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

Jungbeom Lee[1] Jihun Yi[1] Chaehun Shin[1] Sungroh Yoon[1,2,]
[1] Department of Electrical and Computer Engineering, Seoul National University, Seoul, South Korea
[2] ASRI, INMC, ISRC, and Institute of Engineering Research, Seoul National University
{jbeom.lee93, t080205, chaehuny, sryoon}@snu.ac.kr

S1. Implementation details

TV norm. To suppress the artifacts in the mask $\mathcal{M}$, we regularized $\mathcal{M}$ with total variation (TV) norm in Eq. 1 in the main paper, as done in Fong *et al.* [4]. The resulting loss function to find the best $\mathcal{M}^*$ becomes:

$$\begin{aligned}\mathcal{L}_{\mathcal{M}} =& \lambda \left\|\mathcal{M}\right\|_1 + \lambda_{\text{TV}} \left\|\nabla\mathcal{M}\right\|_\beta^\beta \\ &+ \mathbb{1}_{\text{box}} \left\|t^c - f^{\text{box}}(\Phi(I,\mathcal{M}),o)\right\|_1 \\ &+ \mathbb{1}_{\text{cls}} \left\|p^c - f^{\text{cls}}(\Phi(I,\mathcal{M}),o)\right\|_1,\end{aligned} \tag{S1}$$

where λ_{TV} is a balancing factor for TV norm. We set λ_{TV} to 10^{-4} and β to 3. We observed that the resulting mask $\mathcal{M}^*$ has a little dependency on the value of λ_{TV}.

We can find the best $\mathcal{M}^*$ by using gradient descent with respect to $\mathcal{M}$. Letting the mask at iteration t be $\mathcal{M}^t$, the mask at iteration $t+1$ can be expressed as

$$\mathcal{M}^{t+1} = \mathcal{M}^t - \xi\nabla_{\mathcal{M}^t}\mathcal{L}_{\mathcal{M}^t}, \tag{S2}$$

where ξ is a learning rate. Indeed, the update in Eq S2 was implemented through Adam optimizer.

Optimization details for semantic segmentation. We used the default setting provided by [14], except for the batch size, the number of training iterations, and the learning rate. We set the batch size to 8, the number of training iterations to 2.4×10^4, and the learning rate to 2×10^{-4}.

Optimization details for instance segmentation on the PASCAL VOC dataset. Regarding the characteristics of the PASCAL VOC dataset [2], we adjusted the input image size and the anchor size accordingly. We set the max and min size of training images to 800 and 512, respectively, and anchor sizes for each FPN level to [21, 42, 84, 168, 332]. We trained Mask R-CNN [5] with a learning rate 8×10^{-3} for 2×10^4 iterations.

Optimization details for instance segmentation on the MS COCO 2017 dataset. We followed the default settings provided by maskrcnn-benchmark repository [13].

Post-processing of semantic and instance segmentation. CRF [8] is a popular post-processing technique for semantic and instance segmentation [6, 7, 9, 10, 17]. We also used CRFs as a post-processing method for semantic and instance segmentation.

S2. Additional Results

Comparison of per-class mIoU scores. Table S1 shows the per-class mIoU of our method and recently produced methods.

More examples of BBAMs. We present more examples of BBAMs for PASCAL VOC [2] validation images with Faster R-CNN [15] (Figure S1) and for MS COCO 2017 [12] validation images with Faster R-CNN [15] (Figure S2).

Additional mask examples on semantic segmentation. Figure S3 shows more examples of the semantic masks produced by DSRG [7], Shen *et al.* [16], FickleNet [9], Lee *et al.* [10], and our method.

More mask examples on instance segmentation. Figure S4 shows more examples of the instance masks on PASCAL VOC 2012 validation images obtained from IRNet [1], Hsu *et al.* [6], and our method. Figure S5 shows examples of instance masks on MS COCO 2017 validation images obtained by our method.

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *CVPR*, 2019. 1
[2] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 1, 3
[3] Junsong Fan, Zhaoxiang Zhang, and Tieniu Tan. Cian: Cross-image affinity net for weakly supervised semantic segmentation. *AAAI*, 2020. 2
[4] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *ICCV*, 2017. 1
[5] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 1
[6] Cheng-Chun Hsu, Kuang-Jui Hsu, Chung-Chi Tsai, Yen-Yu Lin, and Yung-Yu Chuang. Weakly supervised instance seg-

	bkg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mIOU
Results on validation images:																						
Shen *et al.* [16]	86.8	71.2	32.4	77.0	24.4	69.8	85.3	71.9	86.5	27.6	78.9	40.7	78.5	79.1	72.7	73.1	49.6	74.8	36.1	48.1	59.2	63.0
CIAN [3]	88.2	79.5	32.6	75.7	56.8	72.1	85.3	72.9	81.7	27.6	73.3	39.8	76.4	77.0	74.9	66.8	46.6	81.0	29.1	60.4	53.3	64.3
FickleNet [9]	89.5	76.6	32.6	74.6	51.5	71.1	83.4	74.4	83.6	24.1	73.4	47.4	78.2	74.0	68.8	73.2	47.8	79.9	37.0	57.3	64.6	64.9
SSDD [17]	89.0	62.5	28.9	83.7	52.9	59.5	77.6	73.7	87.0	34.0	83.7	47.6	84.1	77.0	73.9	69.6	29.8	84.0	43.2	68.0	53.4	64.9
Lee *et al.* [10]	90.8	82.2	35.1	82.4	72.2	71.4	82.7	75.0	86.9	18.3	74.2	29.6	81.1	79.2	74.7	76.4	44.2	78.6	35.4	72.8	63.0	66.5
AdvCAM [11]	90.0	79.8	34.1	82.6	63.3	70.5	89.4	76.0	87.3	31.4	81.3	33.1	82.5	80.8	74.0	72.9	50.3	82.3	42.2	74.1	52.9	68.1
BBAM (Ours)	92.7	80.6	33.8	83.7	64.9	75.5	91.3	80.4	88.3	37.0	83.3	62.5	84.6	80.8	74.7	80.0	61.6	84.5	48.6	85.8	71.8	73.7
Results on test images:																						
Shen *et al.* [16]	87.2	76.8	31.6	72.9	19.1	64.9	86.7	75.4	86.8	30.0	76.6	48.5	80.5	79.9	79.7	72.6	50.1	83.5	48.3	39.6	52.2	63.9
FickleNet [9]	90.3	77.0	35.2	76.0	54.2	64.3	76.6	76.1	80.2	25.7	68.6	50.2	74.6	71.8	78.3	69.5	53.8	76.5	41.8	70.0	54.2	65.0
SSDD [17]	89.0	62.5	28.9	83.7	52.9	59.5	77.6	73.7	87.0	34.0	83.7	47.6	84.1	77.0	73.9	69.6	29.8	84.0	43.2	68.0	53.4	64.9
Lee *et al.* [10]	91.2	84.2	37.9	81.6	53.8	70.6	79.2	75.6	82.3	29.3	76.2	35.6	81.4	80.5	79.9	76.8	44.7	83.0	36.1	74.1	60.3	67.4
AdvCAM [11]	90.1	81.2	33.6	80.4	52.4	66.6	87.1	80.5	87.2	28.9	80.1	38.5	84.0	83.0	79.5	71.9	47.5	80.8	59.1	65.4	49.7	68.0
BBAM (Ours)	92.8	83.5	33.4	88.9	61.8	72.8	90.3	83.5	87.6	34.7	82.9	66.1	83.9	81.1	78.3	77.4	55.2	86.7	58.5	81.5	66.4	73.7

Table S1: Comparison of per-class mIoU scores.

mentation using the bounding box tightness prior. In *NeurIPS*, 2019. 1

[7] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *CVPR*, 2018. 1, 5

[8] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NeurIPS*, 2011. 1

[9] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *CVPR*, 2019. 1, 2, 5

[10] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Frame-to-frame aggregation of active regions in web videos for weakly supervised semantic segmentation. In *ICCV*, 2019. 1, 2, 5

[11] Jungbeom Lee, Eunji Kim, and Sungroh Yoon. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. *arXiv preprint arXiv:2103.08896*, 2021. 2

[12] 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 *ECCV*, 2014. 1, 4

[13] Francisco Massa and Ross Girshick. maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch. https://github.com/facebookresearch/maskrcnn-benchmark, 2018. 1

[14] Kazuto Nakashima. DeepLab with PyTorch. https://github.com/kazuto1011/deeplab-pytorch. 1

[15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 1, 3, 4

[16] Tong Shen, Guosheng Lin, Chunhua Shen, and Ian Reid. Bootstrapping the performance of webly supervised semantic segmentation. In *CVPR*, 2018. 1, 2, 5

[17] Wataru Shimoda and Keiji Yanai. Self-supervised difference detection for weakly-supervised semantic segmentation. In *ICCV*, 2019. 1, 2

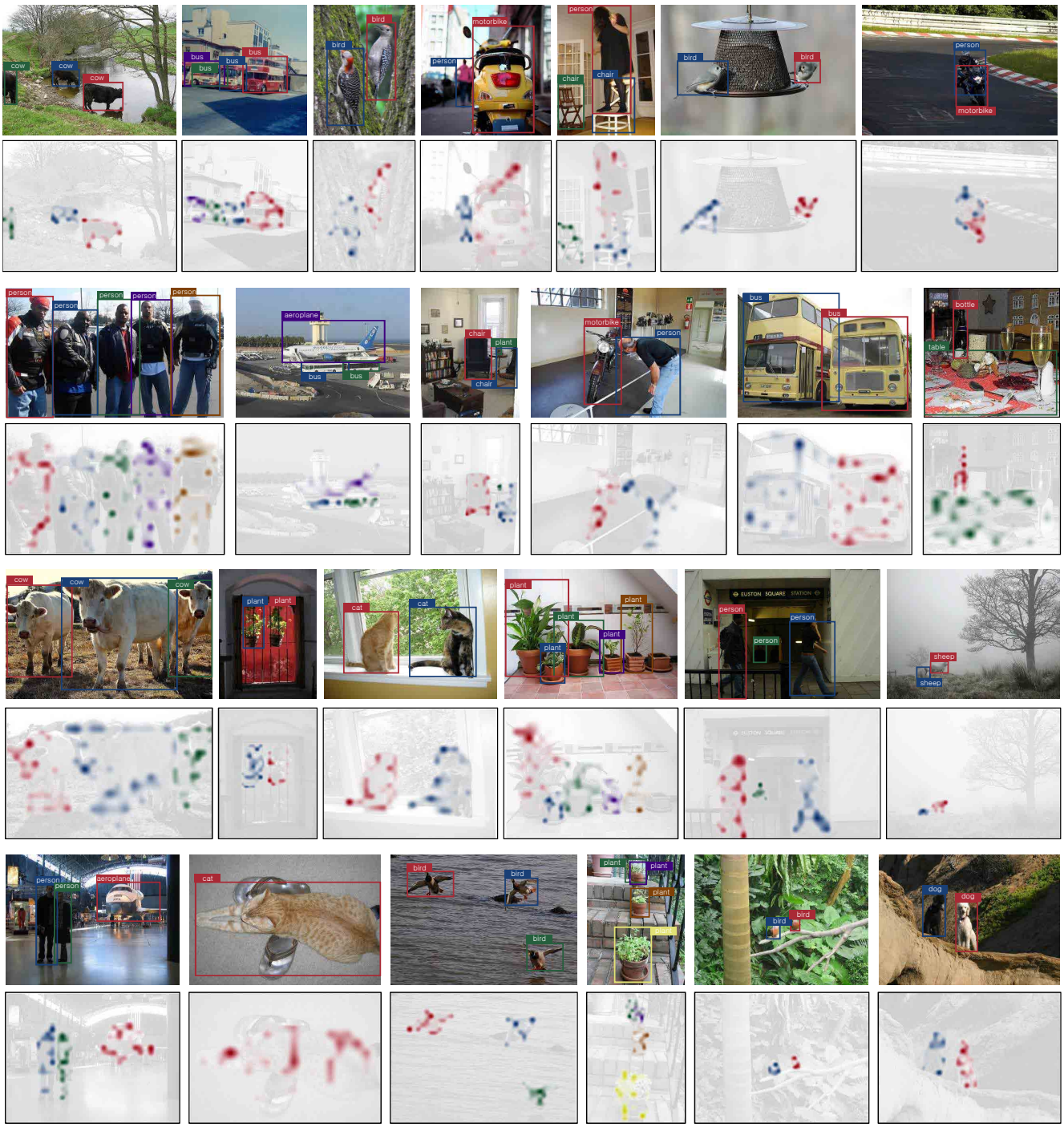

Figure S1: Examples of PASCAL VOC [2] validation images with the results of object detection and corresponding BBAMs, obtained from Faster R-CNN [15].

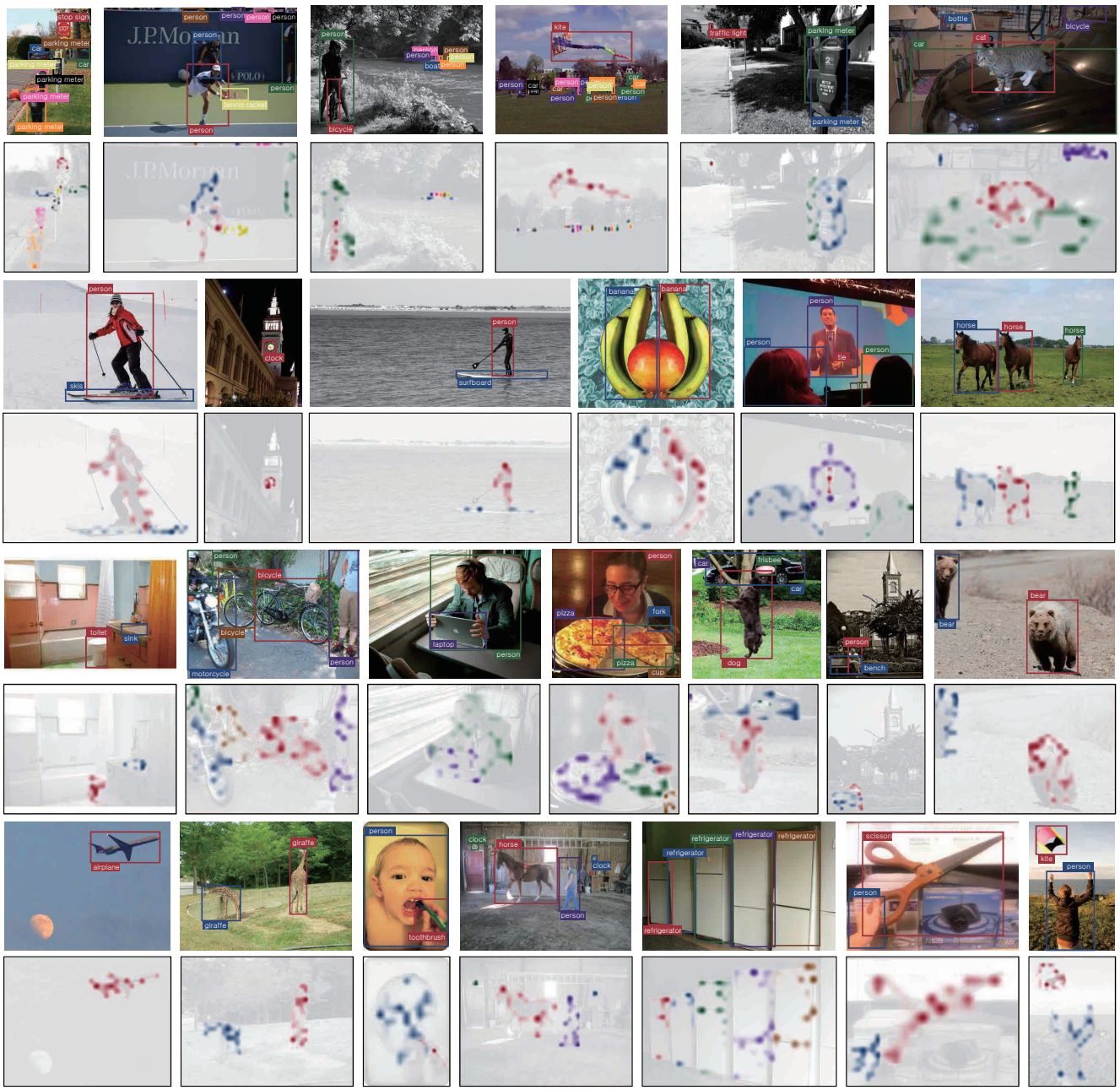

Figure S2: Examples of MS COCO 2017 [12] validation images with the results of object detection and corresponding BBAMs, obtained from Faster R-CNN [15].

Figure S3: Examples of predicted semantic masks for PASCAL VOC validation images of DSRG [7], Shen *et al.* [16], FickleNet [9], Lee *et al.* [10], and our method.

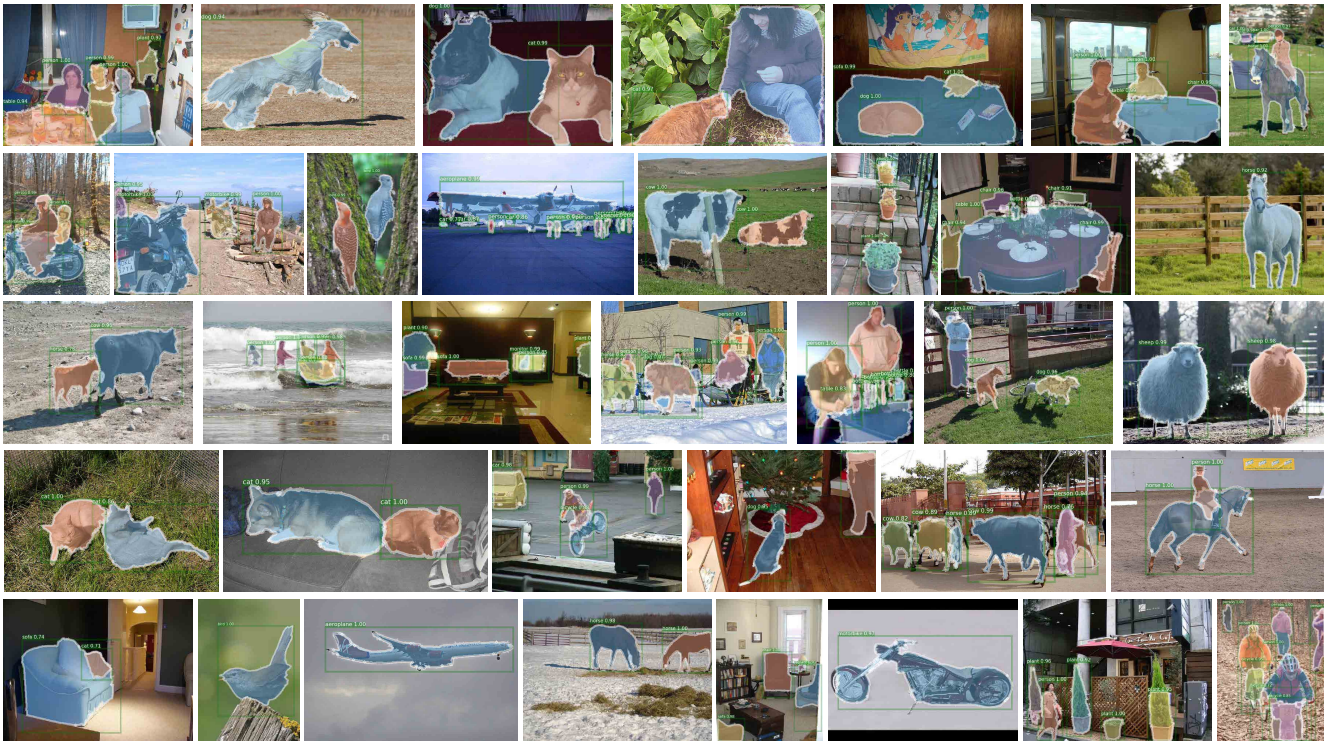

Figure S4: Examples of predicted instance masks for PASCAL VOC validation images of our method.

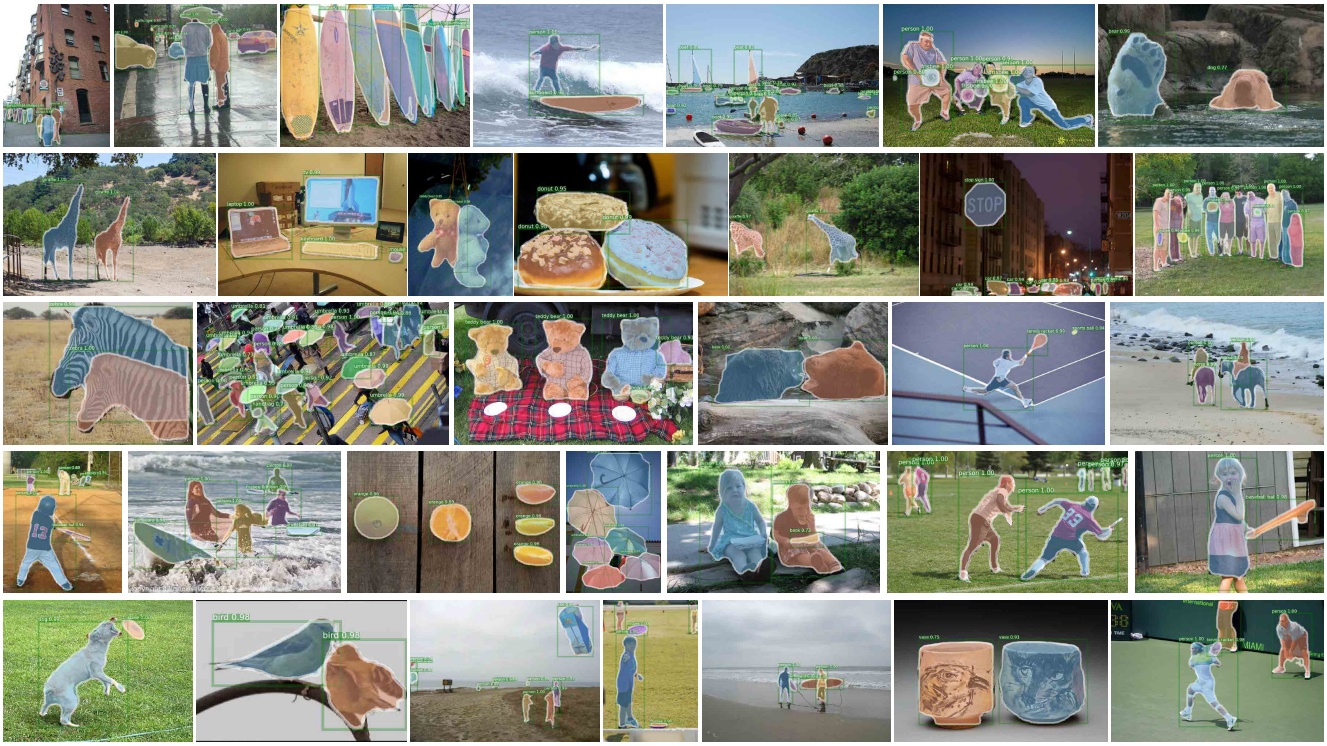

Figure S5: Examples of predicted instance masks for MS COCO 2017 validation images of our method.