Supplementary for "Context Decoupling Augmentation for Weakly Supervised Semantic Segmentation"

Yukun Su[1,2], Ruizhou Sun[1,2], Guosheng Lin[3†], and Qingyao Wu[1,2†]

[1]School of Software and Engineering, South China University of Technology
[2]Key Laboratory of Big Data and Intelligent Robot, Ministry of Education
[3]School of Computer Science and Engineering, Nanyang Technological University
suyukun666@gmail.com, ruizhousun@foxmail.com, gslin@ntu.edu.sg, qyw@scut.edu.cn

A. Appendix

This appendix provides additional experiments and more visualization of different models. Sec. A.1 gives the additional experiments of retraining and shows the performance in more detail, and Sec. A.2 presents more qualitative results and the blended objects during the training process.

A.1. Additional experiments and More results

A.1.1 Retraining

In our CDA framework, it is possible to train many rounds. Namely, once we obtain the improved CAM models, we can get the more precise object instances and utilize them to refine the models in the next training round. From Table 2, we can observe that when round = 2, CDA can achieve the best mIoU performance on CAM and pseudo-mask. However, we consider that constantly increasing the number of training rounds can not bring continuous performance boost for the model. Thus, we set round =1 since it can already achieve relatively high performance and cost less time to train.

A.1.2 Training cost

As shown in Table 3, we compare our proposed CDA on training cost on per image with CONTA [4]. It can be observed that our CDA can achieve better performance than CONTA with less training time, which shows the effectiveness and practicality of our method.

A.1.3 Performance of each class

Table 1 shows the detailed mIoU performance of each class to demonstrate the improvement occurs in objects with high correlation with background.

A.2. More Visualizations

A.2.1 CAM + CDA

As shown in Figure 1 (1^{st} to 4^{th} row), we can observe that the original CAM will only highlight a few parts of the objects, while our CAM+Aug by CDA method can activate most regions of the objects, which is beneficial for mining the object seed areas. As depicted in Figure 1 (5^{th} to 6^{th} row), CAM will mistakenly recognize the contextual background that has a strong coupling relationship with the object to be predicted due to the confounding bias. When we deploy CDA to train CAM, it will focus more on the target areas.

A.2.2 Segmentation Masks

Figure 2 and Figure 3 present the qualitative results of our CDA approach applying on SEAM [3], AffinityNet [2] baseline and compares them to the original methods (The visualization results of IRNet [1] baseline can be seen in Figure 7 of the main text). We can observe that CDA can help to yield respectable segmentation masks on both baselines. In particular, in some semantically coherent areas, CDA can help to distinguish objects more independently and thus these objects have better boundaries.

A.2.3 Online Objects Blending

As we mentioned, our CDA adopts an online data augmentation training strategy and we show an example to illustrate what objects will be blended in the original input images. As shown in Figure 4, in each epoch, the new object instances with different categories from the original images

†Corresponding authors.

Method	bkg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mIoU(%)
AffinityNet [2]	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
AffinityNet+CDA	89.6	71.3	31.8	83.4	51.8	62.7	80.9	69.8	77.9	32.7	69.3	44.8	82.6	65.8	72.5	76.3	42.9	73.2	44.9	70.9	54.1	64.2
SEAM [3]	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
SEAM + CDA	89.1	69.7	34.5	86.4	41.3	69.2	81.3	79.5	82.1	31.1	78.3	50.8	80.6	76.1	72.2	77.6	48.8	81.2	42.5	60.6	54.3	66.1

Table 1. The detail semantic segmentation performance on the PASCAL VOC 2012 validation set. Since the original paper of IRNet [1] does not provide performance on each class, we here only compare SEAM [3] and AffinityNet [2].

Setting	CAM	Pseduo-Mask
Round = 0 (baseline)	48.3	65.9
Round = 1	50.8	67.7
Round = 2	**50.9**	**67.7**
Round = 3	50.6	67.3
Round = 4	50.6	67.4

Table 2. Experiments of retraining on IRNet [1] backbone. Round = 0 indicates training without CDA.

Method	CAM	Pseduo-Mask	Time(s)
SEAM	55.4	63.4	0.6
+ CONTA [4]	56.2	65.4	2.8
+ CDA	**58.4**	**66.4**	**1.2**

Table 3. Comparison of the training cost on per image of the proposed CDA on SEAM [3] backbone with other methods.

are randomly pasted in the images to form the augmented input images. This greatly increases the diversity of combinations of various scenes and object instances, and thus enhance the decoupling capability of the networks. Data can be fully randomized and our simple random method does not need external knowledge. Therefore, context decoupling can be achieved in most cases.

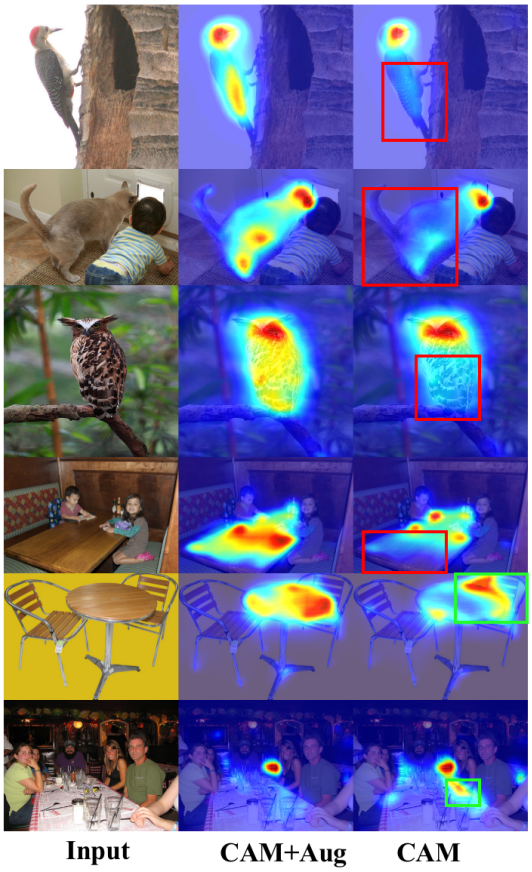

Figure 1. Qualitative visualization of CAMs and our CDA framework (backbone: IRNet [1]). Red Box refers to areas that lack activation. Green Box indicates the over-activated areas or irrelevant objects that are activated. The labels of the images from top to bottom are “bird”, “cat”, “bird”, “table”, “table”, “chair”.

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2209–2218, 2019. 1, 2

[2] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4981–4990, 2018. 1, 2, 3

[3] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12275–12284, 2020. 1, 2, 3

[4] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xiansheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. *arXiv preprint arXiv:2009.12547*, 2020. 1, 2

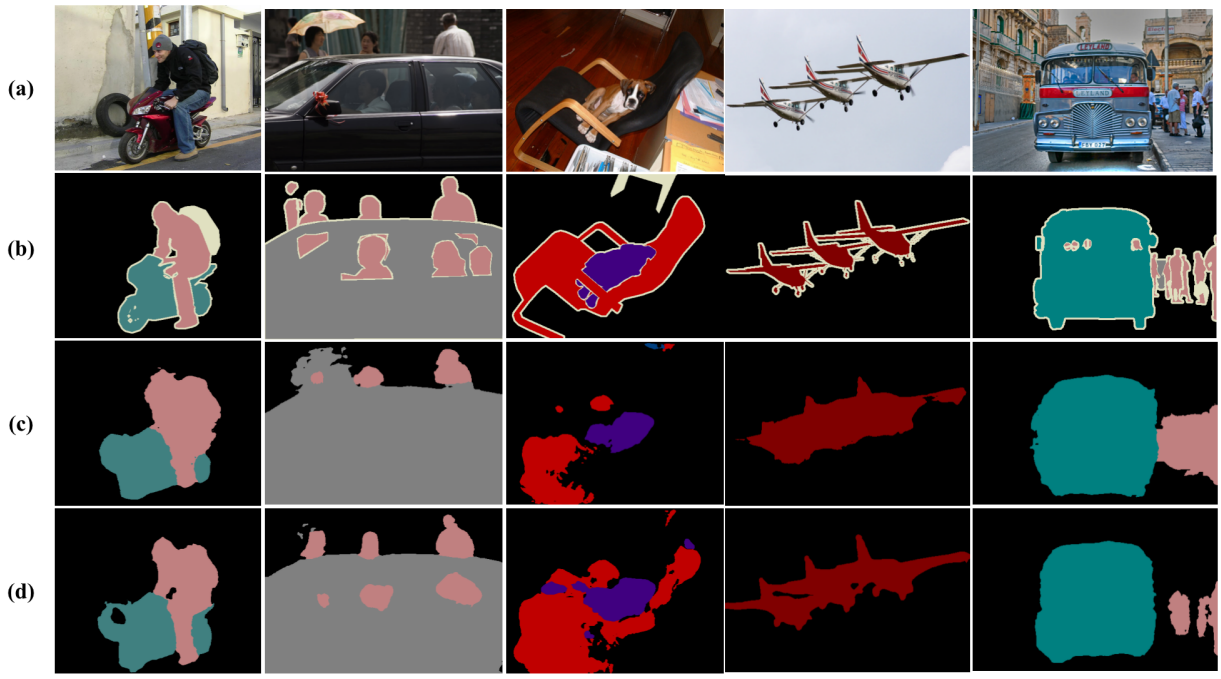

Figure 2. Qualitative results on the PASCAL VOC 2012 *val* set. (a) Input images. (b) Ground-truth labels. (c) Results obtained by SEAM [3] backbone. (d) Results of our SEAM + CDA.

Figure 3. Qualitative results on the PASCAL VOC 2012 *val* set. (a) Input images. (b) Ground-truth labels. (c) Results obtained by AffinityNet [2] backbone. (d) Results of our AffinityNet + CDA.

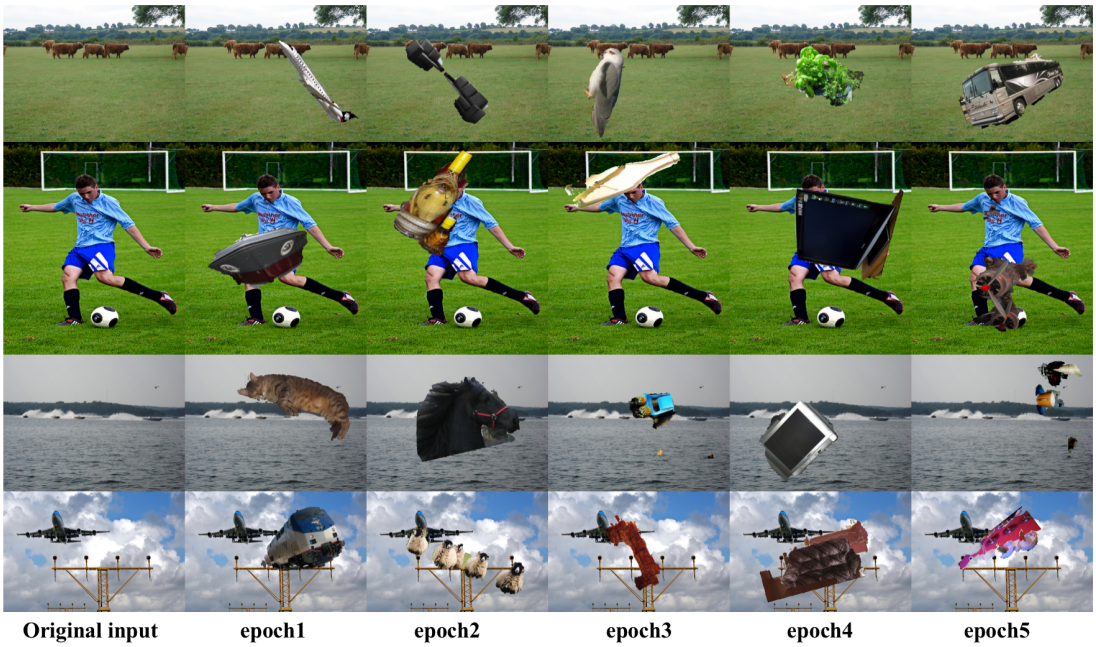

Figure 4. Visualization of object instances to be blended into the original input images during the online training process. Note that we adopt a pairwise manner to train the network, which means in each epoch, the original input images together with the new blended images are input to the network.