# Background-Aware Pooling and Noise-Aware Loss for Weakly-Supervised Semantic Segmentation

Youngmin Oh, Beomjun Kim, and Bumsub Ham
Yonsei University

CVPR VIRTUAL JUNE 19-25

## Summary

### Problem statement

- Object bounding boxes provide a definite background along with the extent of each object, but each box contains a mixture of foreground and background
- Goal: train segmentation models with object bounding boxes

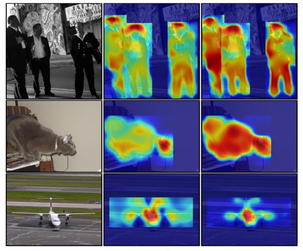
### Motivation

- How can we generate high-quality pseudo pixel-level labels from object bounding boxes?
- How can we train segmentation models with noisy labels?

### Contributions

- Introduce a simple yet effective framework which mainly consists of three stages: (1) train a CNN for image classification; (2) generate pseudo pixel-level labels; (3) train a CNN for semantic segmentation
- Propose a background-aware pooling (BAP) layer that leverages a background prior to separate foreground and background regions inside object bounding boxes
- Introduce a noise-aware loss (NAL) that alleviates the influence of incorrect labels adaptively
- Demonstrate state-of-the-art performance on PASCAL VOC 2012 and MS-COCO

## Comparison between GAP and BAP



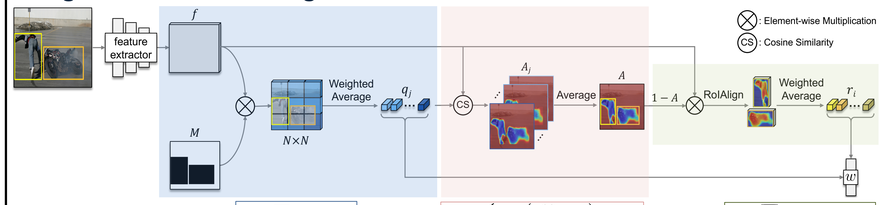
Input. GAP. BAP.

| Method | *train* | *val* |
|---|---|---|
| *Supervision*: Image-level labels (10K) | | |
| AffinityNet (CVPR 2018) | 59.7 | - |
| *Supervision*: Boxes (10K) | | |
| Box | 65.4 | 62.2 |
| GrabCut (SIGGRAPH 2004) | 65.7 | 66.1 |
| MCG (PAMI 2016) | 66.2 | 66.9 |
| WSSL (ICCV 2015) | 69.7 | 71.1 |
| Ours | | |
| GAP | 75.5 | 76.1 |
| BAP: $Y_{\text{crf}}$ w/o $u_0$ | 77.0 | 77.8 |
| BAP: $Y_{\text{crf}}$ | **78.7** | **79.2** |
| BAP: $Y_{\text{ret}}$ | 70.8 | 69.9 |

mIoUs of pseudo labels on PASCAL VOC 2012

## Stage 1

### Image classification using BAP




$G(j)$: $j$-th grid cell, $(1 \le j \le N^2)$
$J$ is the number of valid grid cells
$\mathcal{B} = \{B_1, B_2, \ldots, B_K\}$

$$q_j = \frac{\sum_{\mathbf{p}\in G(j)} M(\mathbf{p}) f(\mathbf{p})}{\sum_{\mathbf{p}\in G(j)} M(\mathbf{p})}$$

<queries>

$$A_j(\mathbf{p}) = \begin{cases} \text{ReLU}\left(\frac{f(\mathbf{p})}{\|f(\mathbf{p})\|} \cdot \frac{q_j}{\|q_j\|}\right) & , \mathbf{p} \in \mathcal{B} \\ 1 & , \mathbf{p} \notin \mathcal{B} \end{cases}$$

$$A(\mathbf{p}) = \frac{1}{J}\sum_j A_j(\mathbf{p})$$

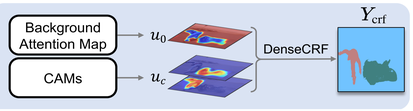
<background attention map>

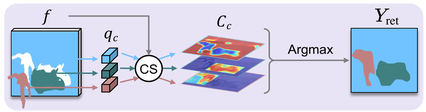
$$r_i = \frac{\sum_{\mathbf{p}\in B_i} (1 - A(\mathbf{p})) f(\mathbf{p})}{\sum_{\mathbf{p}\in B_i} (1 - A(\mathbf{p}))}$$

<foreground vectors>

## Stage 2

### Pseudo label generation




$$u_0(\mathbf{p}) = A(\mathbf{p}) \qquad \text{CAM}_c(\mathbf{p}) = \text{ReLU}(f(\mathbf{p}) \cdot w_c)$$

$$u_c(\mathbf{p}) = \begin{cases} \frac{\text{CAM}_c(\mathbf{p})}{\max_{\mathbf{p}}(\text{CAM}_c(\mathbf{p}))} & , \mathbf{p} \in \mathcal{B}_c \\ 0 & , \mathbf{p} \notin \mathcal{B}_c \end{cases}$$

$$q_c = \frac{1}{|\mathcal{Q}_c|}\sum_{\mathbf{p}\in\mathcal{Q}_c} f(\mathbf{p}) \qquad C_c(\mathbf{p}) = \frac{f(\mathbf{p})}{\|f(\mathbf{p})\|} \cdot \frac{q_c}{\|q_c\|}$$

$Q_c$ : a set of locations labeled as a class $c$ in $Y_{\text{crf}}$

## Stage 3

### The NAL is defined as follows: $\mathcal{L} = \mathcal{L}_{\text{ce}} + \lambda \mathcal{L}_{\text{wce}}$

$$\mathcal{L}_{\text{ce}} = -\frac{1}{\sum_c |\mathcal{S}_c|}\sum_c \sum_{\mathbf{p}\in\mathcal{S}_c} \log H_c(\mathbf{p})$$

$$H_c(\mathbf{p}) = \frac{e^{\tau \frac{\phi(\mathbf{p})}{\|\phi(\mathbf{p})\|} \cdot \frac{W_c}{\|W_c\|}}}{\sum_i e^{\tau \frac{\phi(\mathbf{p})}{\|\phi(\mathbf{p})\|} \cdot \frac{W_i}{\|W_i\|}}}$$

$H_c$ : a probability for a class $c$
$S_c$ : a set of locations where both $Y_{\text{crf}}$ and $Y_{\text{ret}}$ give the same label $c$

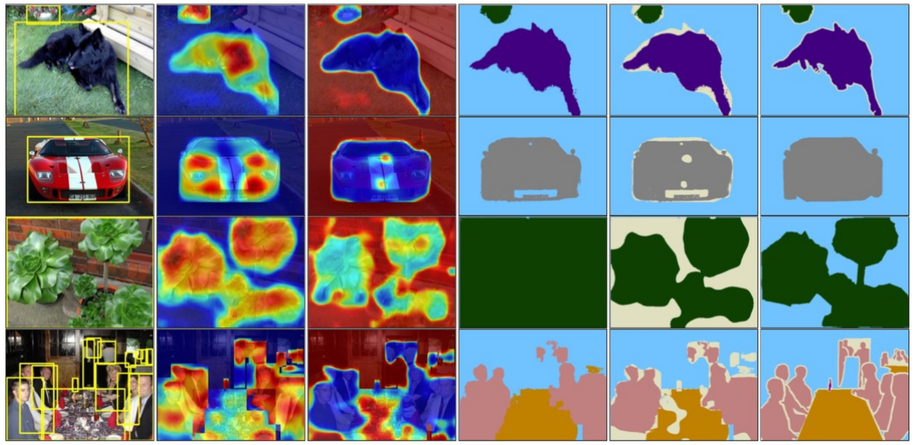
$$\mathcal{L}_{\text{wce}} = -\frac{1}{\sum_c \sum_{\mathbf{p}\in\sim\mathcal{S}_c} \sigma(\mathbf{p})}\sum_c \sum_{\mathbf{p}\in\sim\mathcal{S}_c} \sigma(\mathbf{p}) \log H_c(\mathbf{p})$$

$$D_c(\mathbf{p}) = 1 + \left(\frac{\phi(\mathbf{p})}{\|\phi(\mathbf{p})\|} \cdot \frac{W_c}{\|W_c\|}\right)$$

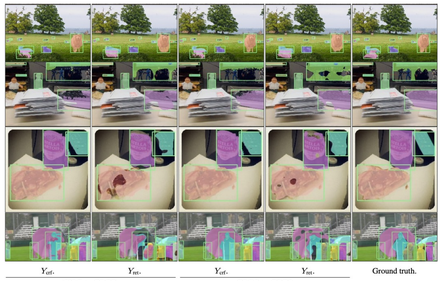
$$\sigma(\mathbf{p}) = \left(\frac{D_{c^*}(\mathbf{p})}{\max_c(D_c(\mathbf{p}))}\right)^{\gamma}$$

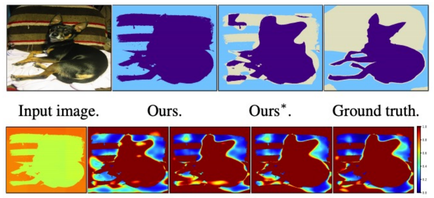
$\sigma$ : a confidence map where values approach to one when the label $c^* = Y_{\text{crf}}(\mathbf{p})$ is confident
$\sim Sc$ : a set of locations where $Y_{\text{crf}}$ and $Y_{\text{ret}}$ give different labels

## Experiments



Input image. $u_c$. $u_0$. Ours. Ours*. Ground truth.



Input image. Ours. Ours*. Ground truth.

Confidence maps $\sigma$.

| Method | *val* | *test* |
|---|---|---|
| *Supervision*: Image-level labels (10K) with Saliency (3K) | | |
| SeeNet (NIPS 2018) | 61.1 | 60.7 |
| FickleNet (CVPR 2019) | 61.2 | 61.9 |
| OAA (ICCV 2019) | 63.1 | 62.8 |
| ICD (CVPR 2020) | 64.0 | 63.9 |
| *Supervision*: Boxes (10K) | | |
| BoxSup (ICCV 2015) | 62.0 | 64.6 |
| WSSL (ICCV 2015) | 60.6 | 62.2 |
| SDI (CVPR 2017) | 65.7 | 67.5 |
| BCM (CVPR 2019) | 66.8 | - |
| Ours | | |
| w/ $Y_{\text{crf}}$ | 67.8 | - |
| w/ $Y_{\text{ret}}$ | 66.1 | - |
| w/ NAL | **68.1** | **69.4** |
| *Supervision*: Boxes (9K) with Masks (1K) | | |
| BoxSup (ICCV 2015) | 63.5 | 66.2 |
| WSSL (ICCV 2015) | 65.1 | 66.6 |
| SDI (CVPR 2017) | 65.8 | 66.9 |
| BCM (CVPR 2019) | 67.5 | - |
| Ours w/ NAL | **70.5** | **71.5** |

mIoUs of DeepLab-V1 on PASCAL VOC 2012