Dissimilarity Coefficient based Weakly Supervised Object Detection

Aditya Arun
CVIT, KCIS
IIIT Hyderabad

C.V. Jawahar
CVIT, KCIS
IIIT Hyderabad

M. Pawan Kumar
University of Oxford,
The Alan Turing Institute

Abstract

We consider the problem of weakly supervised object detection, where the training samples are annotated using only image-level labels that indicate the presence or absence of an object category. In order to model the uncertainty in the location of the objects, we employ a dissimilarity coefficient based probabilistic learning objective. The learning objective minimizes the difference between an annotation agnostic prediction distribution and an annotation aware conditional distribution. The main computational challenge is the complex nature of the conditional distribution, which consists of terms over hundreds or thousands of variables. The complexity of the conditional distribution rules out the possibility of explicitly modeling it. Instead, we exploit the fact that deep learning frameworks rely on stochastic optimization. This allows us to use a state of the art discrete generative model that can provide annotation consistent samples from the conditional distribution. Extensive experiments on PASCAL VOC 2007 and 2012 data sets demonstrate the efficacy of our proposed approach.

1. Introduction

Object detection requires us to localize all the instances of an object category of interest in a given image. In recent years, significant advances in speed and accuracy have been achieved by detection frameworks based on Convolutional Neural Networks (CNNs) [7, 13, 14, 16, 24, 26, 27]. Most of the existing methods require a strongly supervised data set, where each image is labeled with the ground-truth bounding boxes of all the object instances. Given the high cost of obtaining such detailed annotations, researchers have recently started exploring the weakly supervised object detection (WSOD) problem [3, 9, 18, 21, 22, 23, 33, 34, 39, 40, 41, 42]. The goal of WSOD is to learn an accurate detector using training samples that are annotated with image-level labels (which indicate the presence of an object category).

Given the wide availability of image-level labels, WSOD offers a cost-effective and highly scalable learning paradigm. However, this comes at the cost of introducing uncertainty in the location of the object instances during training. For example, consider the task of detecting a car. Given a training image annotated to indicate the presence of a car, we are still faced with the challenge of identifying the bounding box for the car.

In order to effectively model uncertainty in weakly supervised learning, Kumar *et al.* [20] proposed a probabilistic framework that models two distributions: (i) a conditional distribution, which represents the probability of an output conditioned on the given annotation during training; and (ii) a prediction distribution which represents the probability of an output at test time. The parameters of the two distributions are estimated jointly by minimizing the dissimilarity coefficient [25], which measures the distance between any two distributions using a task specific loss function.

The aforementioned dissimilarity coefficient based framework has provided promising results in domains where the conditional distribution is simple to model (that is, consists of terms that depend on a few variables at a time) [1, 20]. However, WSOD presents a more challenging scenario due to the complexity of the underlying conditional distribution. Specifically, given the hundreds or even thousands of bounding box proposals for an image, the annotation constraint imposes a term over all of these bounding box proposals such that at least one of them corresponds to the given image-level label. This leads to a challenging scenario where the distribution is not factorizable over the bounding box proposals. While previous works have approximated this uncertainty as a fully factorized distribution for computational efficiency, we argue that such a choice leads to poor accuracy.

To overcome the difficulty of a complex conditional distribution, we make the key observation that deep learning relies on stochastic optimization. Therefore, we do not need to explicitly model this complex distribution but simply estimate the distribution using samples. This observation opens the door to the use of state-of-the-art deep generative models such as the Discrete DISCO Net [4, 5].

We test the efficacy of our approach on the challenging PASCAL VOC 2007 and 2012 data sets. To generate

the weakly supervised data sets, we use the image-level labels, discarding the bounding box annotations. We achieve 53.6% detection AP on PASCAL VOC 2007 and 49.5% detection AP on PASCAL VOC 2012 data set, significantly improving the state-of-the-art by 1.5% on both data sets.

To summarize, we make the following contributions.

- Efficiently model the complex non-factorizable, annotation aware conditional distribution using the deep generative model, the Discrete DISCO Net.
- Empirically show the importance of modeling the uncertainty in the annotations in a single unified probabilistic learning objective, the dissimilarity coefficient.
- State-of-the art performance for the task of WSOD on challenging PASCAL VOC 2007 and 2012 data sets.

2. Related Work

Conventional methods often treat WSOD as a Multiple Instance Learning (MIL) problem [10] by representing each image as a bag of instances (that is, putative bounding boxes) [2, 6, 31, 36, 38]. The learning procedure alternates between training an object classifier and selecting the most confident positive instances. However, these methods are susceptible to poor initialization. To address this, different strategies have been developed, which aim to improve the initialization [19, 29, 30, 31], regularize the model with extra cues [2, 6], or relax the MIL constraint [38] to make the objective differentiable. These hard-MIL based methods have demonstrated their effectiveness, specially when CNN features are used to represent object proposals [6]. However, these models are not end to end trainable and also do not explicitly model the uncertainty.

A more interesting line of work is to integrate MIL strategy as deep networks such that they are end to end trainable [3, 9, 12, 33, 34, 37, 40, 41, 42]. In their work, Bilen *et al.* [3] proposed a smoothed version of MIL that softly labels object proposals instead of choosing the highest scoring ones. Building on this soft-MIL based approach, Diba *et al.* [9] integrate the MIL strategy with better bounding box proposals into an end-to-end cascaded deep network. Tang *et al.* [33] refine the prediction iteratively through multi-stage instance classifier. Zhang *et al.* [40] add curriculum learning using the MIL framework. As we shall see, our formulation brings out the curriculum learning naturally during training. Other end-to-end trainable frameworks for WSOD employ domain adaptation [22, 31], expectation-maximization algorithm [18, 39] and saliency based methods [21]. Although these methods are end to end trainable, they not only model a single distribution for two related tasks, but also model the complex distribution with a fully factorized one. This makes these approach sub-optimal as what we truly want is to model a distribution which enforces at least one bounding box proposals corresponding to the image-level label.

There have been attempts to further improve the performance of the weakly supervised detectors by combining them with the strongly supervised detectors. Typically, the predicted instances from a trained weakly supervised detector are treated as a pseudo-strong label to train a strongly supervised network [12, 22, 33, 34, 40, 41, 42]. However, there is only a unidirectional connection between the two detectors. In their work, Wang *et al.* [37] train a weakly and strongly supervised model jointly, in a collaborative manner. This is similar in spirit to ours in using two distributions. However, they model their weakly supervised detector with a fully factorized distribution. The improvement in results reported by these papers advocates the importance of modeling two separate distributions. In this work, we explicitly define the two distributions employed during training and test time and jointly train them by minimizing the dissimilarity coefficient [25] based objective function.

3. Model

3.1. Notation

We denote an input image as $\mathbf{x} \in \mathbb{R}^{(H \times W \times 3)}$, where H and W are the height and the width of the image respectively. For the sake of simplifying the subsequent description of our approach, we assume that we have extracted B bounding box proposals from each image. In our experiments, we use Selective Search [35]. Each bounding box proposal, $b^{(i)}$, can belong to one of $C+1$ categories from the set $\{0, 1, \ldots, C\}$, where category 0 is background, and categories $\{1, \ldots, C\}$ are object classes.

We denote an image-level label by $\mathbf{a} \in \{0,1\}^C$, where $\mathbf{a}^{(j)} = 1$ if image $\mathbf{x}$ contains the j-th object. Furthermore, we denote the unknown bounding box labels by $\mathbf{y} \in \{0, \ldots, C\}^B$, where $\mathbf{y}^{(i)} = j$ if the i-th bounding box $b^{(i)}$ is of the j-th category. A weakly supervised data set $\mathcal{W} = \{(\mathbf{x}_i, \mathbf{a}_i) | i = 1, \ldots, N\}$ contains N pairs of images $\mathbf{x}_i$ and their corresponding image-level labels $\mathbf{a}_i$.

3.2. Probabilistic Modeling

Given a weakly supervised data set $\mathcal{W}$, we wish to learn an object detector that can predict the bounding box labels $\mathbf{y}$ of a previously unseen image. Due to the uncertainty inherent in this task, we advocate the use of a probabilistic formulation. Following [1, 20], we define two distributions. The first one is the prediction distribution $\Pr_p(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}_p)$, which models the probability of the bounding box labels $\mathbf{y}$ given an input image $\mathbf{x}$. Here $\boldsymbol{\theta}_p$ are the parameters of the distribution. As the name suggest, this distribution is used to make the prediction at test time.

In addition to the prediction distribution, we also construct a conditional distribution $\Pr_c(\mathbf{y}|\mathbf{x}, \mathbf{a}; \boldsymbol{\theta}_c)$, which

models the probability of the bounding box labels $\mathbf{y}$ given the input image $\mathbf{x}$ and its image-level annotations $\mathbf{a}$. Here $\boldsymbol{\theta}_c$ are the parameters of the distribution. The conditional distribution contains additional information, namely the presence of foreground objects in each image. Thus, we can expect it to provide better predictions for the bounding box labels $\mathbf{y}$. We will use this property during training in order to learn an accurate prediction distribution using the conditional distribution. The details on the modeling of the two distributions are discussed below.

3.2.1 Prediction Distribution

The task of the prediction distribution is to accurately model the probability of the bounding box labels given the input image. Taking inspiration from the supervised models [13, 14, 27], we assume independence between the probability of the output for each bounding box proposal. Therefore, the overall distribution for an image equals the product of the probabilities of each proposal,

$$\Pr_p(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}_p) = \prod_{i=1}^{B} \Pr_p(\mathbf{y}^{(i)}|\mathbf{x};\boldsymbol{\theta}_p). \tag{1}$$

We model this distribution using the Fast-RCNN architecture [13] (see Figure 1(a)). As the prediction distribution is specified by a neural network, we henceforth refer to it as the *prediction net*. In this setting, the parameters of the distribution $\boldsymbol{\theta}_p$ are the weights of the prediction net.

3.2.2 Conditional Distribution

Given B bounding box proposals for an image $\mathbf{x}$ and the image-level label $\mathbf{a}$, the conditional distribution $\Pr_c(\mathbf{y}|\mathbf{x},\mathbf{a};\boldsymbol{\theta}_c)$ models the probability of bounding box labels $\mathbf{y}$ under the constraint that they are compatible with the annotation $\mathbf{a}$. Specifically, there exists at least one bounding box i such that $\mathbf{y}^{(i)} = j$, for every positive image-level label $\mathbf{a}^{(j)} = 1$.

Note that due to the requirement that the bounding box labels $\mathbf{y}$ are compatible with the annotation $\mathbf{a}$, the conditional distribution cannot be trivially decomposed over bounding box proposals. This is in stark contrast to the simple prediction net, which uses a fully factorized distribution. If one were to explicitly model the conditional distribution, then one would be required to compute its partition function during training, which would be prohibitively expensive. To alleviate this computational challenge, we make a key observation that in practice we only need access to a representative set of samples from the conditional distribution. This opens the door to the use of the recently proposed Discrete DISCO Net [4]. In what follows, we briefly describe Discrete DISCO Nets while highlighting their applicability to our framework.

Discrete DISCO Net: Discrete DISCO Net [4] is a deep probabilistic framework that implicitly represents a probability distribution over a discrete structured output space. The strength of the framework lies in the fact that it allows us to adapt a pointwise deep network (a network that provides a single pointwise prediction) to a probabilistic one by the introduction of noise.

In the context of our setting, consider the modified Fast-RCNN network in Figure 1(b) for the conditional distribution. Once again, as we are using a neural network, we will henceforth refer to it as the *conditional net*. The parameters of the conditional distribution $\boldsymbol{\theta}_c$ are the weights of the conditional net. The colored filters in the middle of the network represent the noise that is sampled from a uniform distribution. Each value of the noise filter $\mathbf{z}_k$ results in a different score function[1] $\mathcal{G}_k(\mathbf{y};\mathbf{x},\mathbf{z}_k,\boldsymbol{\theta}_c) \in \mathbb{R}^{B\times C}$. We generate K different score functions using K different noise samples. These score functions are then used to sample corresponding bounding box labels $\hat{\mathbf{y}}_c^k$ such that all ground truth labels are present in it. This enables us to generate samples from the underlying distribution encoded by the network parameters. Note that obtaining a single sample is as efficient as a simple forward pass through the network. By placing the filters sufficiently far away from the output layer of the network, we can learn a highly non-linear mapping from the uniform distribution (used to generate the noise filter) to the output distribution (used to generate bounding box labels).

Inference: For the input pair $(\mathbf{x},\mathbf{z}_k)$, the classification branch of the conditional net outputs a score function $\mathcal{G}_k(\mathbf{y};\mathbf{x},\mathbf{z}_k,\boldsymbol{\theta}_c)$, which is a $B \times C$ matrix. The (i,j)-th element of the matrix, denoted by $\mathcal{G}_k^{(i,j)}$, denotes the score of the bounding box i belonging to the category j. We will now redefine this score function such that it respects the constraints imposed by the annotation $\mathbf{a}$. In other words, for each category j such that $\mathbf{a}^{(j)} = 1$ there must exist at least one bounding box i in $\mathbf{y}$ such that $\mathbf{y}^{(i)} = j$. The joint score for all the bounding box labels $\mathbf{y}$ is given by,

$$S_k(\mathbf{y};\mathbf{x},\mathbf{z}_k,\boldsymbol{\theta}_c) = \sum_{i=1}^{B} \mathcal{G}_k(\mathbf{y}^{(i)};\mathbf{x},\mathbf{z}_k,\boldsymbol{\theta}_c) - H_k(\mathbf{y}), \tag{2}$$

where,

$$H_k(\mathbf{y}) = \begin{cases} 0 & \text{if } \forall j \in \{1,\dots,C\} \text{ s.t. } \mathbf{a}^{(j)} = 1, \\ & \exists i \in \{1,\dots,B\} \text{ s.t. } \mathbf{y}^{(i)} = j, \\ \infty & \text{otherwise.} \end{cases} \tag{3}$$

Given the scoring function in equation (2), we compute the k-th sample as

$$\hat{\mathbf{y}}_c^k = \arg\max_{y\in\mathcal{Y}} S_k(\mathbf{y};\mathbf{x},\mathbf{z}_k,\boldsymbol{\theta}_c). \tag{4}$$

[1]The use of score function in this paper should not be confused with the scoring rule theory, which is used to design the learning objective of DISCO Nets.

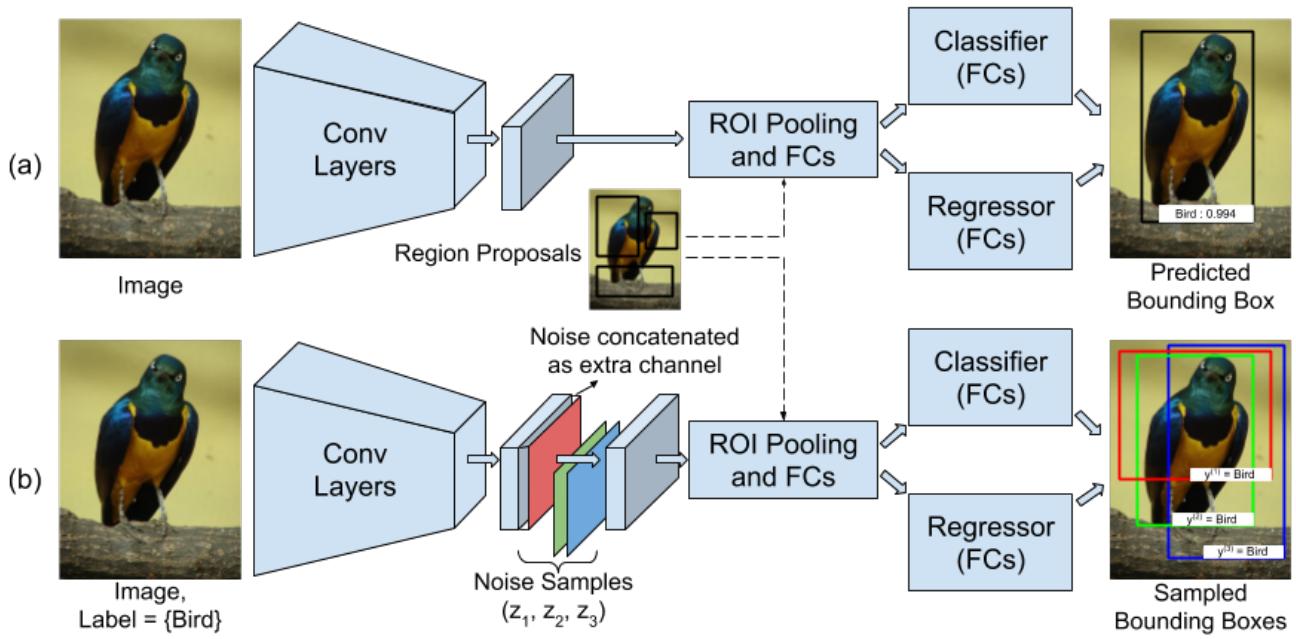

Figure 1. *The overall architecture. (a) Prediction Network: a standard Fast-RCNN architecture is used to model the prediction net. For an input image, bounding box proposals are generated from selective search [35]. Features from each of these proposals are computed by the region of interest (ROI) pooling layers, which are then passed through the classifier and regressor to predict the final bounding box. (b) Conditional Network: a modified Fast-RCNN architecture is used to model the conditional net. For a single input image* $\mathbf{x}$ *and three different noise samples* $\{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3\}$ *(represented as red, green and blue matrix), three different bounding boxes* $\{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{y}^{(3)}\}$ *are sampled for the given image-level label (bird in this example). Here the noise filter is concatenated as an extra channel to the final convolutional layer. For both the networks, the initial conv-layers are fixed during training. Best viewed in color.*

Note that in equation (4) the $\arg\max$ needs to be computed over the entire output space $\mathcal{Y}$. A naïve brute force algorithm for this would be computationally infeasible. However, by using the structure of the higher order term H_k, we can design an efficient yet exact algorithm for equation (4). Specifically, we assign each bounding box proposal i to its maximum scoring object class. If all the ground truth annotations $\mathbf{a}$ are not present in the generated bounding box labels, then we sample the bounding box which has the highest score corresponding to the foreground label, otherwise we sample all bounding boxes which satisfies the constraint.

4. Learning Objective

In order to estimate the parameters of the prediction and conditional distribution, $\boldsymbol{\theta}_p$ and $\boldsymbol{\theta}_c$, we define a unified probabilistic learning objective based on the dissimilarity coefficient [25]. To this end, we require a task specific loss function, which we define next.

4.1. Task Specific Loss Function

We define a loss function for object detection that decomposes over the bounding box proposals as follows:

$$\Delta(\mathbf{y}_1, \mathbf{y}_2) = \frac{1}{B}\sum_{i=1}^{B} \Delta(\mathbf{y}_1^{(i)}, \mathbf{y}_2^{(i)}). \tag{5}$$

Following the standard practice in most modern object detectors [17], $\Delta(\mathbf{y}_1^{(i)}, \mathbf{y}_2^{(i)})$ is further decomposed as a weighted combination of the classification loss and the localization loss. We use λ to denote the loss ratio (ratio of the weight of localization loss to the weight of classification loss). We use a simple $0-1$ loss as our classification loss Δ_{cls}, and $smoothL1$ [13] for our localization loss Δ_{loc}. Formally, the task specific loss is given by,

$$\Delta(\mathbf{y}_1^{(i)}, \mathbf{y}_2^{(i)}) = \Delta_{cls}(\mathbf{y}_1^{(i)}, \mathbf{y}_2^{(i)}) + \lambda\Delta_{loc}(b_1^{(i)}, b_2^{(i)}). \tag{6}$$

4.2. Objective Function

The task of both the prediction distribution and the conditional distribution is to predict the bounding box labels. Moreover, as the conditional distribution utilizes the extra information in the form of the image-level label, it is expected to provide more accurate predictions for the bounding box labels $\mathbf{y}$. Leveraging on the task similarity between the two distributions, we would like to bring the two distributions close to each other, so that the extra knowledge of the conditional distribution can be transferred to the prediction distribution. Taking inspiration from [1, 20], we design a joint learning objective that can minimize the dissimilarity coefficient [25] between the prediction distribution and conditional distribution. In what follows, we briefly describe

the concept of dissimilarity coefficient before applying it to our setting.

Dissimilarity Coefficient: The dissimilarity coefficient between any two distributions $\Pr_1(\cdot)$ and $\Pr_2(\cdot)$ is determined by measuring their diversities. The diversity of a distribution $\Pr_1(\cdot)$ and a distribution $\Pr_2(\cdot)$ is defined as the expected difference between their samples, where the difference is measured by a task-specific loss function $\Delta'(\cdot,\cdot)$. Formally, we define the diversity as,

$$DIV_{\Delta'}(\Pr_1,\Pr_2) = \mathbb{E}_{\mathbf{y}_1\sim\Pr_1(\cdot)}[\mathbb{E}_{\mathbf{y}_2\sim\Pr_2(\cdot)}[\Delta'(\mathbf{y}_1,\mathbf{y}_2)]]. \quad (7)$$

If the model correctly brings the two distribution close to each other, we could expect the diversity $DIV_{\Delta'}(\Pr_1,\Pr_2)$ to be small. Using this definition of diversity, the dissimilarity coefficient of $\Pr_1$ and $\Pr_2$ is given by,

$$\begin{aligned} DISC_{\Delta'}(\Pr_1,\Pr_2) =& DIV_{\Delta'}(\Pr_1,\Pr_2) \\ &- \gamma DIV_{\Delta'}(\Pr_2,\Pr_2) \\ &- (1-\gamma) DIV_{\Delta'}(\Pr_1,\Pr_1), \end{aligned} \quad (8)$$

where $\gamma \in [0,1]$. In other words, the dissimilarity coefficient between $\Pr_1$ and $\Pr_2$ is the difference between the diversity of $\Pr_1$ and $\Pr_2$, and a convex combination of their self-diversities. The self-diversity terms encourages the samples from each of the two distribution to be diverse, thus better representing the uncertainty of the task. In our experiments, we use $\gamma = 0.5$, which results in a symmetric dissimilarity coefficient between two distributions.

Learning Objective for Detection: Given the above definition of dissimilarity coefficient, we can now specify our learning objective for the task specific loss Δ tuned for object detection (6) as

$$\boldsymbol{\theta}_p^*, \boldsymbol{\theta}_c^* = \arg\min_{\boldsymbol{\theta}_p,\boldsymbol{\theta}_c} DISC_{\Delta}(\Pr_p(\boldsymbol{\theta}_p), \Pr_c(\boldsymbol{\theta}_c)), \quad (9)$$

where each of the diversity terms can be derived from equation (7). As discussed in Section 3.2, the conditional distribution is difficult to model directly. Therefore, the corresponding diversity terms are computed by stochastic estimators from K samples $\hat{\mathbf{y}}_c^k$ of the conditional net. Thus, each of the diversity terms can be written as[2]

$$\begin{aligned} &DIV_{\Delta}(\Pr_p,\Pr_c) \\ &= \frac{1}{BK}\sum_{i=1}^{B}\sum_{k=1}^{K}\sum_{\mathbf{y}_p^{(i)}} \Pr_p(\mathbf{y}_p^{(i)};\boldsymbol{\theta}_p)\Delta(\mathbf{y}_p^{(i)},\hat{\mathbf{y}}_c^{k,(i)}), \end{aligned} \quad (10)$$

$$\begin{aligned} &DIV_{\Delta}(\Pr_c,\Pr_c) \\ &= \frac{1}{K(K-1)B}\sum_{\substack{k,k'=1\\k'\neq k}}^{K}\sum_{i=1}^{B}\Delta(\hat{\mathbf{y}}_c^{k,(i)},\hat{\mathbf{y}'}_c^{k',(i)}), \end{aligned} \quad (11)$$

$$\begin{aligned} &DIV_{\Delta}(\Pr_p,\Pr_p) \\ &= \frac{1}{B}\sum_{i=1}^{B}\sum_{\mathbf{y}_p^{(i)}}\sum_{\mathbf{y}'^{(i)}_p} \Pr_p(\mathbf{y}_p^{(i)};\boldsymbol{\theta}_p)\Pr_p(\mathbf{y}'^{(i)}_p;\boldsymbol{\theta}_p)\Delta(\mathbf{y}_p^{(i)},\mathbf{y}'^{(i)}_p). \end{aligned} \quad (12)$$

Here, $DIV_{\Delta}(\Pr_p,\Pr_c)$ measures the diversity between the prediction net and the conditional net, which is the expected difference between the samples from the two distributions as measured by the task specific loss function Δ. Here $\Pr_p$ is explicitly modeled, hence the expectation of its sample can be computed easily. However, as $\Pr_c$ is not explicitly modeled, we compute the required expectation by drawing K samples from the distribution. Likewise, $DIV_{\Delta}(\Pr_c,\Pr_c)$ measures the self diversity of the conditional net. We draw K samples from the distribution to compute the required expectation. Also, the self diversity of the prediction net $DIV_{\Delta}(\Pr_p,\Pr_p)$ can be exactly computed as $\Pr_p$ is explicitly modeled.

5. Optimization

As we employ deep neural networks to model the two distributions, our objective function (9) is ideally suited to be minimized by stochastic gradient descent. While it may be possible to compute the gradients of both the networks simultaneously, in this work we use a simple coordinate descent optimization strategy. In more detail, the optimization proceeds by iteratively fixing the prediction network and learning the conditional network, followed by learning the prediction network for fixed conditional network.

The main advantage of using the iterative training strategy is that it results in an approach similar to the fully supervised learning of each network. This in turn allows us to readily use the algorithm developed in Fast-RCNN [13] and Discrete DISCO Net [4]. The outputs from the fixed network are treated as the pseudo ground truth bounding box labels for the other network. Furthermore, the iterative learning strategy also reduces the memory complexity of learning as only one network is trained at a time.

Figure 2 provides the visualization of the performance of the two networks over the different iterations of the iterative learning procedure for two difficult cases. In Columns 1 and 2, different category objects are present in the same image whereas, in columns 3 and 4, multiple instances of the same category is present. The estimated bounding box labels from the prediction net and those sampled from the conditional net for two images are depicted. For conditional

[2]Details in the supplementary material

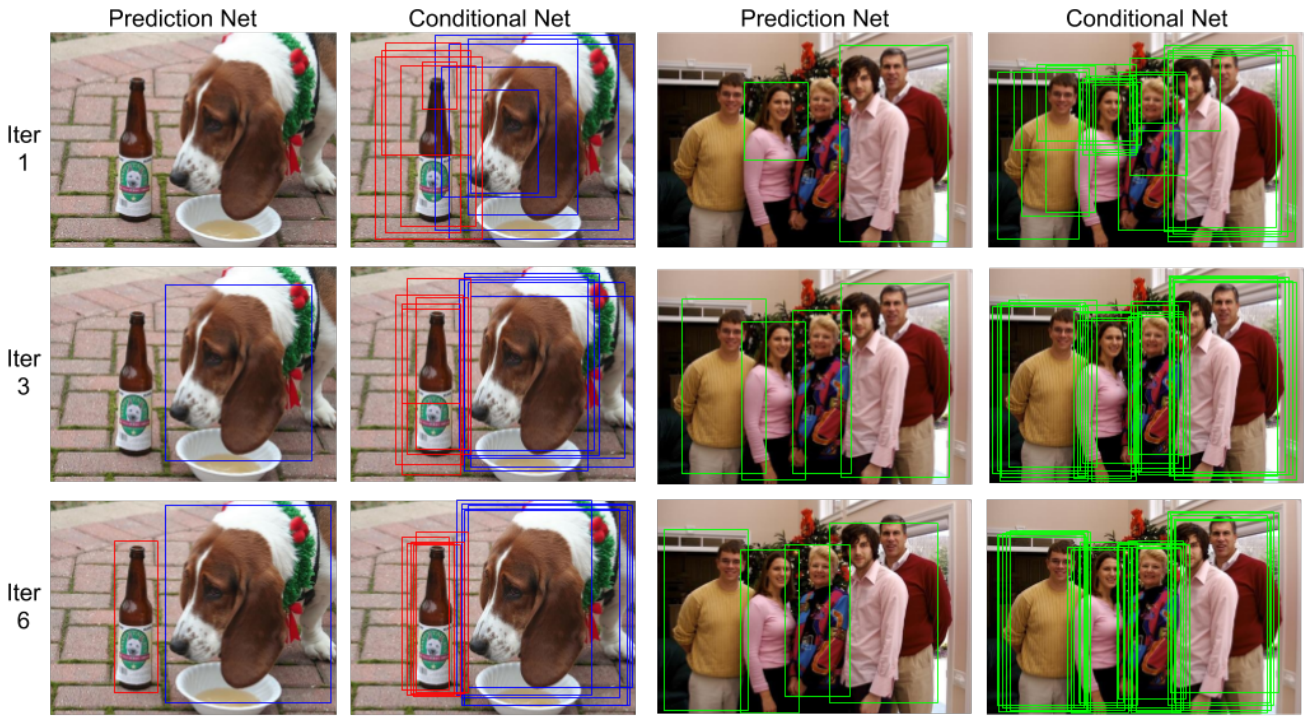

Figure 2. *Example of predictions of prediction net and conditional net. For prediction net, the visualization is after taking standard non maximal suppression using standard score threshold* = 0.7. *Column 1 and 3 are output of the prediction network while column 2 and 4 are output from the conditional network. Row 1 represents prediction of the two networks after first iteration and row 2 and 3 represents prediction of the two networks after third and sixth (final) iteration respectively. Each object class is represented by different colored bounding box, where green box represents the* person *category and red and blue represents the* bottle *and* dog *category respectively.*

net, we superimpose five different samples of bounding box labels. If all the samples agree with each other on bounding box labels, then the bounding boxes will have a high overlap, otherwise they will be scattered across the image. For visualization purposes only, a standard non maximal suppression (NMS) is applied with a score threshold of 0.7 on the output of the prediction net. However, note that the non maximal suppression is not used during training of the prediction net. The two steps of the iterative algorithm are described below in brief. For completeness, the details are provided in the supplementary material.

5.1. Optimization over Prediction Distribution

For a fixed set of parameters $\boldsymbol{\theta}_c$ of the conditional network, the learning objective of the prediction net corresponds to the following:

$$\boldsymbol{\theta}_p^* = \arg\min_{\boldsymbol{\theta}_p} DIV_{\Delta}(\mathrm{Pr}_p, \mathrm{Pr}_p) - (1-\gamma) DIV_{\Delta}(\mathrm{Pr}_p, \mathrm{Pr}_p). \tag{13}$$

Note that, due to the use of dissimilarity coefficient, the above objective differs slightly from the one used for Fast-RCNN [13]. However, importantly, it is still differentiable with respect to $\boldsymbol{\theta}_p$. Hence, the prediction net can be directly optimized via stochastic gradient descent.

In order to visualize the optimization of the prediction net, let us consider Figure 2. The first two columns show the bounding box labels from the prediction and the conditional nets for an image with single foreground object. As the image has a large foreground object with a clean background, both the prediction and the conditional nets have low uncertainty. This represents an easy case where the prediction net already has a high confidence for the bounding box labels in initial iterations, and therefore has little to gain from the conditional net. As expected, we see only a minor improvement in the predicted bounding box labels of the prediction net over the iterations.

The last two columns show bounding box labels from the prediction and conditional nets for a challenging image. The object *dog* presents moderate difficulty to our algorithm, where initially the prediction net is highly uncertain while the conditional net has low uncertainty. After few iterations, the information present in the conditional net is successfully transferred over to the prediction net. This is shown in last row of the third column where the prediction net does a reasonable job at estimating the bounding boxes.

The second object *bottle* in the image is a difficult exam-

ple because of its small scale. We observe high uncertainty in both the networks. In such cases the prediction and the conditional nets will reject the bounding box labels having high diversity. Moreover, the uncertainty in the prediction net also decreases by learning from other easier instances of the object present in the data set.

5.2. Optimization over Conditional Distribution

For a fixed set of parameters $\boldsymbol{\theta}_p$ of the prediction network, the learning objective for the conditional network corresponds to the following,

$$\boldsymbol{\theta}_c^* = \arg\min_{\boldsymbol{\theta}_c} DIV_{\Delta}(\mathrm{Pr}_p, \mathrm{Pr}_c) - \gamma DIV_{\Delta}(\mathrm{Pr}_c, \mathrm{Pr}_c). \tag{14}$$

The above objective function is similar to the one used in [4] for supervised learning of Discrete DISCO Nets. As our conditional net employs a sampling procedure over the scoring function $S_k(\mathbf{y}; \boldsymbol{\theta}_c)$, objective (14) is non-differentiable. However, as observed in [4], it is possible to compute an unbiased estimate of the gradients using the direct loss minimization technique [15, 32]. Therefore, the conditional net can be optimized using stochastic gradient descent. We present the technical details of optimization, which are similar to those in [4], in the supplementary material.

In order to visualize the optimization of the conditional net, let us first consider the easy case in Figure 2 (columns 1-2). Similar to the prediction net in the previous subsection, the uncertainty in the conditional net decreases marginally over the iterations, as it already has high confidence for the bounding box labels. For the challenging objects present in the image of the last two columns, we see that the prediction net has high uncertainty. The improvement in the predictions of the conditional net for these two cases are mainly attributed to the information gained by training on other easier examples of the *dog* and the *bottle* category present in the data set.

6. Experiments

6.1. Data set and Evaluation Metrics

Data set: We evaluate our method on the challenging PASCAL VOC 2007 and 2012 data sets [11] which have $9,962$ and $22,531$ images respectively for 20 object categories. These two data sets are divided into the train, val and test sets. Here we choose trainval set of 5011 images for VOC 2007 and $11,540$ images for VOC 2012 to train our network. The trainval set is further split into $80\% - 20\%$ to create new training and validation sets. We use a non-standard training-validation split in order to maximize the number of training images for our networks, while not overfitting our hyper-parameters on the test set. As we focus on weakly supervised detection, only image-level labels are utilized during training.

Evaluation Metric We use two metrics to evaluate our detection performance. First we evaluate detection using mean Average Precision (mAP) on the PASCAL VOC 2007 and 2012 test sets, following the standard PASCAL VOC protocol [11]. Second, we compute CorLoc [8] on the PASCAL VOC 2007 and 2012 trainval splits. CorLoc is the fraction of positive training images in which we localize an object of the target category correctly. Following [11], a detected bounding box is considered correct if it has at least 0.5 IoU with a ground truth bounding box.

6.2. Implementation Details

We use standard Fast-RCNN [13] to model prediction distribution and a modified Fast-RCNN to model the conditional distribution, as shown in Figure 1(a). We use the ImageNet pre-trained VGG16 Network [28] as the base CNN architecture for both our prediction and conditional nets.

The Fast-RCNN architecture is modified by adding a noise filter in its 5^{th} conv-layer as an extra channel as shown in Figure 1(b). A 1×1 filter is used to bring the number of channels back to the original dimensions (512 channels). No architectural changes are made for the prediction net. The bounding box proposals required for the Fast-RCNN is obtained from the Selective Search algorithm [35]. Results based on the Region Proposal Networks are given in the supplementary material.

Following the standard practice followed in Fast-RCNN, we train and test our method on a single scale. We also construct an ensemble by taking the ImageNet pre-trained VGG11 and VGG13 along with VGG16 and report its results. For all our experiments we choose $K = 5$ for the conditional net. That is, we sample 5 bounding boxes corresponding to 5 noise filters, which are themselves sampled from a uniform distribution. For all other hyper-parameters, we use the same configurations as described in [13].

6.3. Results

In this subsection, we will first compare our method with existing state-of-the-art methods for detection and correct localization tasks on VOC 2007 and 2012 data sets. Then through ablation experiments, see how various terms of our dissimilarity coefficient based objective function contribute towards the accuracy gained. We present further ablation studies in the supplementary material.

6.3.1 Comparison with other methods

We compare our proposed method with other state-of-the-art weakly supervised methods. The detection average precision (AP) and correct localization (CorLoc) on the PASCAL VOC 2007 and 2012 data sets are shown in Table 1, Table 2 and Table 3 respectively. Compared with the other methods, our proposed framework achieves state-of-the-art performance using a single model.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	pson	plant	sheep	sofa	train	tv	mAP
WSDDN [3]	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
WSCCN [9]	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
k-EM [39]	59.8	64.6	47.8	28.8	21.4	67.7	70.3	61.2	17.2	51.5	34.0	42.3	48.8	65.9	9.3	21.1	53.6	51.4	54.7	50.7	46.1
OICR [33]	65.5	67.2	47.2	21.6	22.1	68.0	68.5	35.9	5.7	63.1	49.5	30.3	64.7	66.1	13.0	25.6	50.0	57.1	60.2	59.0	47.0
ZLDN [40]	55.4	68.5	50.1	16.8	20.8	62.7	66.8	56.5	2.1	57.8	47.5	40.1	69.7	68.2	21.6	27.2	53.4	56.1	52.5	58.2	47.6
CL [37]	61.2	66.6	48.3	26.0	15.8	66.5	65.4	53.9	24.7	61.2	46.2	53.5	48.5	66.1	12.1	22.0	49.2	53.2	66.2	59.4	48.3
ML-LocNet [41]	60.8	70.6	47.8	30.2	24.8	64.9	68.4	57.9	11.0	51.3	55.5	48.1	68.7	69.5	28.3	25.2	51.3	56.5	60.0	43.1	49.7
WS-RPN [34]	63.0	69.7	40.8	11.6	27.7	70.5	74.1	58.5	10.0	66.7	60.6	34.7	75.7	70.3	25.7	26.5	55.4	56.4	55.5	54.9	50.4
W2F [42]	63.5	70.1	50.5	31.9	14.4	72.0	67.8	73.7	23.3	53.4	49.4	65.9	57.2	67.2	27.6	23.8	51.8	58.7	64.0	62.3	52.4
Pred Net (VGG)	66.7	69.5	52.8	31.4	24.7	74.5	74.1	67.3	14.6	53.0	46.1	52.9	69.9	70.8	18.5	28.4	54.6	60.7	67.1	60.4	**52.9**
Pred Net (Ens)	67.7	70.4	52.9	31.3	26.1	75.5	73.7	68.6	14.9	54.0	47.3	53.7	70.8	70.2	19.7	29.2	54.9	61.3	67.6	61.2	**53.6**

Table 1. Detection average precision (%) for different methods on VOC 2007 test set.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	pson	plant	sheep	sofa	train	tv	mean
WSCCN [9]	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
WSDDN [3]	68.9	68.7	65.2	42.5	40.6	72.6	75.2	53.7	29.7	68.1	33.5	45.6	65.9	86.1	27.5	44.9	76.0	62.4	66.3	66.8	58.0
ZLDN [40]	74.0	77.8	65.2	37.0	46.7	75.8	83.7	58.8	17.5	73.1	49.0	51.3	76.7	87.4	30.6	47.8	75.0	62.5	64.8	68.8	61.2
OICR [33]	85.8	82.7	62.8	45.2	43.5	84.8	87.0	46.8	15.7	82.2	51.0	45.6	83.7	91.2	22.2	59.7	75.3	65.1	76.8	78.1	64.3
CL [37]	85.8	80.4	73.0	42.6	36.6	79.7	82.8	66.0	34.1	78.1	36.9	68.6	72.4	91.6	22.2	51.3	79.4	63.7	74.5	74.6	64.7
k-EM [39]	79.8	77.8	66.7	50.3	57.0	80.1	89.9	71.5	29.9	75.9	30.5	58.9	73.2	90.2	25.4	51.8	80.2	60.3	72.4	78.9	65.0
WS-RPN [34]	83.8	82.7	60.7	35.1	53.8	82.7	88.6	67.4	22.0	86.3	68.8	50.9	90.8	93.6	44.0	61.2	82.5	65.9	71.1	76.7	68.4
ML-LocNet [41]	81.7	82.9	68.7	44.4	53.9	80.3	88.9	70.5	32.6	74.0	62.7	61.7	81.4	91.6	46.0	60.6	75.2	69.2	78.7	65.8	68.6
W2F [42]	85.4	87.5	62.5	54.3	35.5	85.3	86.6	82.3	39.7	82.9	49.4	76.5	74.8	90.0	46.8	53.9	84.5	68.3	79.1	79.9	70.3
Pred Net (VGG)	88.6	86.3	71.8	53.4	51.2	87.6	89.0	65.3	33.2	86.6	58.8	65.9	87.7	93.3	30.9	58.9	83.4	67.8	78.7	80.2	**70.9**
Pred Net (Ens)	89.2	86.7	72.2	50.9	51.8	88.3	89.5	65.6	33.6	87.4	59.7	66.4	88.5	94.6	30.4	60.2	83.8	68.9	78.9	81.3	**71.4**

Table 2. CorLoc (in %) for different methods on VOC 2007 trainval set.

Method	WSCCN [9]	DSL [18]	OICR [33]	W2F [42]	PredNet(VGG)	PredNet(Ens)
mAP %	37.9	38.3	42.5	47.8	**48.4**	**49.5**
CorLoc %	-	58.8	65.6	69.4	**69.5**	**70.2**

Table 3. Results for different methods on VOC 2012. See supplementary meterial for details.

Method	$\Pr_p, \Pr_c$ (proposed)	$\Pr_p, PW_c$	$PW_p, \Pr_c$	PW_p, PW_c
Mean AP	52.9	50.1	52.6	49.5

Table 4. Detection Average Precision (%) for various ablative settings on VOC 2007 test set

Tables 1 and 2 shows that we significantly outperform methods which only employ a fully factorized distribution in MIL [3, 9]. This empirically demonstrates the usefulness of modeling a complex distribution. Compared to the state-of-the-art method, which trains two separate networks like ours, if we were to only train and test Zhang *et al.* [42] (W2F) using a single scale, where they achieve 49.0% mAP, we get an improvement of 3.9%. We approximate the use of multiple scales by ensembling, which gives us a final improvement over the state-of-the-art method by over 1.2% when compared on multiple scales.

The weakly supervised detector employed in W2F models the annotation constraint using a fully factorized distribution. We argue that our choice of modeling the annotation aware conditional distribution exactly but efficiently, using Discrete DISCO Net, gives us the improved performance. Moreover, unlike W2F, our method combines the weakly supervised and the strongly supervised detectors with a single learning objective instead of training them in a non-end-to-end, cascaded fashion.

6.3.2 Effect of the diversity coefficient terms

In order to understand the effect of various diversity coefficient terms in our objective (8), we remove the self-diversity term in one or both of our probabilistic networks ($\Pr_c$ and $\Pr_p$). In order to obtain a single sample from our conditional network, we feed a zero noise vector (denoted by PW_c). The prediction network still outputs the probability of each bounding box belonging to each class. However, by removing the self-diversity term, we encourage it to output a peakier distribution (denoted by PW_p). Table 4 shows that both the self-diversity terms are important to obtain the maximum accuracy. Relatively speaking, it is more important to include the self-diversity in the conditional network in order to deal with the difficult examples (example, bottle in figure 2). Moreover, this enforces a diverse set of outputs from the conditional network, which helps the prediction network to avoid overfitting the samples during training.

7. Discussion

We presented a novel framework to train an object detector using a weakly supervised data set. Our framework employs a probabilistic objective based on dissimilarity coefficient to model the uncertainty in the location of objects. We show that explicitly modeling the complex non-factorizable conditional distribution is a necessary modeling choice and present an efficient mechanism based on a discrete generative model, the Discrete DISCO Nets, to do so. Extensive experiments on the benchmark data sets have shown that our framework successfully transfers the information present in the image-level annotations for the task of object detection.

In future, we would like to investigate the use of active learning, to further benefit our network in terms of the accuracy of the fully supervised annotations. This will help bridge the performance gap between the strongly supervised detectors and detectors trained using low-cost annotations.

8. Acknowledgements

This work is partially funded by a CEFIPRA grant. Aditya is supported by Visvesvaraya Ph.D. fellowship.

References

[1] Aditya Arun, C V Jawahar, and M Pawan Kumar. Learning human poses from actions. In *BMVC*, 2018.

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

[3] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *CVPR*, 2016.

[4] Diane Bouchacourt. *Task-Oriented Learning of Structured Probability Distributions*. PhD thesis, University of Oxford, 2017.

[5] Diane Bouchacourt, M Pawan Kumar, and Sebastian Nowozin. Disco nets: Dissimilarity coefficients networks. In *NIPS*, 2016.

[6] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. Weakly supervised object localization with multi-fold multiple instance learning. *TPAMI*, 2017.

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

[8] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. Weakly supervised localization and learning with generic knowledge. *IJCV*, 2012.

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

[10] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 1997.

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010.

[12] Weifeng Ge, Sibei Yang, and Yizhou Yu. Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning. In *CVPR*, 2018.

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

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

[15] Tamir Hazan, Joseph Keshet, and David A McAllester. Direct loss minimization for structured prediction. In *NIPS*, 2010.

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017.

[17] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, and Kevin Murphy. Speed/accuracy trade-offs for modern convolutional object detectors. In *CVPR*, 2017.

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

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

[20] M Pawan Kumar, Ben Packer, and Daphne Koller. Modeling latent variable uncertainty for loss-based learning. In *ICML*, 2012.

[21] Baisheng Lai and Xiaojin Gong. Saliency guided end-to-end learning for weakly supervised object detection. In *IJCAI*, 2017.

[22] Dong Li, Jia-Bin Huang, Yali Li, Shengjin Wang, and Ming-Hsuan Yang. Weakly supervised object localization with progressive domain adaptation. In *CVPR*, 2016.

[23] Siyang Li, Xiangxin Zhu, Qin Huang, Hao Xu, and C-C Jay Kuo. Multiple instance curriculum learning for weakly supervised object detection. In *BMVC*, 2017.

[24] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single shot multibox detector. In *ECCV*, 2016.

[25] C Radhakrishna Rao. Diversity and dissimilarity coefficients: a unified approach. *Theoretical population biology*, 1982.

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

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

[28] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

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

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

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

[32] Yang Song, Alexander Schwing, Raquel Urtasun, et al. Training deep neural networks via direct loss minimization. In *ICML*, 2016.

[33] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *CVPR*, 2017.

[34] Peng Tang, Xinggang Wang, Angtian Wang, Yongluan Yan, Wenyu Liu, Junzhou Huang, and Alan Yuille. Weakly supervised region proposal network and object detection. In *ECCV*, 2018.

[35] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *IJCV*, 2013.

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

[37] Jiajie Wang, Jiangchao Yao, Ya Zhang, and Rui Zhang. Collaborative learning for weakly supervised object detection. In *IJCAI*, 2018.

[38] Xinggang Wang, Zhuotun Zhu, Cong Yao, and Xiang Bai. Relaxed multiple-instance svm with application to object discovery. In *ICCV*, 2015.

[39] Ziang Yan, Jian Liang, Weishen Pan, Jin Li, and Changshui Zhang. Weakly-and semi-supervised object detection with expectation-maximization algorithm. *arXiv preprint arXiv:1702.08740*, 2017.

[40] Xiaopeng Zhang, Jiashi Feng, Hongkai Xiong, and Qi Tian. Zigzag learning for weakly supervised object detection. In *CVPR*, 2018.

[41] Xiaopeng Zhang, Yang Yang, and Jiashi Feng. ML-Locnet: Improving object localization with multi-view learning network. In *ECCV*, 2018.

[42] Yongqiang Zhang, Yancheng Bai, Mingli Ding, Yongqiang Li, and Bernard Ghanem. W2F: A weakly-supervised to fully-supervised framework for object detection. In *CVPR*, 2018.