Appendix for Learning to Detour: Shortcut Mitigating Augmentation for Weakly Supervised Semantic Segmentation

This appendix provides further explanation of the additional experiments, and visualizations. Section A describes the ablative results for the hyperparameters and additional experiments. The images selected for the experiments and additional qualitative results are presented in section B.

A. Additional Experiments

Hyperparameter analysis. We analyze the effect of the hyperparameters t_{aug} and λ on the initial seed. t_{aug} is the point in the training epoch at which the disentangled representation is shuffled and the ablative results can be seen in Table A.1. The experimental results show that suboptimal representations are synthesized when shuffled in a state that is not sufficiently disentangled, and no significant performance improvement is observed even after further training.

λ is a balancing scalar of contrastive loss that determines the degree of separation between object-relevant features and background features. Table A.2 shows the results of initial seeds obtained by varying the value of λ. The experimental results confirm that $\lambda = 0.5$ achieved the highest mIoU.

t_{aug}	mIoU (%)
w/o shuffle	49.8
2	46.4
4	51.7
6	**52.4**
8	52.0

Table A.1. Experimental result on the relationship between localization map performance and t_{aug}.

λ	mIoU (%)
0.1	50.9
0.3	51.6
0.5	**52.4**
0.7	51.8
1	51.2

Table A.2. Effect of values of λ on localization map performance.

Experimental results for bias-conflicting samples. A classifier that exploits the shortcut feature to predict class labels leverages the background attribute and is less dependent on target object related attributes. Therefore, the classifier fails to identify the target object for samples outside the general context. To analyze this, we selected images that did not contain a top-3 background, which is a background with a high co-occurring ratio with a specific class. That is, a bias-conflicting sample [2], which is an object-background combination that does not often appear in the training dataset, was selected. We evaluated the localization map generated by IRN [1] and our method; the results of these evaluations can be seen in Table A.3. The proposed method produced better results for bias-conflicting samples of the selected class, was less affected by the background, and more accurately captured the target object.

Experiments on interpolation-based feature shuffling. We experimented with different combining strategies when synthesizing disentangled features. Inspired by previous mix-based methods [6, 7], novel representations are produced by interpolating separated features between instances. The representations z_i and z_j are obtained by concating the object-relevant(z_i^o, z_j^o) and the features(z_i^b, z_j^b) of the i-th and j-th samples, randomly extracted from the training data.

$$z_i = [z_i^o, z_i^b], z_j = [z_j^o, z_j^b] \tag{1}$$

We produce novel feature-target pair as follows:

$$\tilde{z} = \delta z_i + (1-\delta) z_j,\ \tilde{y} = \delta y_i + (1-\delta) y_j, \tag{2}$$

where $\delta \sim Beta(\alpha, \alpha)$, for $\alpha \in (0, \infty)$. y_i and y_j are the target vector corresponding to z_i, z_j, and $\delta \in [0, 1]$. The hyperparameter α determines the strength of the interpolation between two pairs.

We only change the combining strategy, and all other experimental settings and hyperparameters are the same as our proposed method. The quality of localization maps generated by an interpolation-based method is measured by

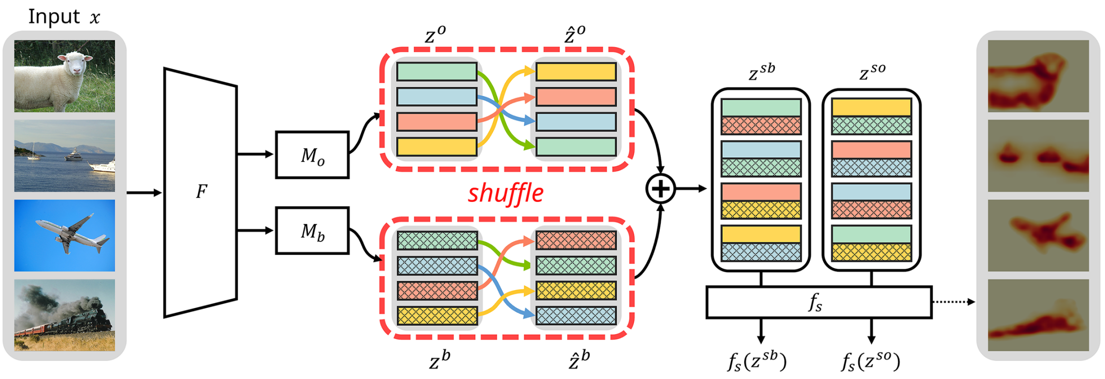


Figure B.1. A detailed overview of SMA when representations are shuffled and combined in a two-way manner.

Class	w/o background (co-occurrence ratio)			Method IRN [1]	 SMA
aeroplane	sky (0.23)	building (0.11)	grass (0.10)	56.45	**57.83**
sheep	grass (0.22)	tree (0.11)	fence (0.08)	61.38	**63.26**
cow	grass (0.20)	tree (0.14)	sky (0.13)	60.15	**62.36**
boat	water (0.18)	sky (0.16)	building (0.10)	59.62	**63.33**
train	track (0.11)	ground (0.11)	sky (0.11)	53.66	**58.82**

Table A.3. We selected an image excluding the specified object with frequently appearing top-3 background from the Pascal VOC 2012 dataset and evaluated the mIoU (%) of the localization map for those samples.

α	mIoU (%)
0.2	51.4
0.4	51.3
0.6	51.5
0.8	51.2
1.0	51.4
SMA (Ours)	**52.4**

Table A.4. Performance of localization map (mIoU) generated by interpolation-based method on various α values.

mIoU. Table A.4 summarizes the results of an interpolation-based method on various α values. As a result of the experiment, it was confirmed that the performance was limited compared to SMA. We hypothesize that the blending of background attributes hinders the classifier's ability to accurately distinguish the foreground.

B. More Visualizations

An overview of SMA. Figure B.1 shows the overall process of our proposed SMA.

Examples of bias-aligned and bias-conflicting samples. Figure B.2 presents examples of bias-aligned samples [4] which is the images of the object-background combination with the high co-occurrence frequency. Example images of bias-conflicting samples [2], in which certain objects and frequently appearing backgrounds are intentionally excluded, can be seen in Figure B.3.

More qualitative examples. Figure B.4 (a) shows examples of pseudo-masks predicted by IRN [1], CDA [5], and our proposed method SMA on Pascal VOC 2012 dataset, and Figure B.4 (b) presents comparison of qualitative results of pseudo-masks produced by AMN [3] and our method on MS COCO 2014 dataset.

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2209–2218, 2019. 1, 2, 4

[2] Jungsoo Lee, Eungyeup Kim, Juyoung Lee, Jihyeon Lee, and Jaegul Choo. Learning debiased representation via disentangled feature augmentation. *Advances in Neural Information Processing Systems*, 34:25123–25133, 2021. 1, 2

[3] Minhyun Lee, Dongseob Kim, and Hyunjung Shim. Threshold matters in wsss: Manipulating the activation for the robust and accurate segmentation model against thresholds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4330–4339, 2022. 2, 4

[4] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020. 2

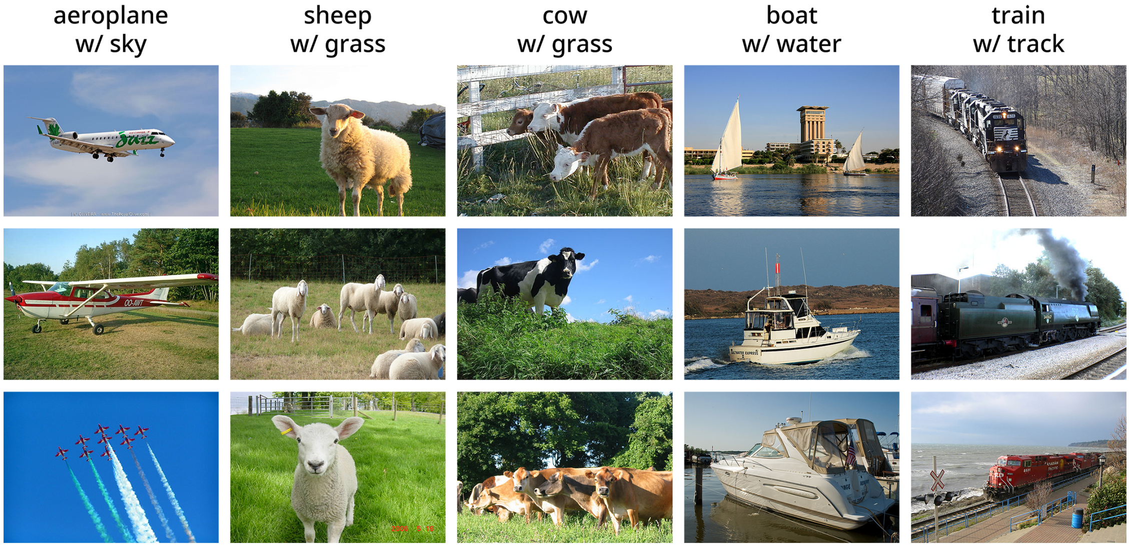


Figure B.2. Examples of **bias-aligned** samples, which refers to images that include a background that frequently appears with a specific object in the Pascal VOC 2012 dataset.

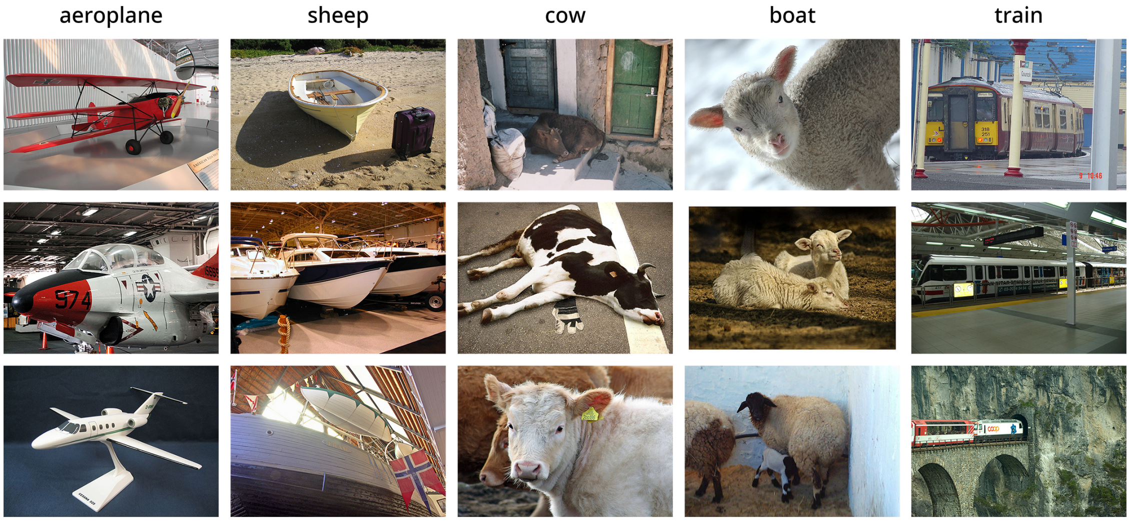
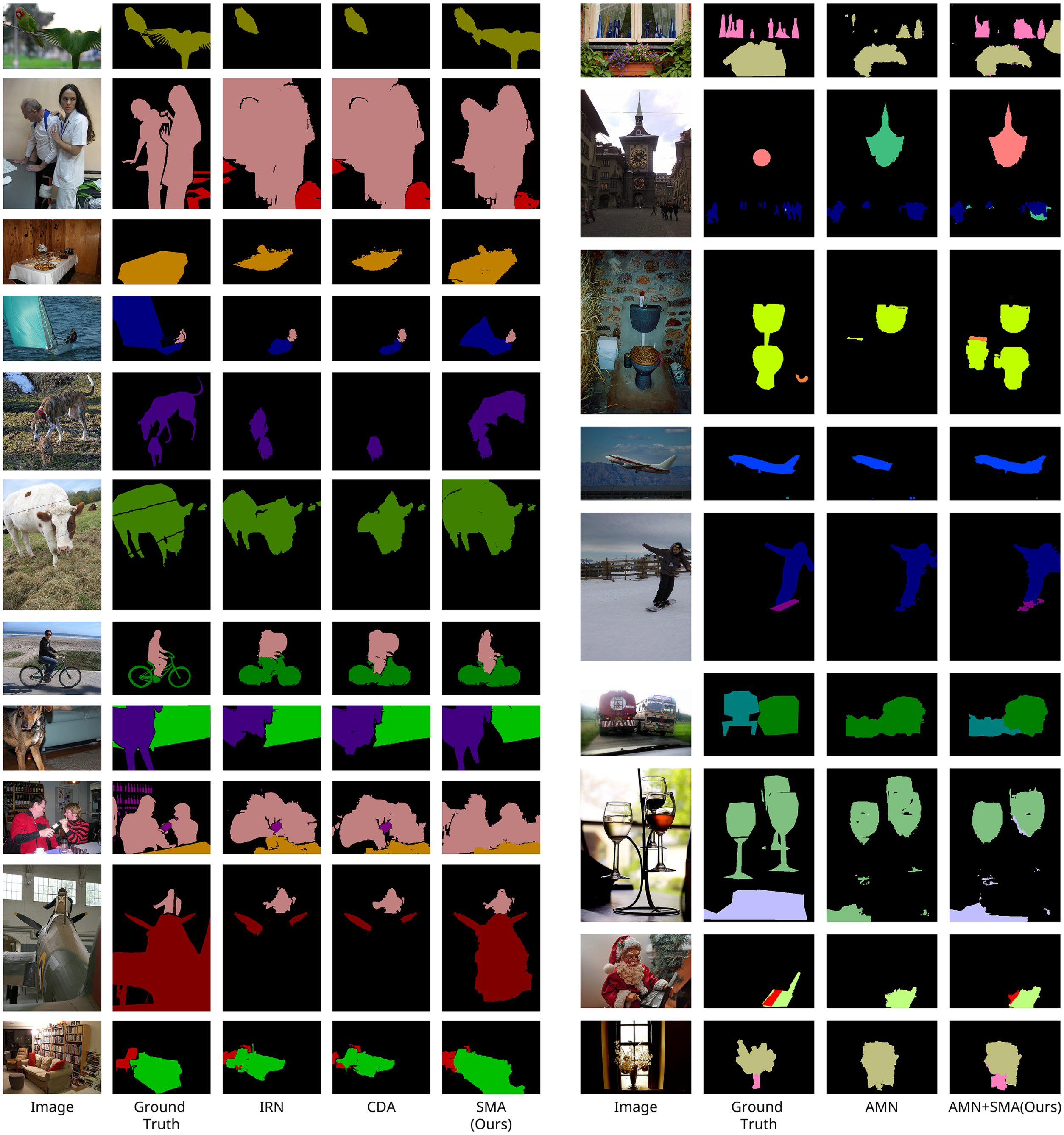


Figure B.3. Example of **bias-conflicting** samples, which is the images that do not include a background that usually appears with a specific object in the Pascal VOC 2012 dataset.

[5] Yukun Su, Ruizhou Sun, Guosheng Lin, and Qingyao Wu. Context decoupling augmentation for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7004–7014, 2021. 2, 4

[6] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 1

[7] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *ICLR*, 2021. 1



(a)

(b)

Figure B.4. Examples of pseudo-masks from (a) IRN [1], CDA [5] and SMA (Ours) on Pascal VOC 2012 (b) AMN [3] and SMA (Ours) on MS COCO 2014 datasets.