Inferring the Class Conditional Response Map for Weakly Supervised Semantic Segmentation

Weixuan Sun, Jing Zhang, Nick Barnes
Australian National University
{weixuan.sun, jing.zhang, nick.barnes}@anu.edu.au

1. Introduction

In the supplementary material, we summarize the contents as follows: 1) We show more qualitative results of our proposed method on PASCAL VOC. 2) We add an algorithm graph of our proposed method. 3) We perform an ablation about how different thresholds affect the performance of the iterative inference module. 4) We report results on the more challenging COCO dataset [1] to further validate effectiveness and generality of our method.

2. Further Qualitative Results

Object Activation Map
Refer to Fig. 1, we demonstrate more qualitative results of our proposed method. In the main paper, we argue that the baseline classifier not only focuses on the discriminative areas, but also generates activation on other object areas. However, the activation distribution is very uneven, and highly discriminative areas will suppress the activation of the other areas. We investigate this issue, and our proposed method can effectively shift the activation to densely cover larger object areas. For example, as shown in the first column of Fig. 1, the activation on the cat head is effectively shifted to the entire cat body.

Object Activation for Each Class.
In addition, due to page limit and for the convenience of read, we combine object activation of all classes together in our qualitative results. Our method can generate better object activation on all classes. As shown in Fig. 2. We generate dense activation covering larger objects areas on both horse and human. Note that we can even detect very inconspicuous objects in the background.

Qualitative Results of Activation Aware Mask Refinement
Refer to Fig. 3, column 2 and column 3 show our semantic segmentation results with/out our activation aware mask refinement respectively. In our method, we adaptively adopt saliency maps in our semantic segmentation training process to supervise the background channel. As shown, we obtain segmentation predictions with better object boundaries and successfully remove false positive predictions in the background.

3. Algorithm Graph

To further demonstrate the simplicity of our method, we show the algorithm graph of our inferring object response map method here. It is worth noting that iterative inference on each split can be implemented to run in parallel on GPUs.

Algorithm 1 Inferring Object Response Maps from a Baseline Classifier

```
Input: Image $i$, image-level label $C$, a fixed baseline classifier $f$;
Output: Object response map $A$;
Feed image $I$ and label $C$ into the fixed classifier $f$ to get baseline CAM and calculate the mass center. Split the image $i$ by the mass center to get 4 splits: $s_1$, $s_2$, $s_3$, $s_4$;
for $i$ in range 4 do:
    Iterative inference on $s_i$.
    Size of the $s_i$: $w$, $h$
    High activation areas $a_i = 0$.
    while True do:
        feed the split and label $C$ into the classifier to obtain response map:$m_i = f(s_i, C)$;
        New high activation region $\Delta a_i = (m_i > 0.7)$;
        Get high activation area $a_i = a_i + \Delta a_i$;
        if $\Delta a_i < (0.01 * w * h)$ then:
            New high activation area is too small;
            Break (Stop iteration);
        else
            Get new split image by removing the high activation region $s_i = s_i - a_i$ ;
        end if
    end while
end for
Combine $a_1$, $a_2$, $a_3$, $a_4$ to obtain object activation map $A$ of the image $i$.
```

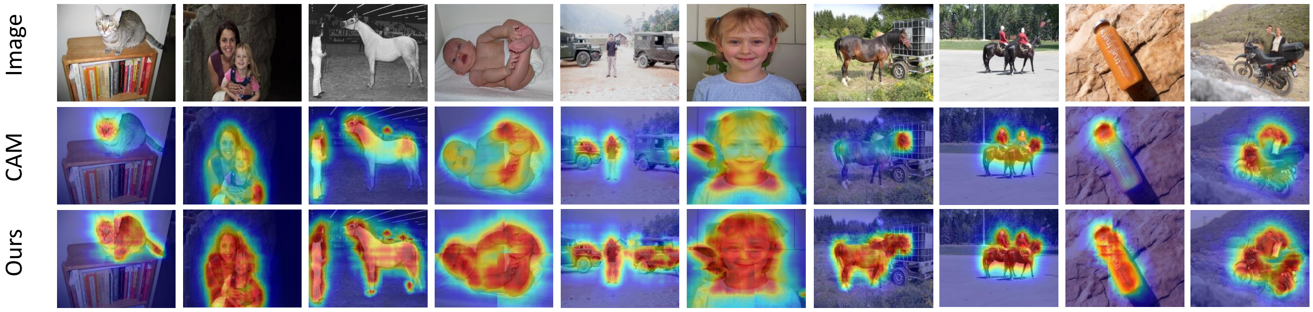

Figure 1. Further sample results of initial response maps on the PASCAL VOC dataset.

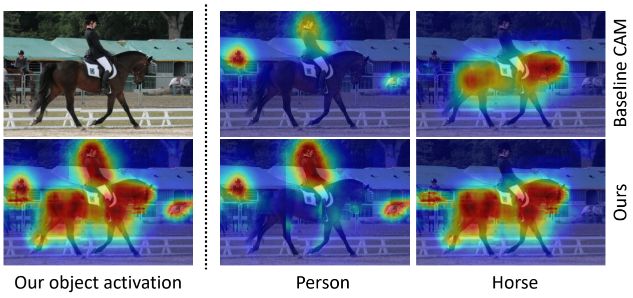

Figure 2. Our method can achieve better object activation for all classes in images with multiple classes. We show refined object activation for each class separately.

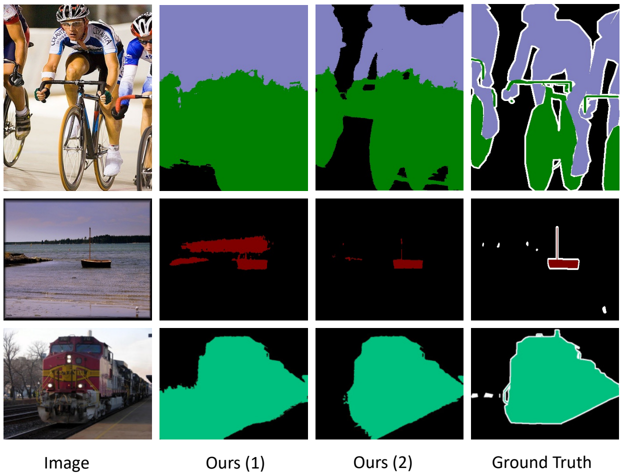

Figure 3. Ours(1) and ours(2) show our semantic segmentation results with/out activation aware mask refinement loss. By adopting activation aware mask refinement, our semantic segmentation predictions have more precise object shapes.

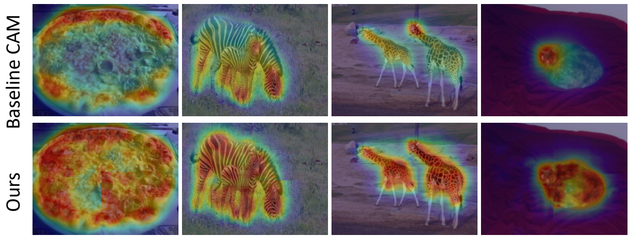

Figure 4. Sample results of initial response maps on COCO dataset.Our approach helps balance object activation across different object parts and densely cover larger object areas.

	baseline	0.4	0.5	0.6	0.7	0.8	0.9
mIoU	48.0	50.8	51.6	52.1	52.2	51.6	50.4

Table 1. Ablation study for high activation threshold for iterative inference.

4. Ablation: Iterative Inference Threshold

In our method section, we propose iterative inference to expand activation to a larger area of object. We select a hard threshold for high activation, *i.e.*, areas with activation higher than the threshold are removed from the original image to be used for the next iteration. In Table 1, we quantitatively evaluate how different thresholds affect the performance of our response map. As shown, our method performs robustly with a consistent improvement over the baseline, and achieves the best result with threshold equal to 0.7. We also investigate different methods to fill the high activation regions in the augmented image and find that using the mean value of the pixels in the original image gives best result. As a future work, we will explore how different types and sizes of removed areas influence the classification prediction and object activation.

5. Results on the COCO dataset

To further validate the effectiveness and generality of our proposed method. We test our method on the more challenging COCO dataset [1], which contains more complex scenes and more classes (80 classes and 1 background class) with 80k images for training and 40k images for validation. We follow the same the process we introduced in the main paper to obtain object response maps on COCO. Refering to the qualitative results in Fig. 4, our method can shift ob-

	Training Set (mIoU)
Baseline	29.4
Ours	29.8

Table 2. Performance comparison in mIoU(%) of the initial response maps on the COCO dataset.

ject activation to densely cover more object areas. In Table 2, we show quantitative results of our method on COCO, we compare the generated object response maps against the semantic segmentation groundtruth of the COCO dataset. Since most recent methods do not report results on COCO, we compare with the baseline CAM [2], and show a performance improvement, demonstrating that our method does generalize to other datasets.

In addition, since the COCO dataset has more challenging scenes with more classes, there are more images with many classes, and hence most objects in these images are smaller as a fraction of image size than on VOC. We observe a performance drop on these images and that is also the reason why the object response maps' performance on COCO is lower than that on the PASCAL VOC dataset. Meanwhile, most current methods aim to address the partial activation issue of the CAM on PASCAL VOC only, since most objects in the PASCAL VOC dataset are relatively large-scale, so CAMs on these objects only focus on discriminative regions. However, it raises a interesting question that CAMs [2] cannot perform well on the small-scale objects and complex scenes, instead of partial activation, small objects are always overly activated. We will explore this issue in the future work.

References

[1] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Eur. Conf. Comput. Vis.*, pages 740–755, 2014.

[2] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2921–2929, 2016.