Box2Seg: Attention Weighted Loss and Discriminative Feature Learning for Weakly Supervised Segmentation

Viveka Kulharia[⋆2], Siddhartha Chandra[*1],
Amit Agrawal[1], Philip Torr[2], Ambrish Tyagi[1]

[1]Amazon Lab126, [2]University of Oxford

We provide supplementary qualitative and quantitative results for our weakly supervised Box2Seg approach, as well as implementation details and results for three additional Grabcut based baselines. In Sec. 1, we show additional qualitative results demonstrating superior performance of our Box2Seg method compared to Box and Grabcut baselines. In Sec. 2, we provide implementation details and results for three Grabcut based baselines, which out-perform some of the previously published approaches. In Sec. 3, we show examples of noisy bounding-box annotations from the OpenImages dataset [2] which was used to pretrain one of our methods, **Box2Seg**(OI), in Table 1 of the main manuscript.

1 Qualitative results

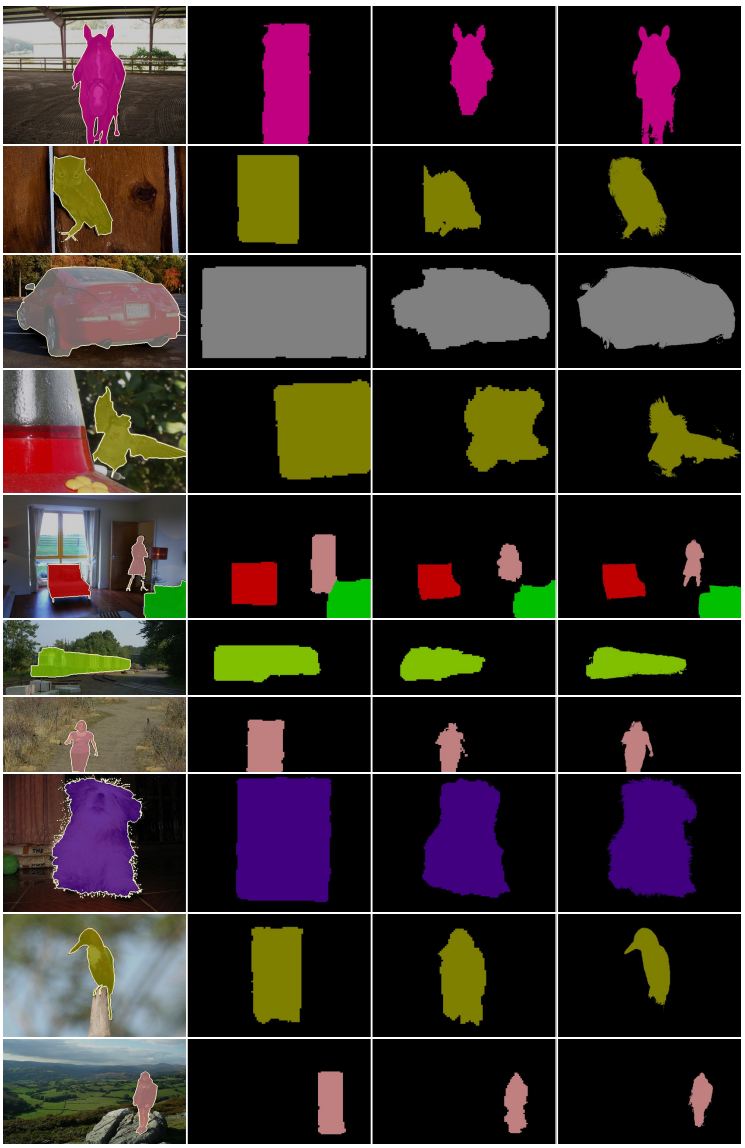
We show more qualitative results in Figure 1 comparing the outputs of our *Box2Seg* model against several baselines we considered in Table 2 of the main manuscript.

2 GrabCut Baselines: Implementation Details and Quantitative Results

We use the GrabCut method [3] to generate the pseudo ground-truth M (ref. to Sec. 3.1 in the manuscript for a more detailed description) for an input image I with bounding box annotations B_{box}. As described in Sec. 3.1 of the manuscript, $B_{box} \in \mathbb{R}^{K\times 5}$, where K is the number of bounding boxes in the image, each comprising of 4 coordinates and a class label. We use $grabCut$ function of the python library $cv2$ as mentioned in the Algorithm 1 to generate the pseudo ground-truth.

In Table 1, we evaluate the accuracy of this GrabCut algorithm itself, without any training, against the segmentation ground-truth. Interestingly, GrabCut output on *ground truth* bounding boxes (*GrabCut-NoTrain-GT*) results in a strong weakly-supervised baseline in itself with 71.6% mIoU. This is better compared to previous weakly supervised methods as shown in Table 1 of the main manuscript. However, since the ground truth bounding boxes are not available at inference, a more practical baseline is to obtain bounding boxes on the validation set using an object detector (we used SNIPER [4]) and then run GrabCut on those. This baseline is referred to as *GrabCut-NoTrain-Det* and obtains 68.5% mIoU. Furthermore, training our segmentation network with *GrabCut* masks as supervision ($\mathcal{L}_{GC}$ loss only, refer to Eqn. 2 in the main manuscript) without

⋆ Authors contributed equally. V. Kulharia was an intern at Amazon Lab126.



(a) Image+GT (b) Box BaseLine (c) GrabCut Baseline (d) Box2Seg

Fig. 1: Additional qualitative results: (a) Original Image with overlayed ground truth (b) Results with Box baseline (c) Results with GrabCut baseline (d) Box2Seg (with CRF). Please refer to Table 2 in the main manuscript for description of the baselines.

Algorithm 1: Python code to create pseudo ground-truth using GrabCut.

Data: I, H, W, B_{box}: image, image height, width, and bounding-boxes
Result: pseudo_ground_truth: M
Denote by $sorted\_bounding\_boxes \leftarrow B_{box}$ sorted from large to small;
pseudo_ground_truth = $0_{H \times W}$;
for $bbox$ *in* $sorted\_bounding\_boxes$ **do**
    x,y,w,h,$category\_id$ = $bbox$;
    mask = $0_{h \times w}$;
    mask[y : y+h , x : x+w] = 2;
    mask[y+int(h/10) : y+int(9*h/10) , x+int(w/10) : x+int(9*w/10)] = 3;
    mask[y+int(h/4) : y+int(3*h/4) , x+int(w/4) : x+int(3*w/4)] = 1;
    bgdModel = $0_{1 \times 65}$;
    fgdModel = $0_{1 \times 65}$;
    cv.grabCut(I,mask,None,bgdModel,fgdModel,5,cv.GC_INIT_WITH_MASK);
    mask = numpy.where((mask==2)||(mask==0),0,1).astype('uint8');
    mask[:y , :] = 0;
    mask[y+h: , :] = 0;
    mask[: , :x] = 0;
    mask[: , x+w:] = 0;
    pseudo_ground_truth[mask>0] = $category\_id$;
end

Method	**mIoU**
GrabCut-NoTrain-GT	71.6
GrabCut-NoTrain-Det	68.5
GrabCut	72.7

Table 1: Comparison of different methods using GrabCut pseudo ground-truths obtained with our Algorithm. We consider two *evaluation-only* baselines, *GrabCut-NoTrain-GT* and *GrabCut-NoTrain-Det*, where we evaluate accuracy of Grabcut based Segmentation (refer to Algorithm 1) on Ground Truth and Detected boxes respectively. We also consider a learning based baseline, *Grabcut* which is a CNN baseline (using our ResNet-101 architecture described in Sec. 4.1 of the manuscript) trained with our loss $\mathcal{L}_{GC}$ described in Eqn. 2 in the manuscript.

affinity or attention losses results in 72.7% mIOU. We expected the *GrabCut* baseline to be more accurate than the *NoTrain* baselines, as the *NoTrain* baselines do not harness the generalization capabilities of CNNs. We were quite surprised to find that previous SOTA papers ([1]) neglected to compare their work with such a trivial, albeit strong baseline. [5] has shown GrabCut results for comparison on the same dataset in its Figure 6 and they also seem quite competitive.

3 Erroneous Annotations in the OpenImages Dataset

We show examples of erroneous bounding box annotations present in the OpenImages dataset [2] in Fig. 2 and 3. We randomly sampled images from the dataset and picked a few examples demonstrating the three most common error modes: (i) bounding boxes which cover multiple object instances, (ii) non-exhaustive bounding boxes which miss objects of interest, and (iii) coarse bounding boxes which are not aligned with object boundaries, and either underestimate or overestimate object size. Case (i) causes inaccuracies in the outputs of the Grabcut algorithm [3] which is targeted for single object foreground extraction. Cases (ii) and (iii) invalidate our assumption that the pixels outside all bounding boxes can be considered *definite background*. Due to the presence of such errors in the bounding box annotations, our **A**ttention **W**eighted **L**oss, described in Sec. 3.3 of the main manuscript, has reduced efficacy. This also explains why we get a relatively minor improvement in performance when we pretrain our method with the OpenImages dataset [2] in Table 1 (of the main manuscript).

Fig. 2: Erroneous bounding box ground truth in the OpenImages dataset: Col. 1 shows ground truth bounding boxes that contain multiple instances inside. Col. 2 shows missing bounding boxes for the person class. Bounding-box annotations are not exhaustive in OpenImages.

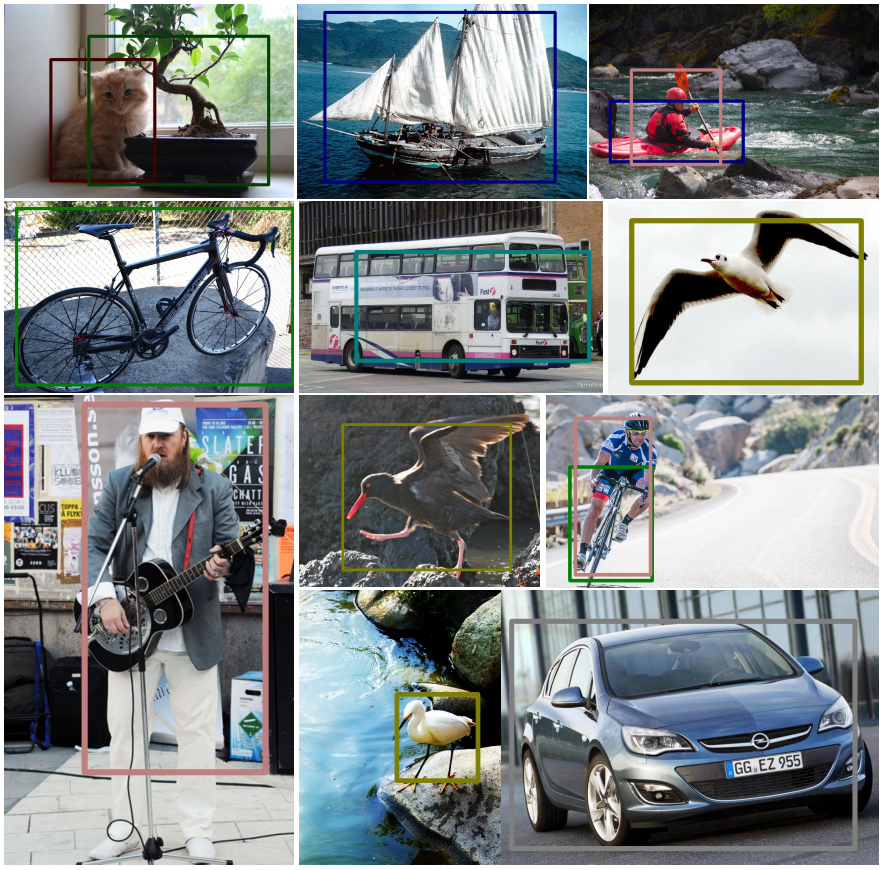


Fig. 3: Erroneous bounding box ground truth in the OpenImages dataset: representative examples from the dataset where bounding-boxes are not aligned with the object boundaries. Eg. the bounding box of the bicycle in row 2 col. 1 overestimates the object boundary on the upper-right corner, and the bounding box underestimates the legs of the person in row 3 col. 1.

References

1. Khoreva, A., Benenson, R., Hosang, J., Hein, M., Schiele, B.: Simple does it: Weakly supervised instance and semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 876–885 (2017) 3
2. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Duerig, T., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. arXiv preprint arXiv:1811.00982 (2018) 1, 4
3. Rother, C., Kolmogorov, V., Blake, A.: Grabcut: Interactive foreground extraction using iterated graph cuts. In: ACM transactions on graphics (TOG). vol. 23, pp. 309–314. ACM (2004) 1, 4
4. Singh, B., Najibi, M., Davis, L.S.: SNIPER: Efficient multi-scale training. Neural Information Processing Systems (NIPS) (2018) 1
5. Xu, N., Price, B., Cohen, S., Yang, J., Huang, T.: Deep grabcut for object selection. In: Proceedings of the British Machine Vision Conference (BMVC). pp. 182.1–182.12 (September 2017), https://dx.doi.org/10.5244/C.31.182 3