# Note: BBAM: Bounding Box Attribution Map for Weakly Supervised Semantic and Instance Segmentation

sgc

May 29, 2022

## 1 Introduction

This paper proposed a bounding box attribution map (BBAM), which can draw on the rich semantics learned by an object detector to produce pseudo GT for training semantic and instance segmentation networks. This method outperforms previous SOTA(2021) of WSSS and WSIS.

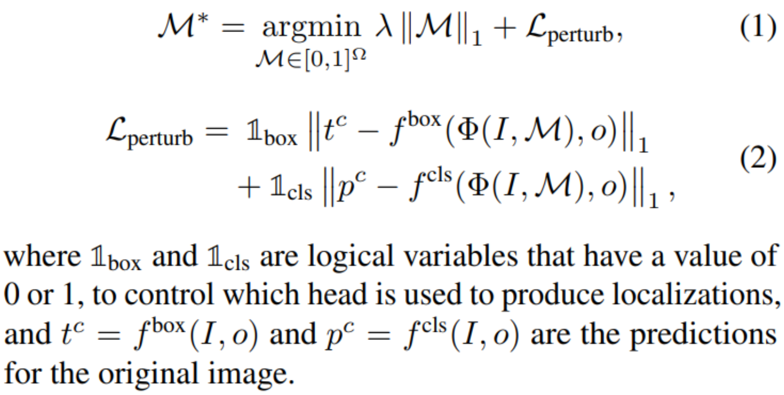
## 2 Method

### 2.1 Revisiting object detector

This paper concentrate on 2-stage detectors. The 2 stages are region proposal and box refinement. A region proposal network(RPN) generates candidate object proposals in the form of bounding boxes. And then refinement step is introduced. RoIpooling then the pooled feature map is passed to classification head and bounding box regression head.

### 2.2 Bounding Box Attribution Map

This paper implemented a mask which captures a subset of the image that produces almost the same prediction as the original image(indicates that these pixels are quite helpful). First construct a perturbation function: $\Phi(I, \mathcal{M}) = I \circ \mathcal{M} + \mu \circ (1 - \mathcal{M})$ where $\mu$ is the per-channel mean of the training data with the same size as $\mathcal{M}$.
For each proposal $o$, the best mask $\mathcal{M}$ is obtained by optimizing the following function using gradient descent with respect to $\mathcal{M}$:



$$\mathcal{M}^* = \underset{\mathcal{M}\in[0,1]^{\Omega}}{\operatorname{argmin}} \lambda \left\|\mathcal{M}\right\|_1 + \mathcal{L}_{\text{perturb}}, \tag{1}$$

$$\begin{aligned}\mathcal{L}_{\text{perturb}} = \; & \mathbb{1}_{\text{box}} \left\|t^c - f^{\text{box}}(\Phi(I,\mathcal{M}),o)\right\|_1 \\ & + \mathbb{1}_{\text{cls}} \left\|p^c - f^{\text{cls}}(\Phi(I,\mathcal{M}),o)\right\|_1 ,\end{aligned} \tag{2}$$

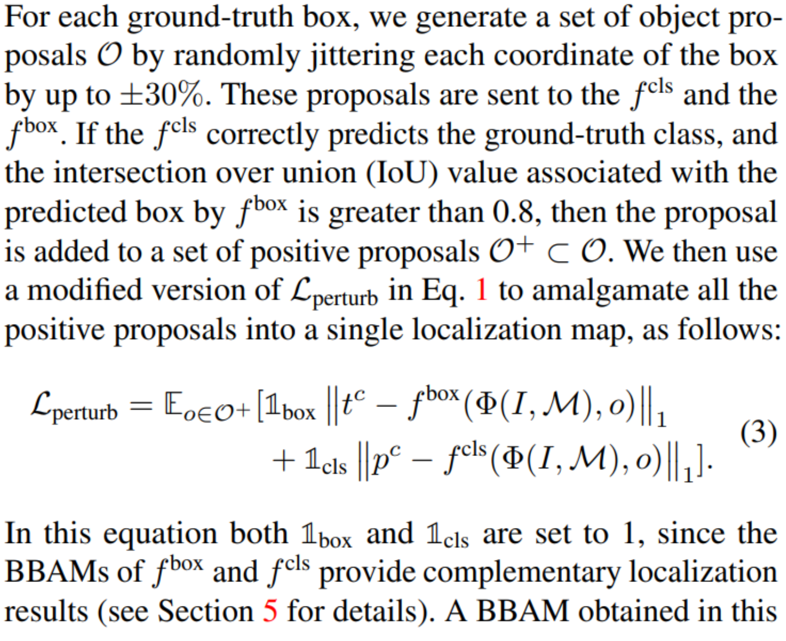
where $\mathbb{1}_{\text{box}}$ and $\mathbb{1}_{\text{cls}}$ are logical variables that have a value of 0 or 1, to control which head is used to produce localizations, and $t^c = f^{\text{box}}(I,o)$ and $p^c = f^{\text{cls}}(I,o)$ are the predictions for the original image.

And to deal with the problem that using a mask of the same spatial size as the input image incurs undesirable artifacts due to the adversarial effect, a stride $s$ is introduced to address it so multiple pixels are disturbed.(Upsample then use the perturbation function)

And authors show that a fixed stride fail to match the sizes of objects, they proposed an adaptive stride.(small stride for small object and vice versa)

## 2.3 Generating Pseudo GT

**Creating masks**



For each ground-truth box, we generate a set of object proposals $\mathcal{O}$ by randomly jittering each coordinate of the box by up to $\pm 30\%$. These proposals are sent to the $f^{\text{cls}}$ and the $f^{\text{box}}$. If the $f^{\text{cls}}$ correctly predicts the ground-truth class, and the intersection over union (IoU) value associated with the predicted box by $f^{\text{box}}$ is greater than 0.8, then the proposal is added to a set of positive proposals $\mathcal{O}^+ \subset \mathcal{O}$. We then use a modified version of $\mathcal{L}_{\text{perturb}}$ in Eq. 1 to amalgamate all the positive proposals into a single localization map, as follows:

$$\begin{aligned}\mathcal{L}_{\text{perturb}} = \mathbb{E}_{o\in\mathcal{O}^+}[ & \mathbb{1}_{\text{box}} \left\|t^c - f^{\text{box}}(\Phi(I,\mathcal{M}),o)\right\|_1 \\ & + \mathbb{1}_{\text{cls}} \left\|p^c - f^{\text{cls}}(\Phi(I,\mathcal{M}),o)\right\|_1].\end{aligned} \tag{3}$$

In this equation both $\mathbb{1}_{\text{box}}$ and $\mathbb{1}_{\text{cls}}$ are set to 1, since the BBAMs of $f^{\text{box}}$ and $f^{\text{cls}}$ provide complementary localization results (see Section 5 for details). A BBAM obtained in this

Then create pseudo instance-level ground-truth masks by considering the pixels in each BBAM with values greater than a threshold $\theta$ to be foreground. The proposals $\mathcal{O}$ is generated through jittering the GT boxes (Given or from RPN). Notably in this paper authors use 2 thresholds. pixels whose attribution values are higher than $\theta_{fg}$ are considered to be part of the foreground, and pixels whose values are lower than $\theta_{bg}$ are considered to be part of the background. The remaining pixels are ignored in the loss computations during training segmentation networks.

**Refine with MCG proposals** $\mathcal{T}$ is the proposal mask generated in the GT generating. First select the mask proposal that has the highest IoU with $\mathcal{T}$. However, that proposal may partially cover the target object. therefore consider other proposals that are completely contained within $\mathcal{T}$.

### 2.4 Training the Segmentation Network

## 3 Results

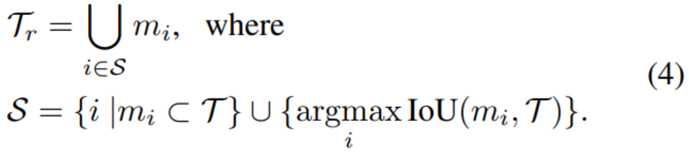
In weakly supervised instance segmentation the proposed method achieved better performance that previous best performing method.
And in WSSS, if fair compared authors think their proposed method work better than other methods.

## 4 Ablation Study

1, MCG proposals helps
2, in BBAM, either the box head ($\nVdash_{box} = 1\textbf{and}\nVdash_{cls} = 1$) or the cls head ($\nVdash_{box} = 0\textbf{and}\nVdash_{cls} = 1$) shows competent performance, but the best performance is achieved when the two heads are used together.

## 5 Summary

This paper built a method based on the 2-stage object detector using cls and box heads. Experiments shows that the BBAM achieves SOTA performance.



$$\mathcal{T}_r = \bigcup_{i \in \mathcal{S}} m_i, \quad \text{where} \qquad (4)$$
$$\mathcal{S} = \{i \,|m_i \subset \mathcal{T}\} \cup \{\underset{i}{\operatorname{argmax}} \, \mathbf{IoU}(m_i, \mathcal{T})\}.$$

## References