A Dual-Network Progressive Approach to Weakly Supervised Object Detection

Xuanyi Dong[1], Deyu Meng[2], Fan Ma[2], Yi Yang[1]*
[1]CAI, University of Technology Sydney, [2]MEKLINNS, Xi'an Jiaotong University
dongxuanyi888@icloud.com; dymeng@mail.xjtu.edu.cn
flower.fan@foxmail.com; Yi.Yang@uts.edu.au

ABSTRACT

A major challenge that arises in Weakly Supervised Object Detection (WSOD) is that only *image-level* labels are available, whereas WSOD trains *instance-level* object detectors. A typical approach to WSOD is to 1) generate a series of region proposals for each image and assign the *image-level* label to all the proposals in that image; 2) train a classifier using all the proposals; and 3) use the classifier to select proposals with high confidence scores as the *positive instances* for another round of training. In this way, the *image-level* labels are iteratively transferred to *instance-level* labels.

We aim to resolve the following two fundamental problems within this paradigm. First, existing proposal generation algorithms are not yet robust, thus the object proposals are often inaccurate. Second, the selected positive instances are sometimes noisy and unreliable, which hinders the training at subsequent iterations. We adopt two separate neural networks, one to focus on each problem, to better utilize the specific characteristic of region proposal refinement and positive instance selection. Further, to leverage the mutual benefits of the two tasks, the two neural networks are jointly trained and reinforced iteratively in a progressive manner, starting with easy and reliable instances and then gradually incorporating difficult ones at a later stage when the selection classifier is more robust. Extensive experiments on the PASCAL VOC dataset show that our method achieves state-of-the-art performance.

CCS CONCEPTS

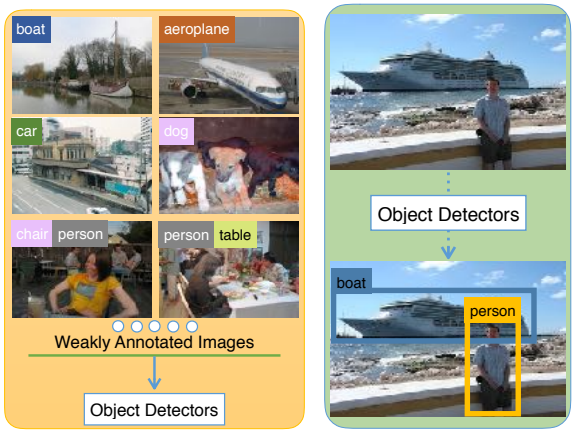
•Computing methodologies → Object detection;

KEYWORDS

Weakly Supervised Object Detection; Dual-Network; Progressive

1 INTRODUCTION

Object detection in images is one of the most fundamental and widely studied problems in computer vision and multimedia. With the significant progress in deep convolutional neural networks (CNN), modern deep CNN-based object detection algorithms have recently been successfully applied to consumer products, such as [9, 16, 28, 33, 34, 50]. These object detection algorithms require a large number of training data with instance-level labels. The instance-level labels, which are very expensive to attain, are indicated by a bounding box surrounding the object within an image. As shown in Figure 1, it is much easier to have numerous images with image-level labels, *e.g.*, ImageNet images or web images from search engines. Weakly Supervised Object Detection (WSOD), which utilizes weak labels at image level to train object detectors, has gradually become imperative in the field [5, 6, 30].



(a) Training Phase (b) Testing Phase

Figure 1: In a typical WSOD setting, only *image-level* labels are available for training. When testing, an algorithm generates *instance-level* labels indicated by bounding boxes.

A typical approach to WSOD starts with image-level labels and usually consists of three steps. First, a series of region proposals (or proposals for short) are generated for each image and an *image-level* label is assigned to all the proposals in that image. Second, a classifier is trained using all the proposals. Finally, the classifier is used to select proposals with high confidence scores as the *positive instances* for another round of training. This type of approach iteratively updates the classifier and then uses the updated classifier to select positive instances to refine the classifier, whereby the *image-level* labels are transferred to *instance-level* labels. One appealing feature of the iterative approach is that any component in each step can be readily replaced by a new method with better performance; for example, a better proposal generation algorithm can always be adopted in step one. Earlier work usually adopts Support Vector Machines (SVM) to generate pseudo labels at step

*To whom all correspondence should be addressed.

MM'17, October 23–27, 2017, Mountain View, CA, USA.

DOI: https://doi.org/10.1145/3123266.3123455

two, *e.g.* [5, 8, 35, 38, 51]. In contrast, Li *et al.* [27] use a CNN classifier along with the mask-out strategy to replace the SVM, resulting in a significant improvement in performance.

It has been empirically demonstrated that a better proposal generation algorithm will improve the accuracy of the classifier in object detection [9, 34, 36]. As the input information of the following iteration, the selected positive instance plays a key role in the classifier update. A more accurate classifier which selects positive instances always improves the performance in the loop. Since an object often has its own shape, a better classifier will generate more accurate labels for proposals, which in turn will provide shape information for proposal generation. Therefore, proposal generation and instance selection (for classifier updates) can be reinforced by each other. However, existing WSOD works usually take proposal generation and positive instance selection for classifier update as two independent subtasks, largely ignoring the complementary nature and mutual benefits of the two. When there are only weak supervisions, it becomes critical to leverage all information available for WSOD.

There are two major problems in the existing iterative algorithms for WSOD. First, these algorithms rely heavily on the proposals generated by an existing algorithm which is not yet robust. Thus the object proposals are often inaccurate, which may degrade the performance of WSOD. Second, given the weak image-level annotations, the selected *positive instances* with high confidence scores can be noisy and unreliable. As the subsequent iterations take the classifier-selected positive instances as input, such noise will hinder the training at a later stage. A straightforward way to ameliorate the limitations is to leverage the mutual benefit of proposal generation and classifier update in one neural network. However, each of the subtasks is quite challenging and has its characteristics. It is a non-trivial task to design a network architecture which is capable of effectively capturing the information of the two different challenging subtasks, especially when only weak supervision is available.

In this paper, we propose a dual-network progressive approach to WSOD. It carefully addresses the two challenges by integrating proposal refinement and classifier update into a joint framework. Rather than simply using a one-pass network architecture, we use two neural networks to preserve the individual characteristics of each subtask. We use the R-FCN architecture [9] with position sensitive score maps to fine-adjust the bounding boxes of the target object. The position sensitive score maps explicitly encode shape information, making it particularly appropriate for proposal shape refinement. We use the Fast R-CNN architecture [16] to update the classifier for positive instance selection, given the superior performance of the architecture in object classification. The two networks, namely the proposal refinement network, and the instance selection network, collaboratively adjust the proposal bounding boxes and refine proposal classifiers by refining the positive instance. Each network focuses on one subtask. The individual characteristics of each subtask are well preserved while the two networks also coordinate to mutually reinforce two training procedures.

If we use all training data at the beginning, initialization could be too noisy due to the weak supervision nature of WSOD, which might impede the subsequent training. Instead, we propose to start with easy and reliable examples and gradually incorporate difficult examples when the proposal classifier is confident. The proposed method is built upon all training samples, but in different phases, it progressively adds more labeled data into the training pool as the instance selection network and the proposal refinement network become stronger. We add a regularization term in our objective function to enable the two networks to share certain information when the training samples for the next iteration are updated, which can be finely interpreted as a self-paced curriculum learning (SPCL) procedure. Each network not only selects more training data to update its structure but also recommends/accepts reliable training data for/from the other network. The communication mechanism coordinates the dual-network framework to progressively and collaboratively converge to an optimal solution when no more training data can be added to the training pool.

In summary, the main contribution of this work is threefold:

- We propose a new dual-network approach to WSOD which optimizes proposal generation and instance selection in a joint framework. The two subtasks work collaboratively and progressively, during which process their individual characteristics and shared information are well exploited and reinforced.
- The proposed approach is a generic framework with high flexibility and extendability. Each component can be readily improved by a better existing or future substitution; for example, context [23], cascade structure [36] or a future new network architecture.
- The combination of the two networks is formulated as a concise regularization term and curriculum regime with mathematical rigor. The learning framework along with the optimization approach is a general one and applicable to a broad range of applications.

2 RELATED WORK

2.1 Weakly Supervised Object Detection.

Multiple Instance Learning (MIL) The majority of existing methods often solve WSOD via MIL [4, 5, 8, 10, 22, 35, 38, 41, 42]. These methods usually select region proposals in an image and interpret the image as a bag of candidate objects, and the bag-level labels are then used to learn the object appearance model. Some methods [8, 10, 38–42] focus on developing better initialization models, and others [4, 5, 8, 41] study the optimization of the poor local minimum problem in MIL. Bilen *et al.* [5] enforce similarities between the possible location of an image and clusters of objects to help optimize the solution into the global minimum. Song *et al.* [41] leverage quasi-Newton optimization techniques to discover object sets by a smoothed latent SVM formulation. These MIL-based methods usually initialize and train detectors from many noisy object candidates. Due to the negative effect of noisy training data, it is difficult to obtain a robust object detector.

Deep CNN for WSOD. Since CNNs greatly improve performance in various computer vision tasks, many researchers take advantage of CNNs to improve their detectors in WSOD problems. Some works focus on utilizing off-the-shelf CNN features, *e.g.*, [2, 4, 5, 32, 41, 42, 46]. Others design new CNN architectures which learn object information from the classification task and transform a classifier into a detector, *e.g.*, [6, 12, 27, 31]. Wang *et al.* [46] use

CNN features for region representation and propose latent category learning to help discover objects. Oquab *et al.* [31] propose a weakly supervised CNN architecture, trained from image-level labels, to predict the location of objects. Bilen *et al.* [6] improve the previous WSOD architecture by a two-stream CNN, in which one stream performs classification of the individual regions and the other performs detection by scoring regions relative to one another. Dong *et al.* [27] tackle the WSOD problem in two progressive adaptation steps. They train an image-level classifier, followed by complex procedures (*e.g.*, Mask-out strategy and MIL).

Most of these CNN-based methods do not explicitly contain region proposal refinement in their architectures, thus their detected objects are limited in the pre-computed proposals and lack the flexibility to refine proposals to suit the true objects. Even though some methods contain region proposal refinement to improve localization accuracy, it is usually set as the post process, *e.g.*,[12, 27]. It is thus not possible to take full advantage of the appearance and localization information during the training process.

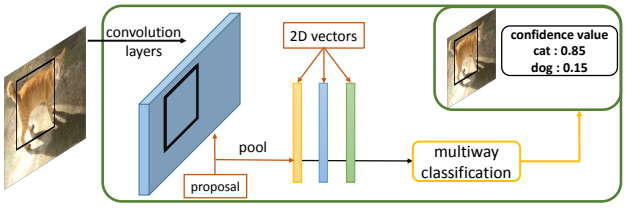
2.2 SPCL and Co-Training

Our approach adopts a progressive strategy with dual-network to address the weak supervision problem. In this sense, it is related to the SPCL and co-training [7]. There are also many applications that show co-training is capable of boosting performance by a large margin [19, 26]. For example, Levin *et al.* [26] propose the basic idea of how to use co-training to train object detectors. Most of these algorithms do not have an explicit model to provide theoretical guide, *e.g.*, [1, 7, 26, 47]. However, we embed this dual-property as a concise regularization term in our objective function, providing more insight to help the practice.

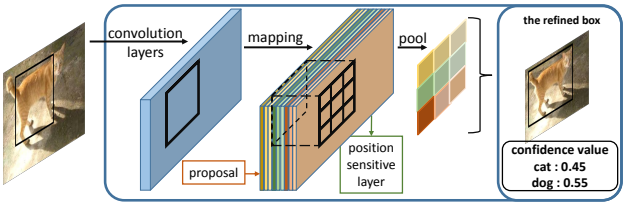
Bengio *et al.* [3] first proposed a general learning strategy: curriculum learning (CL). CL organizes the training examples in a meaningful order to acquire knowledge across a range of concepts, from simple to complex. Kumar *et al.* [25] proposed the self-paced learning (SPL) framework, automatically expanding the training pool in an easy-to-hard manner by converting the curriculum mechanism into a concise regularization term. CL uses human design to organize the examples, and SPL can automatically choose training examples according to loss values. Some other researchers, *e.g.*, [15, 29], further explore the theoretical analysis of the combination of SPL and Co-Training, and successfully improve the performance on some specific tasks. Yang *et al.* [49] found that leveraging the shared information among multiple tasks would improve the performance of multimedia content analysis. Supancic *et al.* [43] utilize SPCL to automatically select the track-right frame. Jiang *et al.* [20] propose Self-Paced Reranking for multimedia event detection. Compared to the existing SPCL algorithms, we incorporate the dual-property into the SPCL model on the detection task, which can also be easily generalized to other applications.

3 THE PROPOSED APPROACH

We introduce the CNN architectures of the Region Proposal Refinement (RPR) and Positive Instance Selection (PIS) networks in Section 3.1. In Section 3.2, we describe the initialization procedure of our framework. Lastly, we explain the objective function of our approach and demonstrate its solution algorithm in Section 3.3.



(a) The Positive Instance Selection Network



(b) The Region Proposal Refinement Network

Figure 2: A comparison between the Region Proposal Refinement (RPR) network and the Positive Instance Selection (PIS) network. The PIS network minimizes the classification error given a bounding box. The RPR network maximizes the IoU (intersection-over-union) between the pseudo bounding box and ground truth box.

3.1 Preliminaries and Motivations

We perform Fast R-CNN [16] without the box regression branch as the Positive Instance Selection (PIS) network, and R-FCN [9] as our basic detector, as shown in Fig 2. Both adopt an unsupervised proposal generation method, Selective Search (SS) [45], to obtain proposal bounding boxes. The PIS network leverages a Region of Interest (RoI) pooling layer to aggregate the features of all proposals into 2D vectors, followed by a classification function. The RPR network uses the position-sensitive layer instead of RoI and simultaneously performs refinement and classification.

Why unsupervised region-based methods? If we use supervised region proposal generation algorithms [34] or proposal-free detection algorithms [28, 33] to generate region proposals, a large amount of training data are required to achieve a high recall with a limited number of proposals. Weakly problems, however, are short of strong annotations, which are inadequate for obtaining good proposals. Further, poor proposals will seriously degrade detection performance. Some methods propose the use of object saliency to extract region proposals in weak settings *e.g.*, [10, 38, 40]. They can generate high-precision proposals, but they have much lower recall than SS or Edge Box [52] algorithms, limiting the final performance. Therefore, we use the state-of-the-art unsupervised proposal generation methods to provide reliable proposal bounding boxes. The objects have high possibilities to be located in these boxes.

Explanation of the CNN architectures. These two CNN architectures have their characteristics and can benefit the two fundamental problems: proposal refinement and instance selection. For example, as shown in Fig.3, the generated object bounding box results from each network are inaccurate, thus can not be used as proper instance-level labels. If we can combine their strengths

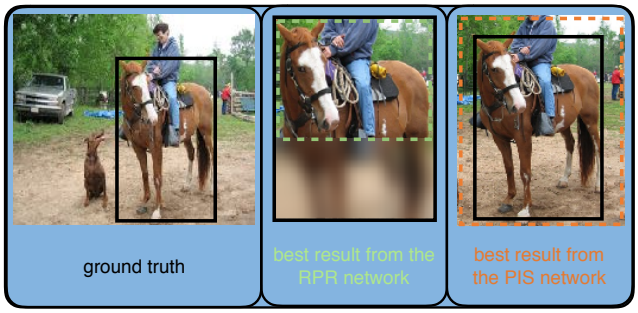


Figure 3: The two networks have different strengths which complement each other. The black box is the ground truth. The green box, generated by the Region Proposal Refinement (RPR) network, misses the lower part because it favors the bounding box which is occupied by the largest part of the object. The orange box, generated by the Positive Instance Selection (PIS) network focus on classification, and does not align well with the ground truth.

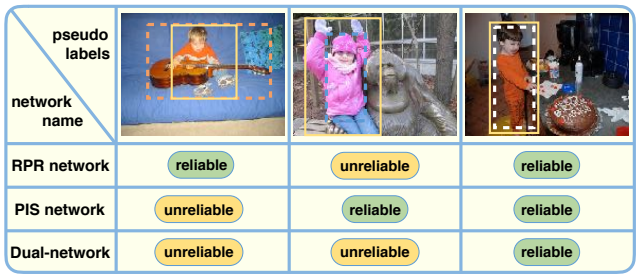


Figure 4: Dual-Network outperforms single network for computing the reliability of images with pseudo-labels. The solid yellow boxes indicate ground truth. The dashed-line boxes indicate pseudo-labels generated in the previous step of the same iteration.

to generate the boxes collaboratively, however, it is possible the results will align well with the ground truth. After our approach has generated the bounding boxes to annotate all images, we aim to select reliably annotated images as our training examples. As shown in Fig.4, each single network regards some wrongly annotated images as good, but our dual-network leverages the strengths of both networks to successfully assess the reliability of each image.

Notations. The PIS network $\boldsymbol{P}$ and RPR network $\boldsymbol{R}$ take an image $\boldsymbol{I}$ and a set of proposals as inputs. Region proposals are generated from the SS algorithm, denoted as :

$$B(I) = \{(left, top, right, bottom)_i | 1 \le i \le n\}, \tag{1}$$

where the results of $B(I)$ are n 4-dimension vectors representing the coordinate of n proposals. Thus the formulations of the PIS and RPR networks can be written as :

$$R(I, B(I)) = \{(rect, score)_{(i,j)} | 1 \le i \le n, 1 \le j \le C\}, \tag{2}$$

$$P(I, B(I)) = \{(score)_{(i,j)} | 1 \le i \le n, 1 \le j \le C\}, \tag{3}$$

where C is the number of object classes, $rect$ is the coordinate of the refined box and $score$ is the confidence score for the corresponding class ($rect \subset \mathbb{R}^4$).

3.2 Initialization

In this section, we introduce how to generate the box-level annotations, in the first epoch, to initialize our dual-network. First, we use SS to generate region proposals for each image. Then, we train a classifier with image-level labels to mine confident proposals as the box-level annotations. In our approach, we adopt a recent technique [23] to train this classifier, in which the basic idea is to leverage the CNN network to aggregate the confidence scores of multiple regions into an image-level classification probability, and to optimize it via the stochastic gradient descent (SGD) algorithm. After training, the confidence scores for each proposal can be obtained from the median results of the network.

Three post-procedures are used to eliminate the noisy proposal. We apply the non-maximum suppression (NMS) operation on the proposals with a threshold of 0.3 [1] for each class. A class-specific confidence threshold is used to further reduce the number of undesired proposals. The proposal classifier trained on weak labels is not robust for complex images; for example, images that contain multiple objects. We, therefore, prune those images that contain many generated bounding boxes to avoid the possibility of unreliable annotated images. Lastly, the reserved annotated images are used as training examples to initialize our dual-network. More specific parameters are described in the experiments section.

3.3 Dual-Network Model and Solutions

Standard SPCL. The progressive strategy in our approach is formulated as an improved SPCL model, which is traditionally used to handles the classification problem. Given a set of training data $\{(x_i, y_i) | 1 \le i \le n\}$, a decision function $f(x_i, \mathbf{w})$ and a loss function $L(y_i, f(x_i, \mathbf{w}))$, SPCL aims to learn the model parameter $\mathbf{w}$ and the latent variable $\mathbf{v} = [v_1, ..., v_n]$ by minimizing the learning objective Eq.4, where v_i indicates whether to select the i^{th} sample into the training pool. λ is a hyper-parameter to control the learning rate, and Ψ is the feasible region representing the predetermined curriculum guide.

$$\min_{\mathbf{w},\mathbf{v}} \mathbb{E}(\mathbf{w}, \mathbf{v}; \lambda) = \sum_{i=1}^{n} v_i L(y_i, f(x_i, \mathbf{w})) - \lambda \sum_{i=1}^{n} v_i, \quad s.t.\ \mathbf{v} \in \Psi,\ \mathbf{v} \in [0,1]^n, \tag{4}$$

Dual-Network. Our approach to the WSOD problem is to leverage two mutually beneficial networks. To incorporate this property, a concise regularization term is embedded to optimize the SPCL model. The label y is also a learning variable, which is a set of bounding boxes ($y \subset [\mathbb{R}^4, C]$) rather than the simple image-level class label in the standard SPCL model. Thus, we reformulate the objective function as Eq.5, where i denotes the index of the i^{th} image, and $j = 1, 2$ indicates the classifier and the detector, respectively. $\mathbf{w}^j$ represents the parameters for each model. $\mathbf{v}^j$ is the SPL regularization term, where $\mathbf{v}^j$ is all v_i^j for the j^{th} model. Similar to Eq. 4, v_i^j determines whether to select the i^{th} image for training the j^{th} model. $(\mathbf{v}^1)^T \mathbf{v}^2$ is the bimodal regularization term, encoding

[1] 0.3 is a practical value commonly used in many papers [9, 16, 34]

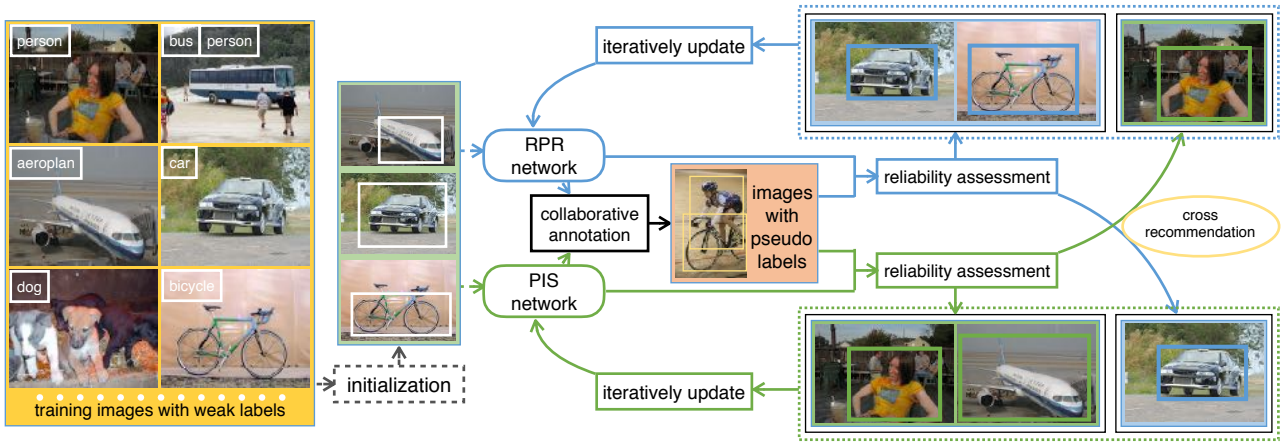


Figure 5: Overview of our Dual-Network Approach. Pseudo bounding boxes are initialized by an object detection algorithm, the results of which are fed into our dual-network algorithm to fine-tune the RPR and PIS networks. The two networks collaboratively generate a pool of the images with pseudo labels (bounding boxes). Each network assesses the reliability of images in the pool to select images to use for training in the next iteration. Each network also recommends/accepts images to/from the other.

the recommendation credibility of annotated images from the opposite network. When the parameter γ is small, each network will tend to ignore the recommendation from the opposite network. As γ grows, the recommendation from the opposite network becomes more important.

$$\min_{\mathbf{w},\mathbf{v},\mathbf{y}} \mathbb{E}(\mathbf{w},\mathbf{v},\mathbf{y};\lambda,\gamma,\Psi) = \sum_{j=1}^{2}\sum_{i=1}^{n} v_i^j L^j(y_i, I_i, B(I_i), w^j) - \sum_{j=1}^{2}\sum_{i=1}^{n} \lambda^j v_i^j - \gamma(\mathbf{v}^1)^T\mathbf{v}^2 \quad (5)$$
$$s.t.\ v_i^j \in \{0,1\},\ \mathbf{v} \in \Psi_v,\ \mathbf{y} \in \Psi_y$$

Curriculum Learning for Ψ_y. It is very difficult to directly optimize $\mathbf{y}$ in Eq.5 for many practical reasons, thus we apply several specifically designed procedures to obtain $\mathbf{y}$. Some researchers generate pseudo boxes $\mathbf{y}$ by using the highest confident proposal of each object class. We cannot well utilize the prior information in this way. In our approach, we use the NMS and threshold filtration[2] after we obtain the bounding box results from the network to choose the reliable generated boxes, which are much more likely to align with the objects. We also have weak image-level labels for each image, which can also help with bounding box generation. If the generated bounding boxes are inconsistent with the image-level labels, these boxes will simply be eliminated and we will only retain the generated boxes with the correct class labels.

Collaborative Annotation. We evaluate several strategies to extend the procedures in Ψ_y, which fuse the results of the PIS and the RPR networks, as follows: a) directly average the confidence scores for each region proposal, and ignore the bounding box refinement; b) similar to a) but contain the box refinement results before NMS operation; c) perform the pseudo box filtration Ψ_y on the proposals and do box refinement on the remained proposals. We analyze the performance of different strategies in the experiment section.

Curriculum Learning for Ψ_v. Real-world images are complex. For example, the 2nd image in Fig.6 contains a collection of 12 potted plants, making it quite difficult to detect all objects. Thus we apply a *image-pruning* strategy to improve the precision and recall of the selected training images. If an image contains too many generated bounding boxes, it will be eliminated, as shown in Fig.6, because the generated bounding boxes in such images have a high probability of being inaccurate. Our networks may not distinguish well between the overlapping objects of the same class due to weak supervision, *e.g.*, the bounding box in the 1st image encloses two close people but the ground truth is two boxes aligned with two separate people. Thus, we compare each generated bounding box with the nearby cropped boxes in that image. Assuming one generated box is $[x, y, w, h]$ [3], the nearby cropped boxes are $[x, y, \frac{w}{2}, h]$, $[x+\frac{w}{2}, y, \frac{w}{2}, h]$, $[x, y, w, \frac{h}{2}]$, $[x, y+\frac{h}{2}, w, \frac{h}{2}]$, $[x+\frac{w}{6}, y+\frac{h}{6}, \frac{2w}{3}, \frac{2h}{3}]$ and $[x-\frac{w}{6}, y-\frac{h}{6}, \frac{4w}{3}, \frac{4h}{3}]$. We calculate the confidence score for these boxes and compare them with the original score. If any nearby box has a higher confidence than the original bounding box, it implicitly indicates that we have not generated a well-aligned box. In this situation, these images are eliminated, and the retained images are used as training examples.

Solutions. The alternative search algorithm is used to solve our dual-network model, *i.e.*, Eq.5. We alternately optimize $\mathbf{y}$, $\mathbf{v}$ and $\mathbf{w}$, and their detailed solutions are described below.

Update $\mathbf{y}$: Fixing $\mathbf{v}$ and $\mathbf{w}$, we directly compute $\mathbf{y}$ via our designed procedure. As described in the curriculum learning for Ψ_y, $\mathbf{y}$ should be solved by several steps: collaborative bounding box generation by the PIS and RPR networks; NMS and thresholding; weak image-level label correction.

[2] 0.7 for the threshold of NMS and 0.2 for the confidence threshold

[3] x, y are the x- and y-coordinate of the top-left corner of this box, and w, h are the width and height.

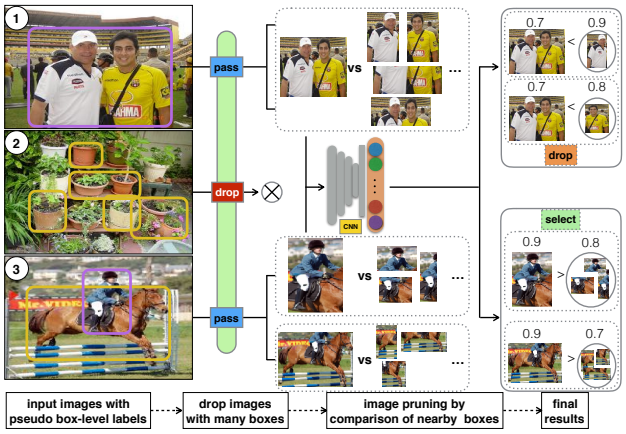


Figure 6: Annotated Image Pruning. When we select images, annotated by our approach, for the training pool, two steps are applied to prune unreliably annotated images. First, an image with many bounding boxes indicates that the image contains many objects and that it is relatively hard to detect objects well. Second, the comparison strategy is used to eliminate noisy annotated images. The 1st image is dropped, because the generated purple box has lower confidence than the left-half and right-half cropped boxes. In contrast, the 3rd image is selected for training.

Update $\mathbf{v}$: $\mathbf{v}$ determines which images will be used to train the two networks. By calculating the derivative of Eq.5 with respect to v_i^j, we obtain the following:

$$\frac{\partial E}{\partial v_i^j} = L^j(y_i, I_i, B(I_i), w^j) - \lambda^j - \gamma v_i^{3-j}, \ s.t. \ \mathbf{v} \in \Psi_v \tag{6}$$

The closed-form solution is

$$v_i^j = \begin{cases} 1 & \text{if } Loss_i^j < \lambda^j + \gamma v_i^{3-j} \\ 0 & \text{if } Loss_i^j \geq \lambda^j + \gamma v_i^{3-j} \end{cases} \tag{7}$$

After calculating $\mathbf{v}$, we follow the instructions $\mathbf{v} \in \Psi_v$ to set the $v_i^j = 0$ for the dropped images, as shown in Fig.6.

Update $\mathbf{w}$: In this step, we update the PIS network ($j = 1$) and the RPR network ($j = 2$) based on the selection indicator $\mathbf{v}$. Both can be solved by following the standard process, as described in [9, 16]. More specific parameters are shown in the experiment section.

The above updating processes are iterated by the sequence $y, v^1, w^1, y, v^2, w^2, y, \ldots$ until no more data are available or the maximum number of iteration is reached. We illustrate the overall framework in Fig.5. We first initialize the two networks according to Section.3.2. For each iteration, we collaboratively generate the annotation for each image, and use two networks to assess the reliability of each annotated image. Each network not only selects training examples by the self-assessment, but also accepts recommendations from the opposite network. After each network has selected their own training examples, the two networks will be updated. As we gradually improve the two networks, we are able to generate more reliable annotated images, and these images will be used to further improve the networks.

Algorithm 1 Algorithms for Solving the Dual-Network Model

Input: image with weak label $Data = \{(x_i, l_i)\}$
region proposals $PRs = \{(x_i, B(x_i))\}$
PIS network w^1/L^1 and RPR network w^2/L^2
hyper-parameter λ, γ, regime Ψ_v, Ψ_y

```
Initialize w^1 and w^2 according to Sec.3.2
Initialize v^1 = O, v^2 = O, iter = 1
while iter < max do
    Generate annotations D = {(x_i, y_i)}
    Set unsatisfied v = 0 based on v ∈ Ψ_v
    Compute loss L^1 for the PIS network.
    Update v^1 and re-train the parameter w^1.
    Compute loss L^2 for the RPR network.
    Update v^2 and re-train the parameter w^2.
end while
```

Output: the PIS with w^1 and the RPR with w^2

4 EXPERIMENTAL STUDIES

4.1 Benchmark Datasets

We evaluate our method on the PASCAL VOC 2007 and 2012 [14] datasets, which are the most widely-used benchmark in the WSOD problem. The VOC 2007 dataset contains 10,022 images annotated with bounding boxes for 20 object categories. It is officially split into 2,501 training, 2,510 validation and 5,011 test images. The VOC 2012 dataset is similar to VOC 2007 but is approximately twice as large. We use the official training and validation splits, and report the evaluation results on the testing split.

4.2 Experimental Setting

Network. Our PIS network and RPR network are built based on ResNet-50 [18] and VGG-16 [37], and other CNN models [13, 24, 44, 48] are also available. In the last iteration, we add the box refinement branch for the PIS network [17]. For the classification network in the initialization step, we choose ContextlocNet [23] with the VGG-16 base model, an improved WSDDN [6]. We use selective search [45] to extract about two thousand region proposals for each image, following the standard processes and configurations, as commonly used in [9, 16].

Initialization. The initial model parameters for each network are pre-trained on the ImageNet ILSVRC 2012 [24]. For simplification, we directly use the official released pre-trained models, downloaded from Caffe Model Zoo[4]. At the initialization step, we train the classification network via the standard procedure, as described in [23]. We empirically eliminate the complex images which have more than four objects with the same class. The class-specific threshold is then used to select high confidence proposals as pseudo box-level annotations. These threshold values are selected from $\{1.0, 2.0, 3.0, 4.0, 5.0\}$ according to the validation performance on the validation set.

Parameters. We choose the parameter γ from $\{0.1, 0.2, ..., 0.5\}$ based on the validation set, and use $\gamma = 0.3$ in our experiments. The number of images selected for training is determined on the basis of the validation set in the first iteration, and the number increases by

[4]https://github.com/BVLC/caffe/wiki/Model-Zoo

Table 1: Quantitative comparison of our dual-network approach on PASCAL VOC 2007 with the state-of-the-art in terms of AP in the test set. PIS denotes the method when only the PIS network is used, and RPR denotes the method when only the RPR network is used. Ensemble indicates simply fusing the results of the two networks. Dual-Network denotes our approach, which achieves state-of-the-art performance on mean AP.

Method	aero	bike	bird	boat	botle	bus	car	cat	chair	cow	table	dog	hors	mbik	pers	plnt	shp	sofa	train	tv	mean
Bilen *et al.* [5]	46.2	46.9	24.1	16.4	12.2	42.2	47.1	35.2	7.8	28.3	12.7	21.5	30.1	42.4	7.8	20.0	26.8	20.8	35.8	29.6	27.7
Wang *et al.* [46]	48.8	41.0	23.6	12.1	11.1	42.7	40.9	35.5	11.1	36.6	18.4	35.3	34.8	51.3	17.2	17.4	26.8	32.8	35.1	45.6	30.9
Zhang *et al.* [51]	47.4	22.3	35.3	23.2	13.0	50.4	48.0	41.8	1.8	28.9	27.8	37.7	41.6	43.8	**20.0**	12.0	27.8	22.9	48.9	31.6	31.3
Kantorov *et al.* [23]	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 *et al.* [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
Li *et al.* [27]	54.5	47.4	41.3	20.8	**17.7**	51.9	63.5	46.1	**21.8**	**57.1**	22.1	34.4	50.5	61.8	16.2	**29.9**	40.7	15.9	55.3	40.2	39.5
Jie *et al.* [21]	52.2	47.1	35.0	26.7	15.4	61.3	**66.0**	**54.3**	3.0	53.6	24.7	43.6	48.4	**65.8**	6.6	18.8	**51.9**	43.6	53.6	**62.4**	41.7
Diba *et al.* [12]	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
PIS	58.9	53.3	39.9	14.0	7.1	64.2	64.1	19.0	10.2	32.3	52.7	51.7	61.1	63.7	7.3	11.9	34.0	51.6	51.3	51.5	40.0
RPR	56.3	50.4	34.5	12.6	11.2	61.2	54.6	23.6	10.4	26.6	47.4	48.0	54.9	55.4	12.8	13.9	22.7	49.8	55.6	47.8	37.5
Ensemble	60.1	54.8	44.0	13.1	12.8	65.5	62.4	22.0	11.0	28.7	**52.2**	55.5	61.2	63.6	10.8	15.0	34.5	56.9	57.7	51.1	41.6
Dual-Network	**62.5**	54.6	**44.3**	12.9	12.7	63.8	60.6	25.0	5.4	48.0	49.3	**58.7**	**66.6**	63.5	8.5	17.3	40.7	**59.4**	53.9	51.4	**43.0**

Table 2: Performance on PASCAL VOC 2012 dataset.

	mAP (test)	CorLoc (trainval)
Li *et al.* [27]	29.1	-
Kantorov *et al.* [23]	35.3	54.8
Jie *et al.* [21]	**38.3**	58.8
Dual-Network	36.6	**61.4**

a fixed step for subsequent iterations. During the network training, we train four epochs in total, and set the learning rate as 0.001 for the first two epochs and 0.0001 for the last two. We use a weight decay of 0.0005 and a momentum of 0.9. All images are resized such that the shorter side of the image is 600 following [9, 16, 34]. We only use one GPU for training, setting two images with 256 RoI per batch. Online Hard Example Mining is adopted for the R-FCN network. During inference, we do not use any argumentation trick.

4.3 Evaluation and Comparison

Two standard methods are used to evaluate the efficiency of our Dual-Network approach. First, we calculate the average precision (AP) evaluation metrics on the test set. Second, we evaluate the correct localization (CorLoc) [11] on the trainval set. Both follow the standard evaluation settings.

Eight recently published state-of-the-art algorithms are compared with our Dual-Network approach, *i.e.*, [5, 6, 12, 21, 23, 27, 46, 51]. At first glance, it seems the comparison is unfair because we use two different networks. But it can be observed that others also merge or cascade multiple networks, For example, Diba *et al.* [12] cascade three stages in their CNN architecture, Bilen *et al.* [6] merge three networks with different architectures, and Dong *et al.* [27] perform two networks to make progressive domain adaptation. Thus for a fair comparison, we carefully cite the best performance with the same training images reported in their papers.

4.4 Analysis

Table 1 shows the AP on the PASCAL VOC 2007 test set. The ensemble method fuses the results from the two networks, which are separately trained. Compared with the single network and the ensemble method, our dual-network shows a significant improvement with about 1.5% absolute gain on mean AP. This observation demonstrates that our dual-network model can satisfactorily leverage the mutual benefits of the two complementary networks. In addition, our approach outperforms most of the other WSOD algorithms in terms of the mean AP. Diba *et al.* [12] achieve a similar mAP to our approach, but we obtain much higher mAP on some specific classes, *e.g.*, airplane (about 13%) and dining table (about 20%), and outperform them in terms of CorLoc. Our approach benefits from the progressive strategy and the performance of some classes which usually are usually easy to distinguish and can be identified at first glance can be greatly improved. For example, an airplane or a dining-table usually takes up a large proportion of the images. These objects can be likely to be well-detected and provide good annotations for the next round. Our approach is weak on the cluttered classes with small size; for example, if there are a lot of small bottles scattered on a table, we cannot clearly distinguish between each bottle with weak supervisions. Thus with the progressive manner, we might obtain a worse result due to poor initialization. We will investigate how to improve the performance of the small objects in the future.

Table 3: Ablation studies on PASCAL VOC 2007 trainval set (CorLoc) and test set (mAP). 'a', 'b' and 'c' indicate the different respective collaborative annotation strategies. RPR and PIS with VGG-16 indicates using the VGG-16 as the base model.

	mAP	CorLoc
PIS with VGG-16	39.2	59.0
RPR with VGG-16	32.7	53.2
a) average score without box refinement	**43.0**	**60.9**
b) box refinement before NMS	fail	fail
c) box refinement after NMS	42.6	60.0

Table 4 shows the correct localization on the PASCAL VOC 2007 trainval set, which is evaluated in the positive training images. The CorLoc only focuses on localizing one object on the positive class,

Table 4: Quantitative comparison on PASCAL VOC 2007 trainval set in terms of correct localization.

Method	aero	bike	bird	boat	botle	bus	car	cat	chair	cow	table	dog	hors	mbik	pers	plnt	shp	sofa	train	tv	mean
Bilen *et al.* [5]	66.4	59.3	42.7	20.4	21.3	63.4	74.3	59.6	21.1	58.2	14.0	38.5	49.5	60.0	19.8	39.2	41.7	30.1	50.2	44.1	43.7
Wang *et al.* [46]	80.1	63.9	51.5	14.9	21.0	55.7	74.2	43.5	26.2	53.4	16.3	56.7	58.3	69.5	14.1	38.3	58.8	47.2	49.1	60.9	48.5
Zhang *et al.* [51]	75.7	37.9	**68.3**	**53.2**	11.9	57.1	59.6	**63.7**	16.4	63.9	17.5	62.3	71.6	71.5	**45.6**	14.7	53.1	41.1	75.5	24.4	49.3
Bilen *et al.* [6]	73.1	68.7	52.4	34.3	26.6	66.1	76.7	51.6	15.1	66.7	17.5	45.4	71.8	82.4	32.6	42.9	71.9	53.3	60.9	65.2	53.8
Li *et al.* [27]	78.2	67.1	61.8	38.1	36.1	61.8	78.8	55.2	28.5	68.8	18.5	49.2	64.1	73.5	21.4	47.4	64.6	22.3	60.9	52.3	52.4
Kantorov *et al.* [23]	83.3	68.6	54.7	23.4	18.3	73.6	74.1	54.1	8.6	65.1	47.1	59.5	67.0	83.5	35.3	39.9	67.0	49.7	63.5	65.2	55.1
Jie *et al.* [21]	72.7	55.3	53.0	27.8	35.2	68.6	81.9	60.7	11.6	71.6	29.7	54.3	64.3	**88.2**	22.2	**53.7**	72.2	52.6	68.9	**75.5**	56.1
Diba *et al.* [12]	83.9	**72.8**	64.5	44.1	**40.1**	65.7	**82.5**	58.9	**33.7**	72.5	25.6	53.7	67.4	77.4	26.8	49.1	68.1	27.9	64.5	55.7	56.7
PIS	84.2	70.6	61.9	23.9	20.6	77.7	80.4	39.5	11.7	70.5	**67.3**	66.0	85.0	**87.1**	21.1	41.8	67.0	58.3	68.8	70.3	58.7
RPR	80.8	61.6	58.9	23.9	18.3	71.1	74.1	36.9	12.4	66.4	53.6	58.6	71.8	80.7	22.2	35.9	59.8	50.3	66.5	60.2	53.2
Ensemble	**86.2**	69.0	64.0	26.6	21.8	79.7	79.0	41.0	14.7	71.9	66.5	64.7	84.4	87.1	23.2	45.1	67.0	59.1	**70.0**	68.8	59.5
Dual-Network	85.3	71.9	66.8	27.0	26.5	**81.2**	78.5	36.1	17.2	**80.6**	61.8	**76.1**	86.3	83.6	22.2	43.6	**74.8**	**60.6**	67.6	70.5	**60.9**

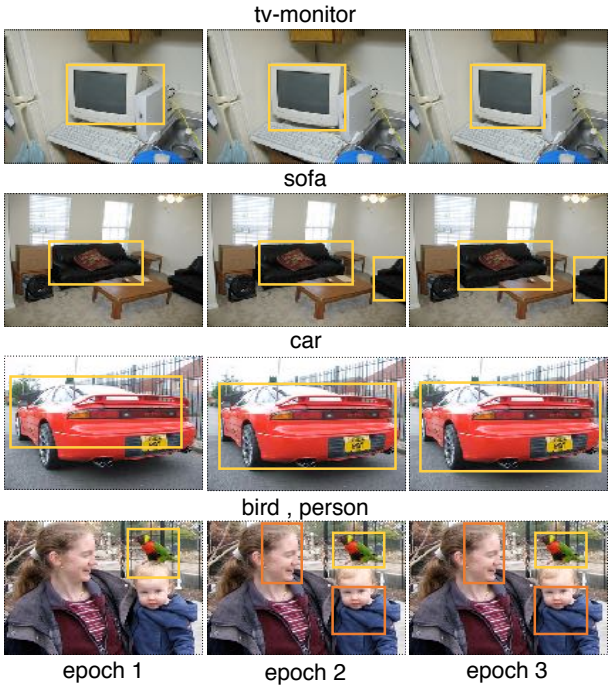


Figure 7: Examples of our generated box-level annotations. The yellow and orange boxes are generated by our dual-network approach and used as annotations for the next round training. Each triplet indicates the boxes generated in the same image from 1st epoch to the 3rd epoch.

which is easier than mAP. Our approach obtains approximately 4% improvement over others. This may be because the localization capability of our PIS network can be improved by collaboration with our RPR network. A similar improvement of CorLoc on PASCAL VOC 2012 can be seen in Fig.2. We also illustrate the success and failure of some annotated images in Fig.7, in which the generated yellow bounding boxes are good and gradually align with the objects. The orange boxes miss some parts of the person, which are bad for training.

4.5 Ablation Studies

We compare the performances of some different configurations, as shown in Table 3. Three collaborative annotation strategies are compared. a) is our baseline method and used in Table 1 and Table 4. As we can observe, the collaborative annotation strategy b) fails. The proposal refinement is unstable and difficult to learn, thus if we refine all the region proposals before NMS, there will be large noisy wrongly-refined proposals, which may harm the performance. Refinement after NMS (c) obtains a slight improvement because the noisy proposals, which have a high possibility of being wrongly refined, are eliminated. In addition, the RPR network with VGG-16 obtains poor performance, because the base VGG-16 contains three full connected layers, which cannot be directly applied in the RPR architecture. As a result, meaningful information is discarded from the RPR architecture with VGG-16 model. This greatly degrades the overall object detection performance, and we, therefore, do not use VGG-16 as our base model.

5 CONCLUSION

In this paper, we propose an effective dual-network approach to solve the WSOD problem. In typical solutions to the WSOD problem, the positive instance selection and the region proposal refinement are fundamental problems, which greatly influence the final detection performance. In contrast, we leverage a PIS network and a RPR network to focus on these two problems, respectively. With the mutual benefits of the characteristics of the two networks, our approach progressively improves both networks and ultimately achieves a better detection performance. Experimental results demonstrate that our dual-network approach achieves state-of-the-art performance compared to other algorithms, achieving 43.0% mAP on the PASCAL VOC 2007 test set. Besides, by substituting more superior modules in each step, our dual-network approach is expected to inspire further strategies to enhance WSOD in future research.

Acknowledgment. Xuanyi Dong and Yi Yang are partially supported by the Google Faculty Research Award and the Data to Decisions Cooperative Research Centre (www.d2dcrc.com.au). Deyu Meng and Fan Ma are partially supported by the China NSFC projects under contract 61373114, 61661166011, 11690011 and 61603292.

REFERENCES

[1] Maria-Florina Balcan, Avrim Blum, and Ke Yang. 2004. Co-training and expansion: Towards bridging theory and practice. In *NIPS*.

[2] Loris Bazzani, Alessandra Bergamo, Dragomir Anguelov, and Lorenzo Torresani. 2016. Self-taught object localization with deep networks. In *WACV*.

[3] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *ICML*.

[4] Hakan Bilen, Marco Pedersoli, and Tinne Tuytelaars. 2014. Weakly supervised object detection with posterior regularization. In *BMVC*.

[5] Hakan Bilen, Marco Pedersoli, and Tinne Tuytelaars. 2015. Weakly supervised object detection with convex clustering. In *CVPR*.

[6] H. Bilen and A. Vedaldi. 2016. Weakly supervised deep detection networks. In *CVPR*.

[7] Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*.

[8] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. 2014. Multi-fold MIL training for weakly supervised object localization. In *CVPR*.

[9] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. 2016. R-FCN: object detection via region-based fully convolutional networks. In *NIPS*.

[10] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. 2010. Localizing objects while learning their appearance. In *ECCV*.

[11] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. 2012. Weakly supervised localization and learning with generic knowledge. *International journal of computer vision* 100, 3 (2012), 275–293.

[12] Ali Diba, Vivek Sharma, Ali Pazandeh, Hamed Pirsiavash, and Luc Van Gool. 2017. Weakly supervised cascaded convolutional networks. *CVPR*.

[13] Xuanyi Dong, Junshi Huang, Yi Yang, and Shuicheng Yan. 2017. More is Less: A More Complicated Network with Less Inference Complexity. In *CVPR*.

[14] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2015. The pascal visual object classes challenge: A retrospective. *International journal of computer vision* 111, 1 (2015), 98–136.

[15] Ma Fan, Meng Deyu, Xie Qi, Zina Li, and Xuanyi Dong. 2017. Self-paced Co-training. In *ICML*.

[16] Ross Girshick. 2015. Fast R-CNN. In *ICCV*.

[17] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.

[19] Omar Javed, Saad Ali, and Mubarak Shah. 2005. Online detection and classification of moving objects using progressively improving detectors. In *CVPR*.

[20] Lu Jiang, Deyu Meng, Teruko Mitamura, and Alexander G Hauptmann. 2014. Easy Samples First: Self-paced Reranking for Zero-Example Multimedia Search. In *ACM MM*.

[21] Zequn Jie, Yunchao Wei, Xiaojie Jin, Jiashi Feng, and Wei Liu. 2017. Deep self-taught learning for weakly supervised object localization. In *CVPR*.

[22] Asako Kanezaki, Yasuo Kuniyoshi, and Tatsuya Harada. 2013. Weakly-supervised multi-class object detection using multi-type 3D features. In *ACM MM*.

[23] Vadim Kantorov, Maxime Oquab, Minsu Cho, and Ivan Laptev. 2016. ContextLocNet: Context-aware deep network models for weakly supervised localization. In *ECCV*.

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *NIPS*.

[25] M Pawan Kumar, Benjamin Packer, and Daphne Koller. 2010. Self-paced learning for latent variable models. In *NIPS*.

[26] Anat Levin, Paul Viola, and Yoav Freund. 2003. Unsupervised improvement of visual detectors using cotraining. In *ICCV*.

[27] Dong Li, Jia-Bin Huang, Yali Li, Shengjin Wang, and Ming-Hsuan Yang. 2016. Weakly Supervised Object Localization with Progressive Domain Adaptation. In *CVPR*.

[28] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, and Scott Reed. 2016. SSD: Single Shot MultiBox Detector. In *ECCV*.

[29] De-yu Meng, Qian Zhao, and Lu Jiang. 2017. A Theoretical Understanding of Self-paced Learning. *Information Sciences* 414, 11 (2017), 319–328.

[30] Pascal Mettes. 2016. Weakly-supervised recognition, localization, and explanation of visual entities. In *ACM MM*.

[31] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. 2015. Is object localization for free? Weakly-supervised learning with convolutional neural networks. In *CVPR*.

[32] Maxime Oquab, Léon Bottou, Ivan Laptev, Josef Sivic, et al. 2014. Weakly supervised object recognition with convolutional neural networks. In *NIPS*.

[33] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *CVPR*.

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*.

[35] Miaojing Shi and Vittorio Ferrari. 2016. Weakly supervised object localization using size estimates. In *ECCV*.

[36] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. 2016. Training region-based object detectors with online hard example mining. In *CVPR*.

[37] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*.

[38] Parthipan Siva, Chris Russell, and Tao Xiang. 2012. In defence of negative mining for annotating weakly labelled data. In *ECCV*.

[39] Parthipan Siva, Chris Russell, Tao Xiang, and Lourdes Agapito. 2013. Looking beyond the image: Unsupervised learning for object saliency and detection. In *CVPR*.

[40] Parthipan Siva and Tao Xiang. 2011. Weakly supervised object detector learning with model drift detection. In *ICCV*.

[41] Hyun Oh Song, Ross B Girshick, Stefanie Jegelka, Julien Mairal, Zaid Harchaoui, Trevor Darrell, et al. 2014. On learning to localize objects with minimal supervision. In *ICML*.

[42] Hyun Oh Song, Yong Jae Lee, Stefanie Jegelka, and Trevor Darrell. 2014. Weakly-supervised discovery of visual pattern configurations. In *NIPS*.

[43] James S Supancic and Deva Ramanan. 2013. Self-paced learning for long-term tracking. In *CVPR*.

[44] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *CVPR*.

[45] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. 2013. Selective search for object recognition. *International journal of computer vision* 104, 2 (2013), 154–171.

[46] Chong Wang, Weiqiang Ren, Kaiqi Huang, and Tieniu Tan. 2014. Weakly supervised object localization with latent category learning. In *ECCV*.

[47] Wei Wang and Zhi-Hua Zhou. 2010. A new analysis of co-training. In *ICML*.

[48] Saining Xie, Ross Girshick, Piotr Dollr, Zhuowen Tu, and Kaiming He. 2017. Aggregated Residual Transformations for Deep Neural Networks. *CVPR*.

[49] Yi Yang, Zhigang Ma, Alexander G Hauptmann, and Nicu Sebe. 2013. Feature selection for multimedia analysis by sharing information among multiple tasks. *IEEE Transactions on Multimedia* 15, 3 (2013), 661–669.

[50] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. 2016. UnitBox: An advanced object detection network. In *ACM MM*.

[51] Dingwen Zhang, Deyu Meng, Long Zhao, and Junwei Han. 2016. Bridging saliency detection to weakly supervised object detection based on self-paced curriculum learning. In *IJCAI*.

[52] C Lawrence Zitnick and Piotr Dollár. 2014. Edge boxes: locating object proposals from edges. In *ECCV*.