symmetry

MDPI

Article

Weakly Supervised Object Detection with Symmetry Context

Xinyu Gu, Qian Zhang and Zheng Lu *

School of Computer Science, University of Nottingham Ningbo China, Ningbo 315100, China
* Correspondence: zheng.lu@nottingham.edu.cn

Abstract: Recently, weakly supervised object detection (WSOD) with image-level annotation has attracted great attention in the field of computer vision. The problem is often formulated as multiple instance learning in the existing studies, which are often trapped by discriminative object parts and fail to localize the object boundary precisely. In this work, we alleviate this problem by exploiting contextual information that may potentially increase object localization accuracy. Specifically, we propose novel context proposal mining strategies and a Symmetry Context Module to leverage surrounding contextual information of precomputed region proposals. Both naive and Gaussian-based context proposal mining methods are adopted to yield informative context proposals symmetrically surrounding region proposals. Then mined context proposals are fed into our Symmetry Context Module to encourage the model to select proposals that contain the whole object, rather than the most discriminative object parts. Experimental results show that the mean Average Precision (mAP) of the proposed method achieves 52.4% on the PASCAL VOC 2007 dataset, outperforming the state-of-the-art methods and demonstrating its effectiveness for weakly supervised object detection.

Keywords: weakly supervised object detection; multiple instance learning; context proposal mining

Citation: Gu, X.; Zhang, Q.; Lu, Z. Weakly Supervised Object Detection with Symmetry Context. *Symmetry* **2022**, *14*, 1832. https://doi.org/10.3390/sym14091832

Academic Editor: Mihai Postolache

Received: 31 May 2022
Accepted: 29 July 2022
Published: 4 September 2022

1. Introduction

Weakly supervised object detection (WSOD) aims to detect multiple object instances in bounding boxes from a given image using only image-level supervision. Compared with fully supervised object detection (FSOD), WSOD significantly reduces human labor cost for instance-level annotation. Despite remarkable progress, WSOD remains as a challenging task in computer vision due to problems such as discriminative region and memory consumption. The performance of existing WSOD approaches [1,2] are still inferior to the fully supervised counterpart [3,4].

Typical WSOD approaches [5,6] formulate the problem as a multiple instance learning (MIL) problem (i.e., there is no negative instance in the negative bag and there is at least one positive instance in the positive bag). Although most existing WSOD methods are based on MIL [7,8], this strategy is known to suffer from the discriminative region problem that tends to detect the most discriminative object parts rather than the whole object. In FSOD approaches, contextual information around the object is widely exploited to facilitate object recognition [9,10]. With such contextual cues, one can decrease the ambiguity when recognizing the object of interest and enrich object local feature representations. This can lead to more accurate object localization. For instance, a table is more likely to be surrounded by chairs instead of bicycles. In WSOD, there are only a few studies [11,12] utilizing contextual information to boost the localization accuracy. Tight box mining with Surrounding Segmentation Context (TS^2C) [11] uses segmentation context to suppress low-quality candidates. ContextLocNet [12] proposes two context-aware guidance models to enforce contextual supervisory, i.e., the external rectangle (context) and the internal rectangle (frame) with a fixed ratio. Instead of simply using a rectangular shape to obtain the surrounding contextual proposals, we seek more informative proposals in the spatial symmetrical areas to diversify context proposal locations alleviating the discriminative region problem when recognizing object instances under difficult circumstances.

In this paper, following [7,13], we also formulate the WSOD problem as MIL and propose a Symmetry Context Module (SCM) and two strategies for context proposal mining capturing context appearance in different spatial areas surrounding an object. Specifically, SCM encourages the detected region proposal to be compatible with its context regions. This is achieved by encoding the region proposal score matrix and the maximum score matrices of its context proposals. Different from other context-aware guidance models such as [12], which adds or subtracts the score matrix of regions with its context or frame, SCM aggregates more comprehensive parts of surrounding context regions than individual spatial sides. Such a design yields an enriched object representation that is better able to capture discriminative information. In order to efficiently select context candidates, we propose two mining strategies for context proposals: 1) naive context proposal mining that uses fixed regions adjacent to the object instance; 2) Gaussian-based context proposal mining that samples context proposals with normal distribution. Our method incorporates rich contextual information in the model that boosts object detection performance.

Due to our contextual mining strategies, our proposed method improves the localization performance of WSOD. Our method performs favorably against some of the state-of-the-art WSOD methods by achieving 52.4% mAP on the PASCAL VOC 2007 dataset. The effectiveness and robustness of the proposed techniques are demonstrated in detailed ablation studies and further qualitative results. The contributions of this work are summarized as follows:

- Two context proposal mining strategies are proposed to better capture the diverse discriminative information for objects of interest.
- A Symmetry Context Module (SCM) is introduced to improve the detection accuracy of our two-stream neural network model.
- Experimental results on the popular PASCAL VOC 2007 and 2012 datasets demonstrate that our method achieves better performance compared with other state-of-the-art approaches.

2. Related Work

2.1. MIL and WSOD

MIL is a classical weakly supervised learning method, which arranges the training data as bags where each bag contains a collection of instances. Supervision is only provided for the entire bag. In other words, the individual label of each instance in the bag is not known. Standard MIL assumes the following: (1) for a positive bag, there is at least one positive instance in the bag; (2) for a negative bag, all its instances are negative. In recent years, much work has adopted MIL approaches successfully in computer vision and other areas [14,15], as MIL inherently targets weakly labeled data.

Different from FSOD, which uses instance-level annotation for training, WSOD only requires image-level annotation, for which MIL fits naturally. In particular, when an image I is annotated with class C, there is at least one positive instance of this object class in this "bag". In addition, this "bag" is negative for other object classes (there is no instances of those classes in this image). Recently, deep neural networks and MIL were used together to improve WSOD performance, such as in [16,17]. Bilen et al. [6] proposed the weakly supervised deep detection network (WSDDN), the first end-to-end MIL based deep neural network in WSOD consisting of predefined proposals, a backbone, and a detection head. Predefined proposals are usually generated by proposal generation techniques such as selective search [18] and edge boxes [19], aiming to cover all possible object locations. A feature representation network pre-trained on large scale datasets (e.g., ImageNet [20]), such as VGG16 [21] or ResNet [22], is used to obtain the feature maps. The predefined proposals and feature maps are then fed into a spatial pyramid pooling (SPP) layer to generate fixed-length feature vectors for each proposal. These feature vectors are then forwarded into the detection head, which contains both detection and classification streams, to locate and classify object instances.

One of the drawbacks in WSDDN is that it tends to focus on the most discriminative parts of the object. This is because the most discriminative regions of an object are more likely to have the highest score compared with other regions that also cover the object of interest. Several techniques are proposed to alleviate this problem. Online Instance Classifier Refinement (OICR) [7] and Proposal Cluster Learning (PCL) [23] try to learn more refined instance classifiers iteratively. Taking a multi-task strategy, Diba et al. [24] jointly train a weakly supervised segmentation network and WSOD to filter object proposals with the aid of a weakly supervised segmentation map. Incorporating low-level features is another way to alleviate the discriminative region problem. For example, WSOD framework with Objectness Distillation (WSOD2) [25] integrates bottom-up and top-down objectness to distill box boundary knowledge. Object symmetry is also exploited in WSOD. For example, posterior regularization is used to enforcethe object and its horizontal mirror version having similar values [26]. Transfer learning is also adopted to leverage an auxiliary dataset to improve WSOD networks [27,28]. Similarly, using contextual information is another way to solve the problem. In this work, we leverage the contextual information in a symmetric spatial area in localization stream in addition to the two stream CNN used in WSDDN. We focus on a context-aware convolutional neural network (CNN) architecture and proposal mining techniques to exploit the discriminative information surrounding the predefined proposals.

2.2. Using Contextual Information in WSOD

Contextual information has been widely employed in object detection [29,30]. Recently, many studies have started to use contextual information in weakly supervised or unsupervised localizations. Hierarchical Context Embedding(HCE) [31] obtains hierarchical contextual region of interest (RoI) features by fusing instance-level and global-level information. Ren et al. [8] use spatial dropout to alleviate part domination and encourage the model to focus more on context. Wei et al. [11] utilize surrounding segmentation context as references to mine high quality object candidates. ContextLocNet [12] proposes additive and contrastive models to help RoI selection. Our context-aware CNN models are inspired by these approaches. Different from ContextLocNet [12], which uses inner and outer frame-shape region as context, we propose symmetry Gaussian-based sampling to mine informative context proposals and devise a Symmetry Context Module to further improve the detection performance.

3. Methodology

3.1. Overall Framework

The overview of our method is illustrated in Figure 1. Similar to many existing studies [6,7,23], we adopt the idea of MIL by a two-stream weakly supervised object detection network. Specifically, given an input image, we first generate object proposals using selective search [18] and context proposals using our proposed context proposal mining strategy; we then extract region features using convolutional layers, a RoI pooling layer [32], and two fully connected layers. Next, the extracted features are branched into our Symmetry Context Module, a context-aware two-stream module. This process encodes the contextual information into the localization stream and obtains instance classifier using weighted pooling strategy. Results of the base instance classifier is further refined by an inter-stream self-training algorithm. The following depicts proposal generation, convolutional and ROI pooling layers, and the detection head used in our framework.

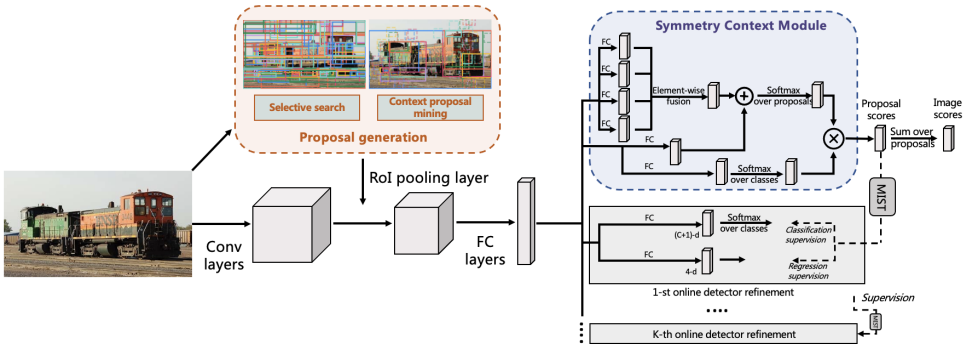


Figure 1. Our framework. The input image is passed through several convolutional layers (using modified VGG-16 [21] architecture pre-trained on ImageNet [20]). The output feature maps, generated region proposals and their corresponding context proposals are passed to RoI Pooling layer to extract region-level feature vectors. The output of fully-connected layers is then branched into several streams: one for our proposed Symmetry Context Module and others for multi-stage online detector refinement.

Proposal generation: Given an input image I and its image label $Y = [y_1, y_2, ..., y_C]^T \in \mathbb{R}^{C\times 1}$, where $y_c = 1$ or 0 indicates the image with or without object c; the pre-computed region proposals R are generated using selective search [18], where C denotes the number of object classes. We obtain context proposals by exploiting the symmetrical property of an object's surrounding environment. We propose two methods to generate the context proposals in Section 3.2, where the first strategy takes adjacent rectangular bounding boxes of a region proposal in four spatial directions as its context proposals, and the second strategy samples the context proposals using Gaussian distribution. We obtain the context proposals P for all the pre-computed proposals R.

Convolutional and ROI pooling layers: Our framework contains 16 convolutional layers from the VGG-16 model [21]. The original penultimate max-pooling layer and the following three convolutional layers are replaced by the dilated convolutional layers [33] to increase feature map size. In addition, we replace the last max-pooling layer by the ROI pooling layer to extract region-level descriptors. The network first takes the entire image I as input and applies a sequence of convolutional layers to obtain feature maps. The precomputed proposals R, context proposals P, and features maps are then forwarded to the ROI pooling layer to obtain the fix-sized feature vectors of proposals and context proposals. Region-level features are further passed to two fully connected layers. We initialize the network layers using the weights of ImageNet [20] pretrained VGG-16 model [21], which is then fine-tuned during our training.

Detection head: We employ the two-stream architecture from Bilen and Vedaldi [6] and propose a context-aware two-stream module. The feature vectors from context proposals and region proposals are branched into our Symmetry Context module to produce score matrices $\mathbf{x}^R$. As explained in Section 3.3, we introduce different fully connected layers to process features from context proposals and region proposals in the localization stream and obtain $\mathbf{x}^d$ by summing up the outputs. The classification stream feeds the features of region proposals to a linear layer and outputs $\mathbf{x}^c$. The details of our Symmetry Context Module are given in Section 3.3. The two matrices are then passed through two softmax layers, each with different directions, as in WSDDN. The value of $\mathbf{x}^c$ is normalized by a softmax layer along the object class direction, and $\mathbf{x}^d$ is passed through the other softmax layer along the proposal direction. The proposal scores $\mathbf{x}^R$ are calculated through element-wise product of the two score matrices. The image score of a specific class ϕ_c is calculated by the sum over all proposals. The basic instance classifier is trained using cross entropy loss is as follows:

$$\mathcal{L}_b = -\sum_{c=1}^{C} \{y_c log\phi_c + (1 - y_c) log(1 - \phi_c)\}. \quad (1)$$

Because the basic two-stream architecture design from WSDDN tends to converge to the most discriminative part of an object, we adopt the online detector refinement (ODR) approach for result refinement. The idea of ODR is a inter-stream self-training training process in which the instances of the latter stream is supervised by the previous ones. Let K be the number of refinement stages, each stage contains a single fully connected layer followed by a softmax over category direction for classification. To improve the localization results as in [7,25], we also use a fully connected layer for localization refinement. The output score vector $\mathbf{x}_j^{Rk}$ of proposal j of the kth instance classifier is $\mathbf{x}_j^{Rk} \in \mathbb{R}^{(C+1)\times 1}$. Note that each instance classifier has $C+1$ categories (the $\{C+1\}$th dimension is for background). To achieve inter-stream self-training progressively, the label for proposal j is $\mathbf{Y}_j^k = [y_{1j}^k, y_{2j}^k, ..., y_{(C+1)j}^k]^T \in \mathbb{R}^{(C+1)\times 1}$. The supervision of refinement stage k comes from the last instance classifier output $\mathbf{x}^{R(k-1)}$. An advanced pseudo groundtruth selection algorithm MIST [8] is used in our experiment to generate diverse pseudo boxes for the training. Pseudo labels are denoted by $\hat{R}$. At each stage, if a region j is highly overlapped with pseudo-box $\hat{r} \in \hat{R}$ for ground-truth class c, we set the classification label y_{cj}^k to 1 and the regression target $\hat{t}_j^k$ by the coordinates of $\hat{r}$. Each refinement stage is trained to minimize the following loss:

$$\mathcal{L}_r^k = \frac{1}{|R|}\sum_{j=1}^{|R|}\lambda_j(-\sum_{c=1}^{C+1} y_{cj}^k log x_{cj}^{Rk} + \mathcal{L}_{smooth-L1}(\hat{t}_j^k, t_j^k)), \tag{2}$$

where λ_j is a scalar per-region weight used in [7]; $\hat{t}_j^k$ and t_j^k are regression targets and predicted coordinates of the jth box; respectively; and $\mathcal{L}_{smooth-L1}$ is a smooth-L1 regression loss function. The overall loss to train our framework is:

$$\mathcal{L} = \mathcal{L}_b + \sum_{k=1}^{K}\mathcal{L}_r^k. \tag{3}$$

3.2. Context Proposal Mining

During proposal generation, a set of region proposals with various sizes and aspect ratios are generated. We observe that many proposals containing discriminative parts of the object often only have high recognition scores, hence decreasing object detection performance. Incorporating the contextual information can help by encouraging the model to select region proposals that contains the whole object instead of discriminative parts only. In this work, we mine additional proposals surrounding object region proposals, namely context proposals, together with region proposals to detect and classify objects. Specifically, we exploit two context proposal mining strategies: naive context proposal mining and Gaussian-based context proposal mining.

3.2.1. Naive Context Proposal Mining

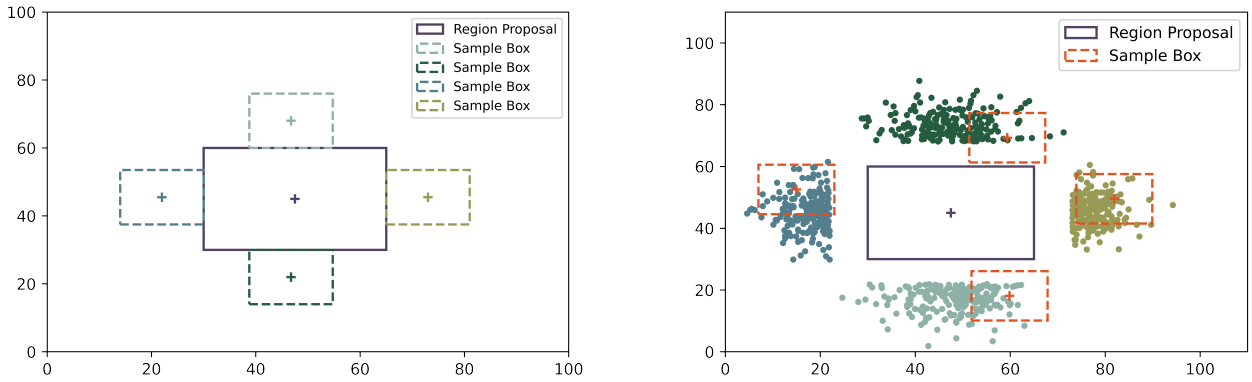
For a given region proposal j, we generate K context proposals. We fix the size of context proposal as s_c for simplicity. As shown in Figure 2a, we take a symmetric approach by mining context proposals from four directions (up, down, left, right). Our naive context proposal mining simply generates context proposals that are adjacent to j, hence K is set to 4. We represent the region proposal by $[x_j, y_j, w, h]$, where x_j and y_j are the coordinates of the region center, and w and h are the width and height of the region. The four context proposals of the region proposal j are generated as $[(x_j - (w+s_c)/2, y_j, s_c, s_c), (x_j + (w+s_c)/2, y_j, s_c, s_c), (x_j, y_j - (h+s_c)/2, s_c, s_c), (x_j, y_j + (h+s_c)/2, s_c, s_c)]$.

3.2.2. Gaussian-Based Context Proposal Mining

Although naive context proposal mining is very simple to implement, relatively fixed context location may not always provide the best information surrounding the object. We seek to further mine more informative context proposals by taking advantage of the symmetrical property of Gaussian distribution. In particular, we sample context proposals

from Gaussian distribution to diversify possible proposal locations. This idea is to mine context proposals that are more informative during training, as those proposals semantically align with object parts. Figure 2b shows example context proposals sampled from Gaussian distribution. Note that we ensure there is no overlapping between the region proposal and context proposals. In this way, contextual information surrounding the object can be incorporated into the model for better performance.

The mean of the context proposal center coordinates for each direction is set slightly away from the region proposal. For example, in our setup, for top and bottom, the mean of the x-coordinate is the same as the x-coordinate of the center of the region proposal. The mean of the y-coordinate is $0.2 \times h$ away from the top or bottom edge, respectively. The standard deviation is set to h. Utilizing the symmetric property of surrounding, we do the same for the left and right directions. Figure 3 illustrates the examples of Gaussian-based context proposals. We can observe that the context proposals sampled by our method capture object surroundings better in most cases.



(**a**) Naive context proposal mining (**b**) Gaussian-based context proposal mining

Figure 2. Illustration of context proposals mining strategies. (**a**) Naive context proposal mining: context proposals adjacent to the region proposals are generated from four different directions. (**b**) Gaussian-based context proposal mining: context proposals are sampled using Gaussian distribution.

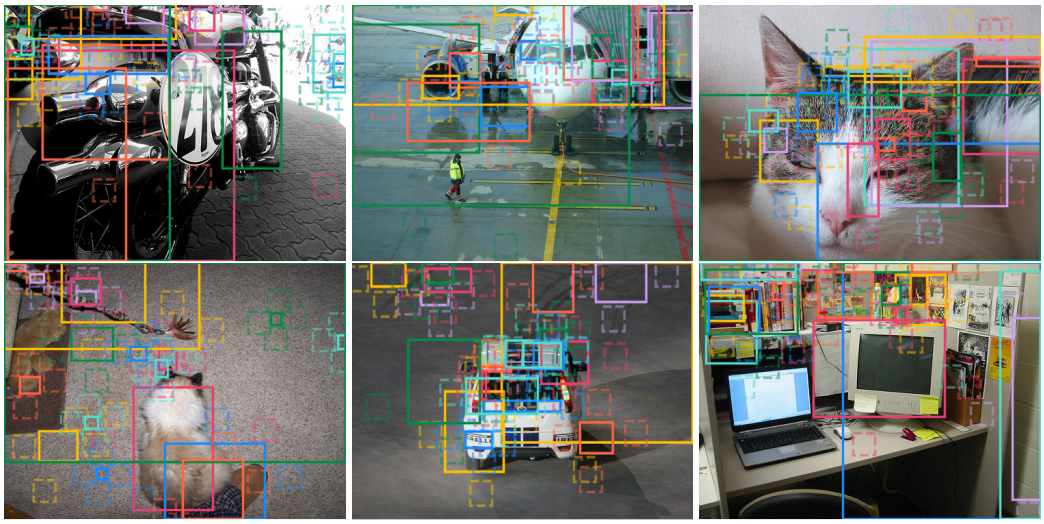


Figure 3. Examples of proposals from Pascal VOC 2007 training set. Proposals with a solid line are region proposals. Proposals with a dashed line are context proposals that share the same color with their corresponding region proposal.

3.3. Symmetry Context Module

Our Symmetry Context Module (SCM) is a context-aware two-stream detection module incorporating contextual information during the detection process. Compared with the detection module in WSDDN [6], our module consists of five different branches, FC_{det} and FC_1-FC_4 for region proposals and its corresponding context proposals in four directions, respectively. We use Gaussian-based context proposal mining to generate context proposals in our experiment. The standard deviation of the context proposals' distance to the region proposal boundary is set to 20% of the respective region proposal size. It should be noted that there is no overlapping area between region proposals and its corresponding context proposals in all the experiment settings.

As shown in Figure 4, the feature vectors from context proposals $F_{context}$ and region proposals F_{RP} are branched into SCM. The four branches for context proposals produce four matrics $\mathbf{x}^{ctx1}, \mathbf{x}^{ctx2}, \mathbf{x}^{ctx3}, \mathbf{x}^{ctx4} \in \mathbb{R}^{C\times|R|}$, each of which contains context proposal scores. We then fuse these context scores into a single matrix $\mathbf{x}^{ctx}$ using the following:

$$\mathbf{x}^{ctx} = max(\mathbf{x}^{ctx1}, \mathbf{x}^{ctx2}, \mathbf{x}^{ctx3}, \mathbf{x}^{ctx4}), \tag{4}$$

where max(·) denotes element-wise max operator. To better illustrate, we show one numerical example of fusion by element-wise max. Assuming $C = 2$ and $|R| = 2$, the score matrices $\mathbf{x}^{ctx1}, \mathbf{x}^{ctx2}, \mathbf{x}^{ctx3}, \mathbf{x}^{ctx4} \in \mathbb{R}^{2\times2}$ are as follows:

$$\begin{aligned} \mathbf{x}^{ctx1} &= \begin{pmatrix} -0.4538 & -2.0972 \\ -1.4895 & 4.5304 \end{pmatrix}, \quad \mathbf{x}^{ctx2} = \begin{pmatrix} -0.6994 & -0.5498 \\ 1.7492 & -0.5544 \end{pmatrix}, \\ \mathbf{x}^{ctx3} &= \begin{pmatrix} -1.8248 & -0.8432 \\ 0.9008 & -0.8110 \end{pmatrix}, \quad \mathbf{x}^{ctx4} = \begin{pmatrix} -4.0172 & -1.8290 \\ 1.6335 & 3.9155 \end{pmatrix}. \end{aligned} \tag{5}$$

The final $\mathbf{x}^{ctx}$ can be obtained from fusion by element-wise max as follows:

$$\mathbf{x}^{ctx} = max(\mathbf{x}^{ctx1}, \mathbf{x}^{ctx2}, \mathbf{x}^{ctx3}, \mathbf{x}^{ctx4}) = \begin{pmatrix} -0.4538 & -0.5498 \\ 1.7492 & 4.5304 \end{pmatrix}. \tag{6}$$

We then obtain the final detection matrix $\mathbf{x}^d$ by element-wise summation over $\mathbf{x}^{ctx}$ and $\mathbf{x}^r$ (scores from region proposal). The classification stream takes the feature vector of region proposals and passes to a fully connected layer obtain classification score matrix $\mathbf{x}^c$. Both $\mathbf{x}^d$ and $\mathbf{x}^c$ are passed through the two softmax layers, each with different directions, as in WSDDN. Proposal scores $\mathbf{x}^R$ are calculated through the element-wise product of the two score matrices.

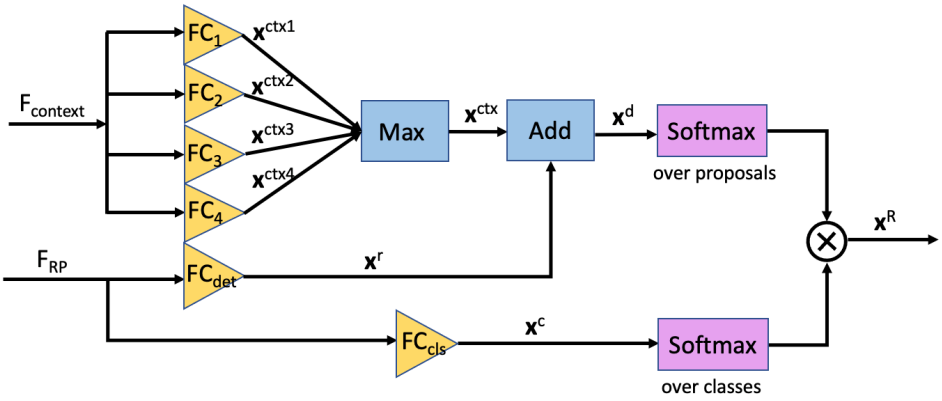


Figure 4. Symmetry Context Module processes the feature maps of context proposals and region proposals in two streams to produce scores for both detection and classification.

4. Experiments

4.1. Datasets and Experimental Setup

We conduct our experiments on the Pascal VOC 2007 [34] and 2012 [35] datasets, the most widely used benchmarks for WSOD. The VOC 2007 dataset has 20 object classes and contains 2501 training images, 2510 validation images, and 4951 testing images. The VOC 2012 dataset contains 5717 training images, 5823 validation images, and 10991 testing images for the same 20 object categories. Following common practice [7,11,23], we use training and validation images with image-level annotation to train our model and testing images for the evaluation. We adopt mean average precision (IoU threshold at 0.5) to evaluate our method, which follows the standard PASCAL VOC protocol [36].

During training, we set batch size as 8 and maximum iteration to 30,000. We use stochastic gradient descent (SGD) for optimization, with an initial learning rate of 0.01 and a weight decay of 0.0001. For each image, 2000 region proposals are generated. For fair comparison, we deploy the similar multi-scale training and testing strategy as in OICR [7] (our baseline model). In particular, we use multiple scales $(480, 576, 688, 864, 1000, 1200)$ with respect to the original aspect ratio to resize the shorter side of each image and capped the longer side to 2000 during training. During evaluation, the shorter side of input images are augmented with scale 800. Horizontal flipping is also used for both training and evaluation. The default p and IoU in MIST are set to 0.15 and 0.2. The total number of refinement branches is set to 3. The mean output of these instance classifiers is used during evaluation. All our experiments are run on a single Tesla V100-PCIE-32GB GPU. Please check the released code for more details (https://github.com/sXZL/WSODSC (accessed on 24 June 2022)).

4.2. Ablation Study

We first conduct an ablation study on the VOC 2007 dataset to demonstrate the effectiveness of different components in our framework, including context proposal mining and the Symmetry Context Module. Experimental results are reported in Tables 1–3.

Table 1. Ablation study of context proposal mining.

Context Proposal Mining	Distance to Region Proposal Boundary	mAP
No context	-	42.26
NCP	0	45.22
NCP	0.9	44.16
GCP	0.1	43.99
GCP	0.2	45.10

Table 2. Ablation study of the number of context proposals.

Context Proposal Mining	Distance to Region Proposal Boundary	Number of Context Proposals per Side	mAP
GCP	0.1	2	43.46
GCP	0.1	1	43.99

Table 3. Ablation study of different ways to fuse proposal scores.

Method	Context Proposal Mining	Distance to Region Proposal Boundary	Fusion Method	mAP
OICR (+MIST + Reg.)	No context	-	-	50.91
Ours	GCP	0.2	mean	51.85
Ours	GCP	0.2	max	52.38

4.2.1. Context Proposals Location

In order to investigate the influence of context proposal location, we evaluate the performance of our framework using different configurations for both strategies. Table 1 shows the experimental results with different sizes and locations. In the table, we denote naive context proposal mining method as NCP and Gaussian-based context proposal mining as GCP. Note that, for naive context proposals, we test context proposals at various distances to the region proposal boundary. For example, 0.9 means the distance is 90% of the distance between the region proposal boundary and the respective image boundary. For Gaussian-based context proposals, we test context proposals sampled using different standard deviations. For example, 0.1 and 0.2 means the standard deviations used are 10% and 20% of the respective region proposal size. We also carry out experiments with a fixed distance to the region proposal boundary, obtaining 2.84% mAP improvement when the distance is 32 pixels.

As shown in Table 1, we can observe that even the proposed naive mining strategies can boost the performance a lot (achieving 1.9–2.96% mAP improvement compared with one of our baseline OICR [7] with 42.3% mAP). This shows the effectiveness and robustness of our context proposal mining. We obtain the best performance with 45.22% mAP when context proposal is just adjacent to the region proposal for NCP. As for the best configuration for GCP, the best configuration is 20% of the respective region proposal size as standard deviation. Note that for this study we set context proposal size as 32.

4.2.2. Effect of Number of Context Proposals

We also study the influence of the number of context proposals for each direction. We choose GCP with 10% as the standard deviation. Table 2 shows that both settings outperform the baseline. Furthermore, single context proposal for each side outperforms two context proposals. This may result from the unweighted fusion of region proposal and context proposal scores. Thus, the contribution of context proposals may overwhelm the corresponding region proposal when the total number of context proposals for each region proposal is 8, as in this case.

In our Symmetry Context Module, four context proposals from one region proposal use four different fully connected layers to generate proposal scores. We opt to conduct experiments to analyze the effect of shared fully connected layers instead of unshared ones. Using individual fully connected layers achieves 45.22% mAP with zero distance in NCP and outperforms the shared version by 0.54% mAP. We also conduct experiments to evaluate different ways to fuse proposal scores. We also add MIST [8] and regression branches (denoted as w/ MIST Reg) for online instance classifier in OICR [7]. Table 3 shows that fusion by element-wise *max* outperforms element-wise average.

4.3. Comparison with Other Baselines

To evaluate our framework, we compare our proposed method with several state-of-the-art methods on VOC 2007 in addition to OICR, as shown in Table 4. It can be observed that the proposed method is very effective and outperforms all the other baselines in most of the categories, leading to a notable improvement on average. In particular, our method performs much better than other baselines on category "Aero", "Bike", "Car", "Train", and "TV". Note that our method outperforms OICR + MIST + Reg. [8] (second best) by 1.5%. We also conduct the experiments on the more challenging dataset VOC 2012, as shown in Table 5. These results show that our proposed method achieves noticeable improvements with other approaches, demonstrating its effectiveness and robustness.

Figures 5 and 6 shows some of our detection results compared with the best baseline OICR + MIST + Reg. [8]. We can see that our method tends to locate the whole object rather than the most discriminative parts. As shown in Figure 6, for animal and person classes, our method is able to effectively locate the whole object thanks to the help of contextual information, whereas the compared baseline results tend to focus on objects' heads. The baseline method tend to ignore small parts of objects of interest such as arms and

hands, leading to false positive results, as shown in Figure 5. We show that incorporating contextual information as proposed in our method can alleviate such problems.

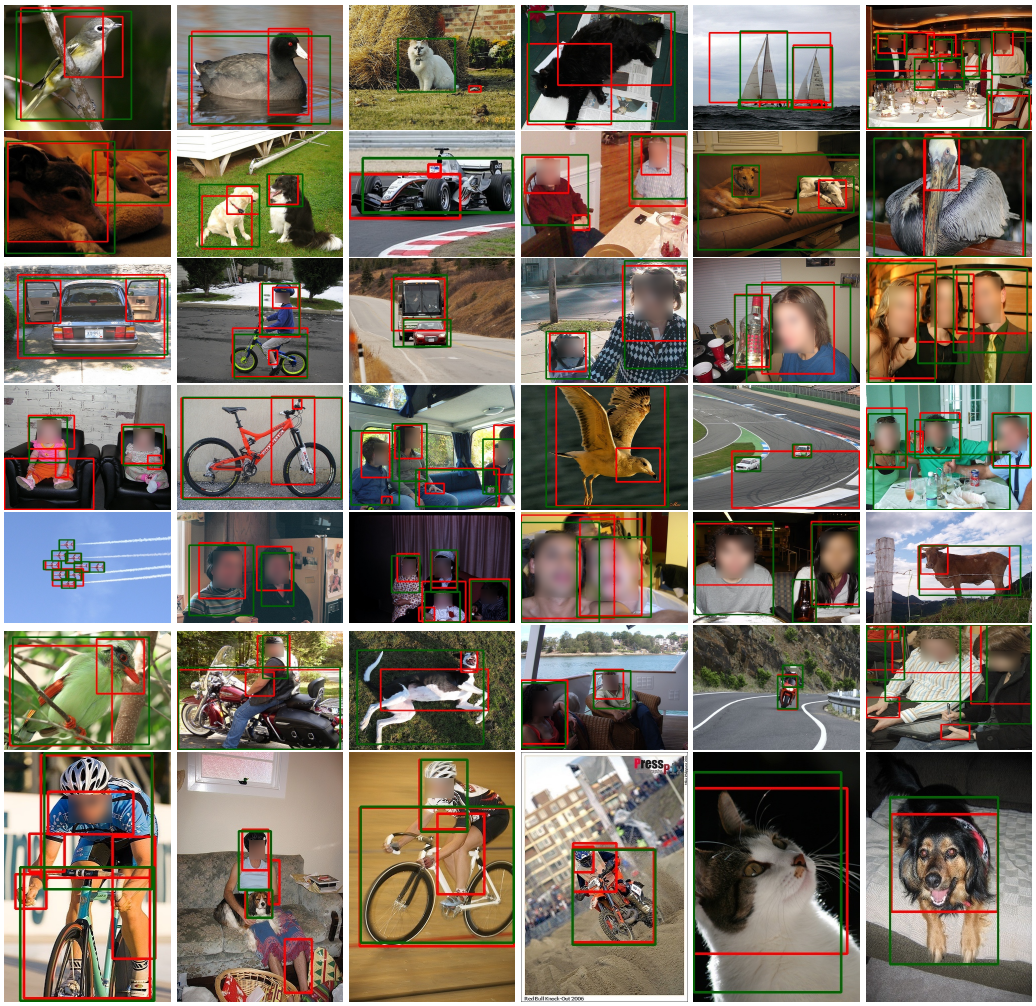


Figure 5. Detection results of our method and the best baseline (OICR + Reg + MIST). Green bounding boxes indicate objects detected by our method, whereas red ones correspond to those detected by the best baseline.

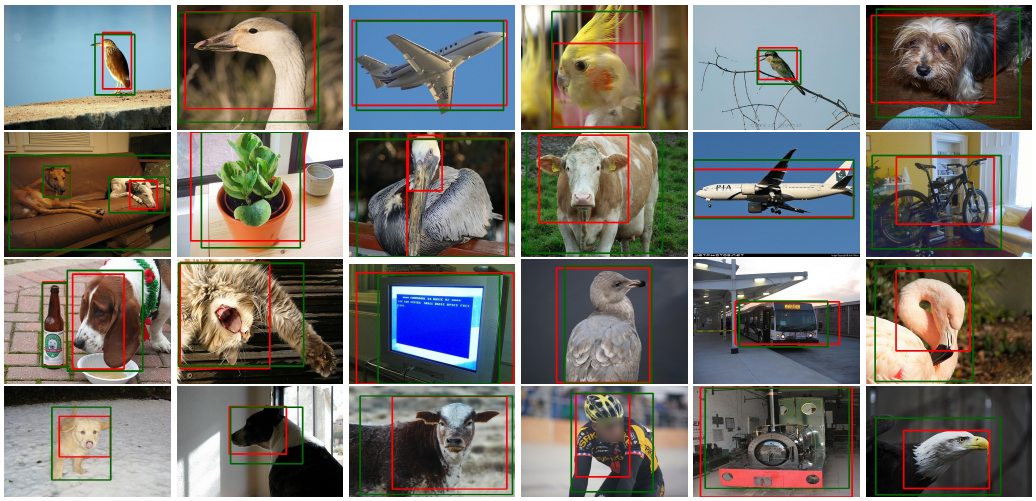


Figure 6. Additional detection results of our method and the best baseline (OICR + Reg + MIST). Green bounding boxes indicate objects detected by our method, whereas red ones correspond to those detected by the best baseline.

Table 4. Comparison of our method on PASCAL VOC 2007 with different baselines (* our implementation).

Method	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Mbike	Persn	Plant	Sheep	Sofa	Train	TV	mAP
ContextLocNet [12]	57.1	52.0	31.5	7.6	11.5	55.0	53.1	34.1	1.7	33.1	49.2	42.0	47.3	56.6	15.3	12.8	24.8	48.9	44.4	47.8	36.3
Bilen [6]	46.4	58.3	35.5	25.9	14.0	66.7	53.0	39.2	8.9	41.8	26.6	38.6	44.7	59.0	10.8	17.3	40.7	49.6	56.9	50.8	39.3
OICR [7]	58.0	62.4	31.1	19.4	13.0	65.1	62.2	28.4	24.8	44.7	30.6	25.3	37.8	65.5	15.7	24.1	41.7	46.9	64.3	62.6	41.2
OICR * [7]	56.1	72.7	40.9	26.7	25.7	66.6	67.1	13.0	24.2	48.4	39.5	16.4	20.3	69.4	8.1	23.9	49.2	47.5	63.9	65.8	42.3
Diba [24]	49.5	60.6	38.6	29.2	16.2	70.8	56.9	42.5	10.9	44.1	29.9	42.2	47.9	64.1	13.8	23.5	45.9	54.1	60.8	54.5	42.8
SGWSOD [37]	48.4	61.5	33.3	30.0	15.3	72.4	62.4	59.1	10.9	42.3	34.3	53.1	48.4	65.0	20.5	16.6	40.6	46.5	54.6	55.1	43.5
TS2C [11]	59.3	57.5	43.7	27.3	13.5	63.9	61.7	59.9	24.1	46.9	36.7	45.6	39.9	62.6	10.3	23.6	41.7	52.4	58.7	56.6	44.3
WSRPN [38]	57.9	70.5	37.8	5.7	21.0	66.1	69.2	59.4	3.4	57.1	57.3	35.2	64.2	68.6	32.8	28.6	50.8	49.5	41.1	30.0	45.3
PCL [23]	62.3	69.3	50.6	28.1	22.1	71.8	68.1	56.8	24.0	61.3	43.1	59.4	45.0	66.2	12.3	23.3	45.3	52.0	65.1	57.2	49.2
SDCN [39]	59.4	71.5	38.9	32.2	21.5	67.7	64.5	68.9	20.4	49.2	47.6	60.9	55.9	67.4	31.2	22.9	45.0	53.2	60.9	64.4	50.2
C-MIL [5]	62.5	58.4	49.5	32.1	19.8	70.5	66.1	63.4	20.0	60.5	52.9	53.5	57.4	68.9	8.4	24.6	51.8	58.7	66.7	63.5	50.5
Yang et al. [40]	57.6	70.8	50.7	28.3	27.2	72.5	69.1	65.0	26.9	64.5	47.4	47.7	53.5	66.9	13.7	29.3	56.0	54.9	63.4	65.2	51.5
OPG [41]	63.0	65.3	49.2	31.7	25.3	70.9	70.9	58.1	27.4	58.6	44.7	47.0	47.2	69.8	13.1	26.1	49.9	51.8	61.7	68.2	50.0
Jiang et al. [42]	60.1	74.5	51.9	29.6	30.2	68.8	72.6	44.6	19.8	66.0	48.8	43.7	63.2	68.2	17.7	25.1	53.7	60.8	56.1	63.1	50.9
OICR + MIST + Reg. [8]	67.9	78.6	55.6	25.6	29.1	69.8	75.4	50.3	27.6	67.2	39.6	28.2	50.2	72.0	15.7	26.1	62.7	52.2	68.0	56.7	50.9
Ours	71.4	79.2	55.5	31.6	22.6	71.5	75.5	52.3	20.4	64.8	44.9	35.2	49.8	71.8	22.3	27.9	59.6	52.3	70.6	68.3	52.4

Table 5. Comparison of our method on PASCAL VOC 2012 with different baselines.

Method	mAP
OICR [7]	37.9
PCL [23]	40.6
SDCN [39]	43.5
Yang et al. [40]	45.6
C-MIL [5]	46.7
OPG [41]	46.2
Jiang et al. [42]	43.8
OICR + MIST + Reg. [8]	47.7
Ours	48.3

5. Discussion and Conclusions

In order to further improve object detection accuracy, in this paper we proposed two context proposal mining approaches and a Symmetry Context Module to incorporate contextual information into the overall WSOD framework. Extensive experiments were conducted on the benchmark datasets Pascal VOC 2007 [36] and Pascal VOC 2012 [35]. We carry out ablation study on the context proposal location for both context mining strategies, achieving 2.9% and 1.95% mAP improvement with fixed distance to the region proposal boundary for NCP and GCP respectively. Experimental results show that the performance can improve further with distance depending on the repective region proposal, demonstrating the generalization ability of our proposed context proposal mining strategies. For fair comparison, we also conducted experiments on different baselines to eliminate the effect of other implementation tricks, achieving 0.9–3.0% overall mAP improvements. Fusing contextual information of symmetrical spatial areas with region proposal scores is effective to increase the object localization accuracy. Fusion with element-wise max of contextual score matrices performs slightly better than fusion by element-wise average, due to covering more discriminative contextual information. The Gaussian-based context proposal mining was more robust at capturing contextual information, further improving the localization accuracy. We also evaluated the effect of the number of context proposals and showed that increasing the number of context proposals may not help the overall performance. Due to the fact that our contextual mining strategies and SCM well exploit and utilize informative contextual information from the surrounding areas of objects, our method has distinct advantages for those categories whose objects have similar surroundings, such as ”Aero”, ”Bike”, ”TV”, etc. Our qualitative results comparing our method with the best performing baseline show more insights on the advantage of our method.

Our future work may consider adjusting SCM to mine more useful information among context proposals. Furthermore, introducing segmentation map as a guidance to mine context proposals online could be another possible improvement. This way could reduce the memory consumption due to a multiple-stream approach.

Author Contributions: Conceptualization, X.G. and Z.L.; methodology, X.G. and Z.L.; validation, X.G., Z.L. and Q.Z.; formal analysis, X.G. and Z.L.; investigation, X.G. and Z.L.; software, X.G.; writing—original draft preparation, X.G.; writing—review and editing, Z.L., X.G. and Q.Z.; data curation, X.G.; visualization, X.G.; supervision, Z.L. and Q.Z.; funding acquisition, Z.L. and Q.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by Ningbo Science and Technology Bureau under Service Industry S&T Programme with project code 2019F1028 and Major Projects Fund with project code 2021Z089.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The Pascal VOC 2007 and 2012 are publicly available from (http://host.robots.ox.ac.uk/pascal/VOC/) accessed on 14 June 2022.

Acknowledgments: We thank the anonymous reviewers for their valuable feedback.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations and Notation Description

Abbreviations and Notation used most frequently in this manuscript:

Abbreviations / Notation	Description
CNN	Convolutional Neural Networks
WSOD	Weakly Supervised Object Detection
FSOD	Fully Supervised Object Detection
MIL	Multiple Instance Learning
WSDDN	Weakly Supervised Deep Detection Network
OICR	Online Instance Classifier Refinement
mAP	Mean Average Precision
MIST	Multiple Instance Self-Training
I	input image
$\mathbf{Y} = [y_1, y_2, ..., y_C]^T \in \mathbb{R}^{C\times 1}$	image labels
$\mathbf{x}^c, \mathbf{x}^d \in \mathbb{R}^{C\times\|R\|}$	score matrix of localization stream and detection stream in SCM
$\mathbf{x}^{ctx} \in \mathbb{R}^{C\times\|R\|}$	fused context proposal score matrix of localization stream
C	number of object classes
K	the number of refinement stages
F_{RP}	feature vectors of region proposals
$F_{context}$	feature vectors of context proposals
ϕ_c	image score of a specific class c
$\mathbf{x}_j^{Rk} \in \mathbb{R}^{(C+1)\times 1}$	output score vector of proposal j of the kth instance classifier
$\mathbf{Y}_j^k \in \mathbb{R}^{(C+1)\times 1}$	label for proposal j of the kth instance classifier

References

1. Wang, H.; Li, H.; Qian, W.; Diao, W.; Zhao, L.; Zhang, J.; Zhang, D. Dynamic Pseudo-Label Generation for Weakly Supervised Object Detection in Remote Sensing Images. *Remote Sens.* **2021**, *13*, 1461. [CrossRef]
2. Huang, Z.; Zou, Y.; Kumar, B.; Huang, D. Comprehensive Attention Self-Distillation for Weakly-Supervised Object Detection. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 16797–16807.
3. Xu, C.; Zheng, X.; Lu, X. Multi-Level Alignment Network for Cross-Domain Ship Detection. *Remote Sens.* **2022**, *14*, 2389. [CrossRef]
4. Zheng, J.; Fu, H.; Li, W.; Wu, W.; Zhao, Y.; Dong, R.; Yu, L. Cross-Regional Oil Palm Tree Counting and Detection via a Multi-Level Attention Domain Adaptation Network. *ISPRS J. Photogramm. Remote Sens.* **2020**, *167*, 154–177. [CrossRef]
5. Wan, F.; Liu, C.; Ke, W.; Ji, X.; Jiao, J.; Ye, Q. C-MIL: Continuation Multiple Instance Learning for Weakly Supervised Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2194–2203. [CrossRef]
6. Bilen, H.; Vedaldi, A. Weakly Supervised Deep Detection Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2846–2854. [CrossRef]
7. Tang, P.; Wang, X.; Bai, X.; Liu, W. Multiple Instance Detection Network with Online Instance Classifier Refinement. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 22–25 July 2017; pp. 3059–3067. [CrossRef]
8. Ren, Z.; Yu, Z.; Yang, X.; Liu, M.Y.; Lee, Y.J.; Schwing, A.G.; Kautz, J. Instance-Aware, Context-Focused, and Memory-Efficient Weakly Supervised Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10595–10604. [CrossRef]
9. Torralba; Murphy; Freeman; Rubin. Context-Based Vision System for Place and Object Recognition. In Proceedings of the IEEE International Conference on Computer Vision, Nice, France, 14–17 October 2003; Volume 1, pp. 273–280. [CrossRef]
10. Gidaris, S.; Komodakis, N. Object Detection via a Multi-region and Semantic Segmentation-Aware CNN Model. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1134–1142. [CrossRef]
11. Wei, Y.; Shen, Z.; Cheng, B.; Shi, H.; Xiong, J.; Feng, J.; Huang, T. TS2C: Tight Box Mining with Surrounding Segmentation Context for Weakly Supervised Object Detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 434–450. [CrossRef]
12. Kantorov, V.; Oquab, M.; Cho, M.; Laptev, I. ContextLocNet: Context-Aware Deep Network Models for Weakly Supervised Localization. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 350–365. [CrossRef]

13. Zhang, D.; Han, J.; Cheng, G.; Yang, M. Weakly Supervised Object Localization and Detection: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *167*, 154–177. [CrossRef] [PubMed]
14. Huang, X.; Xu, K.; Huang, C.; Wang, C.; Qin, K. Multiple Instance Learning Convolutional Neural Networks for Fine-Grained Aircraft Recognition. *Remote Sens.* **2021**, *13*, 5132. [CrossRef]
15. Han, T.; Wang, L.; Wen, B. The Kernel Based Multiple Instances Learning Algorithm for Object Tracking. *Electronics* **2018**, *7*, 97. [CrossRef]
16. Wu, L.; Liu, Q. Weakly Supervised Object Co-Localization via Sharing Parts Based on a Joint Bayesian Model. *Symmetry* **2018**, *10*, 142. [CrossRef]
17. Ali, M.U., Sultani, W., Ali, M. Destruction from Sky: Weakly Supervised approach for Destruction Detection in Satellite Imagery. *ISPRS J. Photogramm. Remote. Sens.* **2020**, *162*, 115–124. [CrossRef]
18. Uijlings, J.R.; Van De Sande, K.E.; Gevers, T.; Smeulders, A.W. Selective Search for Object Recognition. *Int. J. Comput. Vis.* **2013**, *104*, 154–171. [CrossRef]
19. Zitnick, C.L.; Dollár, P. Edge Boxes: Locating Object Proposals from Edges. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 8–14 September 2014; pp. 391–405. [CrossRef]
20. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255. [CrossRef]
21. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
22. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
23. Tang, P.; Wang, X.; Bai, S.; Shen, W.; Bai, X.; Liu, W.; Yuille, A. PCL: Proposal Cluster Learning for Weakly Supervised Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 176–191. [CrossRef]
24. Diba, A.; Sharma, V.; Pazandeh, A.; Pirsiavash, H.; Van Gool, L. Weakly Supervised Cascaded Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5131–5139. [CrossRef]
25. Zeng, Z.; Liu, B.; Fu, J.; Chao, H.; Zhang, L. WSOD2: Learning Bottom-Up and Top-Down Objectness Distillation for Weakly-Supervised Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 8291–8299. [CrossRef]
26. Bilen, H.; Pedersoli, M.; Tuytelaars, T. Weakly Supervised Object Detection with Posterior Regularization. In Proceedings of the BMVC 2014, Nottingham, UK, 1–5 September 2014; pp. 1–12. [CrossRef]
27. Dong, B.; Huang, Z.; Guo, Y.; Wang, Q.; Niu, Z.; Zuo, W. Boosting Weakly Supervised Object Detection via Learning Bounding Box Adjusters. In Proceedings of the IEEE International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 2876–2885.
28. Inoue, N.; Furuta, R.; Yamasaki, T.; Aizawa, K. Cross-Domain Weakly-Supervised Object Detection through Progressive Domain Adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5001–5009.
29. Li, J.; Zhang, C.; Yang, B. Global Contextual Dependency Network for Object Detection. *Future Internet* **2022**, *14*, 27. [CrossRef]
30. Liang, H.; Zhou, H.; Zhang, Q.; Wu, T. Object Detection Algorithm Based on Context Information and Self-Attention Mechanism. *Symmetry* **2022**, *14*, 904. [CrossRef]
31. Chen, Z.M.; Jin, X.; Zhao, B.R.; Zhang, X.; Guo, Y. HCE: Hierarchical Context Embedding for Region-Based Object Detection. *IEEE Trans. Image Process.* **2021**, *30*, 6917–6929. [CrossRef] [PubMed]
32. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448. [CrossRef]
33. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. In Proceedings of the International Conference on Learning Representations, San Juan, Puerto Rico, 2–4 May 2016.
34. Everingham, M.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. Available online: http://host.robots.ox.ac.uk/pascal/VOC/index.html (accessed on 15 June 2022).
35. Everingham, M.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. Available online: http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html (accessed on 15 June 2022).
36. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The PASCAL Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
37. Lai, B.; Gong, X. Saliency Guided End-to-End Learning For Weakly Supervised Object Detection. In Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI'17, Melbourne, Australia, 19–25 August 2017; pp. 2053–2059. [CrossRef]
38. Tang, P.; Wang, X.; Wang, A.; Yan, Y.; Liu, W.; Huang, J.; Yuille, A. Weakly Supervised Region Proposal Network and Object Detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 352–368. [CrossRef]
39. Li, X.; Kan, M.; Shan, S.; Chen, X. Weakly Supervised Object Detection with Segmentation Collaboration. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 9734–9743. [CrossRef]

40. Yang, K.; Li, D.; Dou, Y. Towards Precise End-to-End Weakly Supervised Object Detection Network. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019.
41. Jin, R.; Lin, G.; Wen, C. Online Active Proposal Set Generation for Weakly Supervised Object Detection. *Knowl. Based Syst.* **2022**, *237*, 107726. [CrossRef]
42. Jiang, W.; Zhao, Z.; Su, F.; Fang, Y. Dynamic Proposal Sampling for Weakly Supervised Object Detection. *Neurocomputing* **2021**, *441*, 248–259. [CrossRef]