Spatial Consistency Enhanced Dissimilarity Coefficient based Weakly Supervised Object Detection

Aditya Arun, C.V. Jawahar, M. Pawan Kumar

***Abstract*—We consider the problem of weakly supervised object detection, where the training samples have various types of inexpensive annotations. These annotations can indicate the presence or absence of an object category or include count, point, or scribble annotations. 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, PASCAL VOC 2012, MS COCO 2014, and MS COCO 2017 data sets demonstrate the efficacy of our proposed approach.**

I. 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) [1], [2], [3], [4], [5], [6], [7]. 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 explored the weakly supervised object detection (WSOD) problem [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20]. The goal of Weakly Supervised Object Detection (WSOD) is to learn an accurate detector using training samples that are annotated with more cost-effective labels, such as image-level, count, point, and scribble annotations. Image-level annotations can be as simple as object category labels that indicate the presence of an object, or they can include richer information like per-class object counts, which offer slightly more detailed supervision. Additionally, point and scribble annotations provide a more refined level of guidance by indicating specific object locations (points) or rough object boundaries (scribbles). Although these annotations come at a marginally higher cost than image-level labels, as we shall see, they significantly improve the model's ability to localize objects more accurately during training.

Aditya Arun and C.V. Jawahar are with the CVIT, IIIT Hyderabad.

Given the wide availability of such cheaper-to-obtain 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 with only the presence of a car, we still face the challenge of identifying the precise bounding box for the car. This challenge is somewhat mitigated when additional annotations, such as object counts, points, or scribbles, are available. Object count annotations provide information on the number of instances present, reducing ambiguity about the number of objects to detect. Point annotations, by marking specific locations within the object, help in narrowing down the potential area where the object is located. Scribble annotations, which roughly outline the object, offer even more spatial guidance, making it easier to determine the approximate shape and boundary of the object. Despite these enhancements, WSOD must still contend with the inherent uncertainty introduced by the lack of full supervision as the extent of an object is not known.

In order to effectively model uncertainty in weakly supervised learning, Kumar *et al.* [21] 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 [22], 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) [21], [23]. However, WSOD poses greater difficulty 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 weak labels, such as image-level, count, point, or scribble annotations. 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 appropriate deep generative models such as the Discrete DISCO Net [24], [25].

We test the efficacy of our approach on the challenging PASCAL VOC 2007, 2012, and MS COCO 2014 data sets. To generate the weakly supervised data sets, we discard the bounding box annotations, keeping only the image-level labels and, optionally, keeping the per-class object count, points, or scribbles. Using simple image-level labels we achieve 58.1%, 55.4%, 28.6%, and 28.9% detection mAP@0.5 on PASCAL VOC 2007, 2012, MS COCO 2014 and 2017 data sets respectively, significantly improving the state-of-the-art on all the data sets. Using count supervision provides an average increase of 2.3% detection mAP@0.5 across all data sets. Additionally, using point and scribble annotations we obtain a further increase of 3.3%, and 0.8% detection mAP@0.5 on MS COCO 2014 data set respectively giving state-of-the-art results for WSOD using various types of inexpensive annotations.

This work builds on our previously peer-reviewed research [26] by extending our formulation to incorporate better initialization and spatial cluster regularization, which require a new inference algorithm for the conditional distribution. This allows us to sample more accurate bounding box samples from the conditional distribution. We also introduce a simple curriculum learning based optimization algorithm. Our approach is versatile and can integrate various weak labels, such as image-level, count, point, or scribble annotations. We present additional experiments on the MS COCO 2014 and MS COCO 2017 datasets using different backbone architectures, which lead to a significant improvement in our results.

To summarize, we make the following contributions.

- A unified weakly supervised framework to train object detectors with varying levels of weak labels, such as image-level, count, point, and scribble annotations.
- Efficiently model the complex non-factorizable, annotation aware, spatially consistent 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, PASCAL VOC 2012, MS COCO 2014, and MS COCO 2017 data sets.

II. RELATED WORK

Conventional methods often treat WSOD as a Multiple Instance Learning (MIL) problem [27] by representing each image as a bag of instances (that is, putative bounding boxes) [28], [29], [30], [31], [32]. 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 [30], [33], [34], [35], regularize the model with extra cues [28], [29], or relax the MIL constraint [32] to make the objective differentiable. These hard-MIL based methods have demonstrated their effectiveness, especially when CNN features are used to represent object proposals [29]. However, these models are not end to end trainable and 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 [8], [9], [15], [16], [18], [19], [36], [37]. In their work, Bilen *et al.* [8] proposed a smoothed version of MIL that softly labels object proposals instead of choosing the highest scoring ones. Building on their work, Tang *et al.* [15] refine the prediction iteratively through a multi-stage instance classifier. Gao *et al.* [11] presents a greedy approach to training a WSOD using per-class object count. Ren *et al.* [38] presents a unified framework that can utilize all weakly supervised labels, such as image-level supervision, point supervision, and scribble supervision, but they don't consider count supervision. Chen *et al.* [39] presented their work that leverages point annotations to train object detectors. In contrast, we propose a unified framework that can learn from any weakly supervised labels. Zhang *et al.* [18] add curriculum learning using the MIL framework. In our formulation, we explicitly incorporate curriculum learning based on object instance count. Tang *et al.* [40] proposes to cluster similar object proposals to better distinguish between the object and background noise. In our framework, we cluster object proposals such that the number of clusters are consistent with object count. Other end-to-end trainable frameworks for WSOD employ domain adaptation [13], [30], expectation-maximization algorithm [10], [17] and saliency based methods [12]. 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 design choice makes these approaches sub-optimal as what we truly want is to model a distribution that enforces at least one bounding box proposal corresponding to the given weak label.

To enhance weakly supervised detectors, some approaches combine them with strongly supervised ones, typically using predictions from the weakly supervised detector as pseudo-strong labels to train a strongly supervised network [15], [38], [41], [42], [43], [44], [45]. However, this usually involves a unidirectional connection between the two. Wang *et al.* [37] propose a collaborative training approach for weakly and strongly supervised models, similar in spirit to our use of two distributions, though they fully factorize their weakly supervised detector. Yin *et al.* [45] employ a teacher-student network, using an ensemble of students for diverse pseudo ground truth, but without explicitly modeling uncertainty and using a fully factorized distribution. In contrast, we model uncertainty in the conditional distribution to ensure annotation consistency. The improvements reported in these works highlight the importance of modeling separate distributions. In this work, we explicitly define and jointly train two distributions, minimizing the dissimilarity coefficient [22] based objective function.

III. Model

A. 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 [46] to obtain the aforementioned bounding boxes. 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 the weak annotation by $\mathbf{a} \in 0 \cup \mathbb{Z}^+$. Here, $\mathbf{a}^{(j)} = r$ if image $\mathbf{x}$ contains r instances of the j-th object. We assume $r = 1$ where only object category labels are provided and count information is absent. Furthermore, we denote the unknown bounding box labels by $\mathbf{y} = \{\mathbf{y}^{(i)} \mid \mathbf{y}^{(i)} \in \{0, \ldots, C\}^B \wedge i = 1, \ldots B\}$, where $\mathbf{y}^{(i)} = j$ if the i-th bounding box $b^{(i)}$ is of the j-th category. A weakly supervised data set $\mathcal{W} = \{(\mathbf{x}_i, \mathbf{a}_i) \mid i = 1, \ldots, N\}$ contains N pairs of images $\mathbf{x}_i$ and their corresponding image-level labels $\mathbf{a}_i$. For point and scribble annotations, we retain only those bounding box proposals that fully encompass the annotation. This approach ensures their compatibility with count supervision.

B. 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 [21], [23], 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 weak 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, or optionally object instance count or localization information through point or scribble annotations. 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.

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 [2], [3], [7], 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 [2] (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.

2) Conditional Distribution: Given B bounding box proposals for an image $\mathbf{x}$ and the weak annotation $\mathbf{a}$, 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, we divide the B proposals into multiple clusters. Each cluster of bounding boxes corresponds to a foreground object. The total number of clusters of each foreground class should be equal to their image-level annotation $\mathbf{a}^j = r$.

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 Discrete DISCO Net [24]. In what follows, we briefly describe Discrete DISCO Nets while highlighting their applicability to our framework.

a) Discrete DISCO *Net:* Discrete DISCO Net [24] 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] $\mathcal{F}^k_{u,\mathbf{y}_u}(\boldsymbol{\theta}_c) \in \mathbb{R}^{B \times C}$ for each bounding box proposal u, and the corresponding putative label $\mathbf{y}_u$. We generate K different score functions using K different noise samples. These score functions are then used to sample the corresponding bounding box labels $\hat{\mathbf{y}}^k_c$ such that all ground truth labels are included 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

[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.

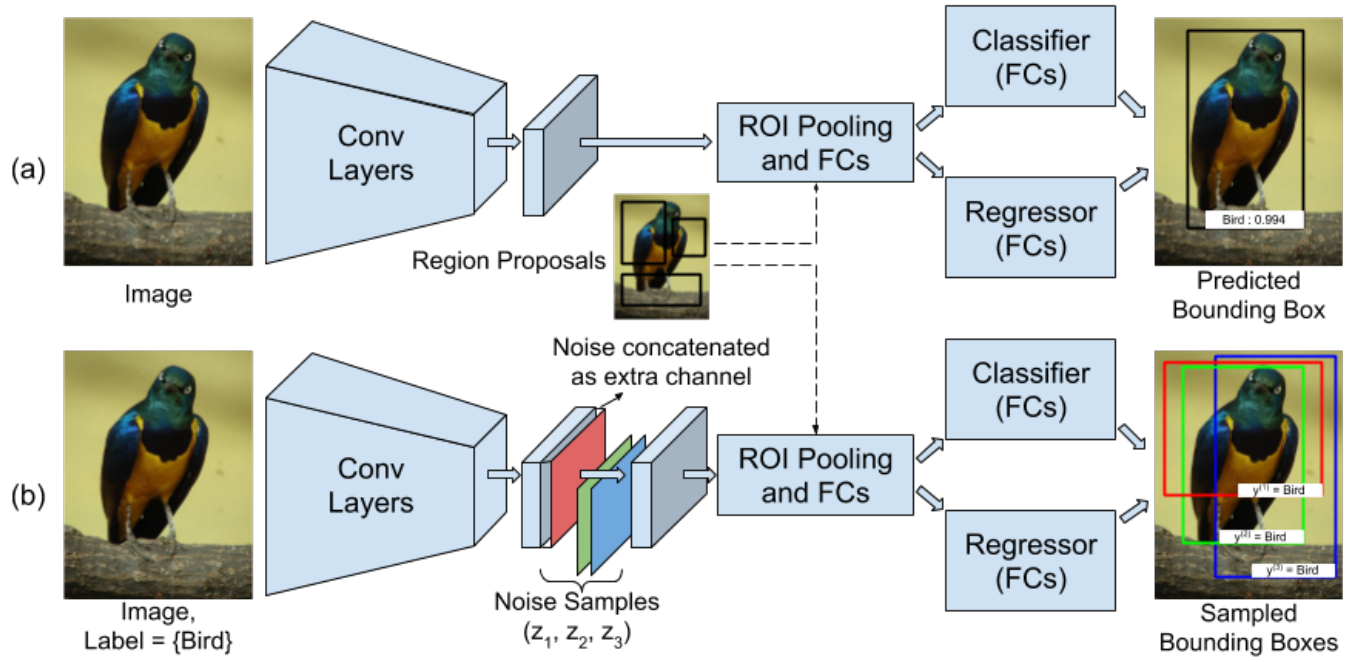

Fig. 1. *The overall architecture. (a) Prediction Network: a standard Fast-RCNN architecture is used to model the prediction net. For an input image, bounding box proposals are generated from selective search [46]. Features from each of these proposals are computed by the region of interest (ROI) pooling 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.*

to generate the noise filter) to the output distribution (used to generate bounding box labels).

In what follows, we will discuss how to redefine the score function $\mathcal{F}^k_{u,\mathbf{y}_u}(\boldsymbol{\theta}_c)$ to obtain a final score function such that it is used to sample the bounding box proposal $\hat{\mathbf{y}}^k_c$.

Initialization by Class Activation Maps In order to incorporate prior knowledge about potential object location, we weigh the score function $\mathcal{F}^k_{u,\mathbf{y}_u}(\boldsymbol{\theta}_c)$ with class activation maps (CAMs) $\mathcal{C}(\mathbf{y_c})$ [47], [48].

$$\mathcal{G}^k_{u,\mathbf{y}_u}(\mathbf{y}_c) = \mathcal{C}(\mathbf{y}_c) \times \mathcal{F}^k_{u,\mathbf{y}_u}(\boldsymbol{\theta}_c). \tag{2}$$

While, we can employ any CAM algorithm, in our experiments, we employ Layer-CAM [49]. When no CAM based algorithm is used, we set $\mathcal{C}(\mathbf{y_c}) = 1$.

Cluster Construction In order to effectively use the count information whenever they are available, we propose to cluster the bounding box proposals such that the number of clusters is equal to the count annotation. To form clusters, the proposals are sorted by their object confidences $\mathcal{G}^k_{u,\mathbf{y}_u}(\mathbf{y}_c)$ and the following steps are iteratively performed:

1) Construct a cluster using the proposal with the highest object confidence for the r non-overlapping instances. This ensures that the number of clusters is consistent with image-level label $\mathbf{a}^{(j)} = r$.
2) Find proposals that overlap with a proposal in the cluster by more than 0.7 and merge them into the cluster.

All object instances not forming part of the foreground objects are considered background boxes. The pseudocode for cluster construction is presented in algorithm 1.

Spatial cluster regularization For each bounding box in a cluster corresponding to the foreground object instance, we can redefine our score function such that highly overlapping proposal bounding boxes should have similar scores and labels

$$\mathcal{G}^{k,n}_{u,\mathbf{y}_u}(\mathbf{y}_c) = \mathcal{G}^{k,n-1}_{u,\mathbf{y}_u}(\mathbf{y}_c) + \sum_{i=1}^{B^r \setminus u} \mathbf{w}_i \mathcal{G}^{k,n-1}_{i,\mathbf{y}_i}(\mathbf{y}_c), \tag{3}$$

where n is the iterator, B^r are the bounding boxes belonging to a particular cluster, and $\mathbf{w}_i = IOU(b_u, b_i)$ is the IOU between the two proposal boxes. Equation (3) is iteratively updated until the scores, weighted by their IOUs, converge. While the algorithm guarantees convergence to a local minimum, in practice, we limit the process to 5 iterations or until the scores stabilize. Empirical evidence shows that 5 iterations are typically sufficient for convergence, providing a good balance between accuracy and speed.

Annotation consistent constraint Finally, we would like to add a constraint such that there must exist at least one bounding box in each clique that satisfies the annotation $\mathbf{a}$.

$$\mathcal{S}^k(\mathbf{y}_c) = \sum_{i=1}^{B^r} \mathcal{G}^{k,n}_{u,\mathbf{y}_u}(\mathbf{y}_c) - \mathcal{H}_k(\mathbf{y}_c), \tag{4}$$

where,

$$\mathcal{H}_k(\mathbf{y}_c) = \begin{cases} 0 & \text{if } \forall j \in \{1,\dots,C\} \text{ s.t. } \mathbf{a}^{(j)} = r, \\ & \exists i \in \{1,\dots,B\} \text{ s.t. } \mathbf{y}^{(i)} = j, \\ \infty & \text{otherwise.} \end{cases} \tag{5}$$

Algorithm 1: Cluster Construction

Input: Bounding boxes B, scores $\mathcal{G}^k_{u,\mathbf{y}_u}(\mathbf{y}_c)$, annotations $\mathbf{a}^{(j)} = r$, IoU threshold τ
Output: Dictionary of class specific clusters with keys $\mathbf{a}^{(j)}$ and values a list of exactly r clusters

Initialize a dictionary 'dict' with keys $\mathbf{a}^{(j)}$ and values an empty list;
foreach *annotation* $\mathbf{a}^{(j)} > 0$ **do**
  `// Initialize variables`
  Initialize an empty list 'clusters';
  Initialize a boolean array 'used_boxes' of length B to track used boxes;
  Sort boxes and scores based on the maximum scores corresponding to $\mathbf{a}^{(j)}$ in descending order;
  `// Create non-overlapping clusters`
  foreach *box b in sorted order* **do**
    if *number of clusters* $\geq r$ **then**
      break;
    if *b is not used* **then**
      Create a new cluster with b;
      Mark b as used;
      foreach *remaining box* b' **do**
        if *b' is not used and IoU(b, b')* $\geq \tau$ **then**
          Add b' to the cluster;
          Mark b' as used;
      Add the cluster to 'clusters';
  `// Ensure exactly r clusters by merging or splitting`
  if *number of clusters* $< r$ **then**
    while *number of clusters* $< r$ **do**
      Split the largest cluster into two smaller clusters;
  else if *number of clusters* $> r$ **then**
    while *number of clusters* $> r$ **do**
      Merge the two most similar or overlapping clusters;
  'dict'[$\mathbf{a}^{(j)}$] = 'clusters';
return *'dict'*;

Given the scoring function in equation (4), we compute the k-th sample as

$$\hat{\mathbf{y}}^k_c = \arg\max_{\mathbf{y}\in\mathcal{Y}} \mathcal{S}^k(\mathbf{y}_c). \tag{6}$$

Note that in equation (6) 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 $\mathcal{H}_k$, we can design an efficient yet exact algorithm for equation (6). Specifically, we assign each bounding box proposal u 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 that has the highest score corresponding to the foreground label. The pseudocode for conditional net inference is presented in algorithm 2.

For point and scribble supervision, we retain only the bounding box proposals that fully contain the annotations. This approach not only narrows the problem's search space but also ensures compatibility with object instance count supervision.

Algorithm 2: Conditional net inference algorithm

Input: A dictionary of class specific clusters 'dict', original scores $\mathcal{S}^k(\mathbf{y}_c)$, annotations $\mathbf{a}^{(j)} = r$
Output: A dictionary Y containing a list of r maximum scoring boxes for each $\mathbf{a}^{(j)}$

foreach *annotation* $\mathbf{a}^{(j)}$ *in 'dict'* **do**
  Initialize an empty list 'max_boxes';
  foreach *cluster* B^r *in 'clusters'* **do**
    `/* Iterative algorithm for spatial cluster regularization */`
    repeat
      for $b_u, b_i \in B^r$ **do**

$$\mathcal{G}^{k,n}_{u,\mathbf{y}_u}(\mathbf{y}_c) = \mathcal{G}^{k,n-1}_{u,\mathbf{y}_u}(\mathbf{y}_c) + \sum_{i=1}^{B^r\setminus u} \mathbf{w}_i \mathcal{G}^{k,n-1}_{i,\mathbf{y}_i}(\mathbf{y}_c)$$

    until $\mathcal{G}^{k,n}_{u,\mathbf{y}_u}(\mathbf{y}_c)$ *has converged*;
    `/* Greedily select the maximum scoring bounding box */`
    $Y[\mathbf{a}^{(j)}] = \arg\max_{y\in\mathcal{B}^r} \mathcal{G}^{k,n}_{u,\mathbf{y}_u}(\mathbf{y}_c)$
return *dictionary* Y;

IV. 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 [22]. To this end, we require a task specific loss function, which we define next.

A. Task Specific Loss Function

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

$$\Delta(\mathbf{y}_1, \mathbf{y}_2) = \frac{1}{B}\sum_{i=1}^{B} \Delta(\mathbf{y}_1^{(i)}, \mathbf{y}_2^{(i)}). \tag{7}$$

Following the standard practice in most modern object detectors [50], $\Delta(\mathbf{y}_1^{(i)}, \mathbf{y}_2^{(i)})$ 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 $smoothL1$ [2] for our localization loss Δ_{loc}. Formally, the task specific loss is given by,

$$\Delta(\mathbf{y}_1^{(i)}, \mathbf{y}_2^{(i)}) = \Delta_{cls}(\mathbf{y}_1^{(i)}, \mathbf{y}_2^{(i)}) + \lambda\Delta_{loc}(b_1^{(i)}, b_2^{(i)}). \tag{8}$$

Here, $b_1^{(i)}$ and $b_2^{(i)}$ are the corresponding bounding box proposals for $\mathbf{y}_1^{(i)}$ and $\mathbf{y}_2^{(i)}$.

B. 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 [21], [23], we design a joint learning objective that can minimize the dissimilarity coefficient [22] 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,

$$DIV_{\Delta'}(\Pr_1,\Pr_2) = \mathbb{E}_{\mathbf{y}_1\sim\Pr_1(\cdot)}[\mathbb{E}_{\mathbf{y}_2\sim\Pr_2(\cdot)}[\Delta'(\mathbf{y}_1,\mathbf{y}_2)]]. \tag{9}$$

If the model correctly brings the two distributions 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,

$$\begin{aligned} DISC_{\Delta'}(\Pr_1,\Pr_2) =& DIV_{\Delta'}(\Pr_1,\Pr_2) \\ &- \gamma DIV_{\Delta'}(\Pr_2,\Pr_2) \\ &- (1-\gamma) DIV_{\Delta'}(\Pr_1,\Pr_1), \end{aligned} \tag{10}$$

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 encourage the samples from each of the two distributions 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 (8) as

$$\boldsymbol{\theta}_p^*, \boldsymbol{\theta}_c^* = \underset{\boldsymbol{\theta}_p,\boldsymbol{\theta}_c}{\arg\min}\, DISC_{\Delta}(\Pr_p(\boldsymbol{\theta}_p), \Pr_c(\boldsymbol{\theta}_c)), \tag{11}$$

where each of the diversity terms can be derived from equation (9). As discussed in Section III-B, 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]

$$\begin{aligned} &DIV_{\Delta}(\Pr_p,\Pr_c) \\ &= \frac{1}{BK}\sum_{i=1}^{B}\sum_{k=1}^{K}\sum_{\mathbf{y}_p^{(i)}} \Pr_p(\mathbf{y}_p^{(i)};\boldsymbol{\theta}_p)\Delta(\mathbf{y}_p^{(i)},\hat{\mathbf{y}}_c^{k,(i)}), \end{aligned} \tag{12}$$

$$DIV_{\Delta}(\Pr_c,\Pr_c) = \frac{1}{K(K-1)B}\sum_{\substack{k,k'=1\\k'\neq k}}^{K}\sum_{i=1}^{B}\Delta(\hat{\mathbf{y}}_c^{k,(i)},\hat{\mathbf{y}'}_c^{k',(i)}), \tag{13}$$

$$\begin{aligned} &DIV_{\Delta}(\Pr_p,\Pr_p) \\ &= \frac{1}{B}\sum_{i=1}^{B}\sum_{\mathbf{y}_p^{(i)}}\sum_{\mathbf{y'}_p^{(i)}} \Pr_p(\mathbf{y}_p^{(i)};\boldsymbol{\theta}_p)\Pr_p(\mathbf{y'}_p^{(i)};\boldsymbol{\theta}_p)\Delta(\mathbf{y}_p^{(i)},\mathbf{y'}_p^{(i)}). \end{aligned} \tag{14}$$

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.

V. Optimization

As we employ deep neural networks to model the two distributions, our objective function (11) is ideally suited to be minimized by stochastic gradient descent. While it may be possible to compute the gradients of both 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 the 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 [2] and Discrete DISCO Net [24]. 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.

For the case where object count labels are present, we employ a simple curriculum-learning based strategy. We first iteratively train the two networks for images with images that have a single object count. Next, we progressively increase the number of objects present in the training image.

[2]Details in Appendix A

A. 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:

$$\boldsymbol{\theta}_p^* = \arg\min_{\boldsymbol{\theta}_p} DIV_{\Delta}(\mathrm{Pr}_p, \mathrm{Pr}_p) - (1-\gamma) DIV_{\Delta}(\mathrm{Pr}_p, \mathrm{Pr}_p). \tag{15}$$

Note that, due to the use of dissimilarity coefficient, the above objective differs slightly from the one used for Fast-RCNN [2]. However, importantly, it is still differentiable with respect to $\boldsymbol{\theta}_p$. Hence, the prediction net can be directly optimized via stochastic gradient descent.

B. 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,

$$\boldsymbol{\theta}_c^* = \arg\min_{\boldsymbol{\theta}_c} DIV_{\Delta}(\mathrm{Pr}_p, \mathrm{Pr}_c) - \gamma DIV_{\Delta}(\mathrm{Pr}_c, \mathrm{Pr}_c). \tag{16}$$

The above objective function is similar to the one used in [24] for supervised learning of Discrete DISCO Nets. As our conditional net employs a sampling procedure over the scoring function $\mathcal{S}^k(\mathbf{y}_c)$, objective (16) is non-differentiable. However, as observed in [24], it is possible to compute an unbiased estimate of the gradients using the direct loss minimization technique [51], [52]. Therefore, the conditional net can be optimized using stochastic gradient descent. We present the technical details of optimization, which are similar to those in [24], in appendix B.

C. Visualization of the learning process

Figure 2 provides the visualization of the performance of the two networks over the different iterations of the iterative learning procedure. Figure 2(a) demonstrates a simple example where single instance of each object is present and only image-level annotations are present during training. Figure 2(b) demonstrates a more complex example where several instances of the same object are present and only image-level annotations are present during training. Figure 2(c) and 2(d) demonstrates the complex example in presence of count annotations and point annotations during training respectively. The estimated bounding box labels from the prediction net and those sampled from the conditional net are depicted. For conditional net, we superimpose five different samples of bounding box labels. If all the samples agree with each other on bounding box labels, 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 the training of the prediction net. The two steps of the iterative algorithm are described below in brief. For completeness, the details are provided in Appendix B.

In order to visualize the learning process, let us first consider the simple example (Figure 2(a)), where only image-level annotations are present during training. We observe that initially (in iteration 1), the conditional net's samples for both *dog* and *bottle* objects have high uncertainty, meaning the samples are spread out and lack consensus. However, they are broadly localized over the object, an information that can be exploited by our algorithm. The same is also reflected in the output of the prediction net, which is unable to detect either object. Over the iterations, the knowledge from the conditional net is transferred to the prediction net, and we see a gradual improvement in the uncertainty of both the prediction net and the conditional net, finally resulting in accurate localization of both the objects.

Figure 2(b) presents a challenging example where multiple instances of the object *person* are present, and only image-level annotations are present during training. We observe that initially the conditional net samples are extremely diverse (and has high uncertainty). Some samples correctly localizes one of the instances of the class *person*, but others span multiple instances of that class. The output of the prediction net also reflects this, with partial localization of one of the object instances and incorrect localization that contains multiple instances or no localization of an instance of the *person* class. Over the iterations, the uncertainty of the prediction and the conditional net reduces, and we see a better localization. Finally, the conditional network has low uncertainty in its samples, even though it misses several instances of the object. The prediction net succesfully localizes several instances of the class *person*, as it also learns from other images containing the *person* class in the data set during iterative training. However, we see that the final output of the prediction net remains imperfect with some instances not localized and some localization containing multiple instances of the same object.

In figure 2(c), we see the sample challenging example where count annotations are present during training. We observe that due to our cluster construction (section III-B2), we can now take multiple samples from the conditional net. Although, initially the uncertainty of the conditional net is high, the samples obtained are better localized than the case where only image-level annotations were present. Over the iterations, we see the uncertainty in both the prediction and conditional nets reducing. We note that in this case, many of the instances are correctly localized but some instances are either partially localized and some localization contains multiple instances.

Finally, in figure 2(d), we consider the challenging example where point annotations are available. We observe that initially the uncertainty of the prediction net is high, but the uncertainty in the conditional net is low. Over the iterations, the information present in the conditional net is successfully transferred to the prediction net, where the final output accurately localizes all instances of the same class.

VI. EXPERIMENTS

A. Data set and Evaluation Metrics

Data set: We evaluate our method on the challenging VOC 2007, and VOC 2012 in PASCAL VOC [53], and COCO 2014 and COCO 2017 in MS COCO [54] data sets. We use the trainval set in VOC 2007 and VOC 2012 data sets that has $5,011$ and $11,540$ images respectively for 20 object

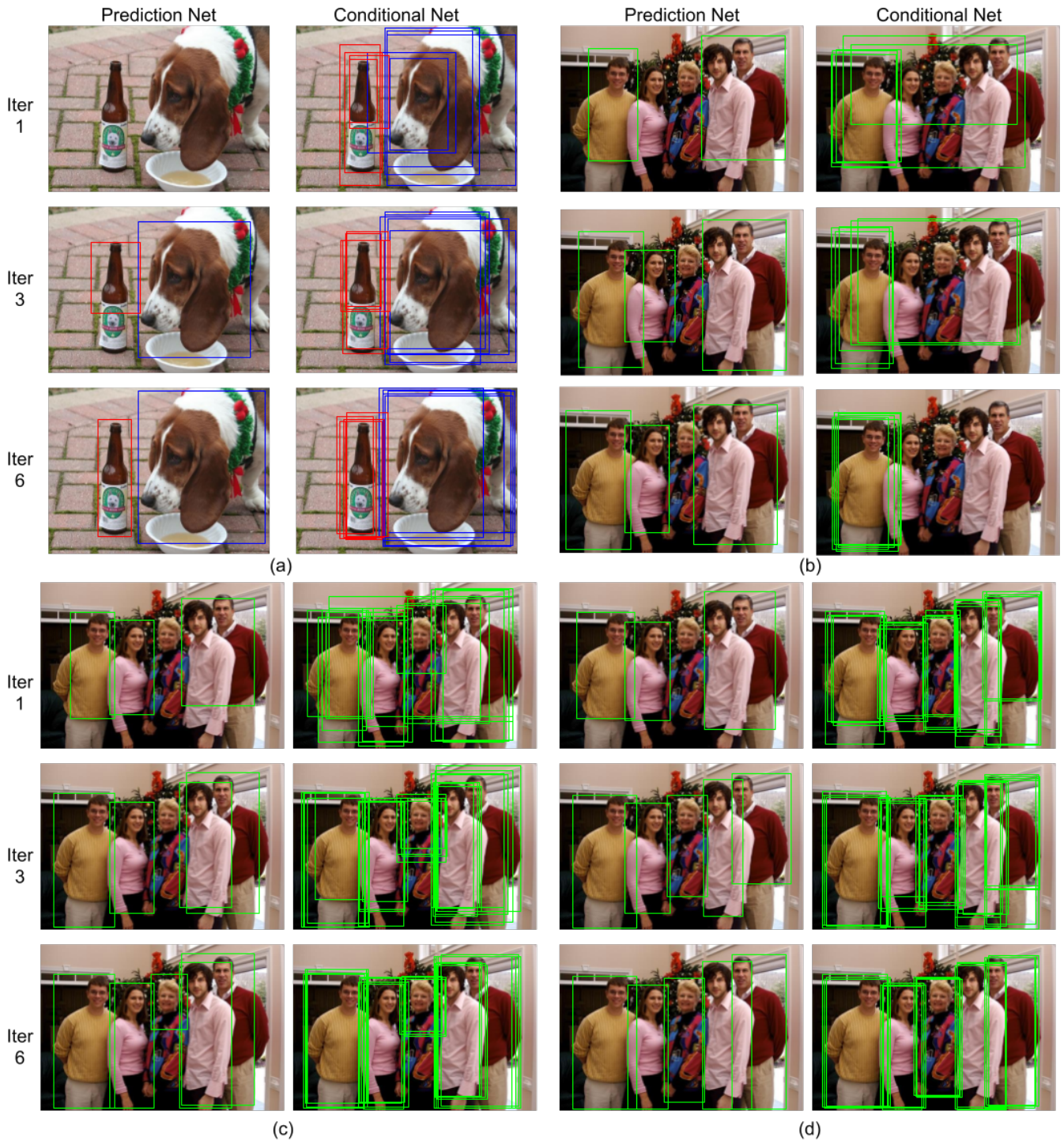

Fig. 2. *Example of predictions of prediction net and conditional net. For prediction net, the visualization is after taking standard non maximal suppression using standard score threshold* = 0.7. *Columns 1 and 3 are outputs of the prediction network while columns 2 and 4 are outputs from the conditional network. Rows 1 and 4 represent the prediction of the two networks after the first iteration and rows 2 and 5 represent the prediction of the two networks after the third iteration. Finally, rows 3 and 6 represent the prediction of the two networks after the sixth (final) iteration. Image set (a) demonstrates a simple example where a single instance of each object type is present and image-level annotations are present during training. Image set (b) demonstrates a complex example where several instances of a single object are present and only image level annotations are present during training. Image sets (c) and (d) demonstrate the complex example in the presence of count and point supervision respectively. Each object class is represented by a different colored bounding box, where the green box represents the* person *category and red and blue represent the* bottle *and* dog *categories respectively. Best viewed in color.*

categories, and the test set contains $4,951$ and $10,991$ images for evaluation. COCO 2014 data includes around $82,783$ images for training and $40,504$ images for validation for 80 object categories. COCO 2017 has $118,287$ images in the train set and $5,000$ images in the validation set.

As we focus on weakly supervised detection, only image-

level labels (I) are utilized during training. We retain instance count information (C) for count supervision. For point annotations (P), we use quasi-center point annotations, where the center of the ground-truth bounding boxes serves as the point annotation. However, if there is an overlap between bounding boxes, we select the nearest non-overlapping point from the center box. In cases where the point annotation falls outside the object or is contained inside other bounding box, we do not make corrections. For scribble (S) supervision, we adopt the setup proposed by Ren *et al.* [38]. Note that Ren *et al.* [43] provide scribble annotations only for COCO 2014 data set. Therefore, for scribble supervision, we only consider COCO 2014 data set.

Evaluation Metric We use two metrics to evaluate our detection performance on the PASCAL VOC data set. First, we evaluate detection using mean Average Precision (mAP) on the PASCAL VOC 2007 and 2012 test sets, following the standard PASCAL VOC protocol [53]. Second, we compute CorLoc [55] 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 [53], a detected bounding box is considered correct if it has at least 0.5 IoU with a ground truth bounding box.

MS-COCO presents a greater challenge compared to PASCAL VOC, as it contains significantly more instances per image (approximately 7 versus 2) and a larger number of classes (80 versus 20). We report mAP results at IoU thresholds of 0.5 and 0.75, along with the more comprehensive AP metric. AP is calculated as the average mAP across 10 IoU thresholds, ranging from 0.5 to 0.95 in 0.05 increments.

B. Implementation Details

We use standard Fast-RCNN [2] 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 [56] and ImageNet pre-trained ResNet network [57] as the base CNN architectures 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 to the prediction net. The bounding box proposals required for the Fast-RCNN are obtained from the Selective Search algorithm [46]. Results based on the Region Proposal Networks are given in the supplementary material.

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 [2].

In order to initiate the training of our proposed framework, we first train the conditional network using the thresholded CAM output as a pseudo bounding box label. Specifically, we threshold the CAM output at 0.7 and create a bounding box that tightly encloses the resulting mask. When count information (C) is available, we ensure that the number of pseudo bounding boxes matches the count annotation. If point (P) or scribble (S) annotations are available, we retain only those bounding box proposals that contain the corresponding point or scribble annotation.

C. Results

In this subsection, we first compare our method with the current state-of-the-art approaches for detection and correct localization tasks on the PASCAL VOC datasets, as well as for detection task on the MS COCO datasets. Next, through ablation experiments, we examine how the different components used to redefine the score function and various terms in our dissimilarity coefficient-based objective function contribute to the improvement in accuracy.

1) Comparison with other methods: We compare our proposed method with other state-of-the-art weakly supervised methods with varying levels of weak supervision. The performance on detection average precision and correct localization metrics for the PASCAL VOC data sets and the detection average precision metrics for the MS COCO data sets are presented in table I. We employ two different backbones for our networks, VGG-16 [56], and ResNet-50 [57]. Compared with the other methods, our proposed framework achieves state-of-the-art performance using a single model and using the selective search for bounding box proposals across varying levels of weak supervision. This demonstrates the efficacy and generalizability of our proposed approach. We also observe a consistent gain of accuracy ($> 1\%$) when using a bigger model that uses ResNet-50, over the baseline model that has VGG-16 as its backbone. Although not surprising, this trends demonstrate that our method is scalable and the accuracies can further improve when using a bigger model that has better representational capacity (such as ResNet-101 or ViT).

Using image level annotations (I), our method significantly outperforms other state-of-the-art methods. Inspired by Bilen *et al.* [8], prior arts [8], [15], [41], [42], [43], [44], [45] employ a fully factorized distribution in MIL objective. We empirically demonstrate the usefulness of modeling a complex distribution. Compared to previous arts [15], [41], [42], [43], [44], [45] that uses two different networks, one for pseudo bounding box generation, and another Fast-RCNN [2] for inference, our iterative training of both the networks using a joint objective enables us to achieve superior performance. Compared to CBL [45], that generates multiple pseudo bounding box labels using an ensemble of student networks to train a teacher network, we get better results by explicitly modeling the uncertainty over the pseudo label generation process and generating unbiased samples using the conditional network.

When we have access to instance count annotations (C), our results improve significantly over the image-level annotation baseline. This is especially noticeable ($+2.4\%$ and $+2.7\%$ AP for COCO 2014 and COCO 2017 respectively) for the MS COCO data sets that have higher instance counts per image compared to the PASCAL VOC data sets. This improvement is attributed to the cluster construction and the use of curriculum learning based on the instance count during training.

TABLE I
COMPARISON WITH THE STATE-OF-THE-ART WSOD METHODS ON PASCAL VOC AND MS COCO DATA SETS.

Method	Sup.	Backbone	VOC 2007		VOC 2012		COCO 2014			COCO 2017		
			mAP	CorLoc	mAP	CorLoc	Avg. Precision, IoU: 0.5:0.95	0.5	0.75	Avg. Precision, IoU: 0.5:0.95	0.5	0.75
WSDDN [8]	*I*	VGG16	34.8	53.5	–	–	9.5	19.2	8.2	–	–	–
OICR [15]	*I*	VGG16	47.0	64.3	42.5	65.6	7.7	17.4	–	–	–	–
WSOD^2 [41]	*I*	VGG16	53.6	69.5	47.2	71.9	10.8	22.7	–	–	–	–
C-MIDN [42]	*I*	VGG16	52.6	68.7	50.2	71.2	9.6	21.4	–	–	–	–
MIST [43]	*I*	VGG16	54.9	68.8	52.1	70.9	11.4	24.3	9.4	12.4	25.8	10.5
OD-WSCL [44]	*I*	VGG16	56.1	69.8	54.6	71.2	14.4	**29.0**	12.4	13.6	27.4	12.2
CBL [45]	*I*	VGG16	57.4	71.8	53.5	72.6	13.6	27.6	–	–	–	–
PredNet (Ours)	*I*	VGG16	**58.1**	**72.4**	**55.4**	**72.9**	**14.8**	28.6	**14.2**	**15.1**	**28.9**	**14.6**
OICR [15]	*I*	R-50	50.1	–	–	–	–	–	–	–	–	–
OD-WSCL [44]	*I*	R-50	56.6	–	–	–	13.9	29.1	11.8	13.8	27.8	12.1
PredNet (Ours)	*I*	R-50	**59.4**	**73.9**	**56.6**	**74.8**	**15.4**	**28.9**	**14.9**	**15.9**	**29.8**	**15.1**
C-WSL [11]	*C*	VGG16	48.2	66.1	45.4	66.9	–	–	–	–	–	–
PredNet (Ours)	*C*	VGG16	**59.6**	**74.1**	**56.8**	**75.1**	**17.2**	**31.6**	**15.5**	**17.8**	**32.1**	**16.4**
PredNet (Ours)	*C*	R-50	**60.7**	**74.9**	**57.0**	**76.3**	**17.6**	**31.9**	**15.7**	**18.3**	**32.3**	**16.7**
UFO^2 [38]	*P*	VGG16	–	–	–	–	12.4	27.0	–	13.5	27.9	–
PredNet (Ours)	*P*	VGG16	**60.1**	**74.4**	**57.2**	**75.4**	**19.0**	**34.9**	**17.9**	**19.6**	**35.2**	**19.3**
P2BNet [39]	*P*	R-50	–	–	–	–	19.4	**43.5**	–	**22.1**	**47.3**	–
PredNet (Ours)	*P*	R-50	**61.0**	**75.4**	**57.4**	**75.7**	**19.9**	36.0	**18.7**	20.7	36.5	**20.1**
UFO^2 [38]	*S*	VGG16	–	–	–	–	13.7	29.8	–	–	–	–
PredNet (Ours)	*S*	VGG16	–	–	–	–	**19.8**	**35.7**	**19.0**	–	–	–
PredNet (Ours)	*S*	R-50	–	–	–	–	**21.1**	**37.8**	**20.2**	–	–	–

Using point annotation (P), our method further improves the baseline based on count supervision and achieves competitive results overall. Again, this is again noticeable in the more complex MS COCO data sets, where several instances of the same object can be cluttered together, thus making the ground truth point annotation more relevant. Using, scribble supervision (S), we further improve the results obtained using count supervision owing to the use of more accurate annotations. Note that due to our use spatial consistency, the improvement achieved after using scribble supervision over point supervision is not as high, thus highlighting the fact that our spatial consistency term effectively captures the extent of an object.

TABLE II
ABLATION EXPERIMENT: DETECTION AVERAGE PRECISION ON COCO 2017 DATA SET WITH COUNT ANNOTATION (C) UNDER DIFFERENT SETTINGS. CAM IS CLASS ACTIVATION MAPS, SR IS SPATIAL REGULARIZATION, AND AC IS ANNOTATION CONSISTENT CONSTRAINT

CAM	SR	AC	COCO 2017 (AP (0.5:0.95))
			12.8
✓			14.9
	✓		14.3
		✓	13.6
✓	✓		16.9
✓		✓	15.6
	✓	✓	15.2
✓	✓	✓	17.8

D. Ablation Experiments

In this section, we examine the impact of applying CAMs, spatial regularization, and annotation consistency constraints to redefine the score function on the COCO 2017 dataset, where instance count information is available. Additionally, we will explore the effects of the diversity coefficient terms and the influence of curriculum learning within the same context.

1) Effect of redefining the score function: To obtain accurate bounding box samples from the conditional network, we redefined the score function by incorporating CAM scores, spatial regularization, and an annotation consistency constraint. Table II illustrates the performance impact of each component and their combinations.

Row 1 represents the baseline scenario, where the highest-scoring bounding box is sampled from the conditional network. While the performance is comparable to other image-level weakly supervised approaches [44].

Incorporating CAM scores yields a significant improvement of 2.1%, underlining the importance of integrating strong priors in the proposed method. Similarly, adding spatial regularization alone leads to a notable performance boost of 1.5%. This improvement can be attributed to spatial regularization's ability to address the common issue where bounding boxes that cover only the most discriminative part of an object are assigned the highest scores, thereby leading to the selection of more accurate bounding boxes. When the model is constrained to select bounding boxes that are consistent with annotations, a further performance gain of 0.8% is observed. This suggests that enforcing annotation consistency encourages more accurate bounding box sampling.

Moreover, the table demonstrates that these three components are complementary to one another. When combined, their performance improves beyond the individual gains, showing an even more substantial boost in accuracy. The best result, with an AP improvement of 5%, is achieved when all three components—CAM scores, spatial regularization, and annotation consistency—are used together to redefine the score function. This indicates that the synergy between these components is crucial for maximizing detection performance.

2) Effect of the diversity coefficient terms: In order to understand the effect of various diversity coefficient terms in our objective (10), we remove the self-diversity term in one or both of our probabilistic networks ($\Pr_c$ and $\Pr_p$). 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 III 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. Moreover, this enforces a diverse set of outputs from the conditional network, which helps the prediction network to avoid overfitting the samples during training.

3) Effect of instance count based curriculum learning: We examine the effect of curriculum learning, which leverages count information (when available) to train the model with increasingly complex images progressively. Implementing curriculum learning results in a performance improvement from 59.4% to 59.6% on the VOC 2007 dataset. A more substantial gain is observed on the more complex COCO 2017 dataset, where the performance increases from 16.7% to 17.8%. Given that COCO 2017 contains an average of 7 instances per image (compared to VOC 2007 that has an average of 2 instances per image), we argue that employing a simple curriculum aids the model in learning better and more discriminative features during the early stages of training. This enables the model to better grasp the concept of an object, ultimately enhancing its performance. These results also show that our proposed approach is amenable to more complex data sets.

E. Additional Comments

Weakly supervised approaches have been shown to improve performance when trained with extra data [38], CLIP alignment [58], or when using better region proposals such as MCG [44] or using Segment Anything Model (SAM) [59]. We consider these approaches to be complementary to our method and can be easily incorporated. However, the scope of our study was to obtain the best performance using diverse weakly supervised data without the need for external data. Additionally, in their paper, Zhou *et al.* [58] uses ground truth bounding boxes during training, violating the weakly supervised setting. A similar issue is present in Seo *et al.* [59] that uses SAM based proposals to obtain superior results. However, SAM [60] itself is partially trained with ground truth segmentation masks, thus violating the weakly supervised setting.

TABLE III
DETECTION AVERAGE PRECISION (%) FOR VARIOUS ABLATIVE SETTINGS ON COCO 2017 WITH INSTANCE COUNT ANNOTATION (C)

Method	$\Pr_p, \Pr_c$ (proposed)	$\Pr_p, PW_c$	$PW_p, \Pr_c$	PW_p, PW_c
AP (0.5:0.95)	17.8	15.2	17.4	14.8

VII. 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.

REFERENCES

[1] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *NIPS*, 2016.
[2] R. Girshick, "Fast R-CNN," in *ICCV*, 2015.
[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014.
[4] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *ICCV*, 2017.
[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *ECCV*, 2016.
[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *CVPR*, 2016.
[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *NIPS*, 2015.
[8] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *CVPR*, 2016.
[9] A. Diba, V. Sharma, A. Pazandeh, H. Pirsiavash, and L. Van Gool, "Weakly supervised cascaded convolutional networks," in *CVPR*, 2017.
[10] Z. Jie, Y. Wei, X. Jin, J. Feng, and W. Liu, "Deep self-taught learning for weakly supervised object localization," in *CVPR*, 2017.
[11] M. Gao, A. Li, R. Yu, V. I. Morariu, and L. S. Davis, "C-wsl: Count-guided weakly supervised localization," in *ECCV*, 2018.
[12] B. Lai and X. Gong, "Saliency guided end-to-end learning for weakly supervised object detection," in *IJCAI*, 2017.
[13] D. Li, J.-B. Huang, Y. Li, S. Wang, and M.-H. Yang, "Weakly supervised object localization with progressive domain adaptation," in *CVPR*, 2016.
[14] S. Li, X. Zhu, Q. Huang, H. Xu, and C.-C. J. Kuo, "Multiple instance curriculum learning for weakly supervised object detection," in *BMVC*, 2017.
[15] P. Tang, X. Wang, X. Bai, and W. Liu, "Multiple instance detection network with online instance classifier refinement," in *CVPR*, 2017.
[16] P. Tang, X. Wang, A. Wang, Y. Yan, W. Liu, J. Huang, and A. Yuille, "Weakly supervised region proposal network and object detection," in *ECCV*, 2018.
[17] Z. Yan, J. Liang, W. Pan, J. Li, and C. Zhang, "Weakly-and semi-supervised object detection with expectation-maximization algorithm," *arXiv preprint arXiv:1702.08740*, 2017.
[18] X. Zhang, J. Feng, H. Xiong, and Q. Tian, "Zigzag learning for weakly supervised object detection," in *CVPR*, 2018.
[19] X. Zhang, Y. Yang, and J. Feng, "ML-Locnet: Improving object localization with multi-view learning network," in *ECCV*, 2018.

[20] Y. Zhang, Y. Bai, M. Ding, Y. Li, and B. Ghanem, "W2F: A weakly-supervised to fully-supervised framework for object detection," in *CVPR*, 2018.
[21] M. P. Kumar, B. Packer, and D. Koller, "Modeling latent variable uncertainty for loss-based learning," in *ICML*, 2012.
[22] C. R. Rao, "Diversity and dissimilarity coefficients: a unified approach," *Theoretical population biology*, 1982.
[23] A. Arun, C. V. Jawahar, and M. P. Kumar, "Learning human poses from actions," in *BMVC*, 2018.
[24] D. Bouchacourt, "Task-oriented learning of structured probability distributions," Ph.D. dissertation, University of Oxford, 2017.
[25] D. Bouchacourt, M. P. Kumar, and S. Nowozin, "Disco nets: Dissimilarity coefficients networks," in *NIPS*, 2016.
[26] A. Arun, C. Jawahar, and M. P. Kumar, "Dissimilarity coefficient based weakly supervised object detection," in *CVPR*, 2019.
[27] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial intelligence*, 1997.
[28] H. Bilen, M. Pedersoli, and T. Tuytelaars, "Weakly supervised object detection with convex clustering," in *CVPR*, 2015.
[29] R. G. Cinbis, J. Verbeek, and C. Schmid, "Weakly supervised object localization with multi-fold multiple instance learning," *TPAMI*, 2017.
[30] H. O. Song, Y. J. Lee, S. Jegelka, and T. Darrell, "Weakly-supervised discovery of visual pattern configurations," in *NIPS*, 2014.
[31] C. Wang, W. Ren, K. Huang, and T. Tan, "Weakly supervised object localization with latent category learning," in *ECCV*, 2014.
[32] X. Wang, Z. Zhu, C. Yao, and X. Bai, "Relaxed multiple-instance svm with application to object discovery," in *ICCV*, 2015.
[33] M. P. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in *NIPS*, 2010.
[34] P. Siva, C. Russell, and T. Xiang, "In defence of negative mining for annotating weakly labelled data," in *ECCV*, 2012.
[35] P. Siva and T. Xiang, "Weakly supervised object detector learning with model drift detection," in *ICCV*, 2011.
[36] W. Ge, S. Yang, and Y. Yu, "Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning," in *CVPR*, 2018.
[37] J. Wang, J. Yao, Y. Zhang, and R. Zhang, "Collaborative learning for weakly supervised object detection," in *IJCAI*, 2018.
[38] Z. Ren, Z. Yu, X. Yang, M.-Y. Liu, A. G. Schwing, and J. Kautz, "Ufo 2: A unified framework towards omni-supervised object detection," in *ECCV*, 2020.
[39] P. Chen, X. Yu, X. Han, N. Hassan, K. Wang, J. Li, J. Zhao, H. Shi, Z. Han, and Q. Ye, "Point-to-box network for accurate object detection via single point supervision," in *ECCV*, 2022.
[40] P. Tang, X. Wang, S. Bai, W. Shen, X. Bai, W. Liu, and A. Yuille, "Pcl: Proposal cluster learning for weakly supervised object detection," *TPAMI*, 2018.
[41] Z. Zeng, B. Liu, J. Fu, H. Chao, and L. Zhang, "Wsod2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection," in *ICCV*, 2019.
[42] Y. Gao, B. Liu, N. Guo, X. Ye, F. Wan, H. You, and D. Fan, "C-midn: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection," in *ICCV*, 2019.
[43] Z. Ren, Z. Yu, X. Yang, M.-Y. Liu, Y. J. Lee, A. G. Schwing, and J. Kautz, "Instance-aware, context-focused, and memory-efficient weakly supervised object detection," in *CVPR*, 2020.
[44] J. Seo, W. Bae, D. J. Sutherland, J. Noh, and D. Kim, "Object discovery via contrastive learning for weakly supervised object detection," in *ECCV*, 2022.
[45] Y. Yin, J. Deng, W. Zhou, L. Li, and H. Li, "Cyclic-bootstrap labeling for weakly supervised object detection," in *ICCV*, 2023.
[46] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *IJCV*, 2013.
[47] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *CVPR*, 2016.
[48] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017.
[49] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, and Y. Wei, "Layercam: Exploring hierarchical class activation maps for localization," *TIP*, 2021.
[50] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy, "Speed/accuracy trade-offs for modern convolutional object detectors," in *CVPR*, 2017.
[51] T. Hazan, J. Keshet, and D. A. McAllester, "Direct loss minimization for structured prediction," in *NIPS*, 2010.
[52] Y. Song, A. Schwing, R. Urtasun *et al.*, "Training deep neural networks via direct loss minimization," in *ICML*, 2016.
[53] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *IJCV*, 2010.
[54] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.
[55] T. Deselaers, B. Alexe, and V. Ferrari, "Weakly supervised localization and learning with generic knowledge," *IJCV*, 2012.
[56] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.
[57] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
[58] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra, "Detecting twenty-thousand classes using image-level supervision," in *ECCV*, 2022.
[59] J. Lin, Y. Shen, B. Wang, S. Lin, K. Li, and L. Cao, "Weakly supervised open-vocabulary object detection," in *AAAI*, 2024.
[60] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *CVPR*, 2023.

Aditya Arun is a Ph.D. candidate in Computer Science and Engineering at IIIT Hyderabad, India. His areas of interest are computer vision, optimization methods, and machine learning.

C.V. Jawahar is an Amazon Chair Professor at IIIT Hyderabad, India. His areas of research include computer vision, machine learning, and document image analysis.

M. Pawan Kumar is a research scientist at DeepMind. Prior to that, he was a faculty member in the Department of Engineering Science at the University of Oxford during 2015 – 2021, where he led the OVAL group which focused on the design and analysis of optimization algorithms for problems arising in computer vision and machine learning. During 2012 – 2015, he was a faculty member in the Center for Visual Computing at Ecole Centrale Paris.