Automatic Weakly Supervised Object Detection From High Spatial Resolution Remote Sensing Images via Dynamic Curriculum Learning

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

***Abstract*— In this article, we focus on tackling the problem of weakly supervised object detection from high spatial resolution remote sensing images, which aims to learn detectors with only image-level annotations, i.e., without object location information during the training stage. Although promising results have been achieved, most approaches often fail to provide high-quality initial samples and thus are difficult to obtain optimal object detectors. To address this challenge, a dynamic curriculum learning strategy is proposed to progressively learn the object detectors by feeding training images with increasing difficulty that matches current detection ability. To this end, an entropy-based criterion is firstly designed to evaluate the difficulty for localizing objects in images. Then, an initial curriculum that ranks training images in ascending order of difficulty is generated, in which easy images are selected to provide reliable instances for learning object detectors. With the gained stronger detection ability, the subsequent order in the curriculum for retraining detectors is accordingly adjusted by promoting difficult images as easy ones. In such way, the detectors can be well prepared by training on easy images for learning from more difficult ones and thus gradually improve their detection ability more effectively. Moreover, an effective instance-aware focal loss function for detector learning is developed to alleviate the influence of positive instances of bad quality and meanwhile enhance the discriminative information of class-specific hard negative instances. Comprehensive experiments and comparisons with state-of-the-art methods on two publicly available data sets demonstrate the superiority of our proposed method.**

***Index Terms*— Dynamic curriculum learning (DCL), instance-aware focal loss, weakly supervised object detection (WSOD).**

Manuscript received October 28, 2019; revised March 2, 2020 and April 26, 2020; accepted April 27, 2020. Date of publication May 18, 2020; date of current version December 24, 2020. This work was supported in part by the National Science Foundation of China under Grant 61701415, Grant 61772425, and Grant 61773315, 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, and in part by the National Key Research and Development Program of China under Grant 2017YFB0502900. *(Corresponding author: Junwei Han.)*

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

Xiaoxu Feng, Junwei Han, Gong Cheng, and Lei Guo are with the School of Automation, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: junweihan2010@gmail.com).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2020.2991407

I. Introduction

With the increasingly rapid development of remote sensing technology, high-resolution optical satellite images are becoming more easily available, which significantly promotes the advancement of understanding the contents of remote sensing images (RSIs) [1]–[3]. Automatic geospatial object detection, as a fundamental yet challenging problem of remote sensing analysis, plays an important role in understanding remote sensing images in an intelligent way [4]–[8].

Automatic object detection aims to determine if a given satellite image contains one or more objects belonging to the class of interest and then locate the position of each predicted object in the image [9]–[13]. As pointed out in the survey of object detection in optical remote sensing images [4], most object detection methods are performed under the fully supervised setting, which often needs a large amount of training data to be manually annotated with a bounding box around each object to be detected. However, with the explosive growth of satellite images in both quantity and quality, it is generally time-consuming, expensive, and sometimes even unreliable to obtain such accurate manual annotation. Recently, a few efforts [14]–[18] have been made to investigate weakly supervised learning (WSL) for object detection from remote sensing images, where the training set needs only binary labels indicating whether an image contains the target object or not. Compared with manual annotation of each object bounding box in training images, providing binary labels for each training image significantly reduces the cost of human labor.

Obviously, it is more difficult to learn robust object detectors from image-level annotated remote sensing images. Currently, most of the aforementioned works follow a typical WSL scheme that firstly generates initial training samples and then uses them to train the detector and annotate training set iteratively until the model converges. In order to improve the detection performance, they either seek better initialization models of training samples [14], [15] or exploit more robust learning strategies [16], [18]. More specifically, the inter-class separability employed in [15] was further integrated with saliency and intraclass compactness in a Bayesian framework

for better initializing training samples [14]. To train the detector robustly, in the works [16], [18], hard/informative negative samples are carefully mined for augmenting the training set and retraining the model in each iteration process. Although promising results are reported, weakly supervised object detection (WSOD) still remains as an open problem for remote sensing image analysis. The above-mentioned approaches fail to provide high-quality initial samples and thus are difficult to obtain optimal object detectors. Moreover, when the wrong annotations of training examples are introduced in the initial iterative learning process, the model is hard to converge.

To overcome the aforementioned limitations of weakly supervised detection, a novel dynamic curriculum learning (DCL) strategy as illustrated in Fig. 1 is proposed to progressively learn the object detectors by feeding training images with increasing difficulty that matches current detection ability. To this end, an entropy-based criterion is firstly designed to evaluate the difficulty for localizing objects in images. Then, an initial curriculum that ranks training images in ascending order of difficulty is generated, in which easy images are selected to provide reliable instances for learning object detectors. With the gained stronger detection ability, the subsequent order in the curriculum for retraining detectors is accordingly adjusted by promoting difficult images as easy ones. In such way, the detectors can be well prepared by training on easy images for learning from more difficult ones and thus gradually improve their detection ability more effectively. Although the proposed method essentially is similar with curriculum learning [19], our method attempts to define the difficulty of an image by considering the detection scores of all possible instances instead of only the top-scored instances, which is more reasonable as remote sensing images are large scale and often contains more than one instance of different classes. Moreover, our dynamic learning strategy is flexible to incorporate prior knowledge into learning in the form of initial curriculum and meanwhile dynamically adjusts the curriculum according to the subsequent learning. Furthermore, we develop an effective instance-aware focal loss for detector learning to alleviate the influence of positive instances of bad quality and meanwhile enhance the discriminative information of class-specific hard negative instances. Comprehensive experiments and comparisons with state-of-the-art methods on two publicly available data sets demonstrate the superiority of our proposed method.

The remainder of this article is organized as follows. The next section briefly reviews the related works. Section III details the proposed DCL-based WSOD method. Comprehensive experimental results are provided in Section IV. Section V concludes this article.

II. Related Works

Object detection has been extensively studied for the last few decades. In this section, we will review the related works about WSL and curriculum learning for remote sensing analysis.

A. Weakly Supervised Learning

Recently, WSL has been gaining more interests for both natural scene images [20]–[26] and remote sensing images analysis [14]–[18], [27]. Although encouraging results have been reported for WSL-based object detection in natural scene images, these approaches cannot be directly applied to remote sensing images. It is especially more challenging for detecting objects from RSIs compared with natural images under weakly supervised settings because objects only occupy a small proportion of the large-scale image with complex backgrounds, and multiple object instances often occur in an image with close arrangement. In addition, large appearance variances in scale and rotation further add to the complexity of the problem. The work [15] is the first attempt to utilize WSL for object detection in high-resolution remote sensing images. It takes full advantage of the rich information from the negative data to heuristically mine the positive samples and meantime evaluate the learned detector. The improved version [14] integrates saliency, intra-class compactness, and inter-class separability based on Bayesian principles for better initializing training samples, and adopts deep Boltzmann machines to build a high-level feature representation for various geospatial objects. In the work [16], a negative bootstrapping scheme is integrated into the iterative learning process to learn a robust detector by selecting the most informative negative samples in each iteration. The work [18] constructs a coupled convolutional neural network (CNN) in a weakly supervised manner to extract the object proposals and simultaneously locate the aircraft from large-scale very high resolution (VHR) images. Although satisfactory results have been reported in the above-mentioned methods, they can only work for a specific object class. More importantly, they often fail to provide a good initialization and are prone to suffer from model drift problem.

B. Curriculum Learning

Curriculum learning has attracted much attention as an emerging machine learning technique in recent years. It can avoid poor local optima in the training of models with non-convex objectives inspired by the human learning mechanism that gradually proceeds from easy to more complex samples in training [19], [28]. In curriculum learning, a curriculum consisting of easy to complex samples is predetermined by exploiting expert knowledge of the data sets. A variant of curriculum learning, namely, self-paced learning, embeds curriculum designing into model learning by introducing a regularization term into the objective. The curriculum is gradually determined by the model itself based on what it has already learned. More recently, self-paced learning has been applied in WSL-based object detection in natural images [22], [25], where self-paced learning is exploited to handle the uncertainty of instances in positive images by considering the instances that are most likely to be correctly detected as easy instances. In [25], a curriculum is designed based on saliency detection to select the salient regions as the initialization of training instances while [22] attempt to build curriculum by selecting a subset of easy classes and easy samples from these classes based on their detection probability scores. Then, the object detectors are iteratively trained under the guidance of the self-paced learning strategy. Compared with curriculum learning and self-paced learning, our method

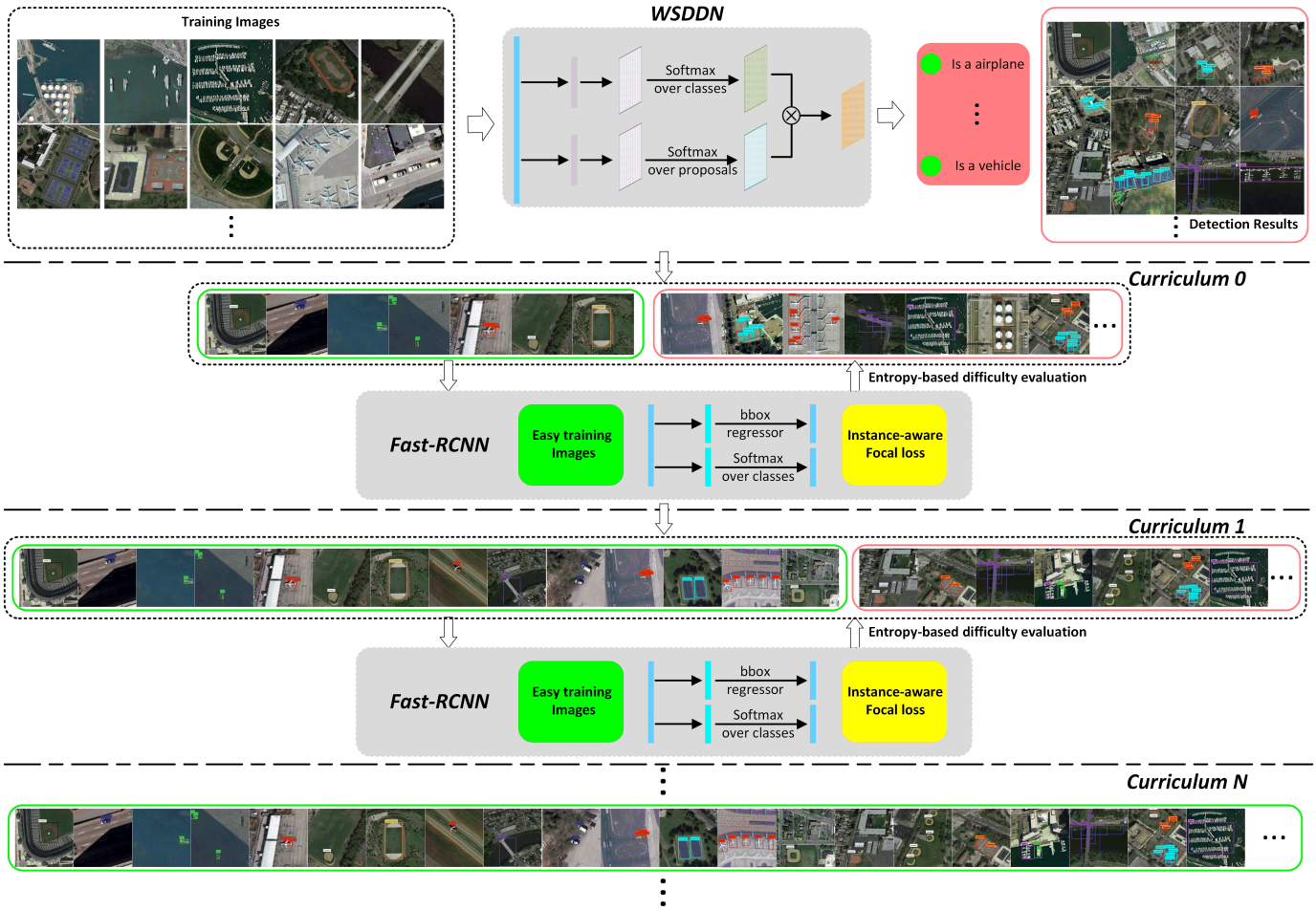


Fig. 1. Flowchart of the proposed DCL based WSOD.

designs an effective measure of image difficulty, which is more practical for remote sensing images. And it is more flexible to incorporate prior knowledge, which can offer better initialization for the subsequent iterative learning.

III. Proposed Method

In this section, we will describe the proposed method in detail. Section III-A describes how to generate the initial curriculum. Then, the proposed DCL-based object detection method is detailed in Section III-B.

A. Initial Curriculum Generation

The initial curriculum is generated to provide an easy-to-difficult order for learning object detector. The key is to evaluate the difficulty of images for accurate detection, which can be influenced by number of categories, characteristics of instances and background clutter, etc. Thus, it is obviously suboptimal to simply employ the probability scores of top-scored proposals as an evaluation criterion. Here, we propose to evaluate the image difficulty by analyzing the distribution of detection scores for all object proposals, through which we can obtain a global understanding of the image and provide a more accurate difficulty evaluation.

1) Weakly Supervised Deep Detection Networks: Without loss of generality, we exploit the widely used weakly supervised deep detection network (WSDDN) [26] as the baseline network to obtain initial localization probability score for each object proposal. In WSDDN, a classification stream is designed in parallel with a localization stream to produce final detection scores through performing elementwise multiplication. Specifically, WSDDN takes as input a set of N training images $\mathcal{I} = \{I_1, I_2, \ldots, I_N\}$ together with their image-level label vectors $\mathcal{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N\}$. For the nth image I_n, let $\boldsymbol{y}_n = [y_n^1, y_n^2, \ldots, y_n^C]$ denotes its image-level label vector, where C is the total number of object categories. $y_n^c = 1$ $(c = 1, 2, \ldots, C)$ if the image is labeled with the cth category, and otherwise $y_n^c = 0$. And a set of object proposals $\mathcal{B}_n = \{B_n^1, B_n^2, \ldots, B_n^{N_n}\}$ are extracted from I_n, where N_n is the total number of object proposals. For each object proposal B_n^m, two kinds of probability scores, namely, classification probability score $\boldsymbol{x}_n^m = [x_n^{m,1}, \ldots, x_n^{m,C}]$ and detection probability score $\boldsymbol{l}_n^m = [l_n^{m,1}, \ldots, l_n^{m,C}]$, are obtained from the classification stream and the localization stream, respectively. Specifically, the classification probability scores $\boldsymbol{x}_n^m$ are obtained by feeding the features of object proposals to a linear mapping followed by a softmax operation. That is

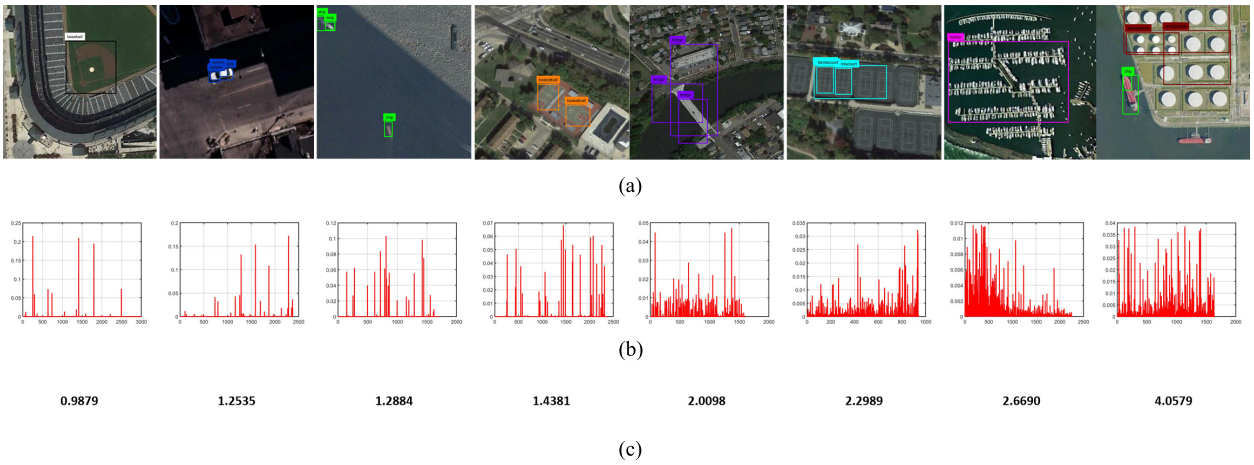


Fig. 2. Illustration of the motivation of the designed entropy-based image difficulty measure criterion.

the features of object proposals are first mapped to a matrix of data by the linear mapping, and then this matrix of data is normalized at the dimension of class to predict which class should be assigned to a proposal by using *the softmax operation*. The detection probability scores $\boldsymbol{l}_n^m$ are obtained by feeding the same features to a second linear mapping followed by a softmax operation but in a class-specific basis. This means that the features of object proposals are first mapped to a matrix of data by the second linear mapping, and then this matrix of data is normalized at the dimension of proposal to select which proposal is more likely to contain the most informative image fragment by using the softmax operation. Then, the final score $\boldsymbol{d}_n^m = [d_n^{m,1}, \ldots, d_n^{m,C}]$ for object proposal B_n^m is obtained by taking element-wise product of $\boldsymbol{x}_n^m$ and $\boldsymbol{l}_n^m$. And the image-level class prediction score $\hat{y}_n^c$ for the cth category can be obtained by summation over object proposals

$$\hat{y}_n^c = \sum_{m=1}^{N^n} d_n^{m,c}. \tag{1}$$

Then, the parameters of WSDDN can be learned by minimizing the binary log image-level loss, denoted as

$$\mathcal{L}(I_n, \boldsymbol{y}_n) = \sum_{c=1}^{C} \log\left(y_n^c\left(\hat{y}_n^c - 1/2\right) + 1/2\right). \tag{2}$$

As can be seen, the (2) is a sum of binary-log-loss terms, one per class. As $\hat{y}_n^c$ is the class prediction score and is in the range of (0, 1), it can be considered as a probability of class c being present in the nth image, i.e. $p(y_n^c = 1)$. When the ground-truth label is positive, the binary log loss becomes $\log(p(y_n^c = 1))$, $\log(1 - p(y_n^c = 1))$ otherwise.

2) Image Difficulty Evaluation: How to measure the image difficulty plays an important role in providing more accurate and suitable curriculums for learning object detectors. Usually, the probability scores of the detected proposals are employed as the criterion to evaluate the image difficulty. However, it is suboptimal to directly measure the difficulty of remote sensing images by using this manner. This is mainly because remote sensing images are large scale with complex backgrounds and often contain more than one instance of different classes. For example, a number of WSDDN detection results along with the final scores are provided in Fig. 2(a). As can be seen, the objects successfully localized by WSDDN usually appear in relatively simple images, which contain a few instances with uncomplicated backgrounds. In addition, object parts instead of whole object are detected with top scores from some images. Thus, instead of only focusing on the top-scored proposals, it is necessary to evaluate the image difficulty based on a global understanding of the image. To achieve this, we attempt to analyze the distribution of detection scores of all proposals to quantify the localization difficulty of each image based on the following key observation.

As shown in Fig. 2(b) of which the x-axis is the object proposals and the y-axis is the probability scores of these object proposals, among object proposals of the images that WSDDN localizes successfully, only a small portion of the object proposals are assigned with high detection scores while a dominantly large portion of them do not cover the object of interest tightly and are assigned with low scores. Inspired by this observation, we propose a distribution-based criterion to evaluate the image difficulty. To be specific, similar with [29], let $\boldsymbol{p}_n^c = [p_n^{1,c}, p_n^{2,c}, \ldots, p_n^{N_n,c}]$ denotes the normalized detection confidence vector of the c-th category over all proposals $\mathcal{B}_n$ in image I_n. The value of the mth element of $\boldsymbol{p}_n^c$ is computed as

$$p_n^{m,c} = \frac{d_n^{m,c} + \varepsilon}{\sum_{m=1}^{N^n}\left(d_n^{m,c} + \varepsilon\right)} \tag{3}$$

where the parameter ε is a small constant to avoid illegal operations in normalizing the detection confidence vector and has no influence on the detection results. According to the aforementioned observation, for a confident detection result, a few of detection scores in $\boldsymbol{p}_n^c$ should be high and most of them should have near-zero values which means that $\boldsymbol{p}_n^c$ is a sparse vector. Here, we adopt the Shannon entropy to measure the sparsity of $\boldsymbol{p}_n^c$

$$e_n^c = -\sum_{m=1}^{N^n} p_n^{m,c} \log p_n^{m,c}. \tag{4}$$

The more sparse s_n^c is, the smaller its entropy e_n^c is. Accordingly, the image difficulty ϑ_n is defined as the mean value of Shannon entropy e_n^c of the annotated classes

$$\vartheta_n = \frac{1}{|\mathcal{C}_n|} \sum_{c \in \mathcal{C}_n} e_n^c \tag{5}$$

where $\mathcal{C}_n = \{c | y_n^c = 1\}$ is the set of annotated classes for image I_n. Thus, the entropy measure is related to the cluttered scene and is designed to measure the difficulty of detecting objects from such scene. Fig. 2(c) presents the example image difficulty scores computed by the proposed distribution-based metric. Basically, the localization difficulty of each image mimics the detection performance of WSDDN.

B. Dynamic Curriculum Learning

Next, we will describe the designed DCL strategy that progressively learns the detection model on easy training images in the given initial curriculum and in turn adjusts the subsequent learning order in the curriculum based on current renewed detection ability to involve more difficult images for further improvement.

1) Fast Region-Based Convolutional Network: In our work, we built upon the widely used fast region-based convolutional network (Fast RCNN) framework to learn the object detector. Fast RCNN is composed of a backbone network for feature extraction and an ROI network with two sibling output layers. The first layer outputs a discrete probability distribution, $d = [d^0, \ldots, d^C]$, for each *region of interest* (ROI) over C categories and backgrounds. While the second layer outputs the bounding-box regression offsets, $t^c = (t_x^c, t_y^c, t_w^c, t_h^c)$, for each of the C categories. Note that t^c follows [30] to specify a scale-invariant translation and log-space height/width shift relative to an object proposal. Assume that each ROI is labeled with a ground-truth class c and a ground-truth bounding-box location, Fast-RCNN can be learned by optimizing the following multitask objective function

$$\mathcal{L}_{\text{ROI}} = \mathcal{L}_{\text{cls}}(d, c) + \lambda [c \geq 1] \mathcal{L}_{\text{reg}}(t^c, v) \tag{6}$$

in which the first term $\mathcal{L}_{\text{cls}}(d, c) = -\log d^c$ is classification loss for true class c. The Iverson bracket indicator function $[c \geq 1]$ equals to 1 when $c \geq 1$ and 0 when $c = 0$ (background class). The parameter λ controls the balance between the first term and the second bounding box regression term, which is set to be 1 as in [30]. The bounding box regression term is defined over a tuple of true bounding box regression targets $v = (v_x, v_y, v_w, v_h)$ and a predicted tuple $t^c = (t_x^c, t_y^c, t_w^c, t_h^c)$, for class c. For background ROIs, there is no notion of a ground-truth bounding box and hence the bounding box regression loss is ignored. For other ROIs, the bounding box regression term is defined as

$$\mathcal{L}_{\text{reg}}(t^c, v) = \sum_{i \in \{x, y, w, h\}} \text{smooth}_{L_1}(t_i^c - v_i) \tag{7}$$

where

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases} \tag{8}$$

is a robust L_1 loss that is less sensitive to outliers than the L_2 loss used in R-CNN. When the regression targets are unbounded (i.e., $|x| \geq 1$), the L_2 loss has higher regression loss value for outliers than the L_1 loss. This is because the L_1 loss is the absolute value of x while the L_2 loss is the square of x. Thus, training with L_2 loss requires careful tuning of learning rates in order to prevent gradient explosion for outliers. Equation (8) eliminates this sensitivity [30].

2) Instance-Aware Focal Loss: The top-scored positive object proposals mined from easy images are considered as the ground-truth in the Fast-RCNN learning. *As same as common practice*, positive instances are sampled from object proposals that have intersection over union (IoU) overlap with a ground truth bounding box larger than 0.5 while negative instances are sampled from object proposals of IoU in the interval [0.1, 0.5). Despite that positive object proposals are mined from the relatively easy image, they cannot be all reliable due to the high variation of image difficulty. Thus, it is suboptimal to directly optimize the loss defined in (6) by using ROIs sampled from an image as described above. To conquer that, an effective instance-aware focal loss is designed to alleviate the influence of positive instances of bad quality and meanwhile enhance the role of discriminative information carried by class-specific hard negative instances. Note that class-specific hard negative instances are top-scored object proposals but are assigned with the wrong class label which is not any of the annotated image-level class labels. These object proposals would carry more discriminative information than the common negatives that have a maximum IoU with ground truth in the interval [0.1, 0.5). Thus, we impose heavier penalty on misclassifying these class-specific hard negative samples in the detector learning. Formally, the instance-aware focal loss is defined as

$$\mathcal{L}_{\text{FL}} = \mathcal{V}(\text{ROI}) \mathcal{L}_{\text{ROI}} \tag{9}$$

where $\mathcal{V}(\text{ROI})$ is a weighting function which assigns weight to an instance (i.e., ROI) as

$$\mathcal{V}(\text{ROI}) = \begin{cases} \gamma^c, & \text{if ROI} \in \mathcal{R}_{\text{Common}} \\ (1 - d^0)^\beta, & \text{if ROI} \in \mathcal{R}_{\text{Specific}}. \end{cases} \tag{10}$$

Note that the common set $\mathcal{R}_{\text{Common}}$ consists of the mined positive and negative object proposals according to their IoUs with ground truth and the specific set $\mathcal{R}_{\text{Specific}}$ consists of the class-specific hard negative object proposals mined from images that do not contain the corresponding class. For ROIs of $\mathcal{R}_{\text{Common}}$, similar to Online Instance Classifier Refinement (OICR) [21], we down-weights the loss by multiplying γ^c that indicates the reliability of the positive ground truth, which is the detection probability score for class c provided by WSDDN or previous learned detection model. As for the ROIs from the specific set $\mathcal{R}_{\text{Specific}}$, we employ the same modulating factor with the common focal loss [31], which would assign larger weights than the value of $1 - d^0$ to penalize misclassifying class-specific hard negatives. In the training of Fast-RCNN-based object detector, each minibatch is constructed from two images randomly chosen from the training subset. In such case, we may select two images that

do not contain the corresponding class-specific hard negatives. To be suitable with Fast-RCNN training, inspired in [32], we exploit a trick that these ROIs of $\mathcal{R}_{\text{Specific}}$ are randomly pasted into certain images that contains the corresponding class under no intersection with other ROIs. Thus, the information from class-specific hard negatives can be effectively used during training.

3) DCL-Based Object Detection: The overall procedure of the proposed DCL-based WSOD is shown in Algorithm 1.

Algorithm 1 Dynamic Curriculum Learning-Based Weakly Supervised Object Detection

Input: Training set $\mathcal{I} = \{I_1, I_2, \ldots, I_N\}$ with image-level labels $\mathcal{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots \boldsymbol{y}_N\}$, initial difficulty threshold τ_0, difficulty threshold increase stepsize θ;

Initial Curriculum Generation:

Learn WSDDN by minimizing the binary log image-level loss in Eqn. (2);

Compute the normalized detection confidence vector $\boldsymbol{p}_n^c$ over all proposals in image I_n with Eqn. (3);

Compute the Shannon entropy to measure the sparsity of $\boldsymbol{p}_n^c$ with Eqn. (4);

Compute the localization difficulty with Eqn. (5) and rank the images in ascending order of difficulty to generate initial curriculum ζ_0;

Dynamic Curriculum Learning:

$k \leftarrow 0$

Repeat

Select images of difficulty score smaller than τ_k from curriculum ζ_k as training subset $\mathcal{S}_k$;

Learn detection model M_k on $\cup_{i=0}^{k}\mathcal{S}_k$ using Fast RCNN with instance-aware focal loss;

Perform M_k on the training images in $\zeta_k - \mathcal{S}_k$ and obtain the re-localization outputs;

Re-evaluate the image difficulty with Eqn. (5) according to the re-localization outputs;

$k \leftarrow k+1$, $\tau_k \leftarrow \tau_k + \theta$, $\zeta_k \leftarrow \zeta_k - \mathcal{S}_k$

Until $\zeta_k = \emptyset$

Output: the learned detection model M_k

Given the initial curriculum ζ_0, images of difficulty scores smaller than the threshold τ are selected as easy training subset $\mathcal{S}_0$. Then, the detection model M_0 is learned on $\mathcal{S}_0$ by following the Fast RCNN framework with the proposed instance-aware focal loss in (9). After training on the easiest images, the detector M_0 gains increased detection ability for better dealing with relatively more difficult images identified in curriculum ζ_0. Thus, the subsequent learning order of training images in $\zeta_0 - \mathcal{S}_0$ is adjusted by re-evaluating their difficulties according to the re-localization outputs of the learned detection model. With better localization performance, some previously identified difficult images are assigned with lower difficulty scores and are considered as easy to be included in the new training subset $\mathcal{S}_1$. Then, a new detection model M_1 is learned on $\mathcal{S}_1 \cup \mathcal{S}_0$. Note that the detection model M_0 is used to mine more reliable instances in $\mathcal{S}_1$ and the localized top-scored instances are used as ground truth for learning M_1. As the iterative learning process proceeds, the curriculum is dynamically adjusted. Note that there is no mechanism to guarantee the convergence of the iteration. The iterative procedure simply stopped when all the training images are selected to be included in the final curriculum.

IV. Experiments

To demonstrate the effectiveness of the proposed method, the experiments are organized as follows. We first describe the benchmark data sets. Next, the implementation details and the parameter analysis are provided. Then, the detection results and comparisons with state-of-the-art methods are presented. We finally present a series of studies to demonstrate the impact of each component on our proposed DCL-based detection system.

A. Data Sets

A comprehensive evaluation of the proposed method is constructed on the NWPU-VHR-10.v2 data set [33] and the benchmark for *object DetectIon in Optical Remote sensing images* (DIOR) [6]. The NWPU-VHR-10.v2 data set contains 1172 images with a size of 400 × 400 pixels which are cropped from 650 images with different sizes (i.e., from 533 × 597 to 1728 × 1028 pixels). This data set still contains the same ten classes of geospatial objects as the NWPU-VHR10 data set, namely, airplane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, and vehicle. In our work, the image set was divided into 58% for training, 17% for validation, and 25% for test, resulting in three independent subsets: a training set containing 679 images, a validation set containing 200 images, and a test set containing 293 images. The DIOR data set contains 23 463 images and 190 288 instances, covering 20 object classes. This data set is much more challenging than the NWPU-VHR-10.v2 data set. We randomly selected 11725 remote sensing images (i.e., 50% of the data set) as trainval set, and the remaining 11738 images are used as test set. The trainval data consists of two parts, the training set and validation set. Please refer to [6] for the number of these three sets for each class. The standard average precision (AP) is used to quantitatively evaluate the performance of the object detection system.

B. Implementation Details and Parameter Analysis

The proposed DCL-based object detection method employed the VGG-16 network pretrained on the ImageNet data set as the backbone network. Each training image is augmented by horizontally mirroring and rotating 90° and 180°. And about 2000 object proposals are extracted from each image by using the selective search method [34].

The initializations of the proposed DCL-based object detection method include the generation of the initial curriculum and the construction of the initial training subset. For the initial curriculum generation, the WSDDN model directly trained on the two benchmark data sets without any parameters being further tuned is adopted to localize object instances in the

training images. Then, the designed entropy-based criterion is used to compute the image difficulty according to the localization results. Afterward, an initial curriculum that ranks training images in ascending order of difficulty is generated. The initial difficulty threshold τ_0 plays an important role in the construction of the initial training subset. It is intuitively set to be 0.4 for both the two data sets, which is much smaller than the average image difficulty score. This will encourage selecting easier images for the initial training subset, which would lead to better initial detectors. Finally, a number of 76 and 632 images are selected as the initial training subset of the NWPU-VHR-10.V2 data set and the DIOR data set, respectively.

The difficulty threshold increase stepsize θ controls the iterative learning process by selecting an appropriate number of training images in each iteration. If we set θ with large value, a large amount of training images is recognized as easy images in early iterations of training. However, the current detector's capability may not match the rapid increase in image difficulty, which can lead to oscillations in training. In contrast, the small θ would include a small number of training images in each iteration, which may result in overfitting in early iterations and further limit the improvement of the detection performance. In the implementation, we optimized the parameter θ on the NWPU-VHR-10.v2 data set and then kept it fixed on the DIOR data set. Fig. 3(a) presents the detailed results for different θ. As can be seen, the best detection performance is achieved when θ is 0.2 and the performance decreased greatly when θ is larger than 0.6. Considering that, we set the difficulty threshold increase stepsize θ to be 0.2.

For training the Fast RCNN-based detectors, the first 10k iterations are trained using the learning rate with 0.001 and then the learning rate decreases to 0.0001 in the next 10k iterations. The momentum and weight decay are set to 0.9 and 0.0005, respectively. Each minibatch of 64 ROIs is constructed from object proposals of two images with 32 ROIs for each image. Among these ROIs, 10 ROIs that have IoU overlap with a ground truth bounding box larger than 0.5 are selected as the positive instances. Twenty ROIs that have a maximum IoU with ground truth in the interval [0.1, 0.5) are employed as common negatives. The remaining two ROIs are class-specific hard negatives, which are two top-scored object proposals mined from images that do not contain the corresponding class. In the proposed instance-aware focal loss, the parameter γ^c is the detection probability score for class c provided by the previous learned detection model. The parameter β is set to assign larger weights than the value of $1 - d^0$ to penalize misclassifying class-specific hard negatives. Thus, the parameter β should be set with values smaller than 1. We optimized the parameter β on the NWPU-VHR-10.v2 data set and then kept fixed on the DIOR data set. To be specific, we varied the value of the parameter β from 0 to 1 with a stride of 0.1 and provided the corresponding results in Fig. 3(b). As can be seen, our method is insensitive to the choice from 0.3 to 0.8. Finally, we set the parameter β to be 0.5. All the other Fast RCNN-specific hyperparameters are the same as used in [49]. Moreover, the Fast RCNN detection model is initialized by the previously learned detection model before training using the newly selected easy images.

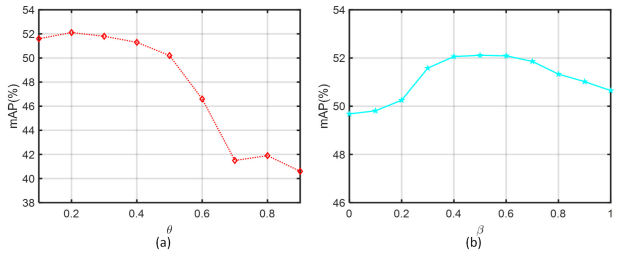


Fig. 3. Detection performance for different values of (a) θ and (b) β on the NWPU-VHR-10.v2 data set.

C. Experimental Results and Comparisons

With the DCL scheme, the learned detection model is used to perform object detection on both two data sets. Figs. 4 and 5 presents a number of object detection results on the NWPU-VHR-10.v2 data set and the DIOR data set, respectively. In Fig. 4, the results are indicated by rectangles together with the class label in different colors. In Fig. 5, the results are indicated by green rectangles together with the class label. As can be seen, the proposed method can successfully detect and locate most of the objects despite their large variations in the sizes and orientations on the two data sets. The failure detection results are shown in Figs. 4 and 5 (bottom row). We can observe that the proposed DCL method demonstrates a less satisfactory detection performance on the object classes of Bridge, Tennis court, Basketball court with some false detection and the object classes of Harbor and Vehicle with some missed detection on the NWPU-VHR-10.v2 data set. The proposed method exhibits similar failure performances on the DIOR data set. Our method performs a false detection for object classes of Expressway toll station, Dam, and Vehicle, and a missed detection for the object class of Harbor. On both two data sets, for the classes of Storage tank, Ship, and Vehicle that the corresponding objects are compactly arranged, our method usually recognizes them as a whole object but fails to detect the individual object. Moreover, due to the lack of instance-level supervision, the proposed method sometimes mistakes the class of Tennis court for Basketball court as these two classes often appear together in the same image.

In addition, to quantitatively evaluate the proposed method, we compare our method with six fully supervised and four weakly supervised state-of-the-art object detection methods. Specifically, the six methods include three popular object detection methods, namely, region-based convolution network (RCNN) [35], Fast RCNN [30], and Faster RCNN [36], for natural images. The other three methods are the collection of part detectors (COPD) [37], a transferred CNN model from AlexNet [33], and rotation-invariant CNN model (RICNN) [5]. The four weakly supervised methods are WSDDN [26], Online Instance Classifier Refinement (OICR) [21], Proposal Cluster Learning (PCL) [23], and Min-Entropy Latent Model (MELM) [24]. Note that the results of the methods of COPD, Transfer CNN, and RICNN on the NWPU-VHR-10.v2 data

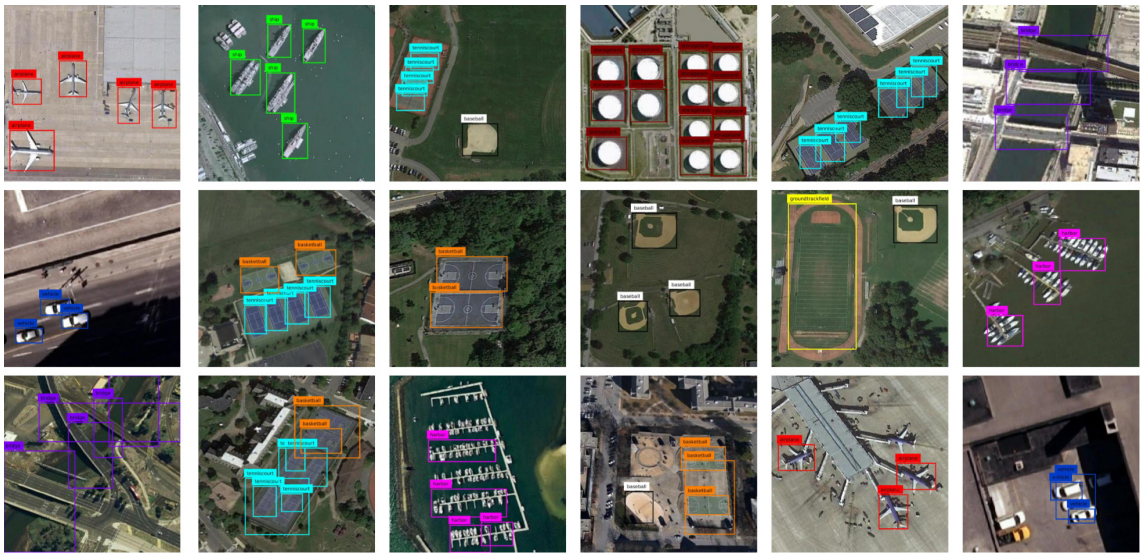


Fig. 4. Number of detection results with the proposed approach on the NWPU-VHR-10.v2 data set.

Fig. 5. Number of detection results with the proposed approach on the DIOR data set.

set are obtained from the report in [33] and the results of the other methods on these two data sets are achieved by using the publicly available codes without any parameters being further tuned.

Table I reports the quantitative comparison results of the ten different methods as measured with AP values on the NWPU-VHR-10.v2 data set. The double underscore in the table is used to distinguish between fully supervised and weakly supervised methods for a clear illustration. We can observe that our method significantly outperforms the other four weakly supervised approaches in terms of mAP by a large margin (about at least 9.82% improvements). For classes of Ship, Harbor, Bridge, and Vehicle, our method achieves better AP results than the four weakly supervised approaches. In addition, our method achieves remarkable success in narrowing the gap between weakly and full supervised methods. Highly competitive performance is achieved compared with fully supervised methods such as Transfer CNN and COPD. For classes of airplane, ship, and baseball diamond, our method demonstrates a superior detection performance than Transfer CNN and COPD methods, which is also competitive compared with the other three fully supervised methods.

Table II presents the quantitative comparison results of the seven different methods in terms of AP on the DIOR data set. As can be seen, our method achieves the best performance among these five weakly supervised methods and outperforms the methods of PCL and MELM with an improvement of 2.00% and 1.53% in terms of mAP, respectively. For classes of Bridge, Expressway toll station, and Stadium, our method achieves better AP results than the other four weakly supervised approaches. As the DIOR data set is much more challenging, the performance of two fully supervised methods is much better than all these five weakly supervised methods. Only for the classes of Baseball field, Chimney, Ground track field, and Stadium, our method achieves competitive or even better results than the two fully supervised methods. This is

TABLE I

PERFORMANCE COMPARISONS OF FIVE DIFFIDENT METHODS IN TERMS OF AP VALUES ON THE DIOR DATA SET

Method	Airplane	Ship	Storage tank	Baseball diamond	Tennis court	Basketball court	Ground track field	Harbor	Bridge	Vehicle	mAP
Fast-RCNN[30]	0.9091	0.9060	0.8929	0.4732	1.0000	0.8585	0.8486	0.8822	0.8029	0.6984	0.8272
RICNN[5]	0.8871	0.7834	0.8633	0.8909	0.4233	0.5685	0.8772	0.6747	0.6231	0.7201	0.7312
RCNN[35]	0.8537	0.8888	0.6278	0.1973	0.9066	0.5823	0.6795	0.7987	0.5422	0.4992	0.6576
Transfer CNN[33]	0.6603	0.5713	0.8501	0.8093	0.3511	0.4552	0.7937	0.6257	0.4317	0.4127	0.5961
COPD[37]	0.6225	0.6937	0.6452	0.8213	0.3413	0.3525	0.8421	0.5631	0.1643	0.4428	0.5488
WSDDN[26]	0.3008	0.4172	0.3498	0.8890	0.1286	0.2385	**0.9943**	0.1394	0.0192	0.0360	0.3512
OICR[21]	0.1366	0.6735	**0.5716**	0.5516	0.1364	0.3966	0.9280	0.0023	0.0184	0.0373	0.3452
PCL[23]	0.2600	0.6376	0.0250	0.8980	**0.6445**	**0.7607**	0.7794	0	0.013	0.1567	0.3941
MELM[24]	**0.8086**	0.6930	0.1048	**0.9017**	0.1284	0.2014	0.9917	0.1710	0.1417	0.0868	0.4229
DCL(Proposed)	0.7270	**0.7425**	0.3705	0.8264	0.3688	0.4227	0.8395	**0.3957**	**0.1682**	**0.3500**	**0.5211**

TABLE II

PERFORMANCE COMPARISONS OF DIFFIDENT METHODS IN TERMS OF AP VALUES ON THE NWPU-VHR-10.V2 DATA SET

Method	Airplane	Airport	Baseball field	Basketball court	Bridge	Chimney	Dam	Expressway service area	Expressway toll station	Golf field	
Fast-RCNN[30]	0.4417	0.6679	0.6696	0.6049	0.1556	0.7228	0.5195	0.6587	0.4476	0.7211	
Faster-RCNN[36]	0.5028	0.6260	0.6604	0.8088	0.2880	0.6817	0.4726	0.5851	0.4806	0.6044	
WSDDN[26]	0.0906	**0.3968**	0.3781	0.2016	0.0025	0.1218	**0.0057**	0.0065	0.1188	0.0490	
OICR[21]	0.0870	0.2826	0.4405	0.1822	0.0130	0.2015	0.0009	0.0065	0.2989	0.1380	
PCL[23]	0.2152	0.3519	0.5980	0.2349	0.0295	0.4371	0.0012	0.0090	0.0149	0.0288	
MELM[24]	**0.2814**	0.0323	**0.6251**	**0.2872**	0.0006	**0.6251**	0.0021	**0.1309**	0.2839	0.1515	
DCL(Proposed)	0.2089	0.2270	0.5421	0.1150	**0.0603**	0.6101	0.0009	0.0107	**0.3101**	**0.3087**	
Method	Ground track field	Harbor	Overpass	Ship	Stadium	Storage tank	Tennis court	Train station	Vehicle	Windmill	mAP
Fast-RCNN[30]	0.6293	0.4618	0.3803	0.3213	0.7098	0.3504	0.5827	0.3791	0.1920	0.3810	0.4998
Faster-RCNN[36]	0.6700	0.4386	0.4687	0.5848	0.5237	0.4235	0.7952	0.4802	0.3477	0.6544	**0.5548**
WSDDN[26]	0.4253	0.0466	0.0106	0.0070	0.6303	0.0395	0.0606	0.0051	0.0455	0.0114	0.1326
OICR[21]	**0.5739**	0.1066	**0.1106**	0.0909	0.5929	0.0710	0.0068	0.0014	**0.0909**	0.0041	0.1650
PCL[23]	0.5636	0.1676	0.1105	0.0909	0.5762	0.0909	0.0247	0.0012	0.0455	**0.0455**	0.1819
MELM[24]	0.4105	**0.2612**	0.0043	0.0909	0.0858	**0.1502**	**0.2057**	**0.0981**	0.0004	0.0053	0.1866
DCL(Proposed)	0.5645	0.0505	0.0265	**0.0909**	**0.6365**	0.0909	0.1036	0.0002	0.0727	0.0079	**0.2019**

TABLE III

STUDIES OF THE COMPONENTS OF THE PROPOSED DCL-BASED WSOD

Difficulty Measure	Confidence-based criterion		√		
	Entropy-based criterion	√		√	√
Learning Strategy	Curriculum learning			√	
	Dynamic curriculum learning	√	√		√
Loss Function	Cross entropy loss				√
	Instance-aware focal loss	√	√	√	
mAP		**0.5211**	0.5029	0.4631	0.4932

mainly because the objects of all these classes usually occupy a large part of the images and have a relatively low probability of co-occurrence with other categories.

D. Analysis of Components

In this section, the contributions of the three key components in our proposed method including entropy-based difficulty measure, DCL strategy, and instance-aware focal loss are further evaluated.

To this end, three baseline experiments are constructed. In the first baseline experiment, a confidence-based criterion is implemented to measure the image difficulty based on only the confidence score of the top-scored object proposals. For the second baseline experiment, we design a curriculum learning strategy, where the initial curriculum is generated as the same with our DCL but remains unchanged in the subsequent learning. In the last baseline experiment, a standard cross entropy loss function is used to train Fast RCNN.

Table III reports the detection results in terms of mAP for these three baseline experiments. We can observe that: 1) the designed entropy-based criterion can offer 1.82% mAP gains compared with the widely used confidence-based criterion. A possible explanation is that the designed entropy-based criterion measure image difficulty by analyzing the distribution of detection scores for all object proposals, through which we can obtain a global understanding of the image and provide a more accurate difficulty evaluation; 2) our DCL strategy leads to a significant improvement of 5.8% in terms of mAP compared to the curriculum learning strategy, which clearly demonstrate the superiority of our learning strategy for training more robust object detectors under weakly supervised settings. Fig. 6 presents a schematic illustration of our learning strategy. As can be seen, the trained object detectors can

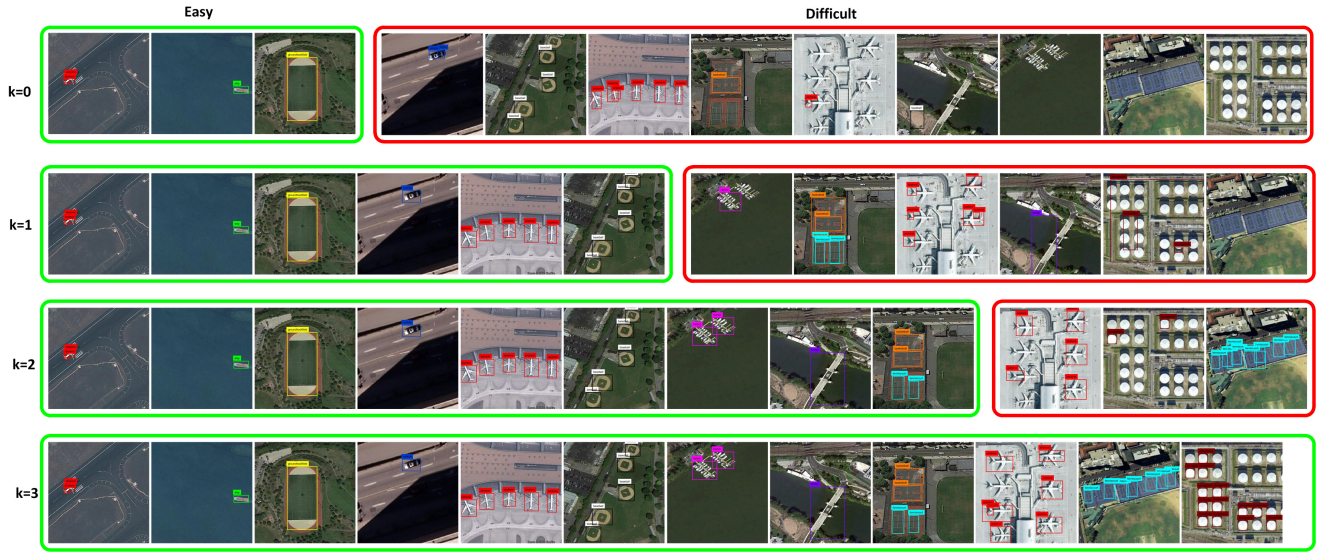


Fig. 6. Schematic illustration of our DCL strategy.

progressively increase their detection skills to localize more accurate instances in more difficult images. Accordingly, the difficult images are reconsidered as easy for the stronger object detectors; 3) the proposed instance-aware focal loss improves 2.79% in terms of mAP compared with cross entropy loss for training Fast RCNN, which illustrates that paying diffident attention on diffident instances in the training benefits learning more discriminative object detectors and further improving the detection performance.

V. Conclusion

In this article, a novel DCL strategy is proposed to perform WSOD from high-resolution remote sensing images, which can progressively learn the object detectors by feeding training images with increasing difficulty that matches current detection ability. To this end, the WSDDN framework is firstly used to generate an initial curriculum that ranks training images in ascending order of difficulty according to the designed entropy-based image difficulty measure criterion. Then, the object detectors based on Fast RCNN are learned by using instances mined from the easy images of the curriculum. In the learning, an effective instance-aware focal loss function is developed to alleviate the influence of positive instances of bad quality and meanwhile enhance the discriminative information of class-specific hard negative instances. With the stronger detection ability obtained from training on easy images, the subsequent order in the curriculum for retraining detectors is accordingly adjusted by promoting difficult images as easy. In such way, the detectors gradually improve their detection ability more effectively. Comprehensive evaluations and comparisons with state-of-the-art methods on two publicly available data sets demonstrate the superiority of our proposed method.

References

[1] J. Liu, Z. Wu, J. Li, A. Plaza, and Y. Yuan, "Probabilistic-kernel collaborative representation for spatial–spectral hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 2371–2384, Apr. 2016.

[2] B. Zhang *et al.*, "Remotely sensed big data: Evolution in model development for information extraction [point of view]," *Proc. IEEE*, vol. 107, no. 12, pp. 2294–2301, Dec. 2019.

[3] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[4] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 11–28, Jul. 2016.

[5] 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.

[6] 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.

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

[8] X. Yao, J. Han, L. Guo, S. Bu, and Z. Liu, "A coarse-to-fine model for airport detection from remote sensing images using target-oriented visual saliency and CRF," *Neurocomputing*, vol. 164, pp. 162–172, Sep. 2015.

[9] X. Chen, S. Xiang, C.-L. Liu, and C.-H. Pan, "Vehicle detection in satellite images by hybrid deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 10, pp. 1797–1801, Oct. 2014.

[10] J. Tang, C. Deng, G.-B. Huang, and B. Zhao, "Compressed-domain ship detection on spaceborne optical image using deep neural network and extreme learning machine," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 3, pp. 1174–1185, Mar. 2015.

[11] Y. Yu, H. Guan, D. Zai, and Z. Ji, "Rotation-and-scale-invariant airplane detection in high-resolution satellite images based on deep-Hough-forests," *ISPRS J. Photogramm. Remote Sens.*, vol. 112, pp. 50–64, Feb. 2016.

[12] X. Han, Y. Zhong, and L. Zhang, "An efficient and robust integrated geospatial object detection framework for high spatial resolution remote sensing imagery," *Remote Sens.*, vol. 9, no. 7, p. 666, 2017.

[13] Y. Xu, Z. Wu, J. Li, A. Plaza, and Z. Wei, "Anomaly detection in hyperspectral images based on low-rank and sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 1990–2000, Apr. 2016.

[14] 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.

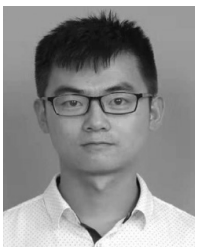
[15] D. Zhang, J. Han, G. Cheng, Z. Liu, S. Bu, and L. Guo, "Weakly supervised learning for target detection in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 4, pp. 701–705, Apr. 2015.

[16] 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.

[17] L. Cao *et al.*, "Weakly supervised vehicle detection in satellite images via multi-instance discriminative learning," *Pattern Recognit.*, vol. 64, pp. 417–424, Apr. 2017.
[18] F. Zhang, B. Du, L. Zhang, and M. Xu, "Weakly supervised learning based on coupled convolutional neural networks for aircraft detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 9, pp. 5553–5563, Sep. 2016.
[19] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 41–48.
[20] 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.
[21] 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. 2843–2851.
[22] E. Sangineto, M. Nabi, D. Culibrk, and N. Sebe, "Self paced deep learning for weakly supervised object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 712–725, Mar. 2019.
[23] 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.
[24] F. Wan, P. Wei, Z. Han, J. Jiao, and Q. Ye, "Min-entropy latent model for weakly supervised object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2395–2409, Oct. 2019.
[25] D. Zhang, J. Han, L. Zhao, and D. Meng, "Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework," *Int. J. Comput. Vis.*, vol. 127, no. 4, pp. 363–380, Apr. 2019.
[26] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2846–2854.
[27] 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.
[28] M. P. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1189–1197.
[29] Y. Li, L. Liu, C. Shen, and A. van den Hengel, "Image co-localization by mimicking a good detector's confidence score distribution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 19–34.
[30] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
[31] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.
[32] 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.
[33] 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.
[34] 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.
[35] 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.
[36] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[37] 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, pp. 119–132, Dec. 2014.

Xiwen Yao 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.



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, Xi'an, China.

His research interests include computer vision and 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.

Gong Cheng 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 are computer vision and pattern recognition.

Lei Guo received the B.S. and M.S. degrees from Xidian University, Xi'an, China, in 1982 and 1986, respectively, and the Ph.D. degree from Northwestern Polytechnical University, Xi'an, in 1993.

He is a Professor with Northwestern Polytechnical University. His recent research interest focuses on image processing.