Threshold Matters in WSSS: Manipulating the Activation for the Robust and Accurate Segmentation Model Against Thresholds (*Supplementary Material*)

Minhyun Lee*, Dongseob Kim*, Hyunjung Shim†
Yonsei University
{lmh315, kou.k, kateshim}@yonsei.ac.kr

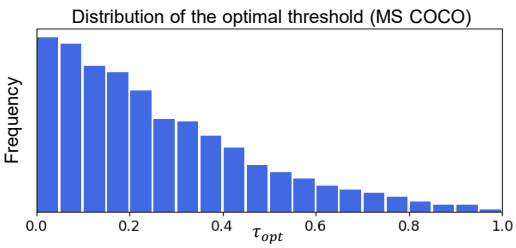

Figure A.1. The distribution of the optimal threshold for 8,278 images randomly sampled from MS COCO 2014 train set. This shows that the optimal threshold per image quite differs from each other even on MS COCO 2014 train set.

A. Implementation Details

Activation manipulation network. For training AMN, we used an Adam [2] optimizer and the learning rate of 5e-6 for updating the backbone parameters and 1e-4 for updating parameters associated with a per-pixel classification head. Both parameter groups adopt the weight decay of 1e-4. The batch size is 16, and the total training epoch is 5. In addition, we adopted label smoothing as a training technique to subside the noise in initial seed, as discussed in [5]. Note that label smoothing strategy has the hyper-parameter ϵ that determines the level of smoothing (i.e., the greater ϵ indicates the stronger effect of label smoothing). In our experiment, we empirically chose $\epsilon = 0.4$ and the same value was applied in all experimental settings. Specifically, given a class label $l_p \in \{0, 1, 2, 3, ..., N\}$ at pixel p of the refined seed S, the target label distribution at p is denoted as S_p and it is rewritten as follows:

$$S_p^c = \begin{cases} 1-\epsilon, & c = l_p \\ \frac{\epsilon}{N-1}, & c \neq l_p, \end{cases} . \tag{1}$$

For a per-pixel classification loss (PCL), we adopted balanced cross-entropy loss [1] as follows:

$$\begin{aligned} \mathcal{L}_{PCL} = &- \frac{1}{\sum_{c \in \mathcal{C}_{fg}} |P_c|} \sum_{c \in \mathcal{C}_{fg}} \sum_{u \in P_c} \log \mathbf{M}_{u,c} \\ &- \frac{1}{\sum_{c \in \mathcal{C}_{bg}} |P_c|} \sum_{c \in \mathcal{C}_{bg}} \sum_{u \in P_c} \log \mathbf{M}_{u,c}, \end{aligned} \tag{2}$$

$$\mathbf{M} = \sigma(g(f(x))), \tag{3}$$

where σ is the softmax function, $\mathbf{M}$ is the activation map from AMN (F_c is the class activation map from the classifier), $\mathcal{C}_{fg}$ is the set of classes that are present in the image (excluding background) and $\mathcal{C}_{bg}$ is the background class. $|P_c|$ denotes the number of pixels belonging to class c.

Segmentation network. For the segmentation network, we adopted DeepLab-v2-ResNet101 and followed the default training settings of AdvCAM [4] for PASCAL VOC 2012. Input images are randomly scaled to $[0.5, 0.75, 1.0, 1.25, 1.5, 1.75, 2.0]$ and cropped to 321×321 (481×481 for MS COCO 2014) for training. We used the SGD optimizer with the batch size of 10 (20 for MS COCO 2014), the momentum of 0.9, and the weight decay of $5e{-}4$. The number of training iterations is 30k and the initial learning rate is $2.5e-4$ with the polynomial learning rate decay $lr_{iter} = lr_{init}(1 - \frac{iter}{max_iter})^{\gamma}$, where γ is set to 0.9. We used balanced cross-entropy loss [1] as in AdvCAM [4].

B. Additional Analysis

Distribution of optimal thresholds on MS COCO 2014. Figure 1 shows the distribution of optimal threshold in the PASCAL VOC 2012. Here, we further investigate whether the same observation holds in MS COCO 2014, which is a large-scale, popular benchmark dataset for semantic segmentation. To efficiently derive the distribution of optimal threshold using MS COCO 2014, we randomly sample 10% of MS COCO 2014 and find the optimal threshold for each image. Figure A.1 shows that the optimal threshold per image is distributed over a wide range from 0 to 1. This result confirms that our observation in PASCAL VOC 2012 con-

*indicates an equal contribution
†Hyunjung Shim is a corresponding author.

	AMN w/o LC	AMN w/ ones	AMN w/ label + noise	AMN
mIoU	58.2%	58.6%	60.5%	62.1%

Table A.1. Accuracy (mIoU) of pseudo-masks from AMN without the boundary refinement on PASCAL VOC 2012 train set. The accuracy depends on the information encoded through the label conditioning module.

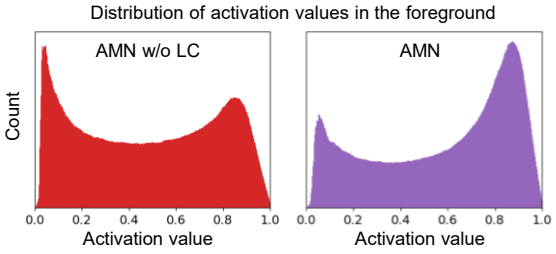

Figure A.2. The distribution of activation values in the foreground on PASCAL VOC 2012 train set. This shows that LC not only reduces non-target activations but also increase the foreground activations of the target objects.

sistently holds in a different dataset; the global threshold is not sufficient to generate the optimal pseudo-masks.

Effects of encoding features. In Section 4.3, we encode label vectors by transforming it into feature vectors for label conditioning. To differentiate the effect of label vector from the effect of encoding any vectors, we conduct additional experiments; 1) encoding a one-vector, 2) encoding the label vector + a random vector and 3) encoding the label vector. Table A.1 compares three cases by reporting the accuracy (mIoU) of pseudo-masks. With a one-vector, no distinct gain is observed over AMN without LC. This implies that the encoding operation itself does not make much difference. In addition, we observe the accuracy gain when encoding noisy labels (i.e., the ground-truth label vector summed up with a noise vector). Since this noisy label also reduces the possible choices, it helps reduce non-target activation to some extent. As expected, the ground-truth image-level labels can lead a noticeable gain, achieving the best accuracy among all.

Effect of label conditioning. Additionally, we observe the histogram of foreground activation values on PASCAL VOC 2012 train set. For this empirical study, we focus on the activation values appearing inside the target objects using ground-truth segmentation mask. As shown in Figure A.2, the effects of LC increase the foreground activations of the target objects–the values within [0.8 1.0] greatly increase and the values within [0.0 0.2] sufficiently decrease. This is coherent with our observation in Figure 4, where LC reduces the horse activation in the cow image and then the cow is correctly activated after applying LC. Overall, we confirm that LC is effective to achieve accurate and robust segmentation performance.

C. Per-class Performance

In Figure 1(a), we showed that the optimal threshold per image quite differs from each other. Herein, Figure A.3 shows the distribution of the optimal threshold per image within the same class on PASCAL VOC 2012 train set. From these results, we find that the distribution of the optimal threshold is widely distributed in most classes and the different class has different tendency; a class-wise global threshold is also largely different from each other.

Figure A.4 shows per-class mIoU of the pseudo-masks according to thresholds on PASCAL VOC 2012 train set. Although the different class exhibits different characteristics in optimal thresholds, AMN tends to generate more accurate and robust pseudo-masks (e.g., the pseudo-mask accuracy of *man* increases a lot, but that of *sofa* is almost same).

Table A.2 shows the per-class mIoU of the pseudo-mask results on PASCAL VOC 2012 train set. For comparison, we report the per-class mIoU of RIB [3]. Since RIB does not present the per-class mIoU of the pseudo-masks, we reproduced their results based on the official implementation of RIB[1]. Table A.2 and Table A.4 show the per-class mIoU of the segmentation results on PASCAL VOC 2012 and MS COCO 2014 datasets, respectively. Specifically, for MS COCO 2014 validation set, we observe the strong gains in several classes; *dining table* / *airplane* are 11.6 / 61.3 with RIB, but 17.2 / 65.5 with ours. These results are consistent with the PASCAL VOC 2012; our method handles the strong imbalance in activation at the pixel-level (*dining table*) and is robust against the threshold choice (*airplane*). This demonstrates that AMN is also effective on large-scale benchmarks.

D. Qualitative Examples

Figure A.5 shows qualitative examples and failure cases of segmentation results from AMN on PASCAL VOC 2012 validation set and MS COCO 2014 validation set. Our method effectively covers the full extent of the objects. Meanwhile, we still have some failure cases: 1) confusing objects (e.g., *sofa* and *chair*), 2) co-occurrence problem (e.g., *railroad* and *train*, 3) shape bias (e.g., *tv/monitor*).

[1] https://github.com/jbeomlee93/RIB

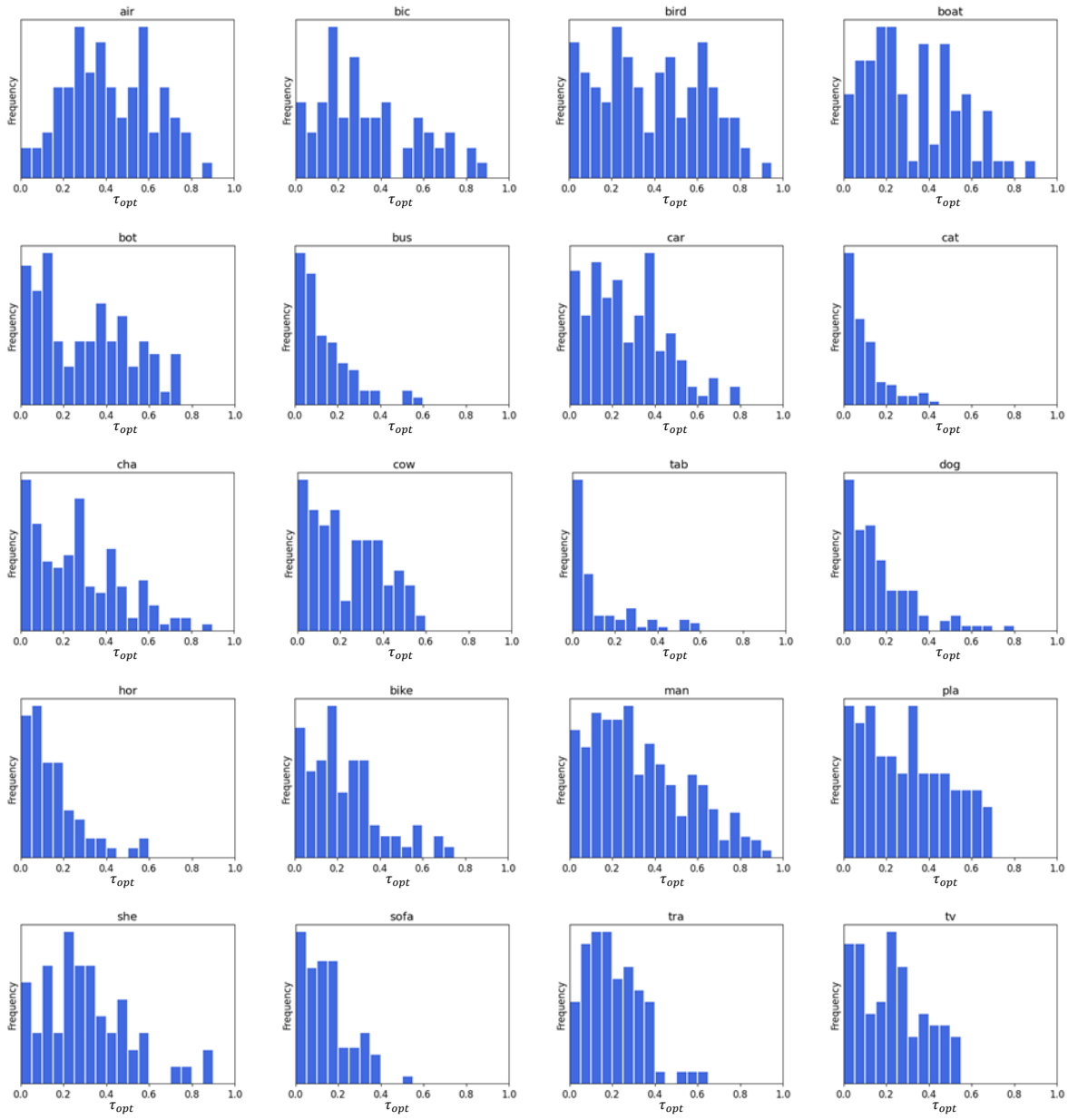

Figure A.3. The distribution of the optimal threshold per class on PASCAL VOC 2012 train set. This shows that the distribution of the optimal threshold per class is quite different.

	bkg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mIOU
RIB*	88.9	70.3	44.5	74.5	62.3	77.8	83.3	73.9	85.9	40.8	82.4	41.9	79.7	83.4	80.6	69.0	59.5	83.7	63.9	60.8	54.2	69.6
AMN (Ours)	90.2	75.3	40.1	77.4	67.9	73.4	85.6	78.9	80.7	36.5	86.1	62.8	78.7	83.4	81.0	74.4	62.4	89.4	62.8	65.3	63.1	72.2

Table A.2. Per-class accuracy (mIoU) of pseudo-masks evaluated on PASCAL VOC 2012 train set. * denotes the reproduced results based on the official implementation of RIB [3].

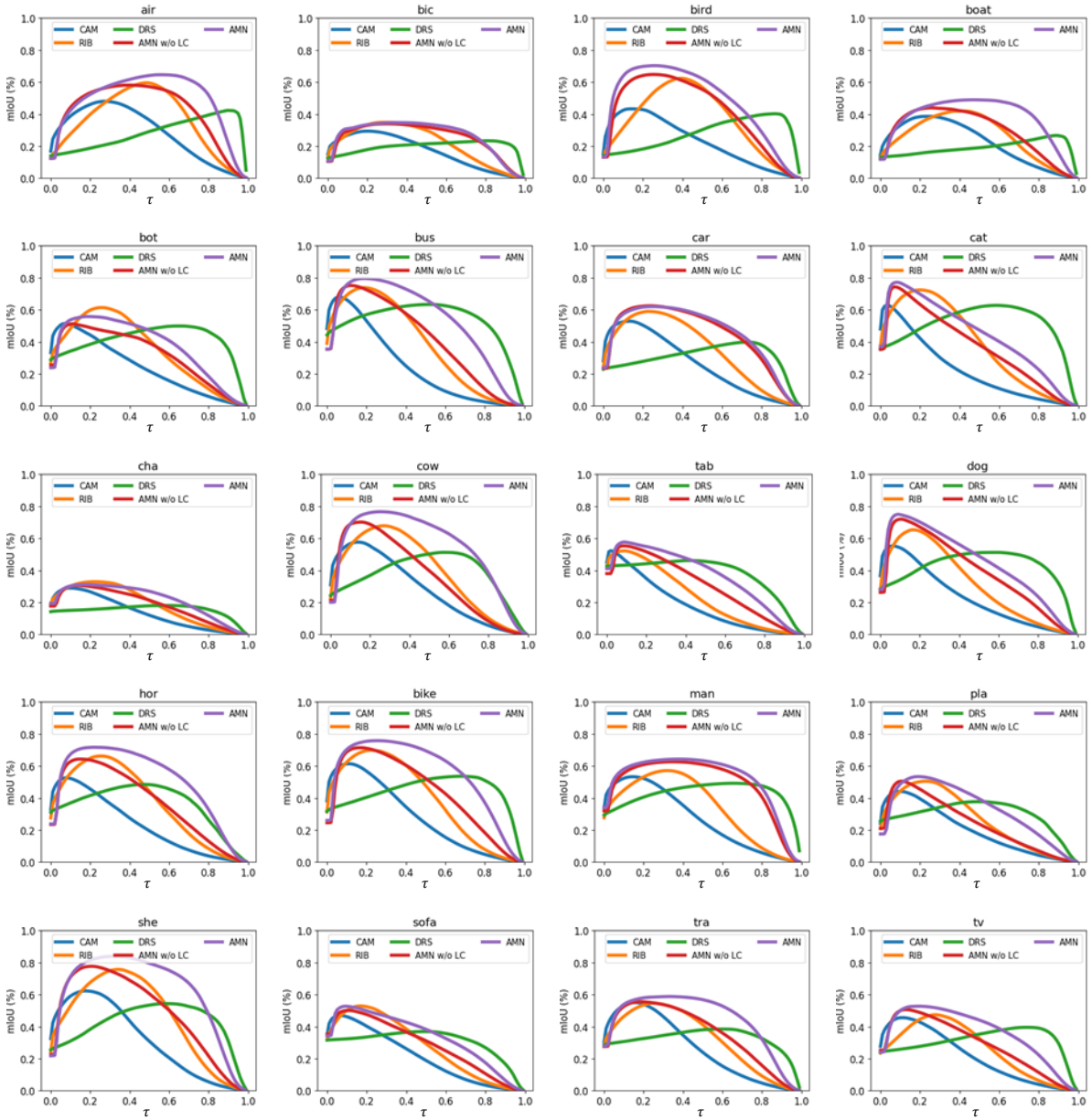

Figure A.4. Per-class mIoU of pseudo-masks according to thresholds on PASCAL VOC 2012 train set. The results are before boundary refinement. AMN shows generally more accurate and robust performance than others.

	bkg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mIOU
Results on PASCAL VOC 2012 val set:																						
AdvCAM	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
RIB	90.3	76.2	33.7	82.5	64.9	73.1	88.4	78.6	88.7	32.3	80.1	37.5	83.6	79.7	75.8	71.8	47.5	84.3	44.6	65.9	54.9	68.3
AMN (Ours)	90.6	79.0	33.5	83.5	60.5	74.9	90.0	81.3	86.6	30.6	80.9	53.8	80.2	79.6	74.6	75.5	54.7	83.5	46.1	63.1	57.5	69.5
Results on PASCAL VOC 2012 test set:																						
AdvCAM	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
RIB	90.4	80.5	32.8	84.9	59.4	69.3	87.2	83.5	88.3	31.1	80.4	44.0	84.4	82.3	80.9	70.7	43.5	84.9	55.9	59.0	47.3	68.6
AMN (Ours)	90.7	82.8	32.4	84.8	59.4	70.0	86.7	83.0	86.9	30.1	79.2	56.6	83.0	81.9	78.3	72.7	52.9	81.4	59.8	53.1	56.4	69.6

Table A.3. Per-class accuracy (mIoU) of segmentation results evaluated on PASCAL VOC 2012.

Class	IRN	RIB	Ours	Class	IRN	RIB	Ours	Class	IRN	RIB	Ours	Class	IRN	RIB	Ours	Class	IRN	RIB	Ours
background	80.5	82.0	82.8	dog	56.2	63.5	67.9	kite	28.8	37.1	43.9	broccoli	52.6	45.4	45.9	cell phone	51.6	54.1	57.7
person	45.9	56.1	53.7	horse	58.1	63.6	65.3	baseball bat	12.6	15.3	16.1	carrot	37.0	34.6	31.3	microwave	42.7	45.2	43.2
bicycle	48.9	52.1	49.3	sheep	64.6	69.1	71.9	baseball glove	7.9	8.1	6.5	hot dog	48.4	49.7	47.0	oven	31.0	35.9	35.5
car	31.3	43.6	38.9	cow	63.8	68.3	70.3	skateboard	27.1	31.8	29.6	pizza	55.9	58.9	57.5	toaster	16.4	17.8	24.3
motorcycle	64.7	67.6	67.1	elephant	79.3	79.5	81.4	surfboard	40.7	29.2	44.6	donut	50.0	53.1	57.3	sink	33.3	33.0	31.4
airplane	62.0	61.3	65.5	bear	74.6	76.7	79.9	tennis racket	49.7	48.9	45.6	cake	38.6	40.7	40.1	refrigerator	40.0	46.0	45.6
bus	60.4	68.5	68.1	zebra	79.7	80.2	82.4	bottle	30.9	33.1	33.0	chair	17.7	20.6	23.6	book	29.9	31.1	29.5
train	51.1	51.3	56.3	giraffe	72.3	74.1	76.5	wine glass	24.3	27.5	31.7	couch	32.6	36.8	36.6	clock	41.3	41.9	47.6
truck	32.2	38.1	38.9	backpack	19.1	18.1	15.5	cup	27.3	27.4	28.8	potted plant	10.5	17.0	19.2	vase	28.4	27.5	30.9
boat	36.7	42.3	41.6	umbrella	57.3	60.1	62.4	fork	16.9	15.9	16.3	bed	33.8	46.2	44.5	scissors	41.2	41.0	39.2
traffic light	48.7	47.8	49.6	handbag	9.0	8.6	7.2	knife	15.6	14.3	16.3	dining table	6.7	11.6	17.2	teddy bear	56.4	62.0	63.9
fire hydrant	74.9	73.4	74.3	tie	24.0	28.6	28.7	spoon	8.4	8.2	8.4	toilet	63.4	63.9	65.4	hair drier	16.2	16.7	21.3
stop sign	76.8	76.3	70.8	suitcase	45.2	49.2	48.6	bowl	17.0	20.7	24.4	tv	35.5	39.7	43.5	toothbrush	16.7	21.0	25.0
parking meter	67.3	68.3	63.2	frisbee	53.8	53.6	56.6	banana	62.4	59.8	61.1	laptop	39.3	48.2	51.8				
bench	31.4	39.7	35.0	skis	8.0	9.7	11.4	apple	43.3	48.5	45.9	mouse	27.9	22.4	30.0				
bird	55.5	57.5	60.0	snowboard	25.5	29.4	30.3	sandwich	37.9	36.9	35.8	remote	41.4	38.0	38.4	mean	41.4	43.8	44.7
cat	68.2	72.4	71.2	sports ball	33.6	38.0	33.9	orange	60.1	62.5	62.9	keyboard	52.9	50.9	48.7				

Table A.4. Per-class accuracy (mIoU) of segmentation results evaluated on MS COCO 2014.

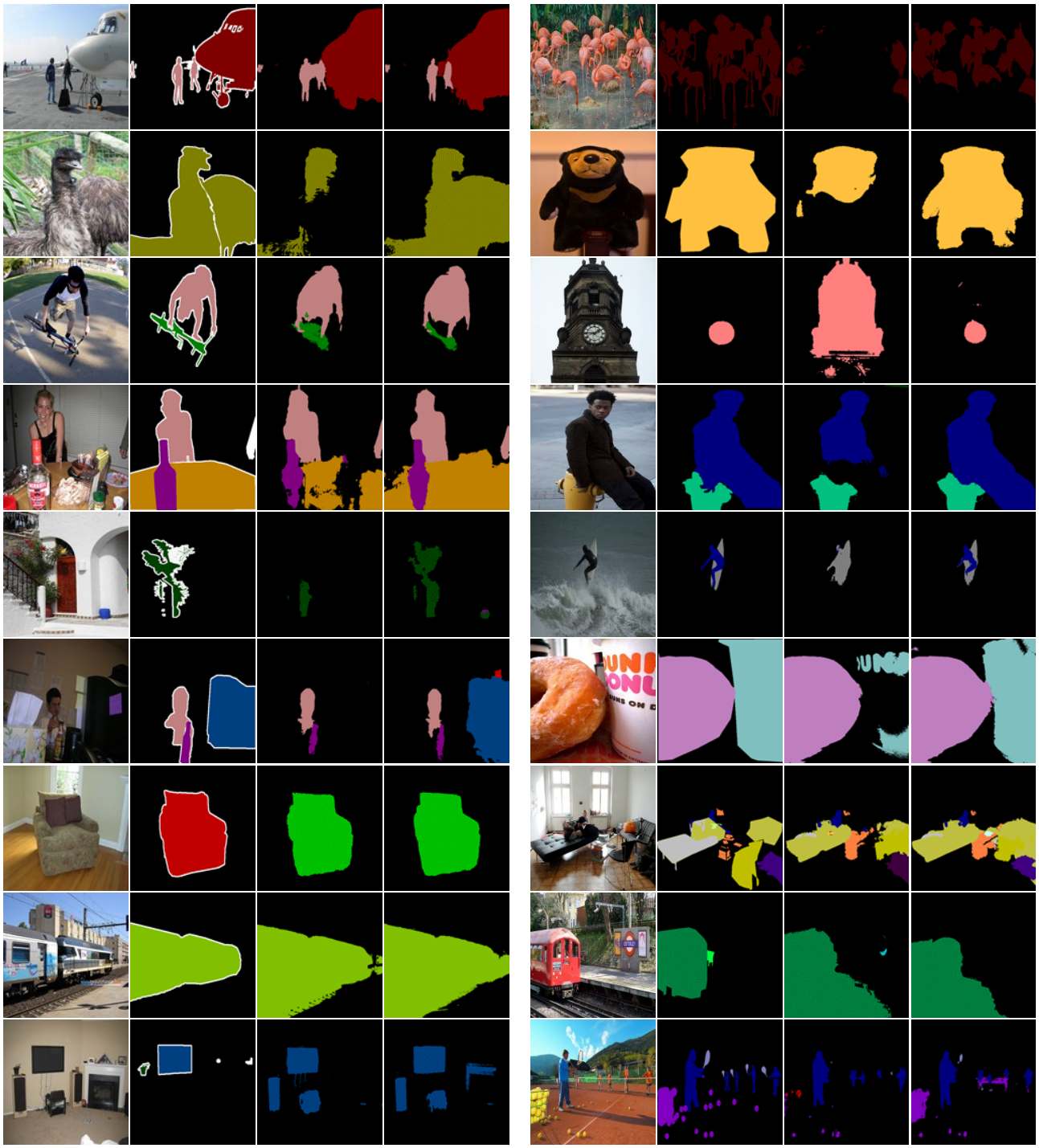

Figure A.5. Qualitative examples of segmentation results on (a) PASCAL VOC 2012 val set and (b) MS COCO 2014 val set.

References

[1] 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

[2] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[3] Jungbeom Lee, Jooyoung Choi, Jisoo Mok, and Sungroh Yoon. Reducing information bottleneck for weakly supervised semantic segmentation. In *NeurIPS*, 2021. 2, 3

[4] Jungbeom Lee, Eunji Kim, and Sungroh Yoon. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In *CVPR*, 2021. 1

[5] Michal Lukasik, Srinadh Bhojanapalli, Aditya Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In *ICML*, 2020. 1