervised approaches in RSIs, the proposed method consistently boosts the state-of-the-arts on average with 60.72% mAP and 73.46% CorLoc, respectively. Compared with the popular weakly supervised approaches, SAENet, respectively, outperformed the weakly supervised deep detection network (WSDDN) [22], OICR [23], proposal cluster learning (PCL) [27], MELM [28], DCL [32], progressive contextual instances refinement (PCIR) [38], and TCANet [35] by 25.6% (60.72% versus 35.12%), 26.2% (60.72% versus 34.52%), 21.31% (60.72% versus 39.41%), 18.43% (60.72% versus 42.29%), 8.61% (60.72% versus 52.11%), 5.75% (60.72% versus 54.97%), and 1.9% (60.72% versus, 58.82%), which were notable margins in terms of mAP. The large improvements are mainly attributed to that our approach can effectively tackle the problems of part domination and potentially ambiguities introduced by multiple spatial transformations in RSIs.

To testify the robustness of our SAENet, we also provide quantitative comparisons in terms of mAP and CorLoc with advanced WSOD works on the larger and more challenging DIOR dataset where the object categories of it are reported in Table V. As indicated in Table VI, our SAENet achieves 27.10% mAP, which further facilitates state-of-the-art TCANet [35] by 1.28% mAP. Specifically, the detection performance of our SAENet surpasses WSDDN [22] (+13.84%) OICR [23] (+10.60%), PCL [27] (+8.91%), MELM [28] (+8.44%), DCL [32] (+6.91%), PCIR [38] (+2.18%), and TCANet [35] (+1.28%) with a large margin. Note that our approach substantially boosts the detection performance of our baseline work on almost all classes, especially for class "Airport" (+16.75%), "Golf filed" (+17.74%),

TABLE III
PERFORMANCE COMPARISONS (AP AND MAP) AMONG DIFFERENT METHODS ON THE NWPU VHR-10.V2 TEST SET

Methods	Airplane	Ship	Storage tank	Baseball Diamond	Tennis court	Basketball court	Ground track field	Harbor	Bridge	Vehicle	mAP
Transferred CNN [47]	0.6603	0.5713	0.8501	0.8093	0.3511	0.4552	0.7937	0.6257	0.4317	0.4127	0.5961
RICNN [16]	0.8871	0.7834	0.8633	0.8909	0.4233	0.5685	0.8772	0.6747	0.6231	0.7201	0.7311
RCNN [48]	0.8537	0.8888	0.6278	0.1973	0.9066	0.5823	0.6795	0.7987	0.5422	0.4992	0.6576
Fast RCNN [49]	0.9091	0.9060	0.8929	0.4732	1.0000	0.8585	0.8486	0.8822	0.8029	0.6984	0.8271
Faster RCNN [46]	0.9090	0.8630	0.9053	0.9824	0.8972	0.6964	1.0000	0.8011	0.6149	0.7814	0.8451
RICO [20]	0.9970	0.9080	0.9061	0.9291	0.9029	0.8013	0.9081	0.8029	0.6853	0.8714	0.8712
WSDDN [22]	0.3008	0.4172	0.3498	0.8890	0.1286	0.2385	0.9943	0.1394	0.0192	0.0360	0.3512
OICR [23]	0.1366	0.6735	0.5716	0.5516	0.1364	0.3966	0.9280	0.0023	0.0184	0.0373	0.3452
PCL [27]	0.2600	0.6376	0.0250	0.8980	**0.6445**	**0.7607**	0.7794	0.0000	0.0130	0.1567	0.3941
MELM [28]	0.8086	0.6930	0.1048	0.9017	0.1284	0.2014	0.9917	0.1710	0.1417	0.0868	0.4229
DCL [32]	0.7270	0.7425	0.3705	0.8264	0.3688	0.4227	0.8395	0.3957	**0.1682**	0.3500	0.5211
PCIR [38]	**0.9078**	**0.7881**	0.3640	0.9080	0.2264	0.5216	0.8851	0.4236	0.1174	**0.3549**	0.5497
TCANet [35]	0.8943	0.7818	**0.7842**	0.9080	0.3527	0.5036	0.9091	**0.4244**	0.0411	0.2830	0.5882
Ours	0.8291	0.7447	0.5020	**0.9674**	0.5566	0.7294	**1.0000**	0.3646	0.0633	0.3189	**0.6072**

TABLE IV
PERFORMANCE COMPARISONS (CORLOC) AMONG DIFFERENT METHODS ON THE NWPU VHR-10.V2 TRAINVAL SET

Methods	Airplane	Ship	Storage tank	Baseball Diamond	Tennis court	Basketball court	Ground track field	Harbor	Bridge	Vehicle	CorLoc
WSDDN [22]	0.2232	0.3681	0.3995	0.9248	0.1796	0.2424	0.9926	0.1483	0.0169	0.0289	0.3524
OICR [23]	0.2941	0.8333	0.2051	0.8176	0.4085	0.3208	0.8660	0.0741	0.0370	0.1444	0.4001
PCL[27]	0.1176	0.5000	0.1282	0.9865	**0.8451**	0.7736	0.9072	0.0000	0.0926	0.1556	0.4506
MELM[28]	0.8596	0.7742	0.2143	0.9833	0.1071	0.4348	0.9500	0.4000	0.1176	0.1463	0.4987
PCIR [38]	**1.0000**	**0.9306**	0.6410	**0.9932**	0.6479	0.7925	0.8969	0.6296	0.1326	0.5222	0.7187
TCANet [35]	0.9691	0.9178	**0.9513**	0.8865	0.6690	0.6283	0.9598	0.5418	**0.1963**	**0.5556**	0.7276
Ours	0.9706	0.9167	0.8781	0.9865	0.4086	**0.8113**	**1.0000**	**0.7037**	0.1481	0.5222	**0.7346**

TABLE V
OBJECT CLASSES IN THE DIOR DATASET

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

"Ground track filed" (+8.08%), "Harbor" (+6.08%), "Stadium" (+37.03%), and "wind mill" (7.47%). Since multiple spatial transformations are leveraged to augment training data, these training data are used to train the WSOD networks in different training iterations without equivariant constraint. It may cause the same instance across different spatial transformations to be distributed as different labels in different training iterations. However, the baseline ignores this challenge. In contrast, our approach aims to alleviate these potential ambiguities and meanwhile pursue the entire object. Besides, the CorLoc results, as shown in Table VII, further reveal the effectiveness of the proposed SAENet. We achieve consistently the state-of-the-art performance by 49.42% CorLoc.

To further establish the superiority of our SAENet, we also provide quantitative comparisons with advanced fully supervised methods on the LEVIR, NWPU VHR-10.v2, and DIOR datasets. As indicated in Table II, Tables III, and VI, we achieve comparable and even superior performance to some fully supervised approaches, such as tiny-networks random access memories (TINY-RAMs) [19], Transferred CNN [47], and regions with CNN features (RCNN) [48]. Meanwhile, the performance gap between the weakly and fully supervised object detection is also further narrowed.

Despite the average good performance, the detection performance for "Chimney" and "Storage tank" is not boosted significantly as their circular shape. Besides, SAENet also has trouble in accurately detecting the classes "bridge," "dam," and "overpass." This is mainly because lacking adequate supervisions misleads the detection model to find the coexisting objects or special background. The detection model often mistakenly identifies the more arresting or large-scale special background as an object (as illustrated in Figs. 4 and 5), i.e., bridges coexisting with more arresting rivers, dams coexisting with large-scale reservoirs, and overpass coexisting with many roads. This is another challenge to be solved for WSOD in RSIs, which drastically reduces the detection performance for those object classes.

TABLE VI

PERFORMANCE COMPARISONS (AP AND MAP) AMONG DIFFERENT METHODS ON THE DIOR TEST SET

Methods	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	mAP
Fast RCNN [49]	0.4417	0.6679	0.6696	0.6049	0.1556	0.7228	0.5195	0.6587	0.4476	0.7211	0.6293	0.4618	0.3803	0.3213	0.7098	0.3504	0.5827	0.3791	0.1920	0.3810	0.4998
Faster RCNN [46]	0.5028	0.6260	0.6604	0.8088	0.2880	0.6817	0.4726	0.5851	0.4806	0.6044	0.6700	0.4386	0.4687	0.5848	0.5237	0.4235	0.7952	0.4802	0.3477	0.6544	0.5548
WSDDN [22]	0.0906	0.3968	0.3781	0.2016	0.0025	0.1218	0.0057	0.0065	0.1188	0.0490	0.4235	0.0466	0.0106	0.0070	0.6303	0.0395	0.0606	0.0051	0.0455	0.0114	0.1326
OICR [23]	0.0870	0.2826	0.4405	0.1822	0.0130	0.2015	0.0009	0.0065	0.2989	0.1380	0.5739	0.1066	0.1106	0.0909	0.5929	0.0710	0.0068	0.0014	0.0909	0.0041	0.1650
PCL [27]	0.2152	0.3519	0.5980	0.2349	0.0295	0.4371	0.0012	0.0090	0.0149	0.0288	0.5636	0.1676	0.1105	0.0909	0.5762	0.0909	0.0247	0.0012	0.0455	0.0455	0.1819
MELM [28]	0.2814	0.0323	0.6251	0.2872	0.0006	0.6251	0.0021	0.1309	0.2839	0.1515	0.4105	**0.2612**	0.0043	0.0909	0.0858	0.1502	0.2057	0.0981	0.0004	0.0053	0.1866
DCL [32]	0.2089	0.2270	0.5421	0.1150	0.0603	0.6101	0.0009	0.0107	0.3101	0.3087	0.5645	0.0505	0.0265	0.0909	0.6365	0.0909	0.1036	0.0002	0.0727	0.0079	0.2019
PCIR [38]	**0.3037**	0.3606	0.5422	0.2660	**0.0909**	0.5859	0.0022	0.0965	**0.3618**	0.3259	**0.5851**	0.0860	**0.2163**	**0.1209**	**0.6428**	0.0909	0.1362	0.0030	0.0909	0.0752	0.2492
TCANet [35]	0.2513	0.3084	**0.6292**	**0.4000**	0.0413	**0.6778**	**0.0807**	0.2380	0.2989	0.2234	0.5385	0.2484	0.1106	0.0909	0.4640	0.1374	**0.3098**	0.0147	**0.0909**	0.0100	0.2582
Baseline [35]	0.1627	0.4566	0.5704	0.2631	0.0734	0.6353	0.0019	**0.3472**	0.2763	0.3764	0.4462	0.1149	0.0939	0.0909	0.1456	**0.1784**	0.0578	0.1129	0.0303	0.0169	0.2226
Ours	0.2057	**0.6241**	0.6265	0.2354	0.0759	0.6462	0.0022	0.3452	0.3062	**0.5538**	0.5270	0.1757	0.0685	0.0909	0.5159	0.1543	0.0169	**0.1441**	0.0141	**0.0916**	**0.2710**

TABLE VII

PERFORMANCE COMPARISONS (CORLOC) AMONG DIFFERENT METHODS ON THE DIOR TRAINVAL SET

Methods	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	CorLoc
WSDDN [22]	0.0572	0.5988	0.9424	0.5594	0.0492	0.2340	0.0103	0.0679	0.4452	0.1275	0.8990	0.0545	0.1000	0.2296	**0.9854**	0.7961	0.1506	0.0345	0.1156	0.0322	0.3244
OICR [23]	0.1598	0.5145	0.9477	0.5579	0.0355	0.2389	0.0000	0.0482	**0.5668**	0.2242	**0.9141**	0.1818	0.1870	0.3180	0.9828	0.8129	0.0745	0.0122	0.1583	0.0198	0.3477
PCL [27]	0.6114	0.4686	0.9539	0.6361	0.0732	0.9507	0.0021	0.0571	0.0514	0.5077	0.8939	0.4212	0.1978	**0.3794**	0.9793	0.8065	0.1377	0.0020	0.1050	**0.0694**	0.4152
MELM [28]	0.7698	0.2894	0.9266	0.6301	0.1300	0.9009	0.0021	0.1696	0.3788	0.4462	0.8808	**0.4939**	0.1565	0.2819	0.9828	0.8297	0.2275	0.1034	0.0462	0.0223	0.4334
TCANet [35]	0.8158	0.5133	**0.9617**	**0.7345**	0.0503	0.9469	**0.1589**	0.3279	0.4595	0.4856	0.8526	0.3891	**0.2017**	0.3063	0.8459	**0.9146**	0.5628	0.0379	0.1045	0.0125	0.4841
Ours	**0.9120**	**0.6937**	0.9548	0.6752	**0.1888**	**0.9778**	0.0021	**0.7054**	0.5432	**0.5143**	0.8828	0.4803	0.0228	0.3356	0.1411	0.8335	**0.6559**	**0.1988**	**0.1641**	0.0285	**0.4942**

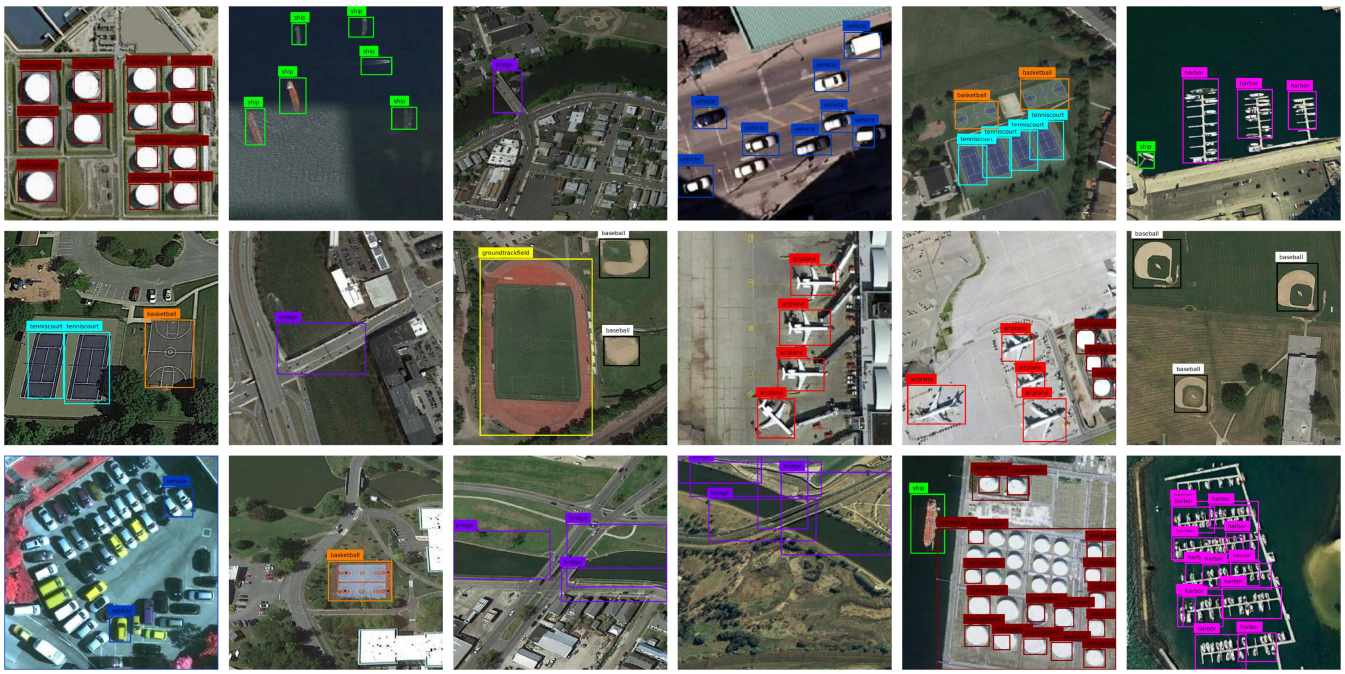


Fig. 4. Sample detection results on the NWPU VHR-10.v2 test split. The first two rows exhibit corrected cases and different colors in rectangle indicate different classes. The last row shows the failed detection results.

E. Complexity Analyses

We further added extra complexity analyses for SAENet in terms of speed versus. All experiments are constructed on ubuntu16.04, NVIDIA RTX 2080 Ti GPU. Compared with our baseline work, the computational efficiency drops from 3.79 frames per second (FPS) to 2.89 FPS. The additional

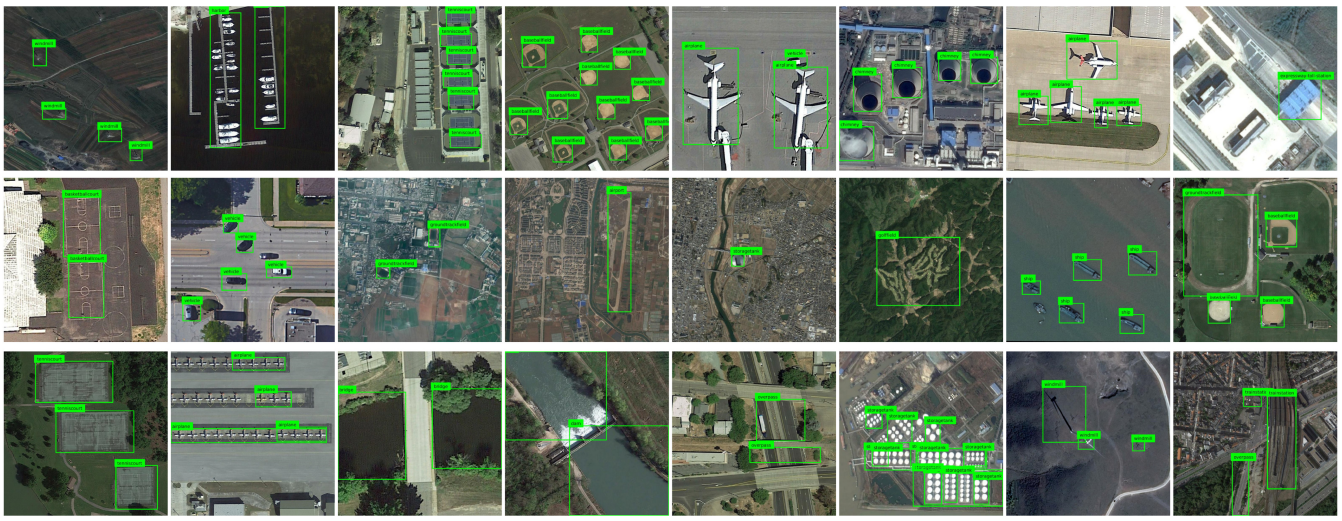


Fig. 5. Example results by SAENet on the DIOR test split. The first two rows show corrected predictions. The last row denotes the failure cases.

calculations are mainly introduced by two ROI-pooling operations. Although the baseline work is slightly faster than our SAENet (3.79 FPS versus 2.89 FPS), its accuracy is reduced by 4.84%. In summary, we achieve a better tradeoff between accuracy and speed compared with the baseline methods.

F. Qualitative Results

Some detection results are visualized in Figs. 4 and 5 to qualitatively analyze the effectiveness of the proposed SAENet on the NWPU VHR-10.v2 and DIOR datasets, respectively. As illustrated in Figs. 4 and 5, the first two rows and third row, respectively, exhibit some successful and failure cases of SAENet. One can see that the proposed method can effectively handle and accurately cover the objects. However, there still remains a challenge to address the problem of small objects and the coexisting scene between objects and special background. The visualization results further validate the effectiveness of the proposed method. As analyzed in comparisons with the state-of-the-arts, the main reason for failing to detect individual classes is that the detection model is predominated by the more arresting and large-scale special coexisting background, such as rivers, reservoirs, and roads. In the future, we will aim at alleviating the aforementioned challenges by introducing causal intervention.

V. Conclusion

In this article, we construct an SAENet to capture the complementary and consistent visual patterns for WSOD in RSIs. Specifically, we first introduce an ADA block. It encourages the detector model to pursue the entire object via adaptively hiding the most discriminative parts and meanwhile highlighting the instance-related regions. Besides, an SSTE mechanism is applied on the positive instance from various transformed images to supply extra self-supervisions. They enforce detection model predicting from transformed images to be consistent. Accordingly, more consistent and exclusive features can be learned to facilitate a more robust object detector. Comprehensive experiments on the challenging NWPU VHR-10.v2 and DIOR datasets demonstrate the superiority of the proposed method in both quantitative and qualitative ways. We significantly boost the state-of-the-art performance.

References

[1] D. Hong *et al.*, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021.

[2] X. Yao, J. Han, G. Cheng, X. Qian, and L. Guo, "Semantic annotation of high-resolution Satellite images via weakly supervised learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3660–3671, Jun. 2016.

[3] P. Zhou, J. Han, G. Cheng, and B. Zhang, "Learning compact and discriminative stacked autoencoder for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4823–4833, Jul. 2019.

[4] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.

[5] S. Zhu, T. Yang, and C. Chen, "Revisiting street-to-aerial view image geo-localization and orientation estimation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 756–765.

[6] X. Yao, J. Han, D. Zhang, and F. Nie, "Revisiting co-saliency detection: A novel approach based on two-stage multi-view spectral rotation co-clustering," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3196–3209, Jul. 2017.

[7] T. Tang, S. Zhou, Z. Deng, H. Zou, and L. Lei, "Vehicle detection in aerial images based on region convolutional neural networks and hard negative example mining," *Sensors*, vol. 17, no. 2, p. 336, 2017.

[8] C. Li, T. Yang, S. Zhu, C. Chen, and S. Guan, "Density map guided object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 190–191.

[9] X. Wu, D. Hong, J. Chanussot, Y. Xu, R. Tao, and Y. Wang, "Fourier-based rotation-invariant feature boosting: An efficient framework for geospatial object detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 302–306, Feb. 2020.

[10] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5966–5978, Jul. 2021, doi: 10.1109/TGRS.2020.3015157.

[11] Y. Shen *et al.*, "Efficient deep learning of nonlocal features for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6029–6043, Jul. 2021.

[12] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[13] T. Yang, S. Zhu, and C. Chen, "GradAug: A new regularization method for deep neural networks," 2020, *arXiv:2006.07989*. [Online]. Available: http://arxiv.org/abs/2006.07989

[14] W. Yu, S. Zhu, T. Yang, C. Chen, and M. Liu, "Consistency-based active learning for object detection," 2021, *arXiv:2103.10374*. [Online]. Available: http://arxiv.org/abs/2103.10374

[15] G. Cheng, J. Han, P. Zhou, and D. Xu, "Learning rotation-invariant and Fisher discriminative convolutional neural networks for object detection," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 265–278, Jan. 2019.

[16] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.

[17] X. Wu, D. Hong, J. Tian, J. Chanussot, W. Li, and R. Tao, "ORSIm detector: A novel object detection framework in optical remote sensing imagery using spatial-frequency channel features," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 5146–5158, Jul. 2019.

[18] W. Yu, T. Yang, and C. Chen, "Towards resolving the challenge of long-tail distribution in UAV images for object detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3258–3267.

[19] Z. Zou and Z. Shi, "Random access memories: A new paradigm for target detection in high resolution aerial remote sensing images," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1100–1111, Mar. 2018.

[20] K. Li, G. Cheng, S. Bu, and X. You, "Rotation-insensitive and context-augmented object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2337–2348, Apr. 2018.

[21] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020.

[22] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2846–2854.

[23] P. Tang, X. Wang, X. Bai, and W. Liu, "Multiple instance detection network with online instance classifier refinement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3059–3067.

[24] Z. Jie, Y. Wei, X. Jin, J. Feng, and W. Liu, "Deep self-taught learning for weakly supervised object localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1377–1385.

[25] V. Kantorov *et al.*, "ContextLocNet: Context-aware deep network models for weakly supervised localization," in *Proc. Eur. Conf. Comput. Vis*, 2016, pp. 350–365.

[26] W. Ren, K. Huang, D. Tao, and T. Tan, "Weakly supervised large scale object localization with multiple instance learning and bag splitting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 405–416, Feb. 2016.

[27] P. Tang *et al.*, "PCL: Proposal cluster learning for weakly supervised object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 176–191, Jan. 2020.

[28] F. Wan, P. Wei, J. Jiao, Z. Han, and Q. Ye, "Min-entropy latent model for weakly supervised object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1297–1306.

[29] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3325–3337, Jun. 2015.

[30] G. Cheng *et al.*, "High-quality proposals for weakly supervised object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 5794–5804, 2020.

[31] D. Li, J.-B. Huang, Y. Li, S. Wang, and M.-H. Yang, "Weakly supervised object localization with progressive domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3512–3520.

[32] X. Yao, X. Feng, J. Han, G. Cheng, and L. Guo, "Automatic weakly supervised object detection from high spatial resolution remote sensing images via dynamic curriculum learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 675–685, Jan. 2021.

[33] G. Cheng, J. Han, L. Guo, and T. Liu, "Learning coarse-to-fine sparselets for efficient object detection and scene classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1173–1181.

[34] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "DropBlock: A regularization method for convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 10727–10737.

[35] X. Feng, J. Han, X. Yao, and G. Cheng, "TCANet: Triple context-aware network for weakly supervised object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6946–6955, Aug. 2021.

[36] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Apr. 2013.

[37] P. Zhou, G. Cheng, Z. Liu, S. Bu, and X. Hu, "Weakly supervised target detection in remote sensing images based on transferred deep features and negative bootstrapping," *Multidimensional Syst. Signal Process.*, vol. 27, no. 4, pp. 925–944, 2016.

[38] X. Feng, J. Han, X. Yao, and G. Cheng, "Progressive contextual instance refinement for weakly supervised object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 8002–8012, Nov. 2020.

[39] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 577–593.

[40] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," 2018, *arXiv:1803.07728*. [Online]. Available: http://arxiv.org/abs/1803.07728

[41] X. Pan *et al.*, "Self-supervised feature augmentation for large image object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 6745–6758, 2020.

[42] K. Wang, X. Yan, D. Zhang, L. Zhang, and L. Lin, "Towards human-machine cooperation: Self-supervised sample mining for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1605–1613.

[43] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen, "Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12275–12284.

[44] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11534–11542.

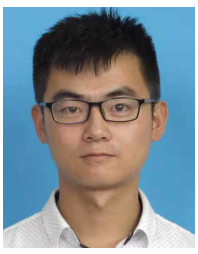
[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[46] S. Ren *et al.*, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural. Inf. Process. Syst.*, 2015, pp. 91–99.

[47] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural. Inf. Process. Syst.*, 2012, pp. 1097–1105.

[48] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[49] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.



Xiaoxu Feng received the B.E. degree from Inner Mongolia University, Hohhot, China, in 2017. He is pursuing the Ph.D. degree with Northwestern Polytechnical University (NWPU), Xi'an, China.

His research interests include computer vision and remote sensing image processing, especially on object detection and scene classification.

Xiwen Yao (Member, IEEE) received the B.S. and Ph.D. degrees from Northwestern Polytechnical University (NWPU), Xi'an, China, in 2010 and 2016, respectively.

He is an Associate Professor with NWPU. His research interests include computer vision and remote sensing image processing, especially on fine-grained image classification and object detection.

Gong Cheng (Member, IEEE) received the B.S. degree from Xidian University, Xi'an, China, in 2007, and the M.S. and Ph.D. degrees from Northwestern Polytechnical University (NWPU), Xi'an, in 2010 and 2013, respectively.

He is a Professor with NWPU. His main research interests are computer vision and pattern recognition.

Jungong Han is a Full Professor and the Chair in computer science with Aberystwyth University, Aberystwyth, U.K. He has authored or coauthored more than 180 articles, including 40+ IEEE Transactions, and 40+ A* conference papers. His research interests span the fields of video analysis, computer vision, and applied machine learning.

Junwei Han (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in pattern recognition and intelligent systems from Northwestern Polytechnical University (NWPU), Xi'an, China, in 1999, 2001, and 2003, respectively.

He is a Professor with NWPU. His research interests include computer vision and brain-imaging analysis.