# Dissimilarity Coefficient based Weakly Supervised Object Detection


Aditya Arun[1], C.V. Jawahar[1], M. Pawan Kumar[2,3]

[1]CVIT, KCIS, IIIT, Hyderabad [2]University of Oxford [3]The Alan Turing Institute


CVPR LONG BEACH CALIFORNIA June 16-20, 2019

## 1. Aim

Localize objects with only image-level annotations at training time

## 2. Previous Works: Multiple Instance Learning (MIL)

**Standard Deep Approach:** Approximates MIL Objective[1]

Input

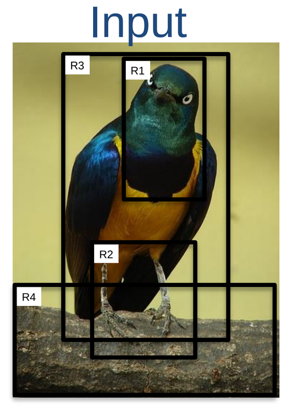


Image: $\mathbf{x}$
Label: $\mathbf{a}$ = {Bird}
Region Proposals: $\mathbf{B}$

CNN ($\theta$)

Classification Stream

| | R1 | R2 | R3 | R4 |
|---|---|---|---|---|
| Bird | 0.6 | 0.5 | 0.8 | 0.2 |
| BG | 0.4 | 0.5 | 0.2 | 0.8 |

Normalize over classes

Detection Stream

| | R1 | R2 | R3 | R4 |
|---|---|---|---|---|
| Bird | 0.1 | 0.1 | 0.7 | 0.1 |
| BG | 0.3 | 0.1 | 0.2 | 0.4 |

Normalize over regions

X

Output: $\mathbf{y}$ = {0, …, C}$^B$
C = Number of classes

| | R1 | R2 | R3 | R4 |
|---|---|---|---|---|
| Bird | 0.06 | 0.05 | 0.56 | 0.02 |
| BG | 0.12 | 0.05 | 0.04 | 0.32 |

+

| | |
|---|---|
| Bird | 0.69 |
| BG | 0.53 |

- Does not explicitly enforce **annotation constraint** - Each image-level annotation should have at least one corresponding region proposal
- Does not model uncertainty in the annotations

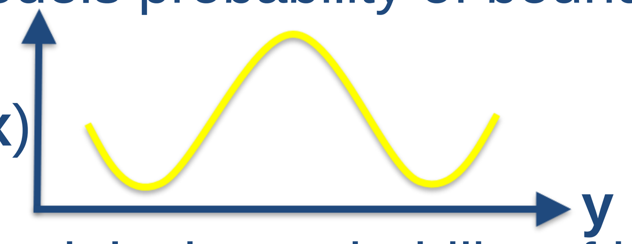
## 3. Overview

**Tasks:**

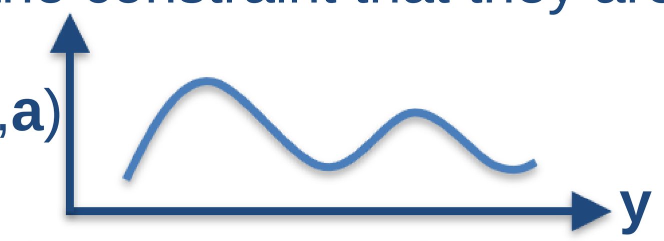
1. During inference, perform object detection
2. During training, model uncertainty over the bounding boxes such that it leverages the image-level annotations

**Two separate distributions for two tasks[6,7]:**

1. A **prediction distribution** that models probability of bounding box labels **y** given the input image **x**

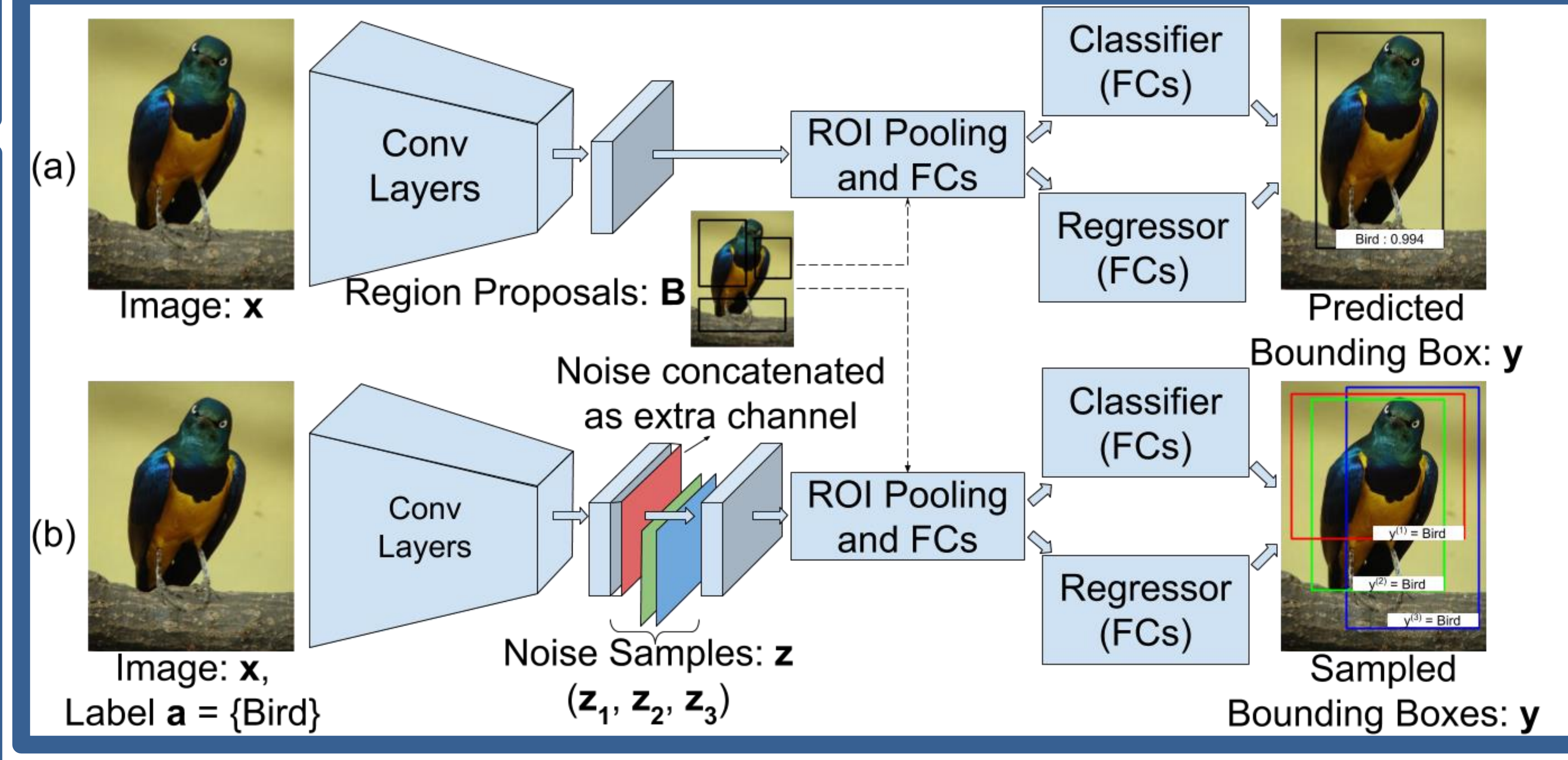



2. A **conditional distribution** that models the probability of bounding box labels **y** under the constraint that they are compatible with the annotation **a**




Ideally, the two distributions must match exactly

## 4. Architecture




**Prediction Net**
$\Pr_p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta}_p)$
$\boldsymbol{\theta}_p$ are the network parameters

**Conditional Net**
$\Pr_c(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{a};\boldsymbol{\theta}_c)$
$\boldsymbol{\theta}_c$ are the network parameters

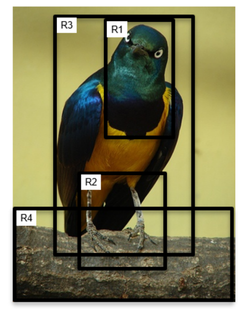
## 5. Modeling Conditional Distribution

**Objective:** Enforce annotation constraint

$$\Pr_c(\mathbf{y}|\mathbf{x},\mathbf{a};\boldsymbol{\theta}_c) = \prod_{i=1}^{|B|} \Pr(\mathbf{y}^{(i)}) \times H(\mathbf{y})$$

where,

$$H(\mathbf{y}) = \begin{cases} 1, \text{iff for each image-level label, there exists} \\ \text{at least one corresponding region proposal} \\ 0, \text{otherwise} \end{cases}$$



| | R1 | R2 | R3 | R4 |
|---|---|---|---|---|
| Bird | 1.3 | 1.5 | 1.8 | 0.2 |
| BG | 2.5 | 2.3 | 1.9 | 1.8 |
| | BG | BG | BG | BG |

Score for sample 1

| | R1 | R2 | R3 | R4 |
|---|---|---|---|---|
| Bird | 1.2 | 0.7 | 2.1 | 1.5 |
| BG | 1.7 | 1.6 | 0.9 | 1.6 |
| | BG | BG | Bird | BG |

R3

Score for sample 2

## 6. Optimization

**Task specific loss function:**

$$\Delta(\mathbf{y}_1,\mathbf{y}_2) = \frac{1}{|B|}\sum_{i=1}^{|B|} \Delta_{cls}(\mathbf{y}_1^{(i)},\mathbf{y}_2^{(i)}) + \Delta_{loc}(\mathbf{r}_1^{(i)},\mathbf{r}_2^{(i)})$$

We use $0-1$ loss for $\Delta_{cls}$ and *smoothL1* for $\Delta_{loc}$. $\mathbf{r}_1^{(i)}$ and $\mathbf{r}_2^{(i)}$ are the region proposal box corresponding to $\mathbf{y}_1^{(i)}$ and $\mathbf{y}_2^{(i)}$ respectively.

**Overall Objective:** Dissimilarity Coefficient Loss

$$DIV_\Delta(\Pr_p,\Pr_c) = \mathbb{E}_{\mathbf{y}_1\sim\Pr_p(\cdot)}[\mathbb{E}_{\mathbf{y}_2\sim\Pr_c(\cdot)}[\Delta(\mathbf{y}_1,\mathbf{y}_2)]]$$

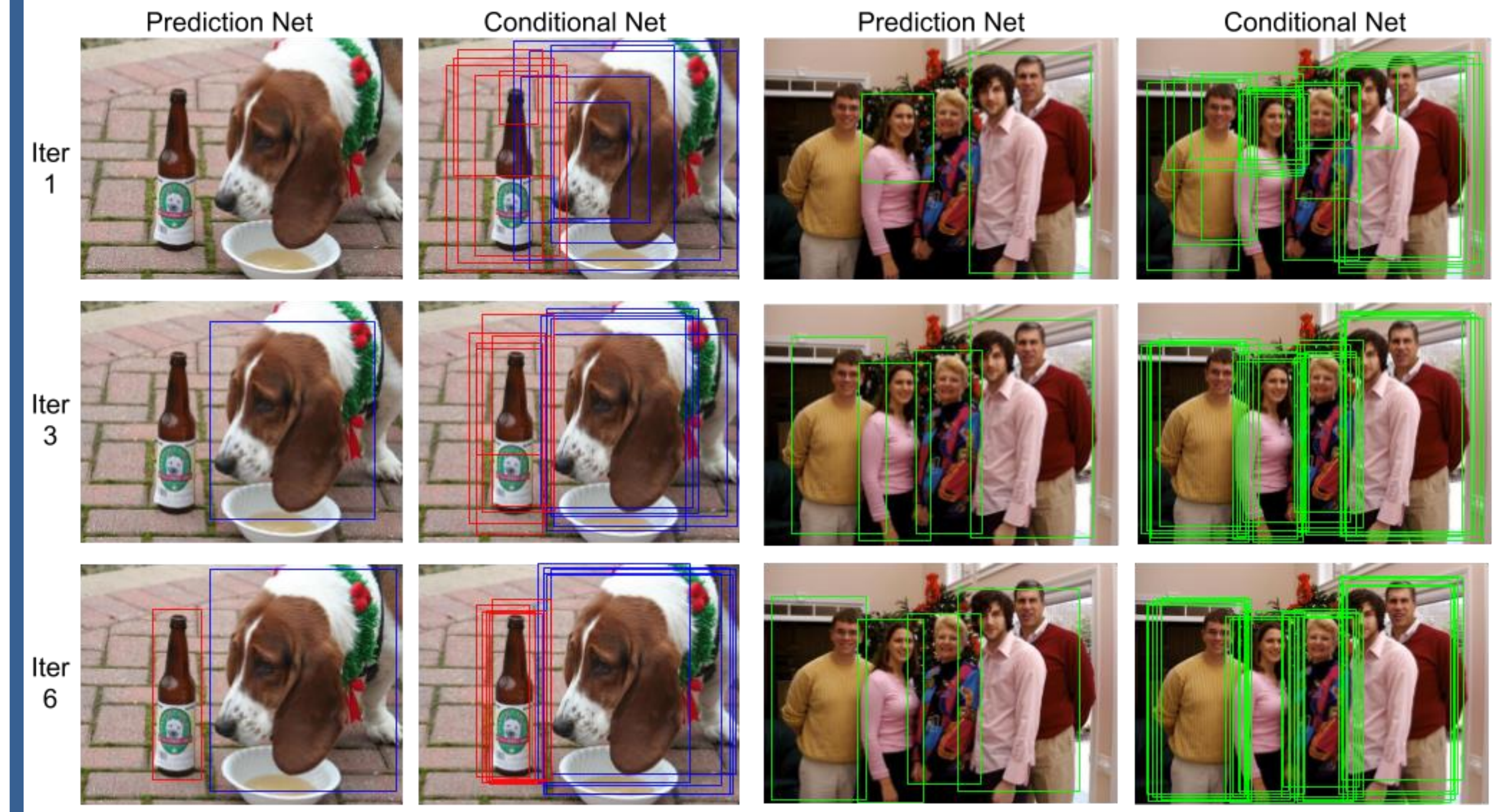
$$DISC_\Delta(\Pr_p,\Pr_c) = DIV_\Delta(\Pr_p,\Pr_c) - \gamma DIV_\Delta(\Pr_c,\Pr_c) - (1-\gamma)DIV_\Delta(\Pr_p,\Pr_p)$$

**Training:** Iterative training

- Fix one network and update the other network using SGD until convergence

## 7. Experiments and Results

### Visualization




### Results

| Method | mAP |
|---|---|
| WSDDN [1] | 39.3 |
| OICR [2] | 47.0 |
| W2F [3] | 52.4 |
| Ours | **53.6** |

VOC 2007

| Method | mAP |
|---|---|
| WSCCN [4] | 37.9 |
| OICR [2] | 42.5 |
| W2F [3] | 47.8 |
| Ours | **49.5** |

VOC 2012

| Method | mAP |
|---|---|
| WSDDN [1] | 11.5 |
| WSCCN [4] | 12.3 |
| ML-LocNet [5] | 16.2 |
| Ours | **17.7** |

COCO 2014

**Ablation Experiments:**

| Method | $\Pr_p, \Pr_c$ | $\Pr_p, PW_c$ | $PW_p, \Pr_c$ | $PW_p, PW_c$ |
|---|---|---|---|---|
| mAP | **52.9** | 50.1 | 52.6 | 49.5 |

## 8. References


[1] Bilen *et al*. Weakly supervised deep detection networks. In CVPR, 2016.
[2] Tang *et al*. Multiple instance detection network with online instance classifier refinement. In CVPR, 2017.
[3] Zhang *et al*. W2F: A weakly-supervised to fully-supervised framework for object detection. In CVPR, 2018.
[4] Diba *et al*. Weakly supervised cascaded convolutional networks. In CVPR, 2017.
[5] Zhang *et al*. ML-Locnet: Improving object localization with multi-view learning network. In ECCV, 2018.
[6] Arun *et al*. Learning human poses from actions. In BMVC, 2018.
[7] Kumar *et al*. Modeling latent variable uncertainty for loss-based learning. In ICML, 2012.