Leveraging Auxiliary Tasks with Affinity Learning for Weakly Supervised Semantic Segmentation

-Supplementary Material-

1. Detailed Quantitative Results

We present per-class segmentation results in terms of IoU on both the *val* and *test* sets of PASCAL VOC, and the *val* set of MS COCO in Table 1, Table 2 and Table 4, respectively. These quantitative results indicate that the proposed AuxSegNet outperforms other state-of-the-art methods on most class categories, further demonstrating the superior performance of the proposed approach.

Table 3 shows ablation studies based on the per-class segmentation IoU on the PASCAL VOC *val* set. We observe that leveraging multi-task auxiliary learning brings a significant improvement of 3.9% in terms of mIoU, compared to the baseline method which uses the raw CAM maps to generate pseudo segmentation labels and trains a single-task model for semantic segmentation. By learning cross-task affinities to refine both task-specific representations and predictions, the proposed AuxSegNet attains a further performance gain of 3.3%. In addition, iteratively learning the entire network with updated pseudo labels refined by the learned cross-task affinities gives another performance boost of 2%. As a result, the accuracy of most class categories is significantly improved, leading to an overall increase of 9.2% in mIoU. This quantitatively demonstrates the effectiveness of our proposed method.

2. Additional Qualitative Results

More examples of qualitative segmentation results of our proposed method on the *val* sets of PASCAL VOC and MS COCO datasets are shown in Figure 1 and 2, respectively. These results show that our method can achieve satisfactory segmentation performance on various challenging scenes with a robust prediction of details.

We present more examples of CAM maps which are used to generate segmentation pseudo labels for each training stage (s=0,1,2,3) on PASCAL VOC and MS COCO *train* sets as shown in Figure 3 and Figure 4, respectively. The raw CAM maps without refinement for stage 0 (s=0) only focus on the local discriminative object regions of large-scale objects, such as the head and hands of the person or wheels of the bus. For small-scale objects, the raw CAM maps tend to over-activate the object regions, resulting in poor boundaries. In contrast, by using the proposed learned cross-task affinities for refinement, the resulted CAM maps (s=1) cover more object regions including those less discriminative ones for large-scale objects. These CAM maps are also better aligned with the boundaries of small-scale objects. Furthermore, with a few more stages of affinity learning and label updating (s=2,3), the refined CAM maps become more complete with more accurate boundaries.

Table 1. Per-class performance comparison with the state-of-the-art WSSS methods in terms of IoUs (%) on the PASCAL VOC *val* set.

	bkg	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbk	person	plant	sheep	sofa	train	tv	mIoU
DSRG [7]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	61.4
MCOF [13]	87.0	78.4	29.4	68.0	44.0	67.3	80.3	74.1	82.2	21.1	70.7	28.2	73.2	71.5	67.2	53.0	47.7	74.5	32.4	71.0	45.8	60.3
AffinityNet [1]	88.2	68.2	30.6	81.1	49.6	61.0	77.8	66.1	75.1	29.0	66.0	40.2	80.4	62.0	70.4	73.7	42.5	70.7	42.6	68.1	51.6	61.7
SeeNet [6]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	63.1
FickleNet [10]	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
OAA^{+} [8]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	65.2
Zeng *et al*. [15]	90.0	77.4	37.5	80.7	61.6	67.9	81.8	69.0	83.7	13.6	79.4	23.3	78.0	75.3	71.4	68.1	35.2	78.2	32.5	75.5	48.0	63.3
CIAN [5]	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
Zhang *et al*. [16]	87.9	75.9	31.7	78.3	54.6	62.2	80.5	73.7	71.2	30.5	67.4	40.9	71.8	66.2	70.3	72.6	**49.0**	70.7	38.4	62.7	58.4	62.6
Luo *et al*. [11]	88.6	64.1	35.4	78.8	50.8	61.0	85.8	77.7	84.6	26.7	75.2	40.8	79.1	77.4	76.0	70.4	48.3	69.2	39.0	69.9	58.3	64.5
Chang *et al*. [3]	88.8	51.6	30.3	82.9	53.0	75.8	**88.6**	74.8	86.6	**32.4**	79.9	**53.8**	82.3	78.5	70.4	71.2	40.2	78.3	**42.9**	66.8	58.8	66.1
ICD [4]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	67.8
Araslanov *et al*. [2]	88.7	70.4	35.1	75.7	51.9	65.8	71.9	64.2	81.1	30.8	73.3	28.1	81.6	69.1	62.6	74.8	48.6	71.0	40.1	68.5	64.3	62.7
SEAM [14]	88.8	68.5	33.3	**85.7**	40.4	67.3	78.9	76.3	81.9	29.1	75.5	48.1	79.9	73.8	71.4	75.2	48.9	79.8	40.9	58.2	53.0	64.5
Zhang *et al*. [18]	90.4	**85.6**	**38.9**	78.9	62.0	73.4	83.7	74.3	82.9	25.8	77.8	30.1	81.1	**79.3**	**76.1**	73.9	38.6	85.0	32.7	72.8	55.7	66.6
Sun *et al*. [12]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	66.2
CONTA [17]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	66.1
AuxSegNet(Ours)	**91.7**	82.5	38.2	84.3	**67.4**	**76.7**	85.0	**79.8**	**90.7**	24.5	**81.2**	22.7	**86.7**	78.7	76.0	**82.2**	37.9	**86.4**	39.3	**75.6**	61.0	**69.0**

Table 2. Per-class performance comparison with the state-of-the-art WSSS methods in terms of IoUs (%) on the PASCAL VOC *test* set.

	bkg	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbk	person	plant	sheep	sofa	train	tv	mIoU
DSRG [7]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	63.2
MCOF [13]	88.2	80.8	31.4	70.9	34.9	65.7	83.5	75.1	79.0	22.0	70.3	31.7	77.7	72.9	77.1	56.9	41.8	74.9	36.6	71.2	42.6	61.2
AffinityNet [1]	89.1	70.6	31.6	77.2	42.2	68.9	79.1	66.5	74.9	29.6	68.7	**56.1**	**82.1**	64.8	78.6	73.5	50.8	70.7	47.7	63.9	51.1	63.7
SeeNet [6]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	62.8
FickleNet [10]	89.8	78.3	34.1	73.4	41.2	67.2	81.0	77.3	81.2	29.1	72.4	47.2	76.8	76.5	76.1	72.9	56.5	82.9	43.6	48.7	**64.7**	65.3
OAA^{+}[8]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	66.4
Zeng *et al*. [15]	90.4	85.4	37.9	77.2	48.2	64.5	**83.9**	74.8	83.4	15.9	72.4	34.3	80.0	77.3	78.5	69.0	41.9	76.3	38.3	72.3	48.2	64.3
CIAN [5]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	65.3
Zhang *et al*. [16]	87.8	77.5	30.8	71.7	36.0	64.2	75.3	70.4	81.7	29.3	70.4	52.0	78.6	73.8	74.4	72.1	54.2	75.2	**50.6**	42.0	52.5	62.9
Luo *et al*. [11]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	64.6
Chang *et al*. [3]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	65.9
ICD[4]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	68.0
Araslanov *et al*. [2]	89.2	73.4	37.3	68.3	45.8	68.0	72.7	64.1	74.1	**32.9**	74.9	39.2	81.3	74.6	72.6	75.4	58.1	71.0	48.7	67.7	60.1	64.3
SEAM[14]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	65.7
Zhang *et al*. [18]	90.7	**85.9**	37.3	**82.5**	50.5	64.8	83.1	77.6	82.8	28.4	**76.8**	34.6	81.2	82.9	**80.5**	73.6	43.9	**85.7**	32.0	71.7	55.2	66.7
Sun *et al*. [12]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	66.9
CONTA [17]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	66.7
AuxSegNet(Ours)	**91.6**	85.1	**39.4**	80.0	**51.4**	**69.9**	81.4	**79.9**	**86.5**	26.6	75.3	29.7	81.7	**83.6**	78.0	**83.1**	56.1	84.5	39.8	**77.2**	60.9	**68.6**

Table 3. Segmentation results of our proposed method in terms of IoUs (%) on the PASCAL VOC *val* set. MT, CT and IL denote the proposed multi-task auxiliary learning, cross-task affinity learning and iterative learning, respectively. No post-processing steps are used in these ablation analysis.

	bkg	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbk	person	plant	sheep	sofa	train	tv	mIoU
Baseline	86.6	60.0	30.8	67.5	50.8	55.4	73.3	70.7	71.8	24.0	63.2	23.1	63.9	60.4	66.9	69.6	33.9	68.4	31.5	64.5	57.8	56.9
AuxSegNet (MT)	88.6	70.3	34.5	74.1	54.9	60.3	77.5	73.1	73.1	25.2	65.8	26.5	69.5	64.3	71.5	72.9	36.3	72.2	32.6	70.4	60.6	60.8
AuxSegNet (MT+CT)	90.1	74.5	38.3	77.5	60.0	66.6	82.0	76.6	82.1	26.7	70.3	23.2	76.1	67.1	76.0	75.6	38.9	77.8	34.5	71.2	61.1	64.1
AuxSegNet (MT+CT+IL)	91.0	80.4	38.2	79.0	65.1	71.5	82.8	78.2	87.5	25.3	73.9	18.3	83.0	71.0	75.4	79.2	38.9	78.1	36.5	73.9	61.4	66.1
Overall Improvement	**+4.4**	**+20.4**	**+7.4**	**+11.5**	**+14.3**	**+16.1**	**+9.5**	**+7.5**	**+15.7**	**+1.3**	**+10.7**	-4.8	**+19.1**	**+10.6**	**+8.5**	**+9.6**	**+5.0**	**+9.7**	**+5.0**	**+9.4**	**+3.6**	**+9.2**

Table 4. Per-class performance comparison with the state-of-the-art WSSS methods in terms of IoU(%) on the MS COCO *val* set.

Class	SEC [9]	DSRG [7]	Luo *et al.* [11]	Ours	Class	SEC [9]	DSRG [7]	Luo *et al.* [11]	Ours
background	74.3	80.6	73.9	**82.0**	wine class	22.3	24.0	27.2	**32.1**
person	43.6	-	48.7	**65.4**	cup	17.9	20.4	21.7	**29.3**
bicycle	24.2	30.4	45.0	43.0	fork	1.8	0.0	0.0	**5.4**
car	15.9	22.1	31.5	**34.5**	knife	1.4	5.0	0.9	1.4
motorcycle	52.1	54.2	59.1	**66.2**	spoon	0.6	0.5	0.0	**1.4**
airplane	36.6	45.2	26.9	**60.3**	bowl	12.5	18.8	7.6	**19.5**
bus	37.7	38.7	52.4	**63.1**	banana	43.6	46.4	52.0	46.9
train	30.1	33.2	42.4	**57.3**	apple	23.6	24.3	28.8	**40.4**
truck	24.1	25.9	36.9	**38.9**	sandwich	22.8	24.5	37.4	**39.4**
boat	17.3	20.6	23.5	**30.1**	orange	44.3	41.2	52.0	**52.9**
traffic light	16.7	16.1	13.3	**40.4**	broccoli	36.8	35.7	33.7	36.0
fire hydrant	55.9	60.4	45.1	**72.7**	carrot	6.7	15.3	29.0	13.9
stop sign	48.4	51.0	43.4	40.3	hot dog	31.2	24.9	38.8	**46.1**
parking meter	25.2	26.3	33.5	**59.8**	pizza	50.9	56.2	69.8	62.0
bench	16.4	22.3	26.3	16.0	donut	32.8	34.2	50.8	43.9
bird	34.7	41.5	29.9	**61.0**	cake	12.0	6.9	37.3	30.6
cat	57.2	62.2	62.1	**68.6**	chair	7.8	9.7	10.7	**11.4**
dog	45.2	55.6	57.5	**66.9**	couch	5.6	17.7	9.4	14.5
horse	34.4	42.3	40.7	**55.6**	potted plant	6.2	14.3	21.8	2.1
sheep	40.3	47.1	54.0	**61.4**	bed	23.4	32.4	34.6	20.5
cow	41.4	49.3	47.2	**60.7**	dining table	0.0	3.8	1.1	**9.5**
elephant	62.9	67.1	64.3	**76.1**	toilet	38.5	43.6	43.8	**57.8**
bear	59.1	62.6	58.9	**73.0**	tv	19.2	25.3	11.5	**36.0**
zebra	59.8	63.2	60.7	**80.8**	laptop	20.1	21.1	37.0	35.2
giraffe	48.8	54.3	45.1	**71.6**	mouse	3.5	0.9	0.0	**13.4**
backpack	0.3	0.2	0.0	**11.3**	remote	17.5	20.6	37.2	23.6
umbrella	26.0	35.3	46.1	35.0	keyboard	12.5	12.3	19.0	17.9
handbag	0.5	0.7	0.0	**2.2**	cellphone	32.1	33.0	38.1	**49.9**
tie	6.5	7.0	15.5	14.7	microwave	8.2	11.2	43.4	**28.7**
suitcase	16.7	23.4	43.6	31.7	oven	13.7	12.4	29.2	13.3
frisbee	12.3	13.0	23.2	1.0	toaster	0.0	0.0	0.0	0.0
skis	1.6	1.5	6.5	**8.1**	sink	10.8	17.8	28.5	21.0
snowboard	5.3	16.3	10.9	7.6	refrigerator	4.0	15.5	23.8	16.6
sports ball	7.9	9.8	0.6	**28.8**	book	0.4	12.3	26.3	8.7
kite	9.1	17.4	14.0	**27.3**	clock	17.8	20.7	13.4	**34.4**
baseball bat	1.0	4.8	0.0	2.2	vase	18.4	23.9	27.1	25.9
baseball globe	0.6	1.2	0.0	**1.3**	scissors	16.5	17.3	37.0	16.6
skateboard	7.1	14.4	7.6	**15.2**	teddy bear	47.0	46.3	58.9	47.3
surfboard	7.7	13.5	17.6	**17.8**	hair drier	0.0	0.0	0.0	0.0
tennis racket	9.1	6.8	38.1	**47.1**	toothbrush	2.8	2.0	11.1	1.4
bottle	13.2	22.3	28.4	**33.2**	**mIoU**	22.4	26.0	29.9	**33.9**

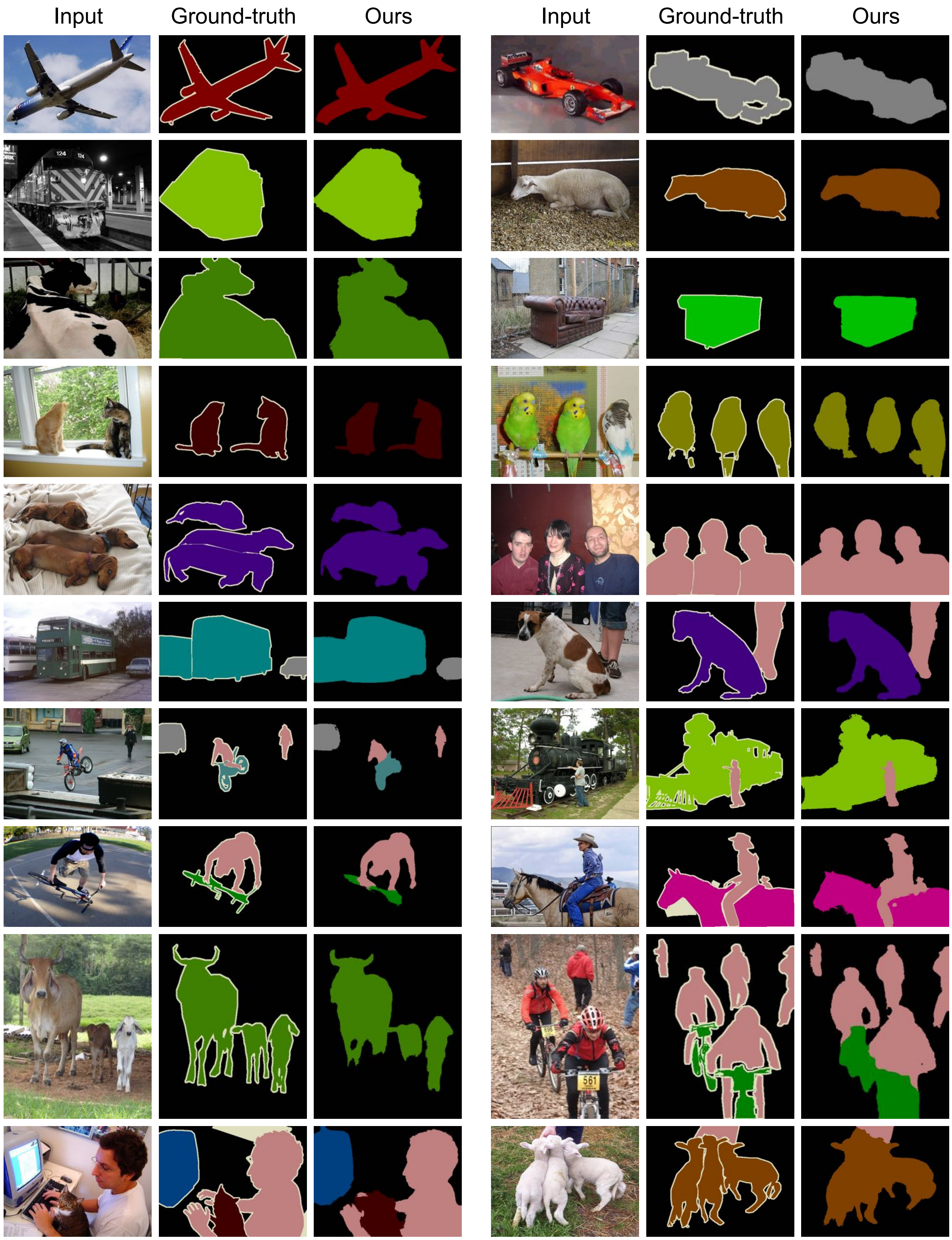

Figure 1. Qualitative segmentation results on the PASCAL VOC *val* set.

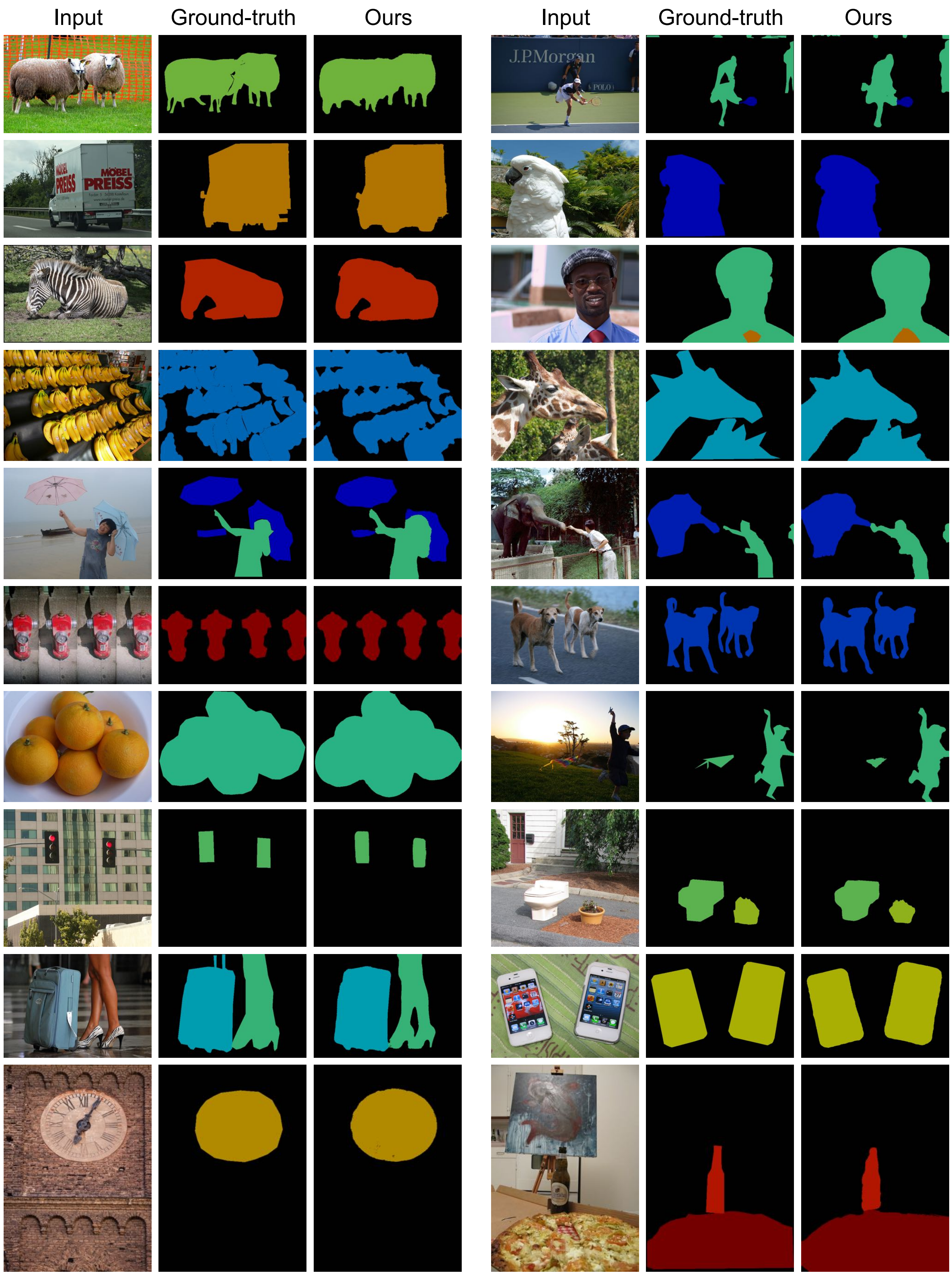

Figure 2. Qualitative segmentation results on the MS COCO *val* set.

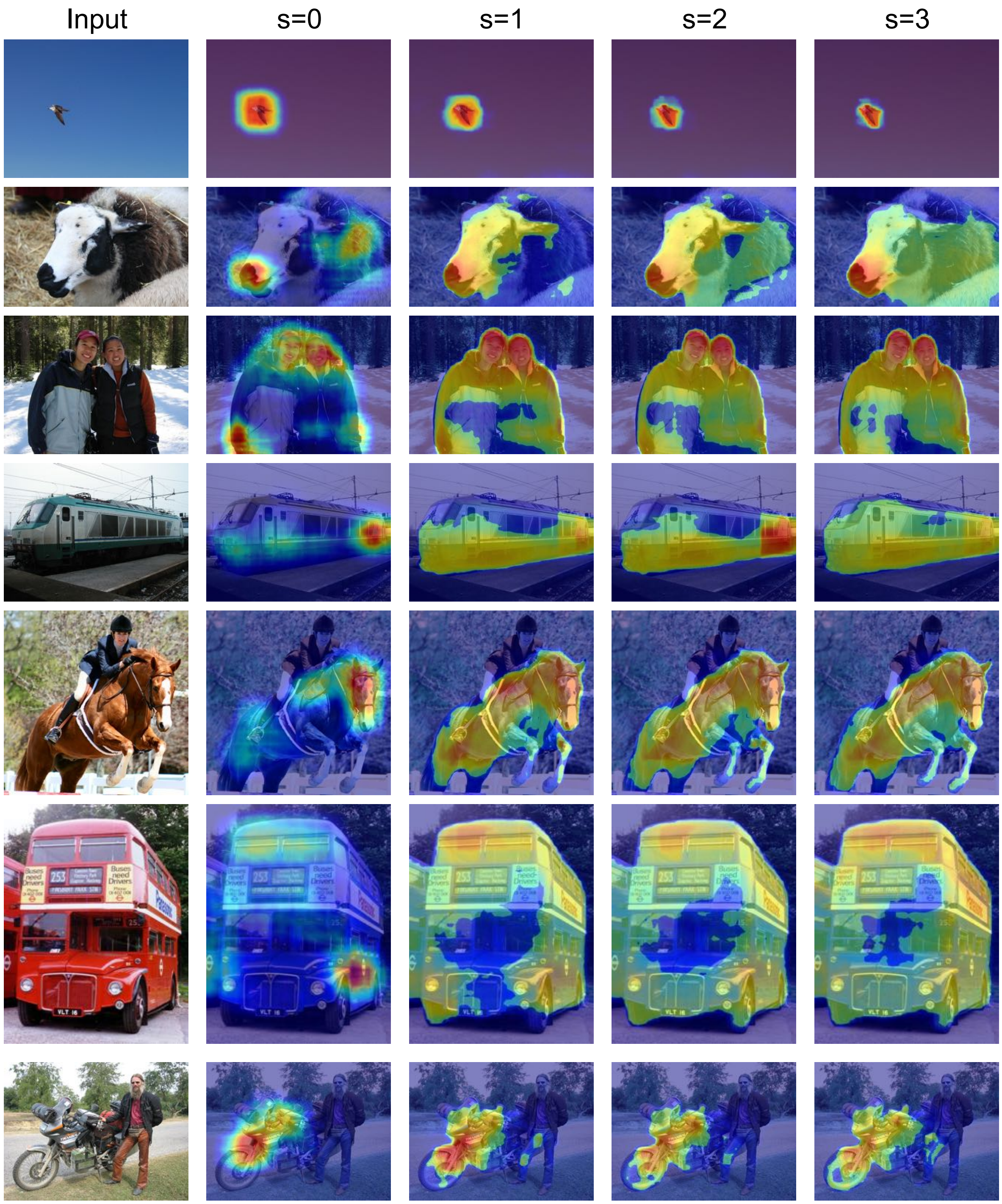

Figure 3. Visualization of CAM maps with iterative improvements on the PASCAL VOC *train* set. (Only the CAM map of the dominant class is shown for each input image.)

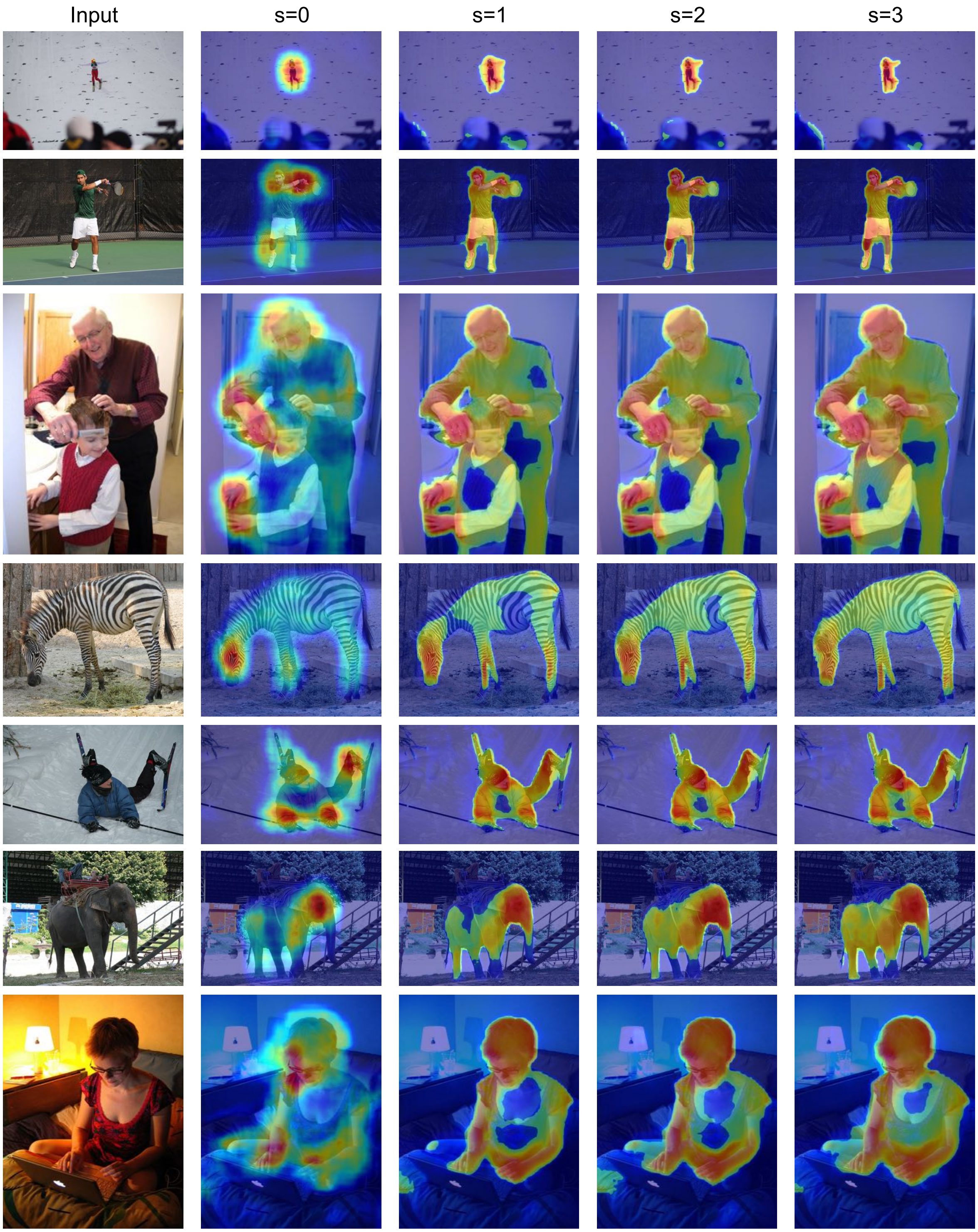

Figure 4. Visualization of CAM maps with iterative improvements on the MS COCO *train* set. (Only the CAM map of the dominant class is shown for each input image.)

References

[1] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *CVPR*, 2018. 2

[2] Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *CVPR*, 2020. 2

[3] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. In *CVPR*, 2020. 2

[4] Junsong Fan, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In *CVPR*, 2020. 2

[5] Junsong Fan, Zhaoxiang Zhang, Tieniu Tan, Chunfeng Song, and Jun Xiao. Cian: Cross-image affinity net for weakly supervised semantic segmentation. In *AAAI*, 2020. 2

[6] Qibin Hou, PengTao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. In *NeurIPS*, 2018. 2

[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. 2, 3

[8] Peng-Tao Jiang, Qibin Hou, Yang Cao, Ming-Ming Cheng, Yunchao Wei, and Hong-Kai Xiong. Integral object mining via online attention accumulation. In *ICCV*, 2019. 2

[9] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *ECCV*, 2016. 3

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

[11] Wenfeng Luo and Meng Yang. Learning saliency-free model with generic features for weakly-supervised semantic segmentation. In *AAAI*, 2020. 2, 3

[12] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. In *ECCV*, 2020. 2

[13] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. In *CVPR*, 2018. 2

[14] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *CVPR*, 2020. 2

[15] Yu Zeng, Yunzhi Zhuge, Huchuan Lu, and Lihe Zhang. Joint learning of saliency detection and weakly supervised semantic segmentation. In *ICCV*, 2019. 2

[16] Bingfeng Zhang, Jimin Xiao, Yunchao Wei, Mingjie Sun, and Kaizhu Huang. Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. In *AAAI*, 2020. 2

[17] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xiansheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. In *NeurIPS*, 2020. 2

[18] Tianyi Zhang, Guosheng Lin, Weide Liu, Jianfei Cai, and Alex Kot. Splitting vs. merging: Mining object regions with discrepancy and intersection loss for weakly supervised semantic segmentation. In *ECCV*, 2020. 2