Built-in Foreground/Background Prior for Weakly-Supervised Semantic Segmentation: Supplementary Material

Fatemehsadat Saleh[1,2], Mohammad Sadegh Ali Akbarian[1,2], Mathieu Salzmann[1,3], Lars Petersson[1,2], Stephen Gould[1], and Jose M. Alvarez[1,2]

[1]The Australian National University (ANU),
[2]Commonwealth Science and Industrial Research Organization (CSIRO),
[3]CVLab, EPFL, Switzerland

1 Foreground/Background Mask Evaluation

Here, we provide a validation and evaluation of our foreground/background masks. To this end, we made use of 10% of randomly chosen training images from the Pascal VOC dataset. We then generated foreground/background masks for these images using our approach, which relies on the activations of the fourth and fifth layers of the segmentation network pre-trained on ImageNet (i.e., before fine-tuning it for semantic segmentation). These masks can then be compared to ground-truth foreground/background masks obtained directly from the pixel level annotations.

We compare our masks with the objectness criterion of [1] and [2], which was employed by [3] and [4, 5] for the purpose of weakly-supervised semantic segmentation. Note that, some approaches such as [6, 2] which have been used for weakly-supervised semantic segmentation [4, 7, 5] require training data with pixel-level/bounding box annotations, and thus are not really comparable to our approach. Note also that a complete evaluation of objectness methods goes beyond the scope of this paper, which focuses on weakly-supervised semantic segmentation.

The objectness methods of [1] and [2] produce a per-pixel foreground probability map. For our comparison to be fair, we further refined these maps using the same dense CRF as in our approach. In Table 1, we provide the results of these experiments in terms of mean Intersection Over Union (mIOU) with respect to the ground-truth masks. Note that our masks are more accurate than those of [1, 2]. In Fig. 1, we show some qualitative results of these three approaches. Note that this further evidences the benefits of our foreground/background masks. In particular, our masks yield a much better object localization accuracy.

2 Evaluation of our CheckMask Procedure

In this section, we evaluate the quality of the masks selected using our CheckMask procedure. Recall that, in our CheckMask procedure, a user selects one

Table 1. Comparison of our foreground/background masks with those obtained using the objectness method of [1] and [2].

	Mean IoU
Masks obtained using [1]	52.34%
Masks obtained using [2]	50.20%
Our masks	**60.08**%

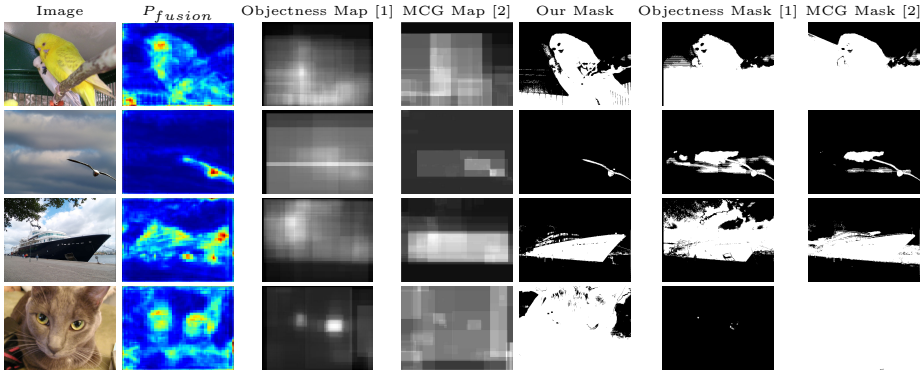

Fig. 1. Qualitative comparison of our masks with those of [1] and [2]. Note that our approach yields much better localization accuracy.

Table 2. CheckMask evaluation: Note that the masks selected by a user with our CheckMask procedure have similar accuracy to the best ones among the M candidates.

	Best Mask	Worst Mask	Random Mask	CheckMask
M=30	66.70	34.37	48.39	64.91

mask out of M candidates, such that, according to the user, the chosen mask best covers the objects of interest. In practice, we used $M = 30$, which was obtained by visually inspecting the candidates for 40 validation images. In Table 2, we report the mIOU of the selected masks w.r.t. the ground-truth masks for the training set. Note that a user might still make mistakes. We therefore also compare the accuracy of the user-selected masks with those obtained by choosing the best and the worst ones among the M candidates (according to the ground-truth) and with those obtained by a uniformly random selection. Note that the accuracy of our CheckMask procedure is very close to that of the best masks.

In practice, it takes the user roughly $2-3$ seconds per image when $M = 30$, as suggested in the paper. While lower values for M would require less annotation time per image, it might also suffer from the fact that none of the candidate accurately covers the objects of interest. To illustrate this, in Fig. 2, we visualize the best of user-selected masks for different values M for a few images. Our visual inspection suggested that $M = 30$ provides a good trade-off between speed and accuracy.

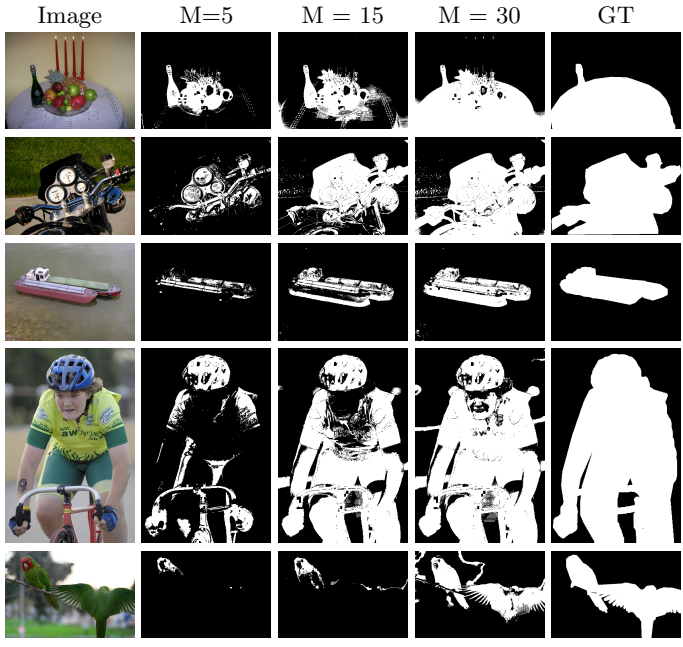

Fig. 2. Effect of M on the selected mask quality. While lower values of M will require less annotation time, the candidate masks do not always accurately cover the objects of interest.

References

1. Alexe, B., Deselaers, T., Ferrari, V.: Measuring the objectness of image windows. Pattern Analysis and Machine Intelligence, IEEE Transactions on **34**(11) (2012) 2189–2202
2. Arbeláez, P., Pont-Tuset, J., Barron, J., Marques, F., Malik, J.: Multiscale combinatorial grouping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2014) 328–335
3. Bearman, A., Russakovsky, O., Ferrari, V., Fei-Fei, L.: What's the point: Semantic segmentation with point supervision. ArXiv e-prints (2015)
4. Pinheiro, P.O., Collobert, R.: From image-level to pixel-level labeling with convolutional networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (June 2015)
5. Wei, Y., Liang, X., Chen, Y., Jie, Z., Xiao, Y., Zhao, Y., Yan, S.: Learning to segment with image-level annotations. Pattern Recognition (2016)
6. Cheng, M.M., Zhang, Z., Lin, W.Y., Torr, P.: Bing: Binarized normed gradients for objectness estimation at 300fps. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (June 2014)
7. Dai, J., He, K., Sun, J.: Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 1635–1643