Weakly Supervised Object Localization with Progressive Domain Adaptation

Dong Li[1], Jia-Bin Huang[2], Yali Li[1], Shengjin Wang[1]*, and Ming-Hsuan Yang[3]
[1]Tsinghua University, [2]University of Illinois, Urbana-Champaign, [3]University of California, Merced

Abstract

We address the problem of weakly supervised object localization where only image-level annotations are available for training. Many existing approaches tackle this problem through object proposal mining. However, a substantial amount of noise in object proposals causes ambiguities for learning discriminative object models. Such approaches are sensitive to model initialization and often converge to an undesirable local minimum. In this paper, we address this problem by progressive domain adaptation with two main steps: classification adaptation and detection adaptation. In classification adaptation, we transfer a pre-trained network to our multi-label classification task for recognizing the presence of a certain object in an image. In detection adaptation, we first use a mask-out strategy to collect class-specific object proposals and apply multiple instance learning to mine confident candidates. We then use these selected object proposals to fine-tune all the layers, resulting in a fully adapted detection network. We extensively evaluate the localization performance on the PASCAL VOC and ILSVRC datasets and demonstrate significant performance improvement over the state-of-the-art methods.

1. Introduction

Object localization is an important task for image understanding. It aims to identify all instances of particular object categories (e.g., person, cat, and car) in images. The fundamental challenge in object localization lies in constructing object appearance models for handling large intra-class variations and complex background clutters. The state-of-the-art approaches typically train object detectors from a large set of training images [11, 14] in a fully supervised manner. However, this strongly supervised learning paradigm relies on *instance-level* annotations, e.g., tight bounding boxes, which are time-consuming and labor-intensive. In this paper, we focus on the *weakly supervised* object localization problem where only binary image-level labels indicating the presence or absence of an object category are available for training. Figure 1 illustrates the problem setting. This particular setting is important for large-scale practical applications because image-level annotations are often readily available from the Internet, e.g., through text tags [15], GPS tags [8], and image search queries [23].

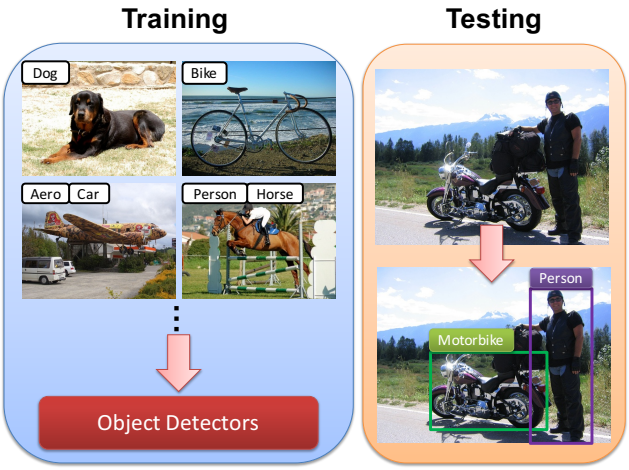

Figure 1. Weakly supervised object localization problem setting. Given a collection of training images with image-level annotations, our goal is to train object detectors for localizing objects in unseen images.

Most existing methods [36, 4, 3, 2, 33, 35, 37, 34] formulate the weakly supervised localization (WSL) task as a multiple instance learning (MIL) problem. Recent efforts include leveraging convolutional neural networks (CNN) to extract discriminative appearance features [41, 36, 37, 2, 3] and transferring knowledge from strongly supervised detectors to other categories without bounding box annotations [27, 16, 17, 31]. While existing methods have shown promising results, these methods have three main drawbacks. First, it's hard to select correct object proposals because the collection of candidate proposals contains too much noise. Typically, only a few out of several thousands of proposals are actual object instances. Second, many approaches use a pre-trained CNN as a feature extractor and do not adapt the weights from whole-image classification to object detection. Third, existing methods often require either auxiliary strongly annotated data or pre-trained detectors for domain adaptation.

*Corresponding author

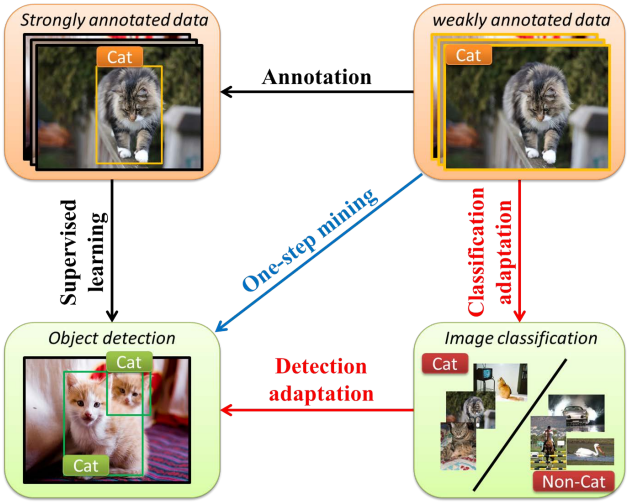

Figure 2. Comparison of our approach with existing object localization methods. Strongly supervised methods use instance-level annotations to train object detectors. Most of the weakly supervised methods use one-step proposal mining to select object instances from a large and noisy candidate pool directly. We propose a two-step progressive domain adaptation approach. We first filter out the noisy object proposal collection and then mine confident candidates for learning discriminative object detectors.

In this paper, we propose a two-step domain adaptation for weakly supervised object localization: classification adaptation and detection adaptation. Figure 2 illustrates the major difference between the proposed algorithm and existing work. Our key observation is that it's hard to train object detectors directly under weak supervisory signals due to the substantial amount of noise in the object proposal collections. Essentially, the main difficulty arises from the large gap between source domain and target domain, as shown in the top-right and bottom-left corner of Figure 2. The goal of our work is to bridge the gap by progressive domain adaptation. In the classification adaptation step, we train a classification network using the given weak image-level labels. We train the classification network to recognize the presence of a certain object category in an image. In the detection adaptation step, we use the classification network to collect class-specific object proposals and apply multiple instance learning to mine confident candidates. We then use the previously selected object candidates to fine-tune all the layers, resulting in a fully adapted detection network.

The proposed algorithm addresses the drawbacks from prior work in three aspects: (1) Our classification adaptation step fine-tunes the network such that it can collect class-specific object proposals with higher precision. This step aims at removing background clutters and potential confusion from similar objects cases, leading to a *purified* collection of object candidates for multiple instance learning. (2) Detection adaptation uses confident object candidates to optimize the CNN representations for the target domain. This step aims at turning image classifiers into object detectors, providing more discriminative feature representations for localizing generic objects (instead of the presence of them) in an image. (3) Our method learns object detectors from weakly annotated data without any strong labels.

We make the following three contributions in this work:

1. We propose to use progressive domain adaptation for weakly supervised object localization. We show that this strategy is crucial for good performance.
2. Our classification adaptation helps filter the object proposal collection, and our detection adaptation helps learn discriminative feature representation for the detection task.
3. We present detailed evaluations on the PASCAL VOC and ILSVRC datasets. Experimental results demonstrate that our progressive domain adaptation algorithm performs favorably against the state-of-the-art methods. Our detector achieves 39.5% mAP on VOC 2007, surpassing the second best performing algorithm by 8 points.

2. Related Work

Weakly supervised learning. Existing methods often treat WSL as an MIL problem [36, 4, 3, 2, 33, 35, 34, 37]. In an MIL framework, each image is considered as a bag of potential object instances. Positive images are assumed to contain *at least* one object instance of a certain object category and negative images do not contain object instances from this category. Using this weak supervisory signal, WSL methods often alternate between (1) selecting the positive object instances from positive images and (2) learning object detectors. However, this results in a non-convex optimization problem. Due to the non-convexity, these methods are sensitive to model initialization and prone to getting trapped into local extrema. Although many efforts have been made to overcome the problem via seeking better initialization models [36, 37, 34, 35, 33] and optimization strategies [4, 3, 2], the localization performance is still limited. We observe that previous MIL-based methods attempt to train object detectors directly from the large and noisy collection of object candidates. In this work, we apply MIL [36] to mine confident candidates. However, unlike existing methods, we apply MIL on a cleaner collection of *class-specific* object proposals (instead of on a large, noisy, category-independent proposals).

Convolutional neural networks for object localization. Recently, convolutional neural networks have achieved great success on various visual recognition tasks [21, 40, 42, 32, 29, 25, 14, 13]. The key ingredient for the success lies in end-to-end training CNN in a *fully supervised* fashion. In

object detection problems, these methods [29, 25, 14, 13] require instance-level supervision. Moving beyond strong supervision, recent work focuses on applying off-the-shelf CNN features [36, 37, 41, 1, 3, 2], learning from weak labels [43, 24] or noisy labels [26, 38]. Our classification adaptation step is related to the method by Oquab et al. [24] in the formulation of multi-label classification. We use a different a multi-label loss that allows us to incorporate negative images during training. Also, we focus on detecting the locations and spatial supports of objects while the method by Oquab et al. [24] only predicts approximate locations of objects. Our work resembles the work by Bazzani et al. [1]. We use a similar mask-out strategy to collect class-specific object proposals. The main differences are three-fold. (1) Our classification adaptation transfers the source classification domain (1000 *single-label* classes for ILSVRC 2012) to the target classification domain (20 *multi-label* classes for PASCAL VOC). (2) We use a contrast-based mask-out strategy for ranking proposals. (3) Instead of training a classifier over pre-trained CNN features, we fine-tune the parameters of all the CNN layers for training object detectors.

Domain adaptation. Some recent approaches use domain adaptation to help learn object detectors or features [27, 16, 17, 31]. Shi et al. [31] learn a mapping relationship between the bounding box overlap and the appearance similarity, and then transfer it to the target domain. Hoffman et al. [16] learn the difference between classification and detection tasks and transfer this knowledge to convert classifiers to detectors using weakly annotated data. Also, MIL is incorporated for joint learning of representation and detector [17]. Rochan et al. [27] transfer existing appearance models of the familiar objects to the unseen object. Existing domain adaptation methods often use strongly annotated source data to improve recognition performance for weakly supervised object localization. Our work differs from these approaches in that we focus on object localization in a weakly supervised manner, i.e., we do not require any instance-level annotation and do not borrow additional strongly annotated data or outside detectors.

Progressive and self-paced learning. Our work is also related to several approaches in other problem contexts. Examples include visual tracking [39], pose estimation [12], image search [20], and object discovery [22]. Progressive methods can decompose complex problems into simpler ones. We find that progressive adaptation is particularly important for the weakly supervised object localization problem.

3. Classification Adaptation

In this section, we introduce the classification adaptation step. This step aims to train the whole-image classification network such that the adapted network is sensitive to the object categories of interest. The original AlexNet [21] is trained for multi-class classification with a softmax loss layer by assuming that only one single object exists per image. In our adapted network, we replace the last classification layer with a multi-label loss layer. Unlike the problem in ImageNet classification, we address a more general multi-label classification problem where each image may contain multiple objects from more than one category.

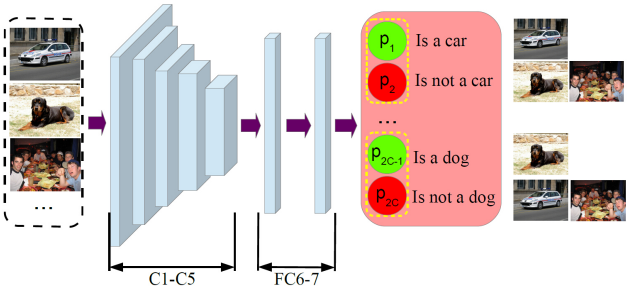

Figure 3. Classification adaptation step. We use the AlexNet architecture [21] and replace the softmax loss layer with the proposed multi-label loss layer. We set the number of nodes in the last fully-connected layer to $2C$ (C is number of object categories). These $2C$ entries are divided into C pairs for representing the *presence* and *absence* of each object category. See Section 3 for details.

Assuming that the object detection dataset has C categories and a total of N training images, we denote the weakly labeled training image set as $\mathcal{I} = \{(\mathbf{I}^{(1)}, \mathbf{y}^{(1)}), \ldots, (\mathbf{I}^{(N)}, \mathbf{y}^{(N)})\}$, where $\mathbf{I}$ is the image data and $\mathbf{y} = [y_1, \ldots, y_c, \ldots, y_C]^\top \in \{0,1\}^C, c \in \{1, \ldots, C\}$ is the C-dimensional label vector of $\mathbf{I}$, in which each entry can be 1 or 0 indicating whether at least one specific object instance exists in the image. In the weakly object localization setting, one image may contain objects from different categories, i.e., more than one entry in $\mathbf{y}$ can be 1. In this case, conventional *softmax* loss cannot be used for this multi-label classification problem. We thus introduce a multi-label loss to handle this problem.

First, we transform the original training label to a new label $\mathbf{t} \in \{0,1\}^{2C}$, where

$$t_{2c-1} = \begin{cases} 1, & y_c = 1 \\ 0, & y_c = 0 \end{cases} \quad \text{and} \quad t_{2c} = \begin{cases} 0, & y_c = 1 \\ 1, & y_c = 0 \end{cases}. \quad (1)$$

In other words, each odd entry of $\mathbf{t}$ represents whether the image contains the corresponding object. Similarly, each even entry represents whether the image *does not* contain the corresponding object.

We then introduce our new loss layer for multi-label classification. We denote the CNN as a function $\mathbf{p}(\cdot)$ that maps an input image $\mathbf{I}$ to a $2C$ dimensional output $\mathbf{p}(\mathbf{I}) \in \mathbb{R}^{2C}$. The odd entry $p_{2c-1}(\mathbf{I})$ represents the probability that the image contains at least one object instance of c-th category.

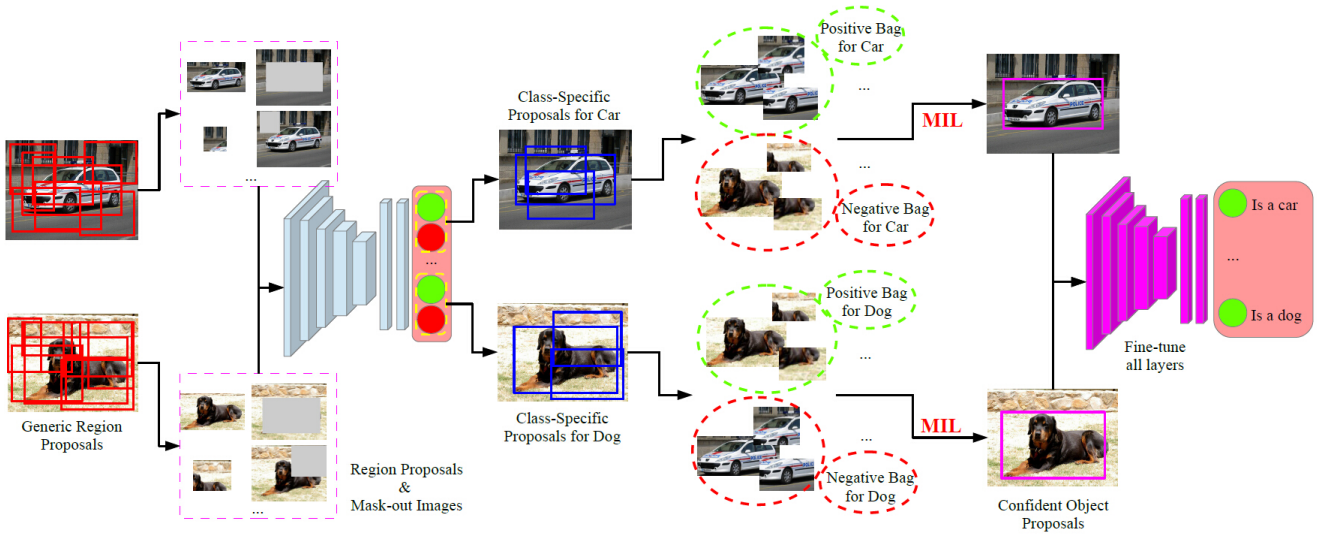

Figure 4. Detection adaptation step. We first use a mask-out strategy to collect class-specific object proposals and apply multiple instance learning to mine confident candidates. We then use these selected object proposals to fine-tune all the layers (marked magenta), resulting in a network that is fully adapted for detection. See Section 4 for details.

Similarly, the even entry $p_{2c}(\mathbf{I})$ indicates the probability that the image does not contain objects of c-th category. We compute the probabilities using a sigmoid for each object class and thus we have $p_{2c-1}(\mathbf{I}) + p_{2c}(\mathbf{I}) = 1$.

We can define negative logarithmic classification loss $L_c(\mathbf{I})$ of one image for category c as,

$$L_c(\mathbf{I}) = -(t_{2c-1} \log p_{2c-1}(\mathbf{I}) + t_{2c} \log p_{2c}(\mathbf{I})). \quad (2)$$

We obtain the final loss function L by summing up all the training samples and losses for all the categories,

$$L = \sum_{i=1}^{N} \sum_{c=1}^{C} L_c(\mathbf{I}^{(i)}) = -\sum_{i=1}^{N} \mathbf{t}^{(i)} \log \mathbf{p}(\mathbf{I}^{(i)}). \quad (3)$$

Here $\log(\cdot)$ is the element-wise logarithmic function. The sum over different categories is done by dot product.

In the classification adaptation network, we substitute the conventional softmax loss layer with our new multi-label loss layer and adjust the number of nodes in the last fully-connected layer to $2C$. We use mini-batch Stochastic Gradient Descent (SGD) for training the CNN. We initialize all the layers except the last layer using the pre-trained parameters on ILSVRC 2012 [6]. We randomly initialize the weights of the modified classification layer. We describe the implementation details in Section 5.1.

4. Detection Adaptation

4.1. Class-specific proposal mining

The goal of detection adaptation step is to transfer image classifiers to object detectors. To train object detectors, we first collect confident object proposals. We use a mask-out strategy to collect class-specific object proposals and apply multiple instance learning to mine confident candidates. The mining procedure offers two key benefits:

- Compared with generic object proposals, class-specific proposals remove substantial noise and potential confusion from similar objects. This helps MIL avoid converging to an undesirable local minimum and reduce computational complexity.
- More precise object proposals can be mined using MIL. These confident object proposals allow us to further fine-tune the network for object detection.

The adapted classification network recognizes whether an image contains a certain object. We use a mask-out strategy to collect object proposals for each class based on the adapted classification network. The idea of masking out the input of CNN has been previously explored in [44, 1]. Intuitively, if the mask-out image by a region causes a significant drop in classification score for the c-th class, the region can be considered discriminative for the c-th class. Inspired by [44, 1], we investigate the contrastive relationship between a selected region and its mask-out image.

Without loss of generality, we take mining object proposals for the c-th category as an example. First, for the image $\mathbf{I}$ labeled with $y_c = 1$, we apply Edge Boxes [46] to generate the initial collection of object proposals. The set of initial proposals is marked as $\hat{\mathbf{B}}_c$. For an initial bounding box region $\hat{\mathbf{b}}$, we denote its corresponding image as $\mathbf{I}_{\text{in}}(\hat{\mathbf{b}})$ and its mask-out image as $\mathbf{I}_{\text{out}}(\hat{\mathbf{b}})$. We generate the mask-out image by replacing the pixel values within $\hat{\mathbf{b}}$ with the fixed mean pixel values pre-computed on ILSVRC 2012. We feed the region image $\mathbf{I}_{\text{in}}(\hat{\mathbf{b}})$ and mask-out image $\mathbf{I}_{\text{out}}(\hat{\mathbf{b}})$ to the adapted classification network. We can then compute the

contrastive score for bounding box region $\hat{\mathbf{b}}$ of image $\mathbf{I}$ as,

$$s_c(\mathbf{I}, \hat{\mathbf{b}}) = p_{2c-1}(\mathbf{I}_{\text{in}}(\hat{\mathbf{b}})) - p_{2c-1}(\mathbf{I}_{\text{out}}(\hat{\mathbf{b}})). \quad (4)$$

Here, if the value of $s_c(\mathbf{I}, \hat{\mathbf{b}})$ is large, it indicates that the region $\hat{\mathbf{b}}$ is likely an object of the c-th category. Note that our mask-out strategy differs from [1], which compute the score difference between the whole image and mask-out image.

With classification adaptation, a bounding box region can achieve higher confidence than the whole image for classification. In Figure 5, we show top 10 class-specific proposals using our mask-out strategy. According to (4), top M (M=50 in our experiments) object proposals of image $\mathbf{I}$ are selected for the c-th category. That is, selected proposals are *category-specific*. We mark the top-ranked proposals as $\mathbf{B}_c$.

Since we set a loose criteria in the previous mask-out step, the top-ranked proposals are still coarse and may contain many false positives. We apply MIL to mine confident candidates for training object detector. In MIL, the label of object candidate is set as a latent variable. During the training, the label is iteratively updated. For candidates set $\mathbf{B}_c$, we set up latent variable $\mathbf{z}_c^k \in \{0,1\}^M, k, c \in 1, \ldots, C$, in which each entry represents whether the corresponding proposal is an object of the k-th category. We make two assumptions for solving $\mathbf{z}_c^{k=c}$.

- For image $\mathbf{I}$ with $y_c = 1$, at least one proposal in $\mathbf{B}_c$ belongs to the c-th category, i.e., $\mathbf{1}^\top \cdot \mathbf{z}_c^{k=c} \geq 1$ where $\mathbf{1}$ is an M-dimensional all-one vector.
- For image $\mathbf{I}'$ with $y_c = 0$, none of proposals in $\mathbf{B}'_{c'\neq c}$ belongs to the c-th category, i.e., $\mathbf{1}^\top \cdot \mathbf{z}_{c'\neq c}^{k=c} = 0$.

Under the two assumptions, we can treat each image with $y_c = 1$ as a positive bag and treat each image with $y_c = 0$ as a negative bag. We then cast the task of solving $\mathbf{z}_c^{k=c}$ as an MIL problem. Note that multiple positive instances can be collected according to the scores of the MIL classifier for each class.

We use the smoothed hinge loss function in [36]. Note that the initialization step in [36] is carried out via a submodular clustering method from the initial object proposals. The noisy collection of proposals limits the performance of clustering process. Also, the initialization step is time-consuming as the similarity measures among all the proposals in all the images need to be computed. Our class-specific proposals not only help filter the object proposal collection but also reduce the training time of MIL.

4.2. Object detector learning

In this step, we aim at adapting the network from multi-label image classification for object detection. We jointly train the detection network with C object classes and a background class instead of training each object detector independently. Similar to [13], we replace the $2C$-dimensional classification layer (for image-level classification) with a randomly initialized (C+1)-dimensional classification layer (for instance-level classification with C object classes and background). We take the top-scoring proposals given by MIL as positive samples for each object category. We collect background samples from object proposals that have a maximum IoU $\in [0.1, 0.5)$ overlap with the mined object proposals by MIL. For data augmentation, we also treat all the proposals that have IoU ≥ 0.5 overlap with a mined object as positive samples.

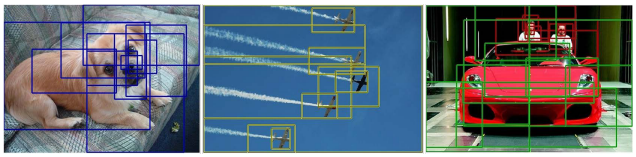

Figure 5. Examples of the mined object proposals using the mask-out strategy. We show top 10 proposals for each category (different colors indicate mined proposals for different categories). Note that the mined object proposals are class-specific.

Given a test image, we first generate object proposals using Edge Boxes [46] and use the adapted detection network to score each proposal. We then rank all the proposals and use non-maximum suppression to obtain final detections.

5. Experiments

5.1. Implementation details

For multi-label image classification training, we use the AlexNet [21] as our base CNN model, initialized with the parameters pre-trained on ImageNet dataset. We train the network with SGD at a learning rate of 0.001 for 10,000 mini-batch iterations. We set the size of mini-batch to 500. For class-specific proposal mining, we use Edge Boxes [46] to generate 2,000 initial object proposals and select top M=50 proposals as the input for multiple instance learning. For object detector training, we use AlexNet [21] and VGGNet [32] as our base models. Similar to Fast-RCNN [13], we set the maximum number iterations to 40k.

We implement our network using Caffe [19]. For the PASCAL VOC 2007 *trainval* set, fine-tuning the AlexNet for classification and detection adaptation takes about 10 hours and 1 hours with a Tesla K40 GPU, respectively. With our mined class-specific proposals, it takes about 3 hours to mine confident object samples on PC wBazzani:WACV16ith a 4.0 GHz Intel i7 CPU and 16 GB memory. The source code, as well as the pre-trained models, are available at the project webpage[1].

5.2. Datasets and evaluation metrics

Datasets. We extensively evaluate the proposed method on the PASCAL VOC 2007, 2010, 2012 datesets [10, 9] and ILSVRC 2013 detection dataset [6, 28]. For VOC 2007, we

[1] https://sites.google.com/site/lidonggg930/wsl

use both *train* and *val* splits as the training set and *test* split as our test set. For VOC 2010 and 2012, we use *train* split as the training set and *val* split as the test set. For the ILSVRC detection dataset, we follow the RCNN [14] in splitting the *val* data into val_1 and val_2. We use val_1 for training object detectors and val_2 for validating the localization performance. Note that we do not use any instance-level annotations (i.e., object bounding boxes) for training in all the datasets.

Evaluation metrics. We use two metrics to evaluate localization performance. First, we compute the fraction of positive training images in which we obtain correct localization (CorLoc) [7]. Second, we measure the performance of object detectors using average precision (AP) in the test set. For both metrics, we consider that a bounding box is correct if it has an intersection-over-union (IoU) ratio of at least 50% with a ground-truth object instance annotation.

5.3. Comparison to the state-of-the-art

We compare the proposed algorithm with state-of-the-art methods for weakly supervised object localization, including the MIL-based methods [33, 4, 36, 37, 2, 3], topic model [30], and latent category learning [41]. For fair comparisons, we do not include methods that use strong labels for training.

Table 1 shows performance comparison in terms of CorLoc on the PASCAL VOC 2007 *trainval* set. Our method achieves 52.4% of average CorLoc for all the 20 categories, outperforming all the state-of-the-art algorithms. Compared to the MIL-based approaches [33, 3, 4], we achieve significant improvements by 10 to 20 points. While these approaches use sophisticated model initialization or optimization strategies for improving MIL, the inevitable noise in the initial collection of category-independent proposals limits the performance of trained object detectors during MIL iterations. Compared to the topic model [30], we incorporate inter-class relations by jointly training CNN with all object classes and background class while they rely on handcrafted features. Wang et al. [41] use a pre-trained CNN for feature extraction. In contrast, we learn feature representations with our classification and detection adaptation, boosting the performance of CorLoc by 3.9 points.

Table 2 shows the detection average precision (AP) performance on the PASCAL VOC 2007 *test* set. Our method achieves 39.5% mAP, outperforming the state-of-the-art approaches by 8 points. Our method using the AlexNet achieves comparable performance with the second best method [41], 31.6% (ours) vs. 31.0% [41]. Most of existing methods [41, 3, 2, 36, 37] use pre-trained networks to extract features for object detector learning and do not fine-tune the network. In contrast, we progressively adapt the network from whole-image classification to object detection. Such domain adaptation helps learn better object detectors from weakly annotated data. Unlike previous work that relies on noisy and class-independent proposals to select object candidates, we mine purer and class-specific proposals for MIL training, which can discard background clutters and confusion with similar objects.

Table 3 shows our detection performance in terms of mean average precision on the PASCAL VOC 2010 and 2012 and ILSVRC 2013 datasets[2]. Using the VGGNet, our method achieves better localization performance. We include the full results in the supplementary materials.

5.4. Ablation studies

To quantify the relative contribution of each step, we examine the performance of our approach using different configurations.

- OM: Using mask-out strategy to mine top M=50 class-specific object proposals.
- MIL: Using MIL to mine confident objects.
- FT: Using the mined object candidates to fine-tune the detection network.

The last four rows of Table 1 show our CorLoc performance on the PASCAL VOC 2007 *trainval* set. We achieve average CorLoc of 31.8% by directly using top-ranked class-specific object proposals. Using MIL for selecting confident objects, we obtain 41.2% with around 10 points improvement. The result demonstrates that MIL iterations help to select better object proposals. The performance boost comes from: (1) the mined object proposals are less noisy and can discard background clutters, and (2) the mined object proposals are class-specific and can discard confusion with similar objects. Furthermore, adding detection network fine-tuning, we obtain 49.8% performance using the AlexNet and 52.4% using a deeper VGGNet [32]. Such network training further boosts the performance by another 10 points. In detection adaptation, we collect confident object proposals and use them to train all the layers. This fine-tuning step helps turn image classifiers to object detectors for modeling object appearance.

The last five rows of Table 2 show our detection AP performance on the PASCAL VOC 2007 *test* set. We refer Song et al. [36] as our MIL baseline. A straightforward approach to train detector uses proposals selected by MIL. However, the simple combination only gives marginal performance improvement from 22.7% to 23.0% because the selected proposals by MIL are too noisy for training object detection network effectively without object mining. Using the top-ranked object proposals based on the adapted classification network, we achieve significant improvement from 23.0% to 31.0%, highlighting the importance of progressive adaptation. Using a deeper network VGGNet [32], we can achieve a large improvement from 26.2% to 39.5%. In ad-

[2]The result of Cinbis et al. [4] is obtained on the VOC 2010 *test* set. The result of Wang et al. [41] is obtained on the ILSVRC 2013 *val* set.

Table 1. Quantitative comparison in terms of correct localization (CorLoc) on the PASCAL VOC 2007 *trainval* set.

Methods	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	Avg.
Siva et al. [33]	45.8	21.8	30.9	20.4	5.3	37.6	40.8	51.6	7.0	29.8	**27.5**	41.3	41.8	47.3	24.1	12.2	28.1	32.8	48.7	9.4	30.2
Shi et al. [30]	67.3	54.4	34.3	17.8	1.3	46.6	60.7	**68.9**	2.5	32.4	16.2	**58.9**	51.5	64.6	18.2	3.1	20.9	34.7	**63.4**	5.9	36.2
Cinbis et al. [4]	56.6	58.3	28.4	20.7	6.8	54.9	69.1	20.8	9.2	50.5	10.2	29.0	58.0	64.9	**36.7**	18.7	56.5	13.2	54.9	59.4	38.8
Bilen et al. [3]	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. [41]	**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
OM	50.4	30	34.6	18.2	6.2	39.3	42.2	57.3	10.8	29.8	20.5	41.8	43.2	51.8	24.7	20.8	29.2	26.6	45.6	12.5	31.8
OM + MIL	64.3	54.3	42.7	22.7	34.4	58.1	74.3	36.2	24.3	50.4	11.0	29.2	50.5	66.1	11.3	42.9	39.6	18.3	54.0	39.8	41.2
OM + MIL + FT-AlexNet	77.3	62.6	53.3	**41.4**	28.7	58.6	76.2	61.1	24.5	59.6	18.0	49.9	56.8	71.4	20.9	44.5	59.4	22.3	60.9	48.8	49.8
OM + MIL + FT-VGGNet	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**

Table 2. Quantitative comparison in terms of detection average precision (AP) on the PASCAL VOC 2007 *test* set.

Methods	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
Cinbis et al. [4]	35.8	40.6	8.1	7.6	3.1	35.9	41.8	16.8	1.4	23.0	4.9	14.1	31.9	41.9	**19.3**	11.1	27.6	12.1	31.0	40.6	22.4
Song et al. [36]	27.6	41.9	19.7	9.1	10.4	35.8	39.1	33.6	0.6	20.9	10.0	27.7	29.4	39.2	9.1	19.3	20.5	17.1	35.6	7.1	22.7
Song et al. [37]	36.3	47.6	23.3	12.3	11.1	36.0	46.6	25.4	0.7	23.5	12.5	23.5	27.9	40.9	14.8	19.2	24.2	17.1	37.7	11.6	24.6
Bilen et al. [2]	42.2	43.9	23.1	9.2	12.5	44.9	45.1	24.9	8.3	24.0	13.9	18.6	31.6	43.6	7.6	20.9	26.6	20.6	35.9	29.6	26.4
Bilen et al. [3]	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. [41]	48.9	42.3	26.1	11.3	11.9	41.3	40.9	34.7	10.8	34.7	18.8	34.4	35.4	52.7	19.1	17.4	35.9	**33.3**	34.8	**46.5**	31.6
OM + MIL	37.2	35.7	25.8	13.8	12.7	36.2	42.4	22.3	14.3	24.2	9.4	13.1	27.9	38.9	3.7	18.7	20.1	16.3	36.1	18.4	23.4
OM + FT-AlexNet	30.4	22.4	15.0	3.5	2.8	26.6	27.3	46.8	0.8	10.8	13.1	34.7	35.8	38.7	12.6	8.4	8.8	12.8	33.6	4.6	19.5
MIL + FT-AlexNet	17.5	50.2	22.5	4.0	9.9	38.8	48.7	39.3	0.3	22.1	10.1	19.8	22.4	49.9	3.4	15.5	32.1	10.8	40.0	1.9	23.0
OM + MIL + FT-AlexNet	49.7	33.6	30.8	19.9	13	40.5	54.3	37.4	14.8	39.8	9.4	28.8	38.1	49.8	14.5	24.0	27.1	12.1	42.3	39.7	31.0
OM+ FT-VGGNet	30.4	25.3	11.1	6.3	1.5	31.3	29.4	**49.1**	1.0	10.6	12.6	**42.0**	38.7	36.7	12.8	10.8	10.3	10.3	34.1	5.0	20.5
MIL + FT-VGGNet	25.6	**58.5**	25.3	1.8	11.7	43.5	53.4	35.7	0.2	32.3	10.7	19.3	32.8	56.5	1.8	15.6	37.3	16.0	43.6	2.9	26.2
OM + MIL + FT-VGGNet	**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**

Table 3. Object detection performance (mAP) on the PASCAL VOC 2010 and 2012 and ILSVRC 2013 datasets.

Methods	VOC 2010	VOC 2012	ILSVRC 2013
Cinbis et al. [4]	18.5	-	-
Wang et al. [41]	-	-	6.0
OM + MIL + FT-AlexNet	21.4	22.4	7.7
OM + MIL + FT-VGGNet	30.7	29.1	10.8

Table 4. Different mask-out strategies in terms of average correct localization from top M proposals.

Mask-out strategy	M=1	M=10	M=50	M=100
In-Out	31.8	**73.8**	**82.9**	**84.2**
Whole-Out	29.6	64.9	76.0	78.5
In	**32.7**	71.0	79.9	81.8

dition, we evaluate the performance using the best proposal (M=1) mined by the mask-out strategy for detection adaptation. The OM+FT method achieves 19.5% mAP using AlexNet and 20.5% using VGGNet. Without the MIL step, the results are poor due to noisy training samples. These experimental results validate the importance of the progressive adaptation steps proposed in this work.

Table 4 shows results using different mask-out strategies. Similar to the top 5 error evaluation for the ImageNet classification protocol, we compute the percentage of positive images in which an object is correctly located by at least one from top M proposals. When M=1, this metric reduces to the commonly used CorLoc. These results show our contrastive score *In-Out* strategy outperforms *Whole-Out*. Only using classification score of the region itself can also collect good proposals because classification adaptation step trains the network to be sensitive to object categories of target datasets. As the classification network is fine-tuned using the whole image, the mask-out image provides additional discriminative power for ranking the object proposals. Considering the trade-off between recall and precision, we set M=50 throughout the experiments.

5.5. Error analysis

In Figure 7, we apply the detector error analysis tool from Hoiem et al. [18] to analyze errors of our detector. Comparing the first and third columns, we achieve significant improvement of localization performance by detection adaptation. Fine-tuning the network for object-level detection helps learn discriminative appearance model for object categories, particularly for animals and furniture classes. Comparing the second and third columns, the importance of class-specific proposal mining step is clear. We attribute the performance boost to the classification adaptation that fine-tunes the network from 1000-way single-label classification (source) to 20-way multi-label classification task (target).

From the error analysis plots, the majority of errors comes from inaccurate localization. We show a few sample results in Figure 8. Our model often detects the correct category of an object instance but fails to predict a sufficiently tight bounding box, e.g., IoU $\in [0.1, 0.5)$. For example, we may detect a human face and a partial train and claim to detect a person or a train. The error analysis suggests that the learned model makes sensible errors. We believe that we can further improve the performance of our model

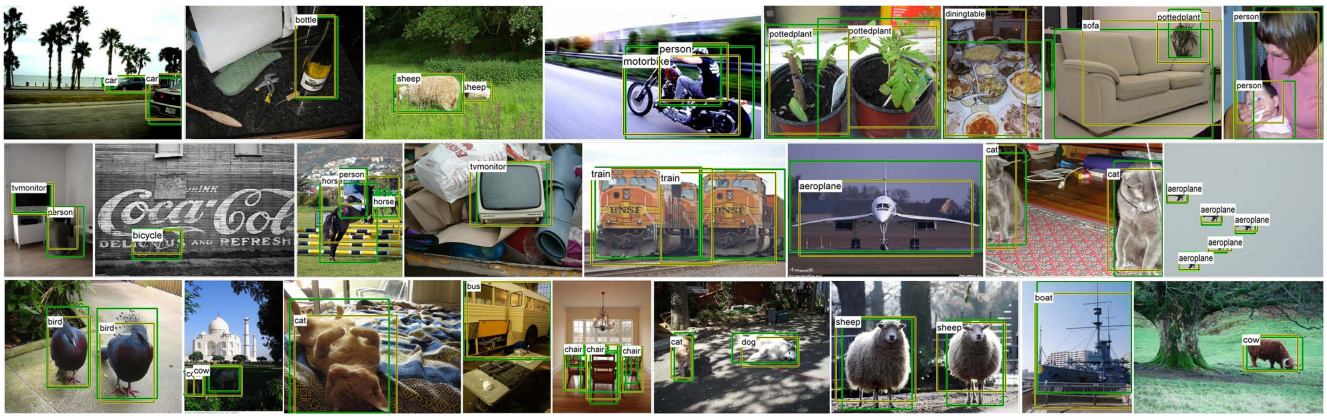

Figure 6. Sample detection results. Green boxes indicate ground-truth instance annotation. Yellow boxes indicate correction detections (with IoU ≥ 0.5). For all the testing results, we set threshold of detection as 0.8 and use NMS to remove duplicate detections.

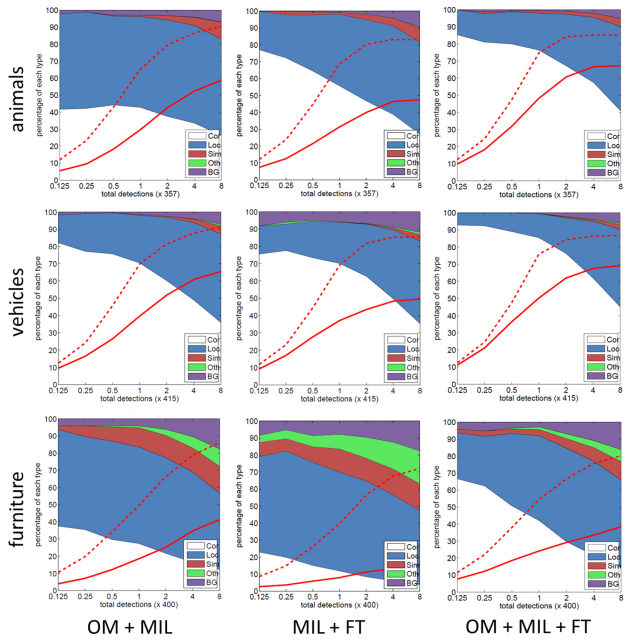

Figure 7. Detector error analysis. The detections are categorized into five types of correct detection (Cor), false positives due to poor localization (Loc), confusion with similar objects (Sim), confusion with other VOC objects (Oth), and confusion with background (BG). Each plot shows types of detection as top detections increase. Line plots show recall as function of the number of objects by IoU ≥ 0.5 (solid) and IoU ≥ 0.1 (dash). The results of "MIL+FT" and "OM+MIL+FT" are obtained using the VGGNet.

by incorporating techniques for addressing the inaccurate localization issues [5, 45].

6. Conclusion

We present a progressive domain adaptation approach to tackle the weakly supervised object localization problem. In classification adaptation, we transfer the classifiers from source to target domains using a multi-label loss function for training a multi-label classification network. In detection adaptation, we transfer adapted classifiers to object detectors. We first use a mask-out strategy to generate class-specific object proposals and apply MIL to mine confident candidates. We then use the selected object proposals to fine-tune all the layers for object detection. Experimental results demonstrate that our algorithm significantly outperforms the state-of-the-art methods. We achieve 39.5% mAP on VOC 2007, surpassing the second best approach by 8 points.

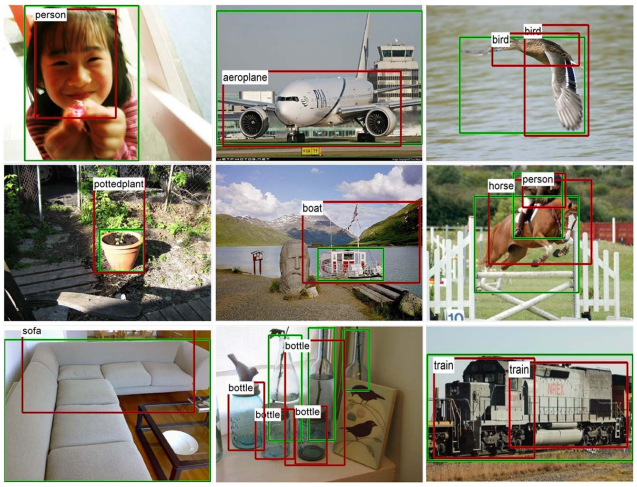

Figure 8. Sample results of detection errors due to imprecise localization.

Acknowledgement. We thank Xu Zhang, Jingchun Cheng and Yali Zhao for helpful discussions. This work is supported in part by the National Science and Technology Support Program under Grant No.2013BAK02B04 and the Initiative Scientific Research Program of Ministry of Education under Grant No.20141081253. This work is also supported in part to Dr. Ming-Hsuan Yang by NSF CAREER Grant (1149783) and NSF IIS Grant (1152576).

References

[1] L. Bazzani, A. Bergamo, D. Anguelov, and L. Torresani. Self-taught object localization with deep networks. In *WACV*, 2016. 3, 4, 5

[2] H. Bilen, M. Pedersoli, and T. Tuytelaars. Weakly supervised object detection with posterior regularization. In *BMVC*, 2014. 1, 2, 3, 6, 7

[3] H. Bilen, M. Pedersoli, and T. Tuytelaars. Weakly supervised object detection with convex clustering. In *CVPR*, 2015. 1, 2, 3, 6, 7

[4] R. G. Cinbis, J. Verbeek, and C. Schmid. Multi-fold mil training for weakly supervised object localization. In *CVPR*, 2014. 1, 2, 6, 7

[5] Q. Dai and D. Hoiem. Learning to localize detected objects. In *CVPR*, 2012. 8

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 4, 5

[7] T. Deselaers, B. Alexe, and V. Ferrari. Weakly supervised localization and learning with generic knowledge. *IJCV*, 100(3):275–293, 2012. 6

[8] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. Efros. What makes paris look like paris? *ACM Transactions on Graphics*, 31(4), 2012. 1

[9] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111(1):98–136, 2014. 5

[10] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010. 5

[11] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 32(9):1627–1645, 2010. 1

[12] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008. 3

[13] R. Girshick. Fast r-cnn. In *ICCV*, 2015. 2, 3, 5

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 1, 2, 3, 6

[15] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV*, 2009. 1

[16] J. Hoffman, S. Guadarrama, E. S. Tzeng, R. Hu, J. Donahue, R. Girshick, T. Darrell, and K. Saenko. Lsda: Large scale detection through adaptation. In *NIPS*, 2014. 1, 3

[17] J. Hoffman, D. Pathak, T. Darrell, and K. Saenko. Detector discovery in the wild: Joint multiple instance and representation learning. In *CVPR*, 2015. 1, 3

[18] D. Hoiem, Y. Chodpathumwan, and Q. Dai. Diagnosing error in object detectors. In *ECCV*, 2012. 7

[19] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*, 2014. 5

[20] L. Jiang, D. Meng, T. Mitamura, and A. G. Hauptmann. Easy samples first: Self-paced reranking for zero-example multimedia search. In *ACM MM*, 2014. 3

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 2, 3, 5

[22] Y. J. Lee and K. Grauman. Learning the easy things first: Self-paced visual category discovery. In *CVPR*, 2011. 3

[23] Q. Li, J. Wu, and Z. Tu. Harvesting mid-level visual concepts from large-scale internet images. In *CVPR*, 2013. 1

[24] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free?–weakly-supervised learning with convolutional neural networks. In *CVPR*, 2015. 3

[25] W. Ouyang, P. Luo, X. Zeng, S. Qiu, Y. Tian, H. Li, S. Yang, Z. Wang, Y. Xiong, C. Qian, et al. Deepid-net: multi-stage and deformable deep convolutional neural networks for object detection. In *CVPR*, 2015. 2, 3

[26] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich. Training deep neural networks on noisy labels with bootstrapping. In *ICLR*, 2014. 3

[27] M. Rochan and Y. Wang. Weakly supervised localization of novel objects using appearance transfer. In *CVPR*, 2015. 1, 3

[28] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3):211–252, 2015. 5

[29] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *ICLR*, 2014. 2, 3

[30] Z. Shi, T. M. Hospedales, and T. Xiang. Bayesian joint topic modelling for weakly supervised object localisation. In *ICCV*, 2013. 6, 7

[31] Z. Shi, P. Siva, T. Xiang, and Q. Mary. Transfer learning by ranking for weakly supervised object annotation. In *BMVC*, 2012. 1, 3

[32] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 2, 5, 6

[33] P. Siva, C. Russell, and T. Xiang. In defence of negative mining for annotating weakly labelled data. In *ECCV*, 2012. 1, 2, 6, 7

[34] P. Siva, C. Russell, T. Xiang, and L. Agapito. Looking beyond the image: Unsupervised learning for object saliency and detection. In *CVPR*, 2013. 1, 2

[35] P. Siva and T. Xiang. Weakly supervised object detector learning with model drift detection. In *ICCV*, 2011. 1, 2

[36] H. O. Song, R. Girshick, S. Jegelka, J. Mairal, Z. Harchaoui, and T. Darrell. On learning to localize objects with minimal supervision. In *ICML*, 2014. 1, 2, 3, 5, 6, 7

[37] H. O. Song, Y. J. Lee, S. Jegelka, and T. Darrell. Weakly-supervised discovery of visual pattern configurations. In *NIPS*, 2014. 1, 2, 3, 6, 7

[38] S. Sukhbaatar and R. Fergus. Learning from noisy labels with deep neural networks. In *ICLR*, 2015. 3

[39] J. S. Supancic and D. Ramanan. Self-paced learning for long-term tracking. In *CVPR*, 2013. 3

[40] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 2

[41] C. Wang, W. Ren, K. Huang, and T. Tan. Weakly supervised object localization with latent category learning. In *ECCV*, 2014. 1, 3, 6, 7

[42] Y. Wei, W. Xia, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan. Cnn: Single-label to multi-label. *arXiv preprint arXiv:1406.5726*, 2014. 2

[43] J. Wu, Y. Yu, C. Huang, and K. Yu. Deep multiple instance learning for image classification and auto-annotation. In *CVPR*, 2015. 3

[44] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014. 4

[45] Y. Zhang, K. Sohn, R. Villegas, G. Pan, and H. Lee. Improving object detection with deep convolutional networks via bayesian optimization and structured prediction. In *CVPR*, 2015. 8

[46] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014. 4, 5