TCANet: Triple Context-Aware Network for Weakly Supervised Object Detection in Remote Sensing Images

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

***Abstract*— Weakly supervised object detection (WSOD) in remote sensing images (RSI) plays an essential role in RSI understanding applications. Currently, predominant works are inclined to first activate the most discriminative region and then pursue the whole object by analyzing the context information of the activated region. However, the most discriminative region usually only covers a small crucial part. Besides, many same-class instances often appear in adjacent locations. In such a case, treating proposals of large spatial overlap as the same-class instances not only introduces potential ambiguities but also misleads the detection model to recognize multiple adjacent instances as one object instance. To address these challenges, a novel triple context-aware network (TCANet) is proposed to learn complementary and discriminative visual patterns for WSOD in RSIs. Specifically, a global context-aware enhancement (GCAE) module is first designed to activate the features of the whole object by capturing the global visual scene context. Then, a dual-local context residual (DLCR) module is further developed to capture the instance-level discriminative cues by leveraging the semantic discrepancy of the local context. Furthermore, an effective adaptive-weighted refinement loss is integrated into the DLCR module to reduce the ambiguities in the label propagating process. The collaboration of GCAE and DLCR formulates a unique TCANet that can be learned in an end-to-end manner. Comprehensive experiments are carried out on the challenging NWPU VHR-10.v2 and DIOR data sets. We achieve a 58.8% mAP and a 25.8% mAP on the NWPU VHR-10.v2 and DIOR data sets, respectively, which both significantly outperform the state of the arts.**

***Index Terms*— Context-aware network, remote sensing images (RSIs), weakly supervised object detection (WSOD).**

Manuscript received May 17, 2020; revised August 25, 2020; accepted October 7, 2020. Date of publication October 26, 2020; date of current version July 22, 2021. This work was supported in part by the National Science Foundation of China under Grant 61701415, Grant 62071388, and Grant 61772425; in part by the Fundamental Research Funds for the Central Universities under Grant 3102019ZDHKY05; in part by the China Postdoctoral Science Foundation under Grant 2018T111094 and Grant 2017M620468; in part by the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2018JQ6025; 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, Junwei Han, and Gong Cheng are with the School of Automation, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: junweihan2010@gmail.com).

Xiwen Yao is with the School of Automation, Northwestern Polytechnical University, Xi'an 710072, China, and also with the Qingdao Research Institute, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: yaoxiwen517@gmail.com).

Digital Object Identifier 10.1109/TGRS.2020.3030990

I. Introduction

OBJECT detection is a fundamental task in the field of remote sensing image (RSI) understanding and plays a key role in many practical applications [1]–[11]. Recently, object detection has achieved breakthrough improvements, which can be attributed to the development of convolutional neural networks (CNNs) [12], [13] and the availability of abundant data sets with subtle manually labeled annotations [14]–[16]. Nevertheless, with the advent of the era of high resolution of remote sensing, the quantity and quality of RSIs increase exponentially so that manually labeling subtle boxes for each instance becomes more laborious and time-consuming. On the contrary, collecting weakly supervised annotations, such as partially supervision, inexact supervision, or inaccurate supervision of instance, is effortless. In this article, we aim to address the object detection problem in RSIs under inexact supervision settings where only image-level annotations can be used to declare the presence or absence of an object category.

Currently, most of the previous works [3], [17]–[29] tackle the weakly supervised object detection (WSOD) problem under an inexact supervision paradigm. These works mainly follow a two-stage manner that first decomposes images into a series of proposals and then iteratively selects the most contributing proposal as the pseudoinstance-level label to train object detectors under multiple instance learning (MIL) constraints [30]. Among them, Bilen and Vedaldi [17] designed a groundbreaking weakly supervised deep detection network (WSDDN) that is the first to integrate the MIL into the WSOD in an end-to-end paradigm. Based on that, a host of research works have been proposed to boost the performance of WSOD via seeking better initialization models [20], [22], [23] or learning strategies [18], [31]–[34]. Our work follows the aforementioned strategy of learning an end-to-end MIL network.

Although remarkable progress has been achieved in the aforementioned methods, there are still two major challenges. First of all, most of the popular WSOD methods are non-convex optimal processes, which leads to the model that tends to discover the most discriminative part rather than the whole object. What is worse, RISs always contain the large-scale cluttered background. Thus, the aforementioned methods

still have difficulties in accurately discovering all the whole objects. This is one of the main reasons why the performance of WSOD is inferior to the fully supervised object detection. To address these issues, most of the popular algorithms [22], [24], [26] motivated by OICR [23] are eager to mine the whole object by propagating image-level label from the most discriminative part to the surrounding regions that have large spatial overlap with it.

Despite its effectiveness, this strategy [23] introduces potential ambiguities by signing different parts of objects as the same label simultaneously. Besides, due to the unavailability of instance-level ground truth, all WSOD methods have the trouble in separating spatially adjacent instances. Yet, many instances of the same class always exist in the adjacent locations and even keep aligned in RSIs, such as tennis court and basketball court. Unfortunately, the detector usually mistakes multiple objects for one goal. Obviously, the aforementioned methods based on OICR [23] exacerbate this problem to some extent. This is another challenge for WSOD in RSIs.

To tackle the first challenge, we elaborately design a novel global context-aware enhancement (GCAE) module to highlight the whole object feature via a self-attention mechanism. More specifically, we first generate the global context feature by explicitly aggregating the features at all positions with a weighted average. Next, the channelwise interdependencies in the obtained global context are captured by conducting an excitation operation. Finally, the instance-containing features are effectively enhanced by fusing the global context feature into features of all positions so that the instance-containing regions are activated and the background regions are suppressed. Accordingly, the module can succeed in discovering the high-quality instances to train more robust object detectors.

To address the second challenge, a dual-local context residual (DLCR) module is further designed to precisely find the object boundaries by leveraging semantic discrepancy of local context. Specifically, the affinity context is introduced to encourage the semantic information of the predicted object region to be similar to its local inside context. Meanwhile, the ambient context is introduced to facilitate the predicted object region different from its outer context. Collaborating with GCAE, increasing the residue of the affinity and ambient contexts not only can highlight the whole instance but also highlight the discrepancy between class-specific instance response and its surroundings. Due to that, the instances that appear in the adjacent locations can be easily distinguished. Accordingly, the low-quality discriminative parts are significantly suppressed. Based on that, an adaptive-weighted refinement loss is introduced to further alleviate ambiguities in the refinement by simultaneously considering the confidence and spatial relationship of region.

The cooperation of GCAE and DLCR formulates a unique end-to-end triple context-aware network (TCANet) where GCAE aims at learning complementary visual patterns and DLCR endeavors to capture discriminative visual patterns. Extensive experiments on the challenging NWPU VHR-10.v2 [15] and DIOR [14] data sets testify the superiority of our method. The main contributions of the proposed TCANet are summarized as follows.

1) We design a novel GCAE module, which can effectively highlight the high-quality whole object by capturing the global context of a visual scene.
2) We introduce a DLCR module collaborating with an adaptive-weighted refinement loss, which can facilitate the object outstanding from its surroundings and significantly reduce the ambiguities at the same time.
3) Comprehensive experimental results clearly testify that the proposed TCANet achieves the state-of-the-art results.

II. Related Work

A. Weakly Supervised Learning (WSL)

This, aiming to directly build the prediction model under partially supervision, inexact supervision or inaccurate supervision settings, has been extensively studied for both field of natural scene and remote sensing as its low cost. Of late, many works [35]–[40] leveraged partially or inaccurate label to address the multilabel classification problem and achieve good performance. For example, Durand *et al.* [39]. introduced a scalable method to learn a ConvNet with partial labels. Although encouraging results have been achieved for multilabel classification task under partially or inaccurate supervision, such supervision cannot be directly used to address object detection problem as this process require solving an optimization problem with the training set in memory. Thus, most of the methods aim to address the WSOD problem under an inexact supervision paradigm where only image-level annotations are employed to declare the presence or absence of an object category. Among them, many methods [17]–[28] in natural scene images leveraged MIL to tackle the WSOD problem and achieved impressive performance. Yet, these approaches cannot be directly applied to RSIs. Obviously, it is more challenging to accurately detect objects in RSIs under WSL manner due to its large-scale cluttered background. Currently, Han *et al.* [28] first attempted to address the WSOD problem in RSIs via iteratively leveraging positive samples that heuristically mined and refined from the negative data to learn object detector. Inspired by that, the work [41] combined a negative bootstrapping scheme with iterative learning to address WSOD in RSIs. Besides, Yao *et al.* [42] designed a dynamic curriculum learning strategy to alleviate the local minimum problem of WSOD in RSIs via simulating the easy-to-difficult learning process of human cognition.

The works [23], [28] are most related to us. Compared with them, we construct a novel end-to-end TCANet that significantly highlights the response of the whole object. Besides, an adaptive-weighted refinement loss is further introduced to reduce the ambiguities in the instance refinement.

B. Contextual Information in Object Detection

Contextual information has been extensively studied in the field of computer vision. It can be applied to enhance the feature representation. Of late, comprehensive studies [17],

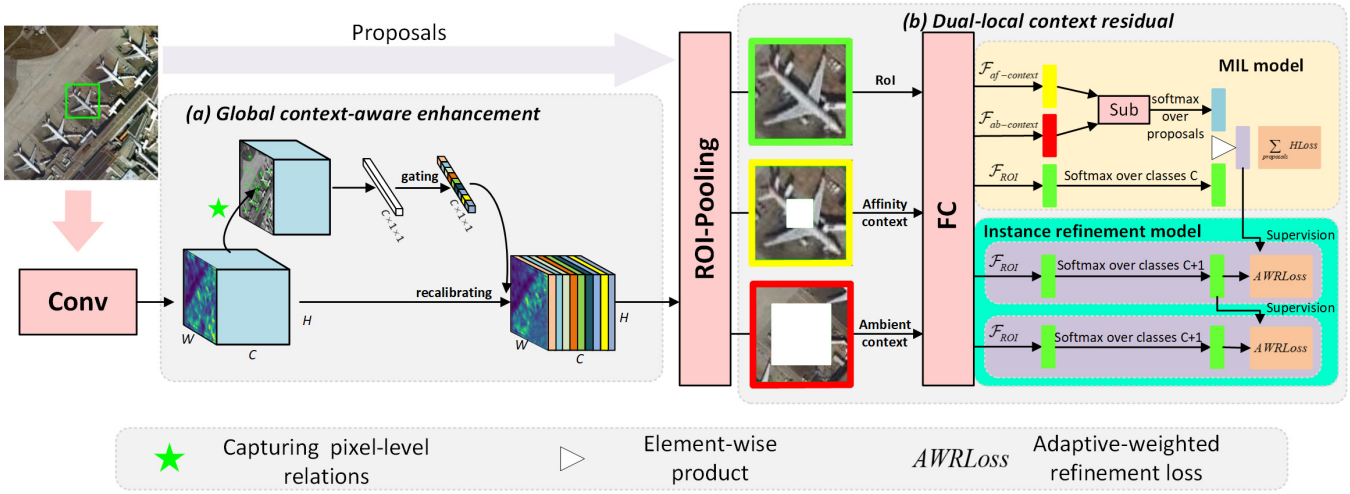


Fig. 1. Illustration of the triple context-aware architecture. To tackle the challenges of nonconvex optimal process and coexisting of many same-class instances in the adjacent locations in RSIs, (a) GCAE can highlight the whole object features via capturing global pixel-level relations and (b) DLCR can capture instance-level discriminative cues so that instance appearing in adjacent locations can be easily distinguished by learning semantic discrepancy of local context.

[19], [43]–[47] successfully leveraged visual context information to facilitate the detection performance. For instance, Bilen and Vedaldi [17] gathered context information by constructing multidirectional recurrent neural network. Li *et al.* [15] aimed to tackle the challenging of appearance ambiguity in RSIs by constructing a local-contextual feature fusion network. Chen *et al.* [44] employed rich context information around the selected region to improve the detection performance. However, the aforementioned methods need to collect subtle annotations and it is very time-consuming and labor-intensive. To this end, some recent works aim to capture the context information under weakly supervised paradigm. Wei *et al.* [24] took full use of the surrounding segmentation context to boost mining the high-quality candidates. Kantorov *et al.* [19] leveraged the surrounding context regions to improve the localization performance. Besides, the approach [48] is introduced to mine the instances with the same class via making the best of both local and global contexts.

In general, there is few attention about how to exploit the context information under weakly supervised manner. Besides, existing weakly supervised methods leverage complicated segmentation network to capture the contextual information or obtain the contextual information from the fixed area. Compared with them, we train the detection model under an inexact supervision paradigm and make full use of the global context information at all positions to enhance mining the high-quality whole object. Besides, the semantic discrepancy of the local context is further leveraged to capture the discriminative visual patterns.

III. Approach

Fig. 1. shows the architecture of the proposed TCANet. TCANet aims to highlight the high-quality object and further distinguish the instances appearing in the adjacent locations via learning complementary and discriminative visual patterns under a weakly supervised paradigm. Specifically, the GCAE module first generates the long-range dependence by capturing the global context of a visual scene and then captures the channelwise interdependencies in the global context so that the instance-specific features can be effectively enhanced. After that, the DLCR module is leveraged to facilitate the object outstanding from its surroundings by increasing the residue of affinity context and ambient context. Meanwhile, an adaptive-weighted refinement loss is further introduced to reduce the ambiguities by simultaneously considering confidence and spatial relationship of region.

A. Basic Weakly Supervised Object Detection Network

Currently, many popular WSOD methods [22]–[24], [26] have clearly demonstrated that generating the pseudoinstance-level annotation for refining the corresponding classifier can effectively boost the performance of WSOD. Inspired by that, we select OICR [20] as our baseline network for its effectiveness.

Given an input image $I \in \mathbb{X}$ and its proposals $\mathcal{B}_x$, this can be generated in [49]. Let $\mathcal{Y} = [Y_1, \ldots, Y_c, \ldots, Y_C] \in \{0, 1\}$ and $\mathbb{Y} = [Y_1, \ldots, Y_c, \ldots, Y_C] \in \{-1, 1\}$ denote its binary-level label to declare whether class-specific object exists in this image. First, a host of fixed-sized features corresponding to each proposal are generated by employing the region-of-interest (RoI) pooling [50]. Then, proposal features are branched into two parallel MIL model and instance refinement model. In the MIL model, the objectness $\Psi_{\text{cr}}^{\text{cls}}$ and spatial recognition score $\Psi_{\text{cr}}^{\text{det}}$ of each proposal are evaluated by feeding the proposal feature into two softmax layers along different directions, where $\Psi_{\text{cr}}^{\text{cls}}$ is the probability of the rth proposal belonging to class c and $\Psi_{\text{cr}}^{\text{det}}$ indicates the contribution of the rth proposal to image being classified to class c. The proposal scores are produced by the elementwise product $\Psi_{\text{cr}}^{\text{cls}} \odot \Psi_{\text{cr}}^{\text{det}}$. Thus, the image score of the Cth class is generated through summation over all proposals. The MIL model is learned by employing the multiclass cross-entropy loss, which is denoted as

$$\text{Loss}_{\text{MIL}} = -\sum_{c=1}^{C} \{\mathcal{Y} \log \Phi^c + (1 - \mathcal{Y}) \log(1 - \Phi^c)\}. \quad (1)$$

In the instance refinement model, the proposal features are mapped as $\{C+1\}$-dimensional score vector $S_c^{\mathcal{B}}$, and $\{C+1\}$ denotes C different object classes and background. Due to the limitation of MIL model, the top-scoring proposal always only covers the discriminative part instead of the whole object. Thus, the instance refinement model first treats the highest scoring proposal as pseudoinstance-level annotation $\mathcal{Y}^c$ and then labels its surrounding proposals to be the same with it and others are labeled as background. Benefiting from that, the weighted softmax loss function is leveraged to train the object detector via treating each labeled proposal as the instance-level supervision and formulates as

$$\text{Loss}_{\text{ref}} = -\frac{1}{|R|}\sum_{r=1}^{|R|}\sum_{c=1}^{C+1}\lambda\mathcal{Y}^c \log S_c^{\mathcal{B}}. \tag{2}$$

B. Global Context-Aware Enhancement

In this section, we aim to strengthen the features of whole object by getting the utmost out of the global context so that the entire object can be easily discovered.

Let $\mathcal{F} = \{x_i\}_{i=1}^{N}$ represent the feature map of an input image, where N denotes the number of the pixels. First, the global context map is generated by estimating each pixel-level relations between the query position and all positions, which is formulates as

$$f(x_i, x_j) = e^{x_i^T x_j} \tag{3}$$

where i is the query position and $f(\cdot)$ is a pairwise function to compute the relationship between position i and j. Here, we adopt the embedded Gaussian function. Thus, the global contextual attention map can be obtained by weighted aggregating the relations of all position together, as in

$$\mathcal{G}(x_i) = \sum_{\forall j}\frac{f(x_i, x_j)}{\mathcal{C}(x)}\cdot\mathcal{U}(x_j) \tag{4}$$

where $\mathcal{C}(\cdot)$ is a normalization factor and $\mathcal{U}(\cdot)$ is a unary function that represents the input signal at the position j. Following the fact that the attention maps are almost the same at different query positions, we obtain the arbitrary query position global context and share it for all positions. Thus, as shown in Fig. 2(a), the matrix multiplication can be applied to simply realize the pairwise computation in (3). Compared with the traditional spatial attention module that only enriches the features with local neighborhood, we can obtain the global context but not require extra steps, such as recurrent CNN [43].

Accordingly, features of all positions are enriched by making full use of the aforementioned global context module. However, simply leveraging the global context module to enrich the feature may activate the background features to some extent. Thus, we hope to effectively capture the global context to highlight instance-specific regions and suppress the background regions by selectively emphasizing the informative context and hiding less useful ones so that they can help to better discover the whole instance.

To this end, the channelwise activations are captured by employing an excitation operation. It can be done by a simple gating mechanism where two conv layers with 1×1 filter

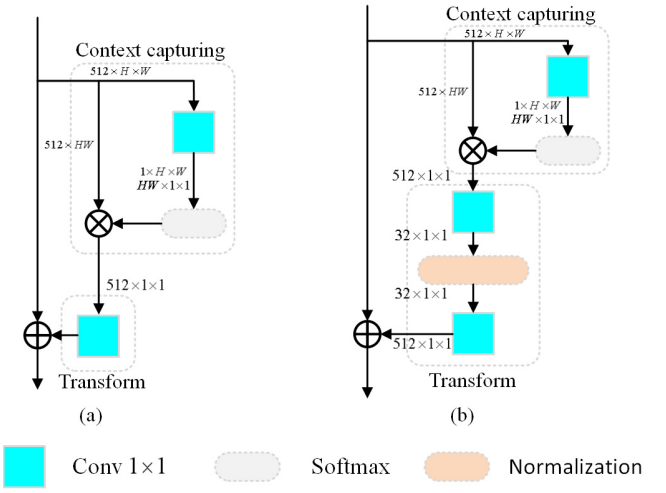


Fig. 2. Illustration of global context modules architecture. (a) Global contextual attention module aims to capture pixel-level relations between the query position and all positions. (b) GCAE module can highlight instance-containing regions and suppress the background regions by adaptively capturing the global context of a visual scene.

formulate a bottleneck. Accordingly, the class-specific features are adaptively emphasized by recalibrating global context features with the learned activation vector. The collaboration of context capturing model and excitation operation formulated our GCAE module, which is shown in Fig. 2(b) and defined as

$$G(x) = \mathcal{F}(x) + W_2\delta(W_1\mathcal{G}(x)) \tag{5}$$

where $\delta(\cdot)$ is the layer normalization function "$W_1 \in \mathbb{R}^{(C/r)\times C}$ and $W_2 \in \mathbb{R}^{C\times(C/r)}$" are the parameter of the dimensionality reduction layer and the dimensionality-increasing layer, respectively.

By leveraging GCAE, each position in the instance-specific region can be effectively highlighted by fully capturing the global context of a visual scene.

C. Dual-Local Context Residual

Benefitted from GCAE, features of the whole instance are effectively highlighted by capturing the global context information. However, many same-class instances frequently appear in the adjacent locations and even keep aligned in RSIs. Simply propagating label from positive instance to its surrounding proposals introduces potential ambiguities. Consequently, the DLCR is further proposed to learn instance-level discriminative cues so that the instances appearing in the adjacent locations can be easily distinguished.

To this end, we introduce the affinity-context block and ambient-context block where the affinity context is designed to highlight the regions that have similar semantic information with their inner regions and the ambient context is proposed to activate regions that are outstanding from their outer context. In other words, the larger the semantic information gap between the region's affinity context and ambient context exists, much better the object is covered. As shown in Fig. 1, the affinity context and ambient context are obtained by feeding the inner region and outer region around each proposal

into the RoI-Pooling where the features of its central area are erased. Next, the extracted features are independently passed through shared fully connected layers to generate the corresponding context feature vectors $\mathcal{F}^r_{afc}$ and $\mathcal{F}^r_{abc}$. To better leverage them, we substitute the detection stream in the MIL model with the semantic information residual between the region's affinity context and ambient context. Therefore, the spatial recognition score of each proposal is generated by performing

$$\Psi^{\text{det}}_{\text{cr}} = \frac{e^{(\mathcal{F}^{\text{cr}}_{afc} - \mathcal{F}^{\text{cr}}_{abc})}}{\sum_{r=1}^{R} e^{(\mathcal{F}^{\text{cr}}_{afc} - \mathcal{F}^{\text{cr}}_{abc})}}. \tag{6}$$

Then, proposal scores are generated by the elementwise product $\Phi_{\text{cr}} = \Psi^{\text{cls}}_{\text{cr}} \odot \Psi^{\text{det}}_{\text{cr}}$. During training, the multiclass cross-entropy function is replaced with hinge loss function to train the MIL model, which is given by

$$\text{Loss}_{\text{MIL}} = \frac{1}{C \times R} \sum_{c=1}^{C} \sum_{r=1}^{R} \max(0, 1 - \mathbb{Y} \cdot \Phi_{\text{cr}}). \tag{7}$$

Accordingly, the scores of low-quality proposals that only cover the small part or background are significantly suppressed. Meanwhile, high-quality proposal is outstanding from the surroundings. Inspired by focal loss [51], we aim to reduce the ambiguities in the refinement by recalibrating the refinement loss weights. Similar to OICR [23], the top-scoring proposal is first selected as pseudoinstance-level ground truth. Then, the proposals that have large spatial overlap with it are defined as positive instances and others are considered as backgrounds.

Although directly propagating label from the top-scoring proposal to its surrounding regions can boost the detection performance, it will introduce potential ambiguities that confuse the model by signing different parts of objects as the same label simultaneously. Meanwhile, simply treating proposals without spatial overlap with the top-scoring proposal as background may sign the true positive instance as background when more than one instance of the same class exists in an image. To alleviate these ambiguities, we first generate the weights of positive instances by simultaneously considering its confidence and spatial relationship, which is formulated as

$$w_{\text{cr}} = \frac{\sum_{\mathbb{I}(\mathcal{B}^r, \mathcal{B}^*) > \varepsilon} (e^{-a(1-\mathbb{I}(\mathcal{B}^r, \mathcal{B}^*))} \cdot S^{\mathcal{B}^r}_c)}{S^{\mathcal{B}^r}_c \cdot \sum_{\mathbb{I}(\mathcal{B}^r, \mathcal{B}^*) > \varepsilon} e^{-a(1-\mathbb{I}(\mathcal{B}^r, \mathcal{B}^*))}} \tag{8}$$

where $\mathbb{I}(\cdot)$ is a function to evaluate the intersect over union (IoU) between the top-scoring proposal $\mathcal{B}^*$ and others. ε is set to 0.5. By leveraging it, when $\mathbb{I}(\mathcal{B}^r, \mathcal{B}^*)$ is large, the Gaussian kernel function $e^{-a(1-\mathbb{I}(\mathcal{B}^r, \mathcal{B}^*))}$ returns a high value; otherwise, it will get a low value. Accordingly, the true positive instances with highly spatial overlap and high confidence are adaptively highlighted by a high loss weight so that ambiguities can be alleviated.

Meanwhile, to prevent instances of the same class from being labeled differently, we only define the proposals with lowly spatial overlap ($0.1 < \text{IoU} < 0.5$) as backgrounds and define the highest score as its loss weight. Besides, we set the loss weights of proposal with minimal spatial overlap (IoU < 0.1) to 0. Accordingly, we modify the refinement loss in (2) as

$$\text{Loss}_{\text{ref}} = -\frac{1}{|R|} \left(\sum_{r \in \mathcal{B}^n_C} w_{\text{cr}} \mathcal{Y}^c \log S^{\mathcal{B}^r}_c + \sum_{m \in \mathcal{B}^n_{C+1}} \lambda \mathcal{Y}^c \log S^{\mathcal{B}^m}_c \right) \tag{9}$$

where $r \in \mathcal{B}^n_C$ denotes positive instances and $m \in \mathcal{B}^n_{C+1}$ denotes negative instances.

Collaborating with context residual module and adaptive-weighted refinement loss tolerates the low-quality proposals to have low confidence so that the ambiguities are reduced.

IV. Experiments

In this section, comprehensive experiments (i.e., data sets, evaluation metrics, and implementation details) are first provided in detail. Next, ablation experiments are elaborately designed to show the contributions of each key component. Finally, we also provide quantitative comparisons with the existing advanced works and qualitative results.

A. Data Sets and Evaluation Metrics

Comprehensive experiments are elaborately conducted on the publicly available challenging NWPU VHR-10.v2 data set and DIOR data set. The NWPU VHR-10.v2 data set contains 1172 images and 2775 instances from ten object categories with the size of 400×400 pixels. In the experiments, the NWPU VHR-10.v2 data set is divided into 75% for training that covers 879 images and 25% for testing that covers 293 images. The DIOR is a newly more challenging data set that contains 23 463 images with the size of 800×800, including 192 472 instances of 20 object categories. The DIOR data set is divided into train, test, and valuation sets where the train set and valuation set (i.e., 11 725 images) are applied for training and the test set (i.e., 11 738 images) is applied for testing.

For evaluation, we still employ two standard evaluation metrics [i.e., the correct location (CorLoc) and the average precision (AP)] to evaluate the performance of WSOD in RSIs. CorLoc is applied to evaluate the localization accuracy on the training set and AP is used to measure the detection performance on the testing set. The two metrics are performed under the PASCAL VOC criteria where IoU > 0.5 is setting to evaluate the results as positive detections.

B. Implementation Details

For the backbone network, we employ the VGG16 [52] where the penultimate max-pooling layer is removed and its subsequent conv layers are substituted with dilated conv layers to protect the object feature of small-sized instances. Besides, the last max pooling is substituted with our GCAE module and DLCR module. The pretrained model on the ImageNet [53] is applied to initialize the backbone network, and the Gaussian distribution with 0 mean and 0.01 standard deviation is employed to initialize other newly added layers. The number of overall iterations for the NWPU VHR-10.v2 data set and

TABLE I
RESULTS ON THE DIOR DATA SET FOR EACH KEY COMPONENT

Architecture	Baseline	√	√	√			
	Dual-local context residual				√	√	√
Loss function	Weighted softmax loss	√	√		√	√	
	Adaptive-weighted refinement loss			√			√
Global context-aware enhancement			√			√	√
mAP(%)		16.50	20.04	19.07	22.26	25.27	25.82
CorLoc(%)		34.77	42.26	41.68	45.12	46.57	48.41

the DIOR data set is set to 30k and 200k, respectively; 0.001 is set as the initial learning rate for both data sets. The step size is set to 10 000 and 100 000. The mini-batch is set to 2 for SGD. The weight decay and momentum are 0.0005 and 0.9, respectively.

Following the predominant methods, the data are augmented by resizing images into five scales {480, 576, 688, 864, 1200} and horizontally flipping images. Different from OICR [23], the refinement times are fixed as 2. In the DICR, the context region is defined as an internal rectangle and external rectangle with a side ratio of 1.8 where the features of its central area are erased. ε in (9) is 0.5. During testing, duplicated bounding boxes are removed by employing the nonmaxima suppression (NMS) [54] with 30% IoU threshold.

C. Ablation Study

To analyze the effectiveness of our TCANet, comprehensive ablation studies are built to evaluate the contributions of different components on the DIOR data set.

1) Influence of GCAE: We first disclose the contribution of GCAE by integrating it into the OICR [23] framework. As shown in Table I, GCAE largely improves the mAP from 16.50% to 20.04% and CorLoc from 34.77% to 42.26%. It is mainly because GCAE effectively activates the features of high-quality proposals by adaptively highlighting the instance-containing region features and suppresses the features of cluttered backgrounds according to the capture of the long-range dependence. Besides, as we can see, collaborating GCAE with DLCR also brings about improvements of 3.01% (25.27% versus 22.26%) mAP, which further demonstrates that GCAE can effectively drive the WOSD to mine the whole instances.

2) Influence of DLCR: As can be seen in Table I, by exploiting DLCR, the performance for both mAP and CorLoc is significantly improved from 16.50% to 22.26% and 34.77% to 45.12%, respectively. Compared with OICR [23] that simply selects the top-scoring proposal, the DLCR module takes full advantages of the local inner–outer context to highlight the regions that have similar semantic information with its inner region and are outstanding from its outer context. In this way, the scores of low-quality proposals that only cover the small part or background are significantly suppressed and the high-quality proposals are outstanding from their surroundings. Accordingly, the instances appearing in the adjacent locations can be easily distinguished and the ambiguities in the refinement can be reduced.

3) Influence of Adaptive-Weighted Refinement Loss: We further study the contribution of adaptive-weighted refinement loss by substituting the weighted loss with it in the OICR [23] framework. As shown in Table I, adaptive-weighted refinement loss increases the mAP from 16.50% to 19.07 and the Corloc from 34.77% to 41.68%, respectively. Joint GCAE, DLCR, and adaptive-weighted refinement loss formulate our TCANet. As can be seen, the adaptive-weighted refinement loss further boosts the mAP from 25.27% to 25.82%, which fully demonstrates that adaptive-weighted refinement loss achieves much better performance. The main reason is that the OICR model labels different parts of the object with the same weight, which hurts the discriminative power of the detector. On the contrary, our method alleviates ambiguities by simultaneously considering the confidence and spatial relationship. Besides, the loss weights of the proposal with minimal spatial overlap (IoU < 0.1) are set to be 0, which successfully prevents the instances of the same class from being labeled differently.

D. Comparisons With Advanced Works

We report the detection performance for each class and provide comparisons with the existing advanced methods under both weakly supervised learning and fully supervised learning.

Tables II and III quantitatively compare the detection performance of TCANet on the NWPU VHR-10.v2 data set with the other popular methods in terms of AP and CorLoc. We can clearly see that the proposed method achieves the state of the arts with 58.8% mAP and 72.76% CorLoc. Compared with the existing predominate WSOD methods, TCANet significantly outperforms the WSOD, OICR, PCL, and MELM by 23.7%, 24.3%, 19.41%, and 16.53% in terms of mAP, respectively. Meanwhile, comparisons with fully supervised methods clearly show that TCANet further bridges the performance gap between the WSOD and the fully supervised object detection.

Table IV illustrates the object categories of DIOR dataset. Quantitative comparisons with both weakly and fully supervised advanced methods in terms of mAP and CorLoc on the more challenging DIOR data set are also shown in Tables V and VI, respectively. As shown in Table V, TCANet far outperforms WSDDN [17], OICR [23], PCL [22], and MELM [26] by 12.56%, 9.32%, 7.63%, and 7.16% mAP, respectively.

Table VI shows the comparisons in terms of CorLoc. It is obvious that the TCANet brings about an improvement of 15.97%, 13.64%, 6.89%, and 5.07% with WSDDN [17], OICR [23], PCL [22], and MELM [26], respectively.

TABLE II
COMPARISONS IN TERMS OF AP FOR 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
COPD [55]	0.6225	0.6937	0.6452	0.8213	0.3413	0.3525	0.8421	0.5631	0.1643	0.4428	0.5488
Transferred CNN [12]	0.6603	0.5713	0.8501	0.8093	0.3511	0.4552	0.7937	0.6257	0.4317	0.4127	0.5961
RICNN [9]	0.8871	0.7834	0.8633	0.8909	0.4233	0.5685	0.8772	0.6747	0.6231	0.7201	0.7311
RCNN [56]	0.8537	0.8888	0.6278	0.1973	0.9066	0.5823	0.6795	0.7987	0.5422	0.4992	0.6576
Fast RCNN [50]	0.9091	0.9060	0.8929	0.4732	1.0000	0.8585	0.8486	0.8822	0.8029	0.6984	0.8271
Faster RCNN [57]	0.9090	0.8630	0.9053	0.9824	0.8972	0.6964	1.0000	0.8011	0.6149	0.7814	0.8451
RICO [15]	0.9970	0.9080	0.9061	0.9291	0.9029	0.8013	0.9081	0.8029	0.6853	0.8714	0.8712
WSDDN [17]	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[22]	0.2600	0.6376	0.0250	0.8980	**0.6445**	**0.7607**	0.7794	0.0000	0.0130	0.1567	0.3941
MELM[26]	0.8086	0.6930	0.1048	0.9017	0.1284	0.2014	0.9917	0.1710	**0.1417**	0.0868	0.4229
Ours	**0.8943**	**0.7818**	**0.7842**	**0.9080**	0.3527	0.5036	0.9091	**0.4244**	0.0411	**0.2830**	**0.5882**

TABLE III
COMPARISONS IN TERMS OF CORLOC FOR DIFFERENT METHODS ON THE NWPU VHR-10.V2 TRAINVAL SET

Methods	Airplane	Ship	Storage tank	Baseball Diamond	Tennis court	Basketball court	Ground track field	Harbor	Bridge	Vehicle	CorLoc
WSDDN [17]	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[22]	0.1176	0.5000	0.1282	**0.9865**	**0.8451**	**0.7736**	0.9072	0.0000	0.0926	0.1556	0.4506
MELM[26]	0.8596	0.7742	0.2143	0.9833	0.1071	0.4348	0.9500	0.4000	0.1176	0.1463	0.4987
Ours	**0.9691**	**0.9178**	**0.9513**	0.8865	0.6690	0.6283	0.9598	**0.5418**	**0.1963**	**0.5556**	**0.7276**

TABLE IV
OBJECT CLASSES IN THE DIOR DATA SET

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

TABLE V
COMPARISONS IN TERMS OF AP FOR 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 [50]	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 [57]	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 [17]	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 [22]	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 [26]	**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
Ours	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**

TABLE VI
COMPARISONS IN TERMS OF CORLOC FOR 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 [17]	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 [22]	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 [26]	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
Ours	**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**

It is obvious that our method significantly boosts the performance of objects that frequently appear in adjacent locations. Specifically, compared with our baseline work, the detection performance for airplane (+75.77%) storage tank (+21.26%), tennis court (+21.63%), basketball court (+10.7%), and harbor (+42.21%) on the NWPU VHR.10.v2 data set is

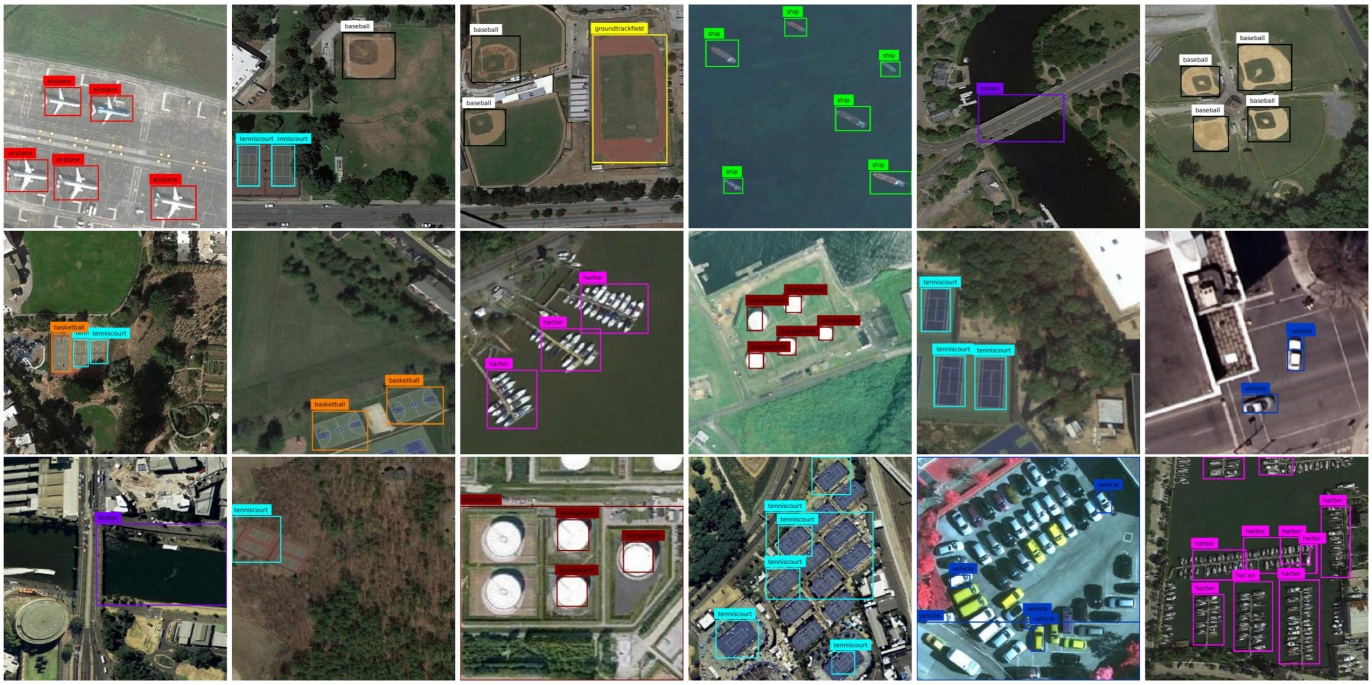


Fig. 3. Some detection results on the NWPU VHR-10.v2 test set (58.82% mAP). The top two rows results are success cases for each class with different colors rectangle. The last row indicates some failure cases.

Fig. 4. Some detection results on the DIOR test set (25.82% mAP). Green rectangle indicates success detections and red rectangle indicates some failure cases.

significantly boosted. Meanwhile, the detection performance for airplane (+16.43%), basketball court (+21.78%), chimney (+47.63%), harbor (+14.18%), storage tank (+6.64%), and tennis court (+30.30%) on the DIOR data set is also significantly improved, which further demonstrates the effectiveness of TCANet. The significant improvement can be attributed to the following aspects: 1) the GCAE effectively highlights the high-quality whole object by capturing the long-range dependence; 2) collaborating GCAE with DLCR highlights the discrepancy of between class-specific instance response and its surroundings so that the instances appearing in the adjacent locations can be easily distinguished; and 3) the adaptive-weighted refinement loss further reduces the ambiguities in the refinement.

The aforementioned comprehensive comparisons fully demonstrate the superiority of our method. Despite better detection performance, our approach fails on individual classes (i.e., bridge, dam, train station, and windmill). The main reasons are as follows.

1) The coexisting of objects and special background hurts the discriminative power of the detector and misleads the detector to identify the special background as object, as the special background is large-scale and cluttered. For example, bridges coexisting with rivers mislead the detector to select rivers as bridges. Reservoirs coexisting with dams lead to the reservoirs being identified as the dams.

2) Different imaging angles and illumination conditions lead to the background features being more prominent than

the object features. For example, the windmills' shadow is more prominent than windmills. Besides, compared with fully supervised methods, the large performance gap still exists in individual classes, e.g., airport, ship, golf filed, storage tank, and vehicle.

E. Qualitative Results

We further provide some detection results on both challenging data sets to qualitatively testify the effectiveness of TCANet. We can clearly observe in Figs. 3 and 4 that each object can be accurately and tightly covered by the predicted bounding boxes. Meanwhile, the objects that appear in the adjacent locations have been accurately distinguished, which further proves the effectiveness of the proposed method. Some failure detection results also are visualized in the last rows of Figs. 3 and 4. As described in the analysis of comparisons with state of the arts, our method sometimes fails to detect the bridges, dams, and windmills by misunderstanding their coexisting special backgrounds as objects, such as rivers, reservoirs, and windmills' shadow. In addition, we also struggle with small objects in the large-scale cluttered background. A new solution is still wanted to address the aforementioned challenging.

V. Conclusion

In this article, we construct a unique and effective end-to-end WSOD network for RSIs, named TCANet, which consists of a GCAE module and a DLCR module. The GCAE is constructed to activate whole object features by extracting the global context of a visual scene. Then, the DLCR model is further designed to learn instance-level discriminative cues so that instances existing in the adjacent locations can be easily distinguished. The cooperation of GCAE and DLCR facilitates the detection model to discover the high-quality instance. Finally, a novel adaptive-weighted refinement loss is proposed to effectively reduce the ambiguities in the refinement. Extensive experiments are elaborately conducted on the challenging NWPU VHR-10.v2 and DIOR data sets. We achieve the state of the arts in the two data sets and significantly outperform the existing predominate works.

References

[1] P. Zhong and R. Wang, "A multiple conditional random fields ensemble model for urban area detection in remote sensing optical images," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 12, pp. 3978–3988, Dec. 2007.

[2] J. Han, G. Cheng, Z. Li, and D. Zhang, "A unified metric learning-based framework for co-saliency detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2473–2483, Oct. 2018.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[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, Feb. 2017.

[8] Q. Li, L. Mou, Q. Liu, Y. Wang, and X. X. Zhu, "HSF-net: Multiscale deep feature embedding for ship detection in optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 7147–7161, Dec. 2018.

[9] 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.

[10] Z. Wu, Y. Li, A. Plaza, J. Li, F. Xiao, and Z. Wei, "Parallel and distributed dimensionality reduction of hyperspectral data on cloud computing architectures," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 6, pp. 2270–2278, Jun. 2016.

[11] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3735–3756, 2020.

[12] 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.

[13] 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, 1998.

[14] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," 2019, *arXiv:1909.00133*. [Online]. Available: http://arxiv.org/abs/1909.00133

[15] 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, Dec. 2017.

[16] G.-S. Xia *et al.*, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3974–3983.

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

[18] 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.*, Jul. 2017, pp. 1377–1385.

[19] V. Kantorov, M. Oquab, M. Cho, and I. Laptev, "ContextLocNet: Context-aware deep network models for weakly supervised localization," in *Proc. Comput. Vis. ECCV*, 2016, pp. 350–365.

[20] 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.*, Jun. 2016, pp. 3512–3520.

[21] 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.

[22] 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.

[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.*, Jul. 2017, pp. 3059–3067.

[24] Y. Wei *et al.*, "TS2C: Tight box mining with surrounding segmentation context for weakly supervised object detection," in *Proc. Comput. Vis. ECCV*, 2018, pp. 434–450.

[25] X. Zhang, J. Feng, H. Xiong, and Q. Tian, "Zigzag learning for weakly supervised object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4262–4270.

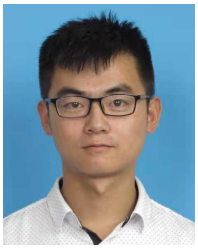
[26] F. Wan *et al.*, "Min-entropy latent model for weakly supervised object detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1297–1306.

[27] R. G. Cinbis, J. Verbeek, and C. Schmid, "Weakly supervised object localization with multi-fold multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 189–203, Jan. 2017.

[28] 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.

[29] G. Cheng, J. Yang, D. Gao, L. Guo, and J. Han, "High-quality proposals for weakly supervised object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 5794–5804, 2020, doi: 10.1109/TIP.2020.2987161.

[30] X. Wang, Z. Zhu, C. Yao, and X. Bai, "Relaxed multiple-instance SVM with application to object discovery," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1224–1232.
[31] W. Ouyang, K. Wang, X. Zhu, and X. Wang, "Learning chained deep features and classifiers for cascade in object detection," 2017, *arXiv:1702.07054*. [Online]. Available: http://arxiv.org/abs/1702.07054
[32] S. Zagoruyko *et al.*, "A MultiPath network for object detection," 2016, *arXiv:1604.02135*. [Online]. Available: http://arxiv.org/abs/1604.02135
[33] Y. Xu, Z. Wu, J. Chanussot, and Z. Wei, "Joint reconstruction and anomaly detection from compressive hyperspectral images using mahalanobis distance-regularized tensor RPCA," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2919–2930, May 2018.
[34] Z. Wu, W. Zhu, J. Chanussot, Y. Xu, and S. Osher, "Hyperspectral anomaly detection via global and local joint modeling of background," *IEEE Trans. Signal Process.*, vol. 67, no. 14, pp. 3858–3869, Jul. 2019.
[35] Y. Li, Y. Zhang, X. Huang, and A. L. Yuille, "Deep networks under scene-level supervision for multi-class geospatial object detection from remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 146, pp. 182–196, Dec. 2018.
[36] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proc. IEEE. Int. Conf. Comput. Vis.*, Oct. 2017, pp. 843–852.
[37] D. Mahajan *et al.*, "Exploring the limits of weakly supervised pretraining," in *Proc. Comput. Vis. ECCV*, 2018, pp. 181–196.
[38] Q. Wang, Y. Li, and Z. Zhou, "Partial label learning with unlabeled data," *Int. J. Conf. Artif. Intell.*, pp. 3755–3761, Aug. 2019.
[39] T. Durand, N. Mehrasa, and G. Mori, "Learning a deep ConvNet for multi-label classification with partial labels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 647–657.
[40] Y. Li, Y. Zhang, and Z. Zhu, "Learning deep networks under noisy labels for remote sensing image scene classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2019, pp. 3025–3028.
[41] 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, Oct. 2016.
[42] 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.*, early access, May 18, 2020, doi: 10.1109/TGRS.2020.2991407.
[43] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2874–2883.
[44] Z. Chen, S. Huang, and D. Tao, "Context refinement for object detection," in *Proc. Comput. Vis. ECCV*, 2018, pp. 71–86.
[45] R. Yu, X. Chen, V. I. Morariu, and L. S. Davis, "The role of context selection in object detection," 2016, *arXiv:1609.02948*. [Online]. Available: http://arxiv.org/abs/1609.02948
[46] X. Zeng *et al.*, "Crafting GBD-net for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 9, pp. 2109–2123, Sep. 2018.
[47] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.
[48] 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.*, early access, Apr. 27, 2020, doi: 10.1109/TGRS.2020.2985989.
[49] 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, Sep. 2013.
[50] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Dec. 2015, pp. 1440–1448.
[51] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2980–2988.
[52] 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
[53] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[54] J. Hosang, R. Benenson, and B. Schiele, "Learning non-maximum suppression," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6469–6477.
[55] G. Cheng, J. Han, P. Zhou, and L. Guo, "Multi-class geospatial object detection and geographic image classification based on collection of part detectors," *ISPRS J. Photogramm. Remote Sens.*, vol. 98, no. 1, pp. 119–132, Dec. 2014.
[56] 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.
[57] 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.



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

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

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, Xi'an, China, in 1999, 2001, and 2003, respectively.

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

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

He is an Associate Professor with Northwestern Polytechnical University. 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, Xi'an, in 2010 and 2013, respectively.

He is a Professor with Northwestern Polytechnical University. His main research interests include computer vision and pattern recognition.