Weakly Supervised Rotation-Invariant Aerial Object Detection Network

Xiaoxu Feng Xiwen Yao* Gong Cheng Junwei Han
School of Automation, Northwestern Polytechnical University, Xi'an, China
fengxiaox@mail.nwpu.edu.cn, yaoxiwen517@gmail.com, {gcheng,jhan}@nwpu.edu.cn

Abstract

Object rotation is among long-standing, yet still unexplored, hard issues encountered in the task of weakly supervised object detection (WSOD) from aerial images. Existing predominant WSOD approaches built on regular CNNs which are not inherently designed to tackle object rotations without corresponding constraints, thereby leading to rotation-sensitive object detector. Meanwhile, current solutions have been prone to fall into the issue with unstable detectors, as they ignore lower-scored instances and may regard them as backgrounds. To address these issues, in this paper, we construct a novel end-to-end weakly supervised Rotation-Invariant aerial object detection Network (RINet). It is implemented with a flexible multi-branch online detector refinement, to be naturally more rotation-perceptive against oriented objects. Specifically, RINet first performs label propagating from the predicted instances to their rotated ones in a progressive refinement manner. Meanwhile, we propose to couple the predicted instance labels among different rotation-perceptive branches for generating rotation-consistent supervision and meanwhile pursuing all possible instances. With the rotation-consistent supervisions, RINet enforces and encourages consistent yet complementary feature learning for WSOD without additional annotations and hyper-parameters. On the challenging NWPU VHR-10.v2 and DIOR datasets, extensive experiments clearly demonstrate that we significantly boost existing WSOD methods to a new state-of-the-art performance. The code will be available at: https://github.com/XiaoxFeng/RINet.

1. Introduction

Object detection is an indispensable task in both computer vision and earth vision with many applications. Recent impressive progress in object detection has been boosted by the boom of powerful deep Convolutional Neural Network (CNN) and the availability of abundant datasets with subtle annotations. However, collecting such subtle annotations is time-consuming and even infeasible, which has seriously impeded the applications of object detection in the real-world. To alleviate the heavy label cost, WSOD, which requires only incomplete image-level annotations to learn the precise object detection model, has been extensively explored and achieved impressive results.

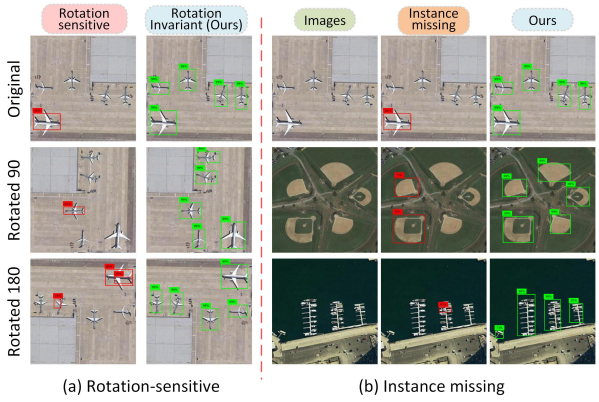

Figure 1. Typical issues and our solutions for WSOD in aerial images. (a) The image and its rotated image produce inconsistent detection results. (b) Existing WSOD methods incline to detect salient objects or object parts, leading to instances missing.

As far as we know, almost all predominated WSOD methods [5,7,8,10,14,17,19,24,27,30,31,33–35,39,45,47] are built on the Weakly Supervised Deep Detection Network (WSDDN) [14] and formulate WSOD as multiple instance learning problems. Based on it, a constructive work, named Online Instance Classifier Refinement (OICR) [31], is proposed to iteratively refine instance classifier in a unified network. More recently, some advanced works [9,17,19,24,25,27,30,33,34,43] are proposed to boost the development of WSOD via adopting novel training strategies [19,27,33,34,44,46], contextual information [17] or extra segmentation networks [9,24,38].

The typical WSOD approaches [7,36,37,42] in aerial images are mainly inspired by the object detection algorithms developed for natural scenes and endeavor to address the sub-optimal problem. Despite their successes, such ill-posed solution ignores the property of aerial images, that is, many object instances with the same category in aerial

*Corresponding author.

images usually appear with arbitrary orientations. It introduces dramatic class-agnostic feature changes, causing sparse feature distribution. Existing predominant WSOD approaches based on regular CNNs which cannot actively encourage such sparse features to be pulled closer without corresponding constraints, causing two typical issues.

(1) Rotation-sensitive. As shown in Figure 1 (a), existing methods incline to detect rotation-insensitive object parts and the detection results are inconsistent after rotation even for the same instance. A natural approach to address it is to use instance-level labels where object rotations come from themselves or rotated transformation, whereas WSOD does not have such annotations. Thus, it is regarded as amongst the hardest challenge of WSOD with no effective solutions.

(2) Instance missing. Most of the existing WSOD works only explore the most discriminative object. Unfortunately, it is common for an aerial image to contain many instances with the same category. This kind of solution leads to seemingly representative yet unstable object detector learning, as it will inevitably introduce class collision problem. Example testifying this issue is illustrated in Figure 1 (b). The ignored lower-scored instances may be regarded as background. A straightforward way to pursue all possible instances is to mine top-ranking instances. However, it is impractical to eliminate uncertainties and trivial solutions for each category under the weakly supervised paradigm.

To tackle the aforementioned issues, in this paper, we propose a novel weakly supervised rotation-invariant aerial object detection network (RINet), and aim at learning rotation-invariant object detectors and pursuing all possible instances. RINet is inspired by human knowledge *i.e.*, the category of object in aerial images remains consistent after arbitrary rotation. It can be treated as an implicit constraint for rotation-invariance learning. Encouraging the detection model to make consistent prediction for the predicted instances before and after the rotation can facilitate rotation-invariant learning online. To this end, RINet is implemented with a flexible multi-branch online detector refinement where the predicted instance labels supervise their arbitrary rotated ones in the latter stream. Finally, the instance-level labels before and after rotation are coupled to generate rotation-consistent annotations for rotation-invariant learning online.

Furthermore, RINet also naturally projects object instances from sparse space to different rotation-aware subspaces, which encourages the same category object instances with similar orientations to be pulled closer on the embedding space. Motivated by this, coupling instances from different rotation-perceptive branches in a complementary manner is conducive to discover all possible instances with the same category. RINet greedily projects predicted labels from different rotation-perceptive branches to an interaction space. Within this interaction space, label propagating is performed over the unlabeled instances under implicit constraint to activate instances in a complementary manner. Integrating all possible instances into the iterative training process can capture abundant intra-class complementary visual patterns to facilitate a more powerful rotation-invariant object detector.

By leveraging category-invariance in rotation, a flexible weakly supervised rotation-invariant object detection network is proposed. It not only bridges the gap existing in the object rotation but also provides the reliable and implicit constraint for instance mining. With an end-to-end learning procedure, as shown in Figure 1, RINet effectively alleviates the aforementioned challenges and generates consistent detection results. The main contributions of this paper are as follows:

- To the best of our knowledge, we are the first attempt to construct a rotation-invariant aerial object detection network under a weakly supervised paradigm, and jointly optimize instance refinement and rotation-invariant object detector in a systematic end-to-end manner.
- We design a rotation-invariant multiple instance mining strategy, coupling instances from different rotation-perceptive branches in a complementary manner, to mine all possible object instances of the same category without introducing additional hyper-parameters.
- Experiments on NWPU VHR-10.v2 [22] and DIOR [23] datasets demonstrate that the proposed RINet significantly updates the performance of state-of-the-art results by a large margin.

2. Related Work

2.1. Weakly Supervised Object Detection

Weakly supervised object detection from both natural scene images and aerial images has been extensively explored and become a well-studied research field in recent years. Most of advanced researches [3, 7, 8, 10, 14, 17, 19, 24, 27, 30, 31, 33, 34, 40, 45] attempt to exploit multiple instance learning (MIL) to address the WSOD task. Following the MIL constraints, the high-scoring positive bags are assigned with the pseudo instance-level label to learn the corresponding object detector. WSDDN [14] is of the first to implement WSOD with MIL in an end-to-end manner and inspires follow-up researches. For example, Tang et al. [31] propose a novel OICR framework by creatively integrating multi-stage classifiers into [14]. In OICR, each stream provides the pseudo instance-level annotation for the next stream learning to perform better detector learning. Based on it, a host of OICR-based WSOD methods [10, 17, 19, 20, 24, 25, 27, 30, 33, 34, 43] are developed

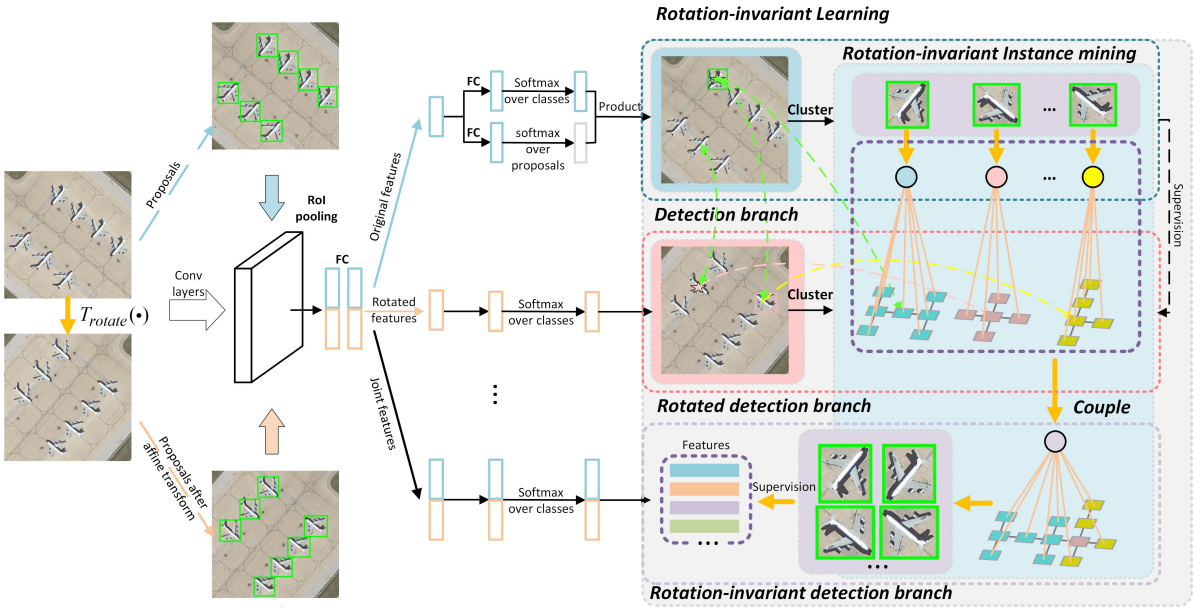

Figure 2. Illustration of the proposed RINet. To address the issues of object rotation and instance missing, RINet first encourages the object detector to make the same prediction for the predicted instances before and after the rotation. Meanwhile, RINet provides implicit constraint for pursuing all possible instances by coupling predicted instance labels among different rotation-perceptive branches.

to further boost the performance of WSOD via introducing more flexible and credible instances mining strategies. However, these methods merely mine the most confident instance while fail to extract other instances of the same class existing in an image. Thus, the above WSOD methods cannot be directly employed to perform object detection from aerial images under weakly supervised settings. This is principal because the aerial images may always contain more than one same class instance. There is no doubt that weakly supervised object detection from aerial images is a more challenging task.

To tackle this challenge, Han et al. [16] design an instance mining strategy from the negative data to refine the WSOD model. More recently, Yao et al. [42] attempt to mine high-quality instances by introducing a dynamic curriculum learning strategy. Wang et al. [36] introduce a multiple instance graph strategy to find high-quality objects via constructing spatial and appearance graph. Feng et al. [8] introduce a triple context-aware network to tackle the issue of grouped instances in aerial images. Despite promising performance, the introduced extra hyper-parameters limited their applications. In contrast, here we attempt to mine all possible instances by taking advantage of the implicit rotation invariance without extra hyper-parameters.

2.2. Rotation-invariant Learning

Object rotation is a major challenge for object detection in aerial images. To this end, existing advanced works [2,4,6,13,15,18,22,41] aim to learn rotation-invariant features by designing learnable rotation-sensitive CNNs. For instance, Cheng et al. [4] construct a rotation-invariant CNN model to learn the rotation-invariant feature representations. Li et al. [22] design a multi-angle anchors based RPN [26] to alleviate the problem of object rotations. The work [6] also introduces a rotation-invariant local binary descriptor so that the orientation for each pattern can be adaptively learned. More recently, a novel oriented detection module [15] is constructed to encode the orientation information and capture rotation-invariant features by adopting active rotating filters. Deepak et al. [13] introduce a rotation-equivariant Siamese network. However, the above methods rely on the subtle manually-labeled annotations. Labeling such subtle annotations is laborious, time-consuming, or even impractical. To the best of our knowledge, the proposed RINet is of the first to address the rotation variation in an end-to-end manner under weakly supervised settings. Moreover, we also flexibly utilize the implicit constraint existing in the object rotation to mine all possible instances.

3. Basic WSOD Framework

In this paper, we choose OICR [31] as our basic framework for WSOD for its expansibility and effectiveness. OICR [31] adopts a flexible instance refinement branch to propagate the binary-level label from the most discriminative region to its adjacent regions. By repeat-

edly implementing the refinement procedure, the latent detector can effectively diagnose the localization of objects. Formally, let $\mathcal{I}$ denote an input image and $Y_i = [y_1, ..., y_c, ..., y_C] \in \{-1, 1\}$ is the image-level label indicating whether an object category appear in an image. $\mathcal{H} \in \mathcal{R}_i$ is the corresponding region proposals which are generated by [32]. We first feed the input image $\mathcal{I}$ and its region proposals $\mathcal{H}$ into a CNN with ROI-pooling [11] to extract corresponding feature vectors $\mathcal{F}_{\mathcal{H}}$. Following WSDDN [14], two parallel branches are employed to generate classification logit $\Psi^{cls}(c, \mathcal{F}_{\mathcal{H}})$ and detection logit $\Psi^{\text{det}}(c, \mathcal{F}_{\mathcal{H}})$ for each region where c denotes the number of image category. Then, two matrices $\Psi^{cls}(c, \mathcal{F}_{\mathcal{H}})$ and $\Psi^{\text{det}}(c, \mathcal{F}_{\mathcal{H}})$ are passed through the softmax operator along the category dimension and proposal dimension to generate corresponding scores $\text{x}^{\text{cls}}=\text{e}^{\Psi^{cls}(c,\mathcal{F}_{\mathcal{H}})} / \sum_{c\in C} e^{\Psi^{cls}(c,\mathcal{F}_{\mathcal{H}})}$, $\text{x}^{\text{det}}=\text{e}^{\Psi^{det}(c,\mathcal{F}_{\mathcal{H}})} / \sum_{\mathcal{H}\in\mathcal{R}_i} e^{\Psi^{det}(c,\mathcal{F}_{\mathcal{H}})}$. These two matrices denote the probability for each proposal belonging to category c and the contribution for each proposal being classified as category c, respectively. The proposal scores are produced via performing an element-wise product: $s(c|\mathcal{H}) = \text{x}^{\text{cls}} \odot \text{x}^{\text{det}}$. Lastly, the image score is generated by the sum over all proposal scores: $\phi(c) = \sum_{\mathcal{H}\in\mathcal{R}_i} s(c|\mathcal{H})$ and the multi-class cross entropy is applied to supervise the model training:

$$\mathcal{L}_{MIL} = -\sum_{c\in C} y_c \log \phi(c) + (1 - y_c) \log(1 - \log \phi(c)). \tag{1}$$

However, WSDDN [14] inclines to discover the most discriminative object parts rather than the full object. To tackle this issue, OICR [31] integrates multi-stage refinement branches into WSDDN [14] where the most confident region and its highly-overlapped adjacent regions in the former refinement branch are treated as pseudo instance-level label $\hat{Y}_c$ to supervise its latter refinement branch learning. Note that, in each refinement branch, the feature vectors of proposal are branched into a $\{C+1\}$-dimensional instance classifier, leading to $s(c|\mathcal{H}_r)$, where the $\{C+1\}^{th}$ dimension is for background. The parameters of the refinement branch are optimized via a weighted softmax loss function:

$$\mathcal{L}_{OICR} = -\frac{1}{|\mathcal{H}|} \sum_{r=1}^{|\mathcal{H}|} \sum_{c=1}^{C+1} \omega \hat{Y}_c \log s(c|\mathcal{H}_r), \tag{2}$$

where ω is an adaptive parameter to alleviate the interference of noise.

4. Weakly Supervised Rotation-invariant Object Detection Network

4.1. Overview

The overview of the proposed RINet is outlined in Figure 2. Building upon OICR [31], RINet addresses the issues of rotation-sensitive and instance missing via encouraging consistent and complementary learning in two modules: rotation-invariant learning and multiple instance mining. Specifically, we first simultaneously feed images before and after rotation transformation into a unified multiple instance detection network which consists of a detection branch, a rotated detection branch, and a rotation-invariant detection branch. The rotation-invariant module generates rotation-consistent labels to encourage the object detector to make the same prediction for the labeled instances before and after the rotation, thereby facilitating the detector to capture rotation-invariant features. Meanwhile, all possible instances are mined by coupling predicted instance labels among different rotation-perceptive branches in a complementary manner.

4.2. Rotation-invariant Learning

Rotation transformation is a common data augmentation and has been widely used in fully supervised object detection in aerial images. Noting that the instance-level labels also keep the same affine transformation. It introduces an implicit constraint to facilitate rotation-invariant learning under a fully supervised paradigm. Yet the unavailability of instance-level labels causes the implicit constraint missing here. Thus, to realize the rotation-invariant learning, the key lies in how to leverage this implicit constraint in a weakly supervised manner.

It is common sense that for the same instance in aerial images, e.g. "airplane" and "baseball field", through arbitrary rotation, their categories are unchanged. Similarly, we can draw the conclusion that is the pseudo instance-level labels obtained by the WSOD model also keep the same affine transformation in the image after rotation. Based on the above analyses, we propose a flexible and effective RINet to drive the detection network to make the same prediction for the labeled instances before and after the rotation.

Given a pair of input images including the original image $\mathcal{I}$ and its rotated image $\mathcal{I}^{rotate} = T_{rotate}(\mathcal{I})$, $\mathcal{H}$ and $\mathcal{H}^{rotate}$ are corresponding region proposals, respectively. We feed them into the same WSOD network to generate the image feature maps and then employ ROI pooling to obtain corresponding feature vectors $\mathcal{F}_{\mathcal{H}}$ and $\mathcal{F}_{\mathcal{H}^{rotate}}$ of proposal, respectively. As illustrated in Figure 2, the proposal feature vectors of original image $\mathcal{F}_{\mathcal{H}}$, rotated image $\mathcal{F}_{\mathcal{H}^{rotate}}$ and its joint proposal feature vectors $\mathcal{F}_{\mathcal{H}^J} = Cat(\mathcal{F}_{\mathcal{H}}, \mathcal{F}_{\mathcal{H}^{rotate}})$ are branched into the detection branch, rotated detection branch and rotation-invariant branch to produce the corresponding classification probability $s(c|\mathcal{H}) \in \mathbb{R}^{|\mathcal{H}|\times(C+1)}$, $s(c|\mathcal{H}^{rotate}) \in \mathbb{R}^{|\mathcal{H}|\times(C+1)}$, and $s(c|\mathcal{H}^J) \in \mathbb{R}^{2|\mathcal{H}|\times(C+1)}$, respectively. For each class existing in the image ($y_c = 1$), we obtain the pseudo instance label $\hat{Y}_c$ by the same way as [31] in the detection branch. According to the implicit constraint, if the r^{th} object pro-

posal $\mathcal{H}_r$ is selected as positive instance or background in the detection branch, the r^{th} rotated proposal $\mathcal{H}_r^{rotate}$ in rotation detection branch also should have the same category with it. Thus, we can leverage the pseudo instance-level label obtained by the original detection branch to supervise the rotated branch. Similarly, the rotated ones also can supervise the original branch. In our experiment, we just need to ensure that different detection branches are fed into the same image with different affine transformations. The multiple instance detection network can be trained by:

$$\mathcal{L}_{rotate} = -\frac{1}{|\mathcal{H}|}\sum_{r=1}^{|\mathcal{H}|}\sum_{c=1}^{C+1}\omega\hat{Y}_c \ \log s(c|\mathcal{H}_r^{rotate}). \quad (3)$$

Iteratively propagating image-level labels from the predicted instances to their rotated instances and highly-overlapped adjacent regions encourages instance classifier to pursue full object extent. Moreover, it also facilitates the multiple instance detection network to discover the same category instances with different orientations in an image. Next, we generate credible and rotation-consistent supervision from the same category with different orientations like full supervised settings via coupling both supervisions from the detection branch and rotated detection branch:

$$\begin{cases}\hat{Y}_c^I = \{\hat{Y}_c\}_r \cup \{\hat{Y}_c\}_{r+|\mathcal{H}|} \\ \hat{Y}_c = \arg\max_c \mathcal{J}(s(c|\mathcal{H}_r), s(c|\mathcal{H}_r^{rotate}))\end{cases}, \quad (4)$$

where $\hat{Y}_c^I \in \mathbb{R}^{2|\mathcal{H}|}$, r denotes the index of pseudo instance-level label, and $\mathcal{J}(\cdot)$ is branch-wise average pooling. The rotation-invariant learning can be realized by enforcing the object detector to make the same prediction for the positive instances before and after the rotation. The parameters of rotation-invariant detection branch are optimized by:

$$\mathcal{L}_{RI} = -\frac{1}{2\,|\mathcal{H}|}\sum_{r=1}^{2|\mathcal{H}|}\sum_{c=1}^{C+1}\omega\hat{Y}_c^I \ \log s(c|\mathcal{H}_r^J). \quad (5)$$

4.3. Multiple Instance Mining

Although selecting the most confident region before and after rotation realizes rotation-invariant learning, such a solution also ignores the important fact that aerial images usually contain many instances with the same category. It has been turned out to be a major reason causing the performance of WSOD inferior to the full supervised object detection. To address this issue, we take advantage of the implicit constrain in RINet to pursue all possible instances without extra hyper-parameters.

To this end, similar to PCL [30], we first adopt K-means to generate a set of clusters according to their proposal scores and then select the proposals from the highest-score cluster as top-ranking proposals in both detection branch and rotated detection branch. Next, we preliminarily construct corresponding undirected unweighted graphs $G_c^o = (V_c^o, E_c^o)$ and $G_c^r = (V_c^r, E_c^r)$ according to their spatial similarity, where vertexes V_c represent these top-ranking proposals, and each edge in E_c correspond to the spatial similarity between vertexes. Vertexes with enough spatial similarity are connected and labeled to the same category. The spatial similarity is generated via computing IoU between vertexes. After that, we project $G_c^o = (V_c^o, E_c^o)$ and $G_c^r = (V_c^r, E_c^r)$ to an interaction space which is more friendly for instance mining according to the implicit constraint. Within this interaction space, we build a graph $G_c^I = (V_c^I, E_c^I)$ to connect the graph from original space and rotated space and perform label propagating over the graph, where $V_c^I = \{G_c^o, G_c^r\}$ and E_c^I is the implicit constraint existing in the rotation. After the propagating, the updated graph is then projected back to original space $\hat{Y}_{G_c^I} \rightarrow \hat{Y}_c^o$ to supervise the rotation-invariant branch learning. Thus, we update the pseudo instance-level label $\hat{Y}_c^I \in \mathbb{R}^{2|\mathcal{H}|}$ via directly propagating the labels of all vertexes in graph G_c^I to rotation-invariant branch:

$$\hat{Y}_c^I = \{\hat{Y}_{G_c^I}\}_r \cup \{\hat{Y}_{G_c^I}\}_{r+|\mathcal{H}|}. \quad (6)$$

In such a way, on one hand, we can employ more instances with the same category to learn the more robust detector. On the other hand, encouraging the more same-category instances with different orientations for training further drives the more robust rotation-invariant learning.

Baseline	RINet	Cluster	MIM	mAP	CorLoc
✓				18.7	43.3
	✓			26.6	48.8
	✓	✓		27.1	51.4
	✓		✓	28.3	52.8

Table 1. Results (%) for different components of RINet on the DIOR trainval and testing set.

5. Experiments

5.1. Datasets and Evaluation Metric

RINet is validated on the commonly used NWPU VHR-10.v2 [22] and DIOR [23] datasets. NWPU VHR-10.v2 [22] is a classical aerial image dataset with the size of 400×400, including 10 object categories. DIOR [23] is a popular public dataset for fully supervised object detection in aerial images but rarely explored with WSOD owing to its hard challenges. It consists of 23463 images including 192472 instances with the size of 800×800 and covers 20 different categories (*i.e.*, Airplane (PL), Airport (AP), Baseball field (BF), Basketball court (BC), Bridge (BR), Chimney (CM), Dam (DA), Expressway service area (ES), Expressway toll station (ET), Golf field (GF), Ground track

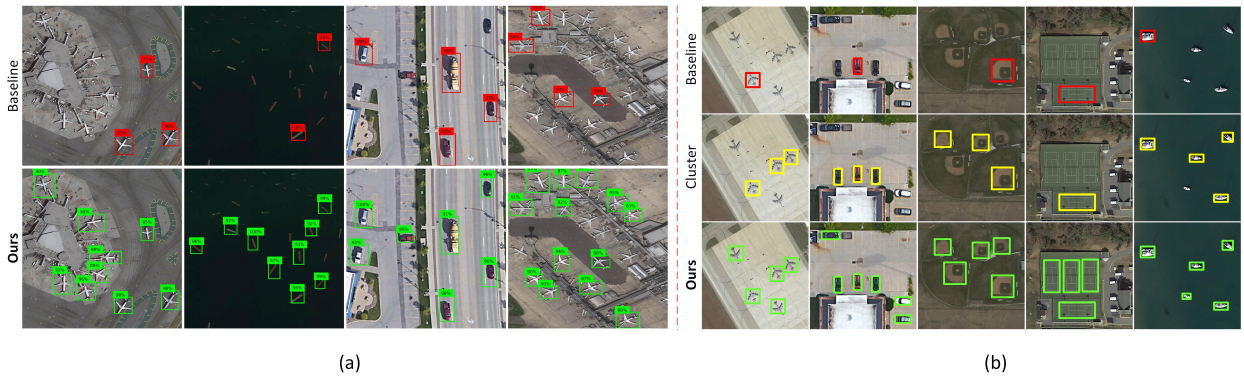

Figure 3. Ablation studies for both rotation-invariant learning and multiple instance mining. (a) Visualization of detection results for rotated objects. (b) Instances discovered by RINet with different instance mining strategies in the training stage.

Methods	Airplane	Ship	Storage tank	Baseball Diamond	Tennis court	Basketball court	Ground track field	Harbor	Bridge	Vehicle	mAP
COPD [1]	62.3	69.4	64.5	82.1	34.1	35.3	84.2	56.3	16.4	44.3	54.9
Transferred CNN [21]	66.0	57.1	85.0	80.9	35.1	45.5	79.4	62.6	43.2	41.3	59.6
RICNN [4]	88.7	78.3	86.3	89.1	42.3	56.9	87.7	67.5	62.3	72.0	73.1
RCNN [12]	85.4	88.9	62.8	19.7	90.7	58.2	68.0	79.9	54.2	49.9	65.8
Fast RCNN [11]	90.9	90.6	89.3	47.3	100.0	85.9	84.9	88.2	80.3	69.8	82.7
Faster RCNN [26]	90.9	86.3	90.5	98.2	89.7	69.6	100.0	80.1	61.5	78.1	84.5
RICO [22]	99.7	90.8	90.6	92.9	90.3	80.1	90.8	80.3	68.5	87.1	87.1
WSDDN [14]	30.1	41.7	35.0	88.9	12.9	23.9	99.4	13.9	1.9	3.6	35.1
OICR [31]	13.7	67.4	57.2	55.2	13.6	39.7	92.8	0.2	1.8	3.7	34.5
PCL [30]	26.0	63.8	2.5	89.8	**64.5**	76.1	77.9	0.0	1.3	15.7	39.4
DCL [42]	72.7	74.3	37.1	82.6	36.9	42.3	84.0	39.6	16.8	35.0	52.1
PCIR [7]	**90.8**	78.8	36.4	90.8	22.6	52.2	88.5	42.4	11.7	35.5	55.0
TCANet [8]	89.4	78.2	78.4	**90.8**	35.3	50.4	90.9	42.4	4.1	28.3	58.8
Ours	90.3	**86.3**	**79.6**	90.7	58.2	**80.4**	**100.0**	**57.7**	**18.9**	**41.6**	**70.4**

Table 2. Average precision (%) for different methods on the NWPU VHR-10.v2 testing set.

field(GTF), Harbor (HB), Overpass (OP), Ship (SH), Stadium (SD), Storage tank (ST), Tennis court (TC), Train station (TS), Vehicle (VH), Wind mill (WM)). Both datasets are split into three subsets, *i.e.*, training set, validation set, and testing set. Noting that almost every image in both datasets contains more than one instance with different orientations. Following the standard routine in WSOD, RINet is trained on the both training set and the validation set, referred to as the trainval set, and evaluated on the testing set. Meanwhile, solely image-level labels are available during the model training. Average Precision (AP) and correct localization accuracy (CorLoc) are employed to evaluate the accuracy of object detection and localization, respectively. All these two metrics are performed on the PASCAL criteria, *i.e.*, IoU threshold at 50%.

5.2. Implementation Details

For a fair comparison, VGG16 [29] pre-trained on the ImageNet [28] is adopted as the backbone and all newly added layers are initialized with a Gaussian distribution with 0-mean and 0.01-standard. Meanwhile, we keep the training settings including learning rate, mini-batch, weight decay, and momentum identical to [30,31,34]. They are set to 0.001, 2, 0.005, and 0.9, respectively. SGD is applied for optimization. The Selective Search [32] is adopted to generate about 2000 proposals per image. We augment training data with three rotation transformations {90°,180°,270°}. During training, RINet performs 20K and 200K iterations and its learning rate will shrink by a factor of 10 every 10K and 100K iterations for NWPU VHR-10.v2 and DIOR datasets, respectively. 0.3 is set as a confidence threshold for NMS to remove duplicated bounding boxes. All experiments are implemented with Pytorch on ubuntu16.04, NVIDIA Tesla V100, cuDNN v5, and CUDA 9.0. Inspired by [17,34], we also set the adaptive weights in OICR loss as $\omega = 0.1$ and generate new baseline (18.7% mAP and 43.3% CorLoc).

5.3. Ablation Studies

Effect of rotation-invariant learning. Our RINet is built upon the OICR [31]. Compared with it, we simultaneously feed images before and after rotation and their corresponding region proposals into the WSOD Network. Meanwhile, we modify its refinement branch as rotated detection branch and rotation-invariant detection branch but remain its instance refinement strategy. The input of the rotation branch and rotation-invariant branch are proposal features after rotation and the joint proposal features before and af-

Methods	PL	AP	BF	BC	BR	CM	DA	ES	ET	GF	GTF	HB	OP	SH	SD	ST	TC	TS	VH	WM	mAP
Fast RCNN [11]	44.2	66.8	67.0	60.5	15.6	72.3	52.0	65.9	44.8	72.1	62.9	46.2	38.0	32.1	71.0	35.0	58.3	37.9	19.2	38.1	50.0
Faster RCNN [26]	50.3	62.6	66.0	80.9	28.8	68.2	47.3	58.5	48.1	60.4	67.0	43.9	46.9	58.5	52.4	42.4	79.5	48.0	34.8	65.4	55.5
WSDDN [14]	9.1	39.7	37.8	20.2	0.3	12.2	0.6	0.7	11.9	4.9	42.4	4.7	1.1	0.7	63.0	4.0	6.1	0.5	4.6	1.1	13.3
OICR [31]	8.7	28.3	44.1	18.2	1.3	20.2	0.1	0.7	29.9	13.8	57.4	10.7	11.1	9.1	59.3	7.1	0.7	0.1	9.1	0.4	16.5
PCL [30]	21.5	35.2	59.8	23.5	3.0	43.7	0.1	0.9	1.5	2.9	56.4	16.8	11.1	9.1	57.6	9.1	2.5	0.1	4.6	4.6	18.2
DCL [42]	20.9	22.7	54.2	11.5	6.0	61.0	0.1	1.1	31.0	30.9	56.5	5.1	2.7	9.1	63.7	9.1	10.4	0.0	7.3	0.8	20.2
PCIR [7]	**30.4**	36.1	54.2	26.6	9.1	58.6	0.2	9.7	36.2	32.6	**58.5**	8.6	**21.6**	**12.1**	**64.3**	9.1	13.6	0.3	9.1	**7.5**	24.9
TCANet [8]	25.1	30.8	**62.9**	**40.0**	4.1	67.8	**8.1**	**23.8**	29.9	22.3	53.9	24.8	11.1	9.1	46.4	13.7	**31.0**	1.5	9.1	1.0	25.8
Ours	26.2	**57.4**	62.7	25.1	**9.9**	**69.2**	1.4	13.3	**36.2**	**51.4**	53.9	**28.6**	4.8	9.1	52.7	**15.8**	20.6	**12.9**	**9.1**	4.7	**28.3**

Table 3. Average precision (%) for different methods on the DIOR testing set.

Methods	PL	AP	BF	BC	BR	CM	DA	ES	ET	GF	GTF	HB	OP	SH	SD	ST	TC	TS	VH	WM	CorLoc
WSDDN [14]	5.7	59.9	94.2	55.9	4.9	23.4	1.0	6.8	44.5	12.8	89.9	5.5	10.0	23.0	98.5	79.6	15.1	3.5	11.6	3.2	32.4
OICR [31]	16.0	51.5	94.8	55.8	3.6	23.9	0.0	4.8	56.7	22.4	**91.4**	18.2	18.7	31.8	98.3	81.3	7.5	1.2	15.8	2.0	34.8
PCL [30]	61.1	46.9	95.4	63.6	7.3	95.1	0.2	5.7	5.1	50.8	89.4	42.1	19.8	37.9	97.9	80.7	13.8	0.2	10.5	**6.9**	41.5
PCIR [7]	81.6	51.3	**96.2**	**73.5**	5.0	94.7	**15.9**	32.8	46.0	48.6	85.3	38.9	**20.2**	30.6	84.6	**91.5**	56.3	3.8	10.5	1.3	48.4
TCANet [8]	91.2	69.4	95.5	67.5	**18.9**	**97.8**	0.2	70.5	54.3	51.4	88.3	48.0	2.3	33.6	14.1	83.4	**65.6**	19.9	**16.4**	2.9	49.4
Ours	**92.7**	**80.9**	92.7	69.5	8.6	90.1	0.2	**71.3**	**62.0**	**65.5**	85.1	**51.4**	15.7	**44.6**	**98.6**	80.3	14.8	**22.7**	6.9	2.6	**52.8**

Table 4. Correct localization (%) for different methods on the DIOR trainval set.

ter rotation, respectively. For a fair comparison, we adopt a consistent instance mining strategy as [31], *i.e.*, only selecting the top-scoring proposal and its high spatial overlapped regions as positive instances, to train our RINet. As shown in Table 1, our RINet achieves 26.6% mAP and 48.8% CorLoc, which significantly boost the baseline by a large margin (+7.9% mAP, and +5.5% CorLoc). Figure 3 (a) further exhibits qualitative comparisons of rotation-invariant learning between baseline and ours. It can be clearly seen that RINet can effectively generate nearly rotation-consistent results for the same category objects with different orientations, compared with baseline method. Comprehensive experiments show that the proposed RINet can effectively alleviate the issue of object rotation.

Effect of multiple instance mining. To disclose the contribution of the proposed instance mining strategy, we first integrate proposal cluster learning [30] into our RINet to mine the same class instances. As presented in Table 1, the performance of detection in terms of mAP and CorLoc are boosted by 0.5% and 2.6%, respectively. Then, we further integrate the proposed multiple instance mining (MIM) strategy into our RINet. We can observe that the proposed approach further brings 1.2% mAP and 1.4% CorLoc improvement, respectively. We also provide qualitative comparisons in Figure 3 (b) for instance mining in the training among our baseline, cluster learning, and proposed approach. It can be seen that RINet successfully discovers all possible instances with different orientations and significantly outperforms the performance of other approaches.

5.4. Comparison with State-of-the-arts.

In this section, we evaluate the proposed RINet on the NWPU VHR-10.v2 and DIOR datasets to provide comprehensive comparisons with the state-of-the-arts.

Table 2 shows quantitative comparisons for each class with existing advanced methods on the NWPU VHR-10.v2 dataset. Among existing weakly supervised approaches in remote sensing images, our RINet achieves the new state-of-the-art mAP of 70.4% and outperforms all others in the most categories. Compared with the baseline, our RINet achieves consistent improvement for each class by a large margin on the testing set. Moreover, Our RINet also outperforms the WSDDN [14], OICR* [31], PCL [30], DCL [42], PCIR [7], TCANet [8] by 35.3%, 35.9%, 31%, 18.3%, 15.4%, 11.6%, respectively which are notable margins in terms of mAP.

On the more challenging DIOR dataset, it can be seen in Table 3 and Table 4 that RINet significantly surpasses existing state-of-the-art with 28.3% mAP and 52.8% CorLoc, respectively, demonstrating the robustness of proposed RINet. This increase of performance mainly comes from the collaboration of rotation-invariant learning and multiple instance mining, which alleviates the issues of object rotation and instances missing.

On both NWPU VHR-10.v2 and DIOR datasets, we also present the results of RINet compared with advanced fully supervised methods. As shown in Table 2 and Table 3, we further narrow the gap between the WSOD and fully supervised methods. Noting that we obtain superior performance to some fully supervised approaches, such as COPD [1], Transferred CNN [21], RCNN [12].

Qualitative visualizations for both successful and failure examples on both NWPU VHR-10.v2 and DIOR datasets are shown in Figure 4 and Figure 5. It can be seen that our RINet can correctly localize multiple discrete instances with different orientations. However, our RINet also has trouble in addressing small objects and scene-ambiguous objects. For example, the detection model tends to discover more salient rivers under image-level labels with the bridge category, as bridges always co-exist with rivers. These remain challenging issues and we can consider introducing causal intervention in the future.

*It is implemented with Caffe

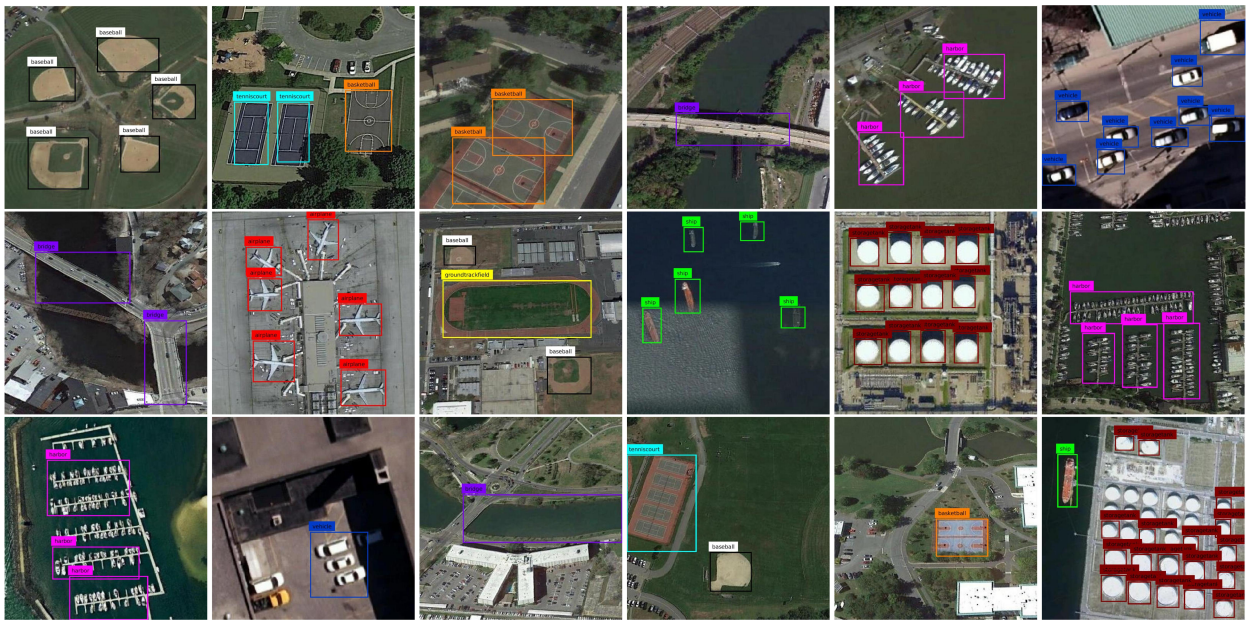

Figure 4. Visualization of detection results on the NWPU VHR-10.v2 testing split (70.4% mAP). The first two rows indicate corrected predictions and different colors rectangle indicates different classes. The third row denotes the failure cases.

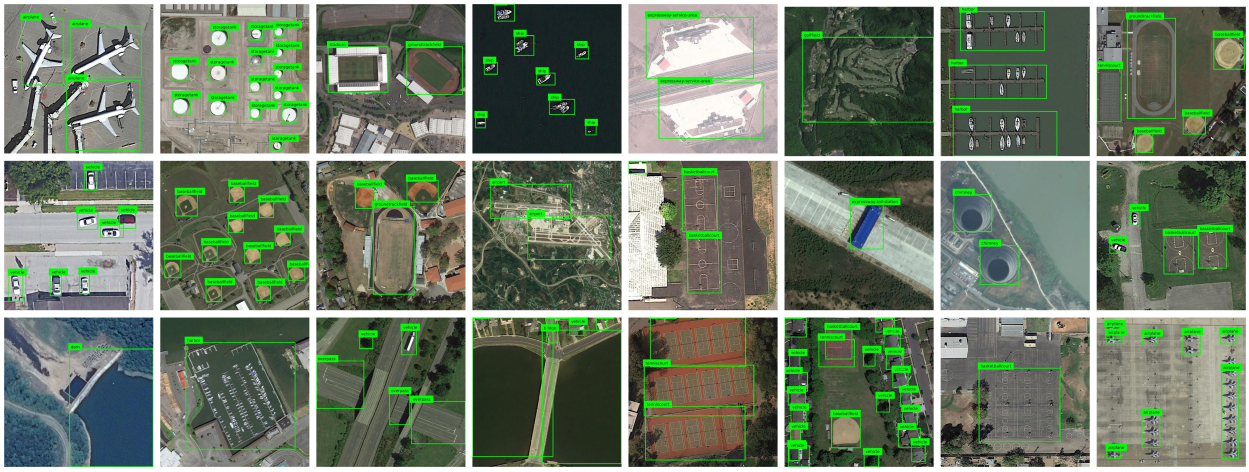

Figure 5. Qualitative results by RINet on the DIOR testing split (28.3% mAP). The first two rows indicate corrected predictions. The last row corresponds to the failure cases.

6. Conclusion

In this paper, we are of the first to address the object rotation issue via constructing a novel and flexible rotation-invariant aerial object detection network (RINet) under a weakly supervised paradigm. RINet is implemented with an online detector refinement with different rotated perceptions. During training, it generates rotation-consistent supervisions and meanwhile pursues all possible instances by coupling predicted instance labels among different rotation-perceptive branches in a complementary way. With all possible rotation-consistent supervisions, RINet jointly optimizes instance refinement and rotation-invariant object detector in an end-to-end manner, leading to rotation-invariant yet diversifying feature learning for WSOD. Comprehensive experiments demonstrate that the proposed RINet outperforms all existing WSOD methods, and produces a new state-of-the-art results.

7. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants 62071388, 62136007, 62036005 and U20B2068, the Key R&D Program of Shaanxi Province under Grant 2021ZDLGY01-08, and the National Key R&D Program of China under Grant 2020AAA0105701.

References

[1] Gong Cheng, Junwei Han, Peicheng Zhou, and Lei Guo. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS Journal of Photogrammetry and Remote Sensing*, 98:119–132, 2014. 6, 7

[2] Gong Cheng, Junwei Han, Peicheng Zhou, and Dong Xu. Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection. *IEEE Transactions on Image Processing*, 28(1):265–278, 2018. 3

[3] Gong Cheng, Junyu Yang, Decheng Gao, Lei Guo, and Junwei Han. High-quality proposals for weakly supervised object detection. *IEEE Transactions on Image Processing*, 29:5794–5804, 2020. 2

[4] Gong Cheng, Peicheng Zhou, and Junwei Han. Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(12):7405–7415, 2016. 3, 6

[5] Bowen Dong, Zitong Huang, Yuelin Guo, Qilong Wang, Zhenxing Niu, and Wangmeng Zuo. Boosting weakly supervised object detection via learning bounding box adjusters. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2876–2885, 2021. 1

[6] Yueqi Duan, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Learning rotation-invariant local binary descriptor. *IEEE Transactions on Image Processing*, 26(8):3636–3651, 2017. 3

[7] Xiaoxu Feng, Junwei Han, Xiwen Yao, and Gong Cheng. Progressive contextual instance refinement for weakly supervised object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 58(11):8002–8012, 2020. 1, 2, 6, 7

[8] Xiaoxu Feng, Junwei Han, Xiwen Yao, and Gong Cheng. Tcanet: Triple context-aware network for weakly supervised object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2020. 1, 2, 3, 6, 7

[9] Yan Gao, Boxiao Liu, Nan Guo, Xiaochun Ye, Fang Wan, Haihang You, and Dongrui Fan. C-midn: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9834–9843, 2019. 1

[10] Yan Gao, Boxiao Liu, Nan Guo, Xiaochun Ye, Fang Wan, Haihang You, and Dongrui Fan. Utilizing the instability in weakly supervised object detection. *arXiv preprint arXiv:1906.06023*, 2019. 1, 2

[11] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 4, 6, 7

[12] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 6, 7

[13] Deepak K Gupta, Devanshu Arya, and Efstratios Gavves. Rotation equivariant siamese networks for tracking. *arXiv preprint arXiv:2012.13078*, 2020. 3

[14] Andrea Vedaldi Hakan Bilen. Weakly supervised deep detection networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2846–2854, June 2016. 1, 2, 4, 6, 7

[15] Jiaming Han, Jian Ding, Jie Li, and Gui-Song Xia. Align deep features for oriented object detection. *arXiv preprint arXiv:2008.09397*, 2020. 3

[16] Junwei Han, Dingwen Zhang, Gong Cheng, Lei Guo, and Jinchang Ren. Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning. *IEEE Transactions on Geoscience and Remote Sensing*, 53(6):3325–3337, 2014. 3

[17] Zeyi Huang, Yang Zou, BVK Kumar, and Dong Huang. Comprehensive attention self-distillation for weakly-supervised object detection. *Advances in Neural Information Processing Systems*, 33, 2020. 1, 2, 6

[18] Ruoqiao Jiang, Shaohui Mei, Mingyang Ma, and Shun Zhang. Rotation-invariant feature learning in vhr optical remote sensing images via nested siamese structure with double center loss. *IEEE Transactions on Geoscience and Remote Sensing*, 2020. 3

[19] Satoshi Kosugi, Toshihiko Yamasaki, and Kiyoharu Aizawa. Object-aware instance labeling for weakly supervised object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6064–6072, 2019. 1, 2

[20] Satoshi Kosugi, Toshihiko Yamasaki, and Kiyoharu Aizawa. Object-aware instance labeling for weakly supervised object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6064–6072, 2019. 2

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 6, 7

[22] Ke Li, Gong Cheng, Shuhui Bu, and Xiong You. Rotation-insensitive and context-augmented object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2337–2348, 2017. 2, 3, 5, 6

[23] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159:296–307, 2020. 2, 5

[24] Xiaoyan Li, Meina Kan, Shiguang Shan, and Xilin Chen. Weakly supervised object detection with segmentation collaboration. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9735–9744, 2019. 1, 2

[25] Chenhao Lin, Siwen Wang, Dongqi Xu, Yu Lu, and Wayne Zhang. Object instance mining for weakly supervised object detection. In *AAAI*, pages 11482–11489, 2020. 1, 2

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015. 3, 6, 7

[27] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Yong Jae Lee, Alexander G Schwing, and Jan

Kautz. Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10598–10607, 2020. 1, 2

[28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 6

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[30] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Yuille. Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–16, 2019. 1, 2, 5, 6, 7

[31] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition*, pages 3059–3067, July 2017. 1, 2, 3, 4, 6, 7

[32] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013. 4, 6

[33] Fang Wan, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, and Qixiang Ye. C-mil: Continuation multiple instance learning for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2199–2208, 2019. 1, 2

[34] Fang Wan, Pengxu Wei, Jianbin Jiao, Zhenjun Han, and Qixiang Ye. Min-entropy latent model for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1297–1306, June 2018. 1, 2, 6

[35] Binglu Wang, Xun Zhang, and Yongqiang Zhao. Exploring sub-action granularity for weakly supervised temporal action localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021. 1

[36] Binglu Wang, Yongqiang Zhao, and Xuelong Li. Multiple instance graph learning for weakly supervised remote sensing object detection. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–1, 2021. 1, 3

[37] Binglu Wang, Yongqiang Zhao, and Xuelong Li. Multiple instance graph learning for weakly supervised remote sensing object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12, 2021. 1

[38] Yunchao Wei, Zhiqiang Shen, Bowen Cheng, Honghui Shi, Jinjun Xiong, Jiashi Feng, and Thomas Huang. Ts2c: Tight box mining with surrounding segmentation context for weakly supervised object detection. In *Proceedings of the European Conference on Computer Vision*, pages 434–450, 2018. 1

[39] Le Yang, Junwei Han, Tao Zhao, Tianwei Lin, Dingwen Zhang, and Jianxin Chen. Background-click supervision for temporal action localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1

[40] Le Yang, Houwen Peng, Dingwen Zhang, Jianlong Fu, and Junwei Han. Revisiting anchor mechanisms for temporal action localization. *IEEE Transactions on Image Processing*, 29:8535–8548, 2020. 2

[41] Qin Yang, Chenglin Li, Wenrui Dai, Junni Zou, Guo-Jun Qi, and Hongkai Xiong. Rotation equivariant graph convolutional network for spherical image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4303–4312, 2020. 3

[42] Xiwen Yao, Xiaoxu Feng, Junwei Han, Gong Cheng, and Lei Guo. Automatic weakly supervised object detection from high spatial resolution remote sensing images via dynamic curriculum learning. *IEEE Transactions on Geoscience and Remote Sensing*, 59(1):675–685, 2020. 1, 3, 6, 7

[43] Zhaoyang Zeng, Bei Liu, Jianlong Fu, Hongyang Chao, and Lei Zhang. Wsod2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8292–8300, 2019. 1, 2

[44] Shiwei Zhang, Wei Ke, Lin Yang, Qixiang Ye, Xiaopeng Hong, Yihong Gong, and Tong Zhang. Discovery-and-selection: Towards optimal multiple instance learning for weakly supervised object detection. *arXiv preprint arXiv:2110.09060*, 2021. 1

[45] Xiaopeng Zhang, Jiashi Feng, Hongkai Xiong, and Qi Tian. Zigzag learning for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4262–4270, June 2018. 1, 2

[46] Yongqiang Zhang, Yancheng Bai, Mingli Ding, Yongqiang Li, and Bernard Ghanem. W2f: A weakly-supervised to fully-supervised framework for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 928–936, June 2018. 1

[47] Yuanyi Zhong, Jianfeng Wang, Jian Peng, and Lei Zhang. Boosting weakly supervised object detection with progressive knowledge transfer. *arXiv preprint arXiv:2007.07986*, 2020. 1