Class Tokens Infusion for Weakly Supervised Semantic Segmentation

Supplementary Material

To show the superiority and effectiveness of the proposed method, we conducted extensive experiments on PASCAL VOC 2012 and MS COCO 2014 datasets. This supplemental material includes visualization results on each dataset and training details on MS COCO dataset.

A. PASCAL VOC 2012

A.1. Visualization Results

CAMs In Fig. A1, we visualize the CAMs on PASCAL VOC 2012 dataset. As mentioned in the main paper, the Baseline [2] exhibits a challenge where class tokens have high correlation and are not distinct. This issue is evident in the fourth and fifth rows of the Fig. A1, where Baseline CAMs for one class encroach upon CAMs for other classes. For instance, in the fourth row of Fig. A1, the CAM of human also activate the region of the boat. In the fifth row, the CAM of the bird also activates the sheep's region. In contrast, our proposed approach fosters distinct representations among class tokens, alleviating the problem of overlapped CAMs. Additionally, we incorporate Background Token (BGT) in our method, guiding ViT to learn BG CAMs to mitigate false activations in ViT. This is illustrated in the second row of the figures, demonstrating a significant reduction in false positive activations.

Background CAM As we mentioned above, we guide the ViT to localize BG CAM in the training phase. Thus, to verify that BG CAM is properly trained, we demonstrate the BG CAMs in the Fig. A2. As shown in the figure, ViT properly localizes the background regions while rejecting the foreground regions. Referring to the figure in the fifth row, the BG CAM localizes well even when multi-class foreground objects exist.

t-SNE As shown in Fig.4 of the main paper, we compare the t-SNE of class tokens $\mathbf{T}^i_{cls}$ in layer $i \in \{2, 7, 11\}$ of Baseline and Ours. In the Fig. A3, t-SNE results of all layers $i \in \{0, 1, \cdots, 11\}$. As the class token interacts with the patch tokens more, the feature space of the class tokens from the baseline gets highly correlated from the third layer. On the contrary, with the proposed CTI, class tokens from our method are less correlated with distinct representation.

Semantic Segmentation By generating high-quality pseudo labels from the proposed methods and utilizing the labels for the training of the semantic segmentation model, we achieved a new state-of-the-art both in *val* and *test* sets. We visualize the semantic segmentation prediction results on the *val* set as in Fig. A4. Our model predicts with high accuracy even if the scene is complex. And, considering the fully supervised semantic segmentation model trained with a 1.4k train set achieves approximately 80% in mIoU, we minimize the gap by only using weak supervision. For a fair comparison with the prior work, we also evaluate our model on *test* set through an online benchmark server. The overall performance of our model can be found in this link.

B. MS COCO 2014

B.1. Training details

CAMs Our model is trained on MS COCO 2014 dataset with a maximum epoch of 60, but converges before 15 epochs. The balancing parameter for CTI is set to 0.1 (PASCAL = 1.0), and the CTI and Bg CAMs losses are computed after two epochs for stable convergence. With a single A6000 GPU, it takes 40 minutes for a single epoch with a batch size of 64, thus costing below 12 hours for the best model. We followed the learning rate and augmentation technique (color-jittering, transformations) used in the prior ViT-based work [1, 2].

Semantic Segmentation For a fair comparison with the prior WSSS work, we also utilize the Deeplab-V2 with ResNet 101 backbone. The initial learning rate is set to 5e-5 and decays with a polynomial scheduler. The power for the scheduler is set to 0.9. We used the SGD optimizer with weight decay 5e-4. During the training, the image is cropped to 321×321 with a batch size of 10.

B.2. Visualization Results

CAMs To show the effectiveness of the proposed method, we visualize the CAMs from MS COCO 2014 dataset in Fig. A5. The MS COCO dataset is more challenging than the PASCAL VOC dataset due to a greater variety of complex scenes and a larger number of existing classes. While ViT-based methods outperform CNN-based methods significantly on the PASCAL VOC dataset, it shows inferior performance on COCO, mainly due to the characteristic high false activation of ViT. However, our proposed method mitigates false activations and yields class-distinct CAMs, demonstrating high performance on COCO and surpassing CNN-based WSSS methods. Multi-class CAMs are illustrated in the top four rows of Fig. A5 and single-class CAMs are shown in the below.

Semantic Segmentation By training the semantic segmentation model with precise pseudo-ground-truth, our model achieves a new state-of-the-art. The visualized prediction results on MS COCO dataset are shown in Fig. A6.

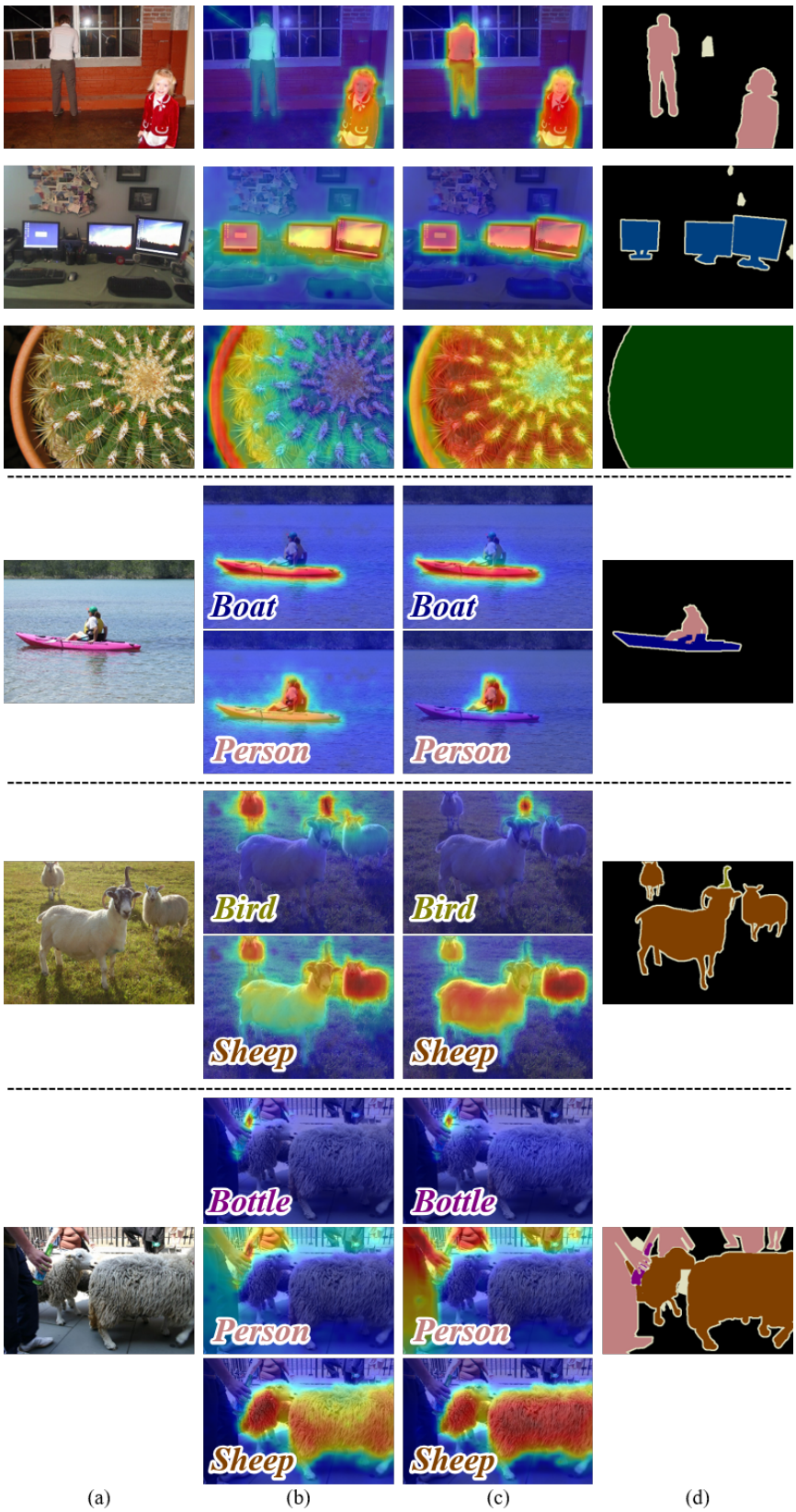


Figure A1. CAMs visualization results on *train* set. (a) Image, (b) Baseline CAMs, (c) Our CAMs, (d) Ground-truth.

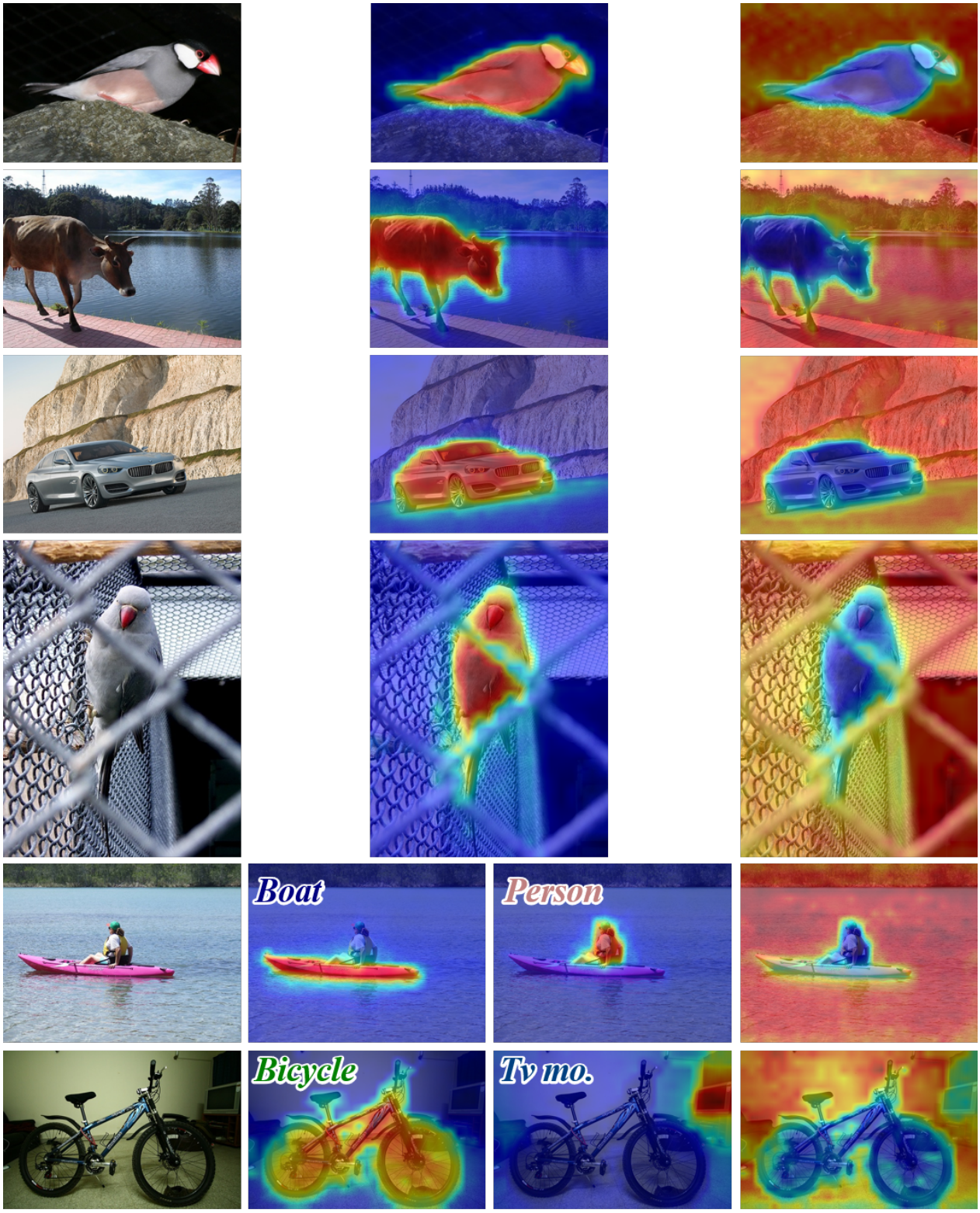


Figure A2. Visualization of background CAM and foreground CAMs on PASCAL VOC *train* set. (a) Image, (b) foreground CAMs (ours), (c) background CAM (ours).

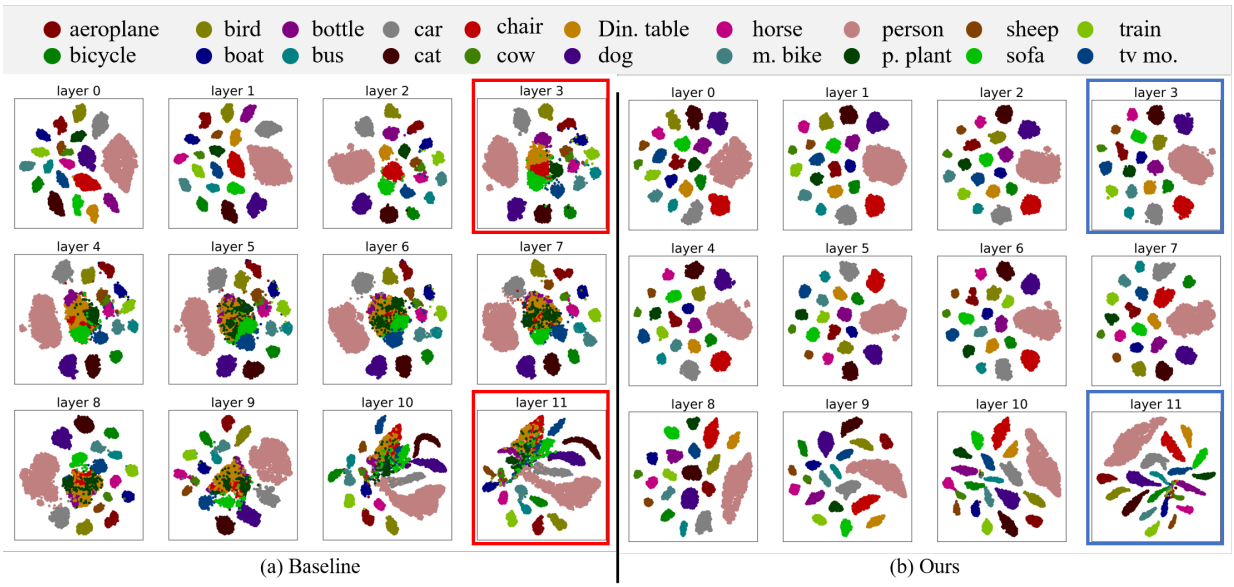
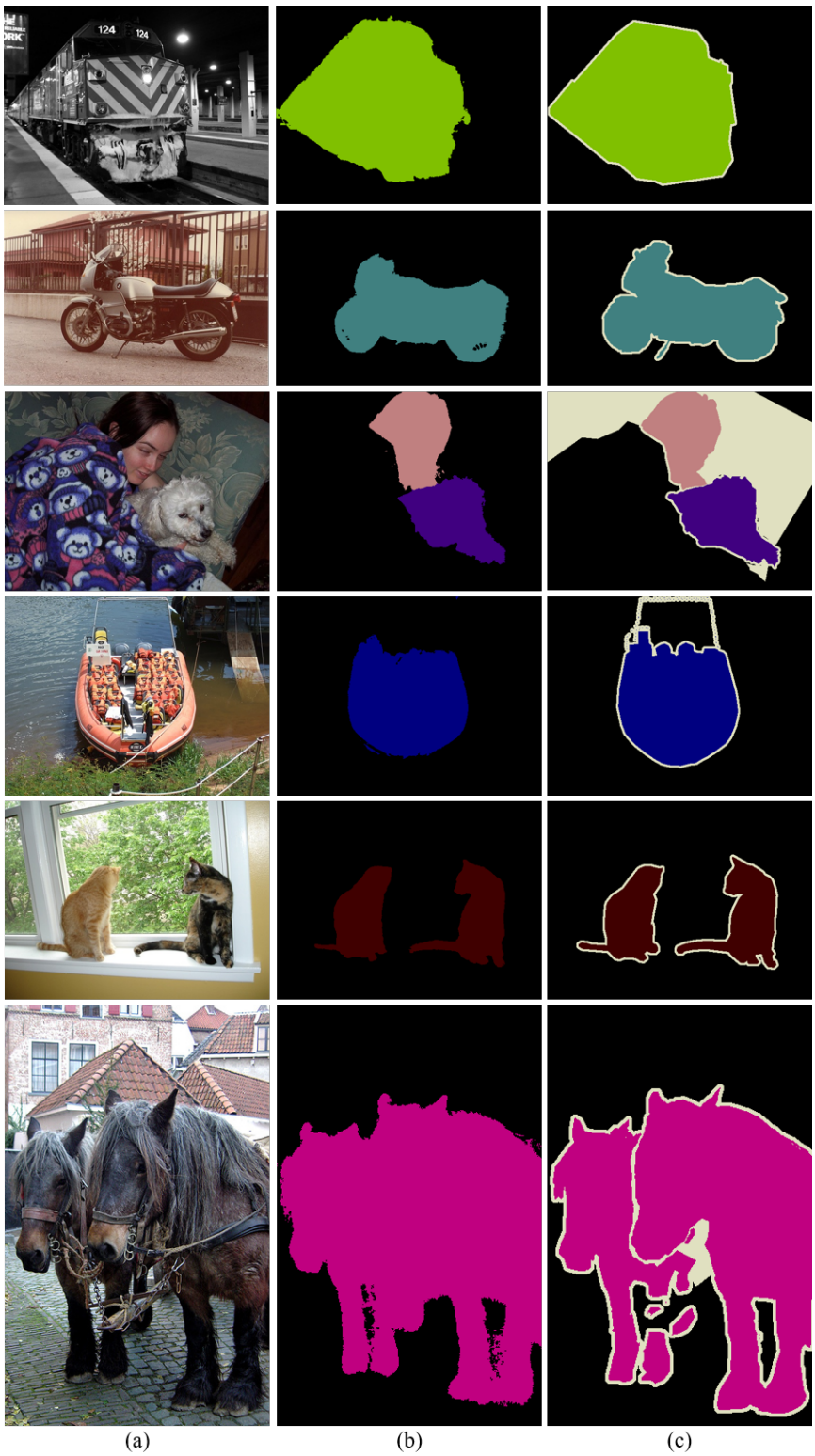


Figure A3. t-SNE comparison result between Baseline (left) and Ours (right). Class tokens from all layers are visualized and sampled from the PASCAL VOC *train* set (10,582 images).

References

[1] Wei Gao, Fang Wan, Xingjia Pan, Zhiliang Peng, Qi Tian, Zhenjun Han, Bolei Zhou, and Qixiang Ye. Ts-cam: Token semantic coupled attention map for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2886–2895, 2021. 1

[2] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. Multi-class token transformer for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4310–4319, 2022. 1



(a) (b) (c)

Figure A4. Semantic segmentation prediction results on the PASCAL VOC 2012 *val* set. From left to right: (a) Image, (b) Ours, (c) Ground-truth.

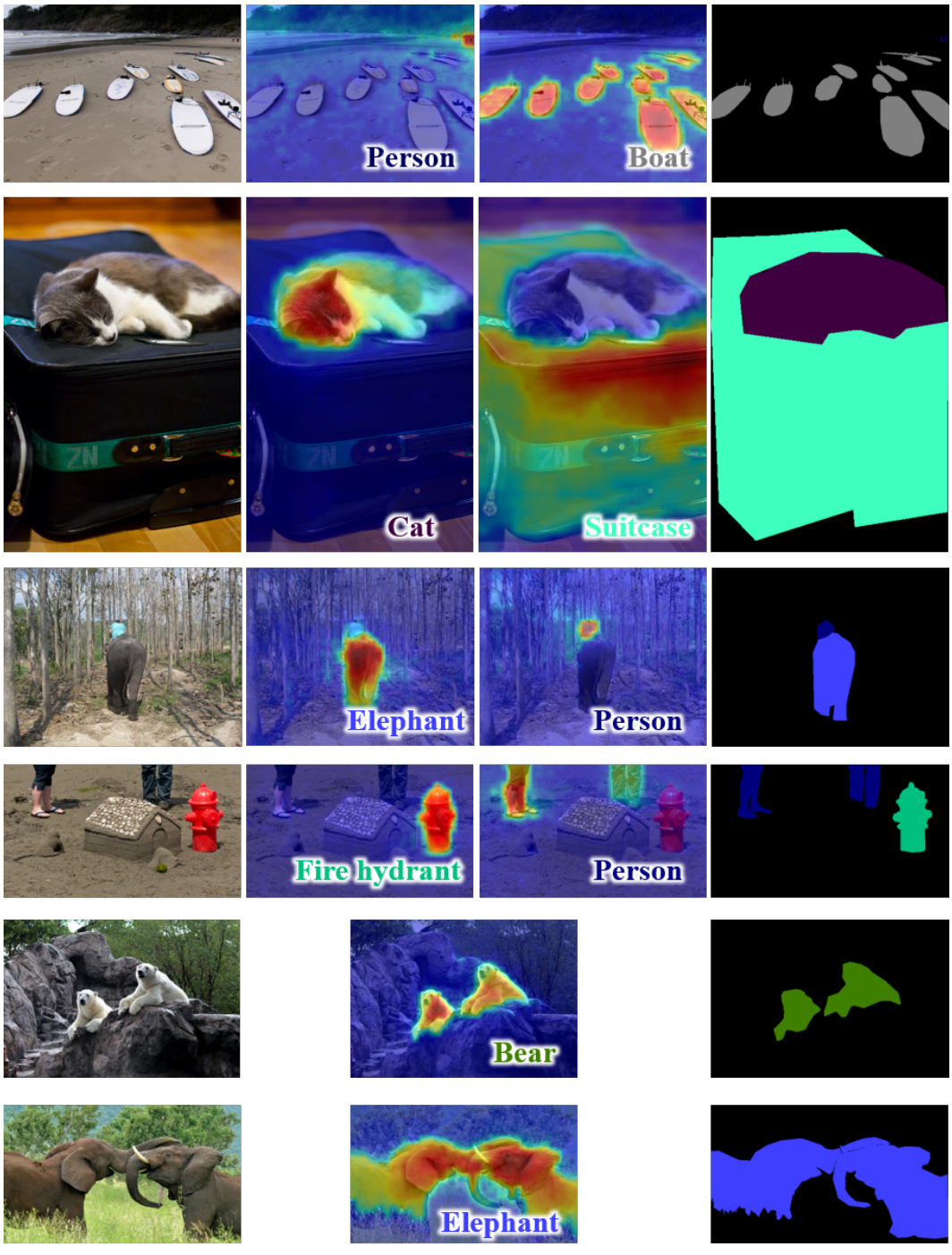
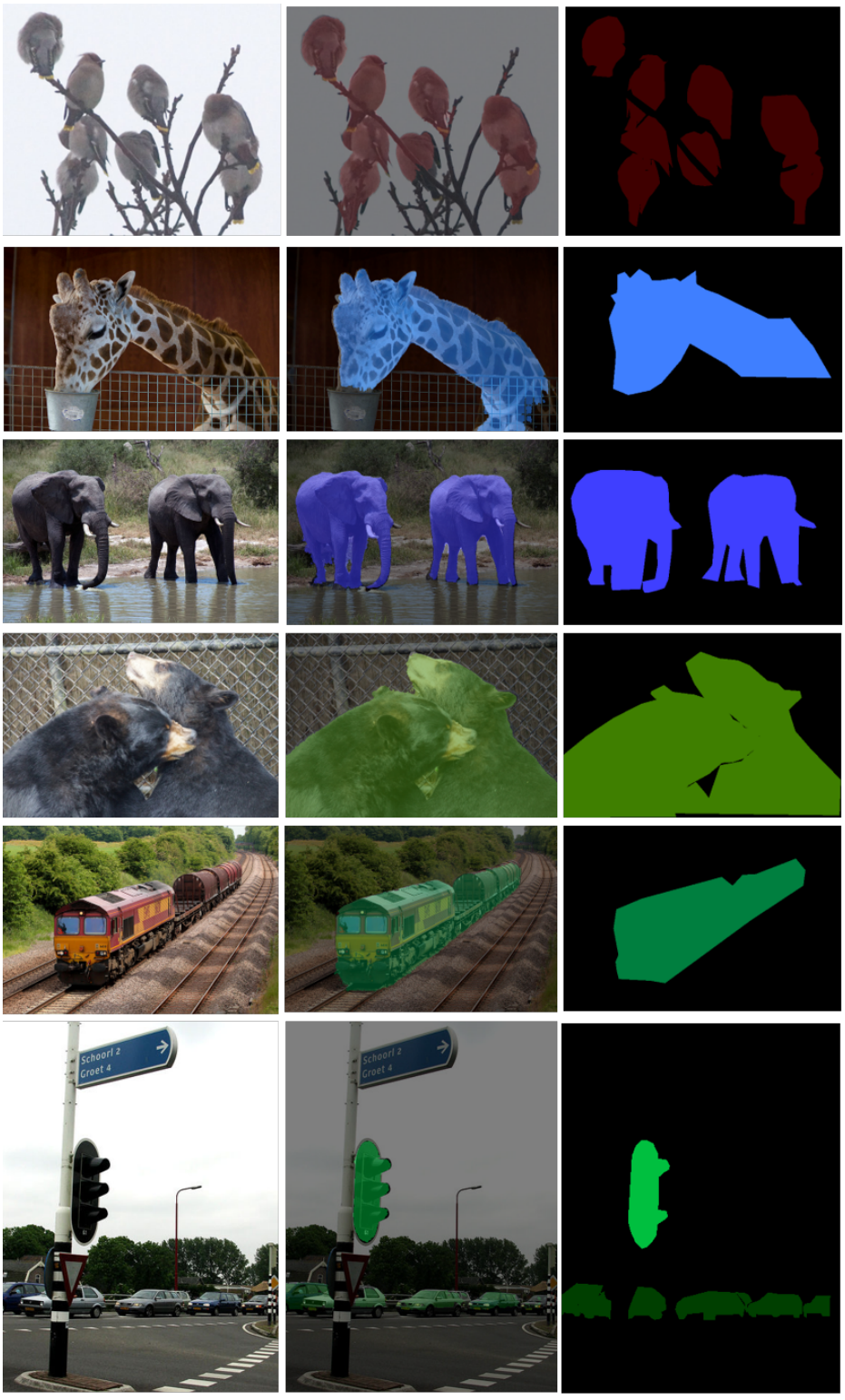


Figure A5. CAMs results on the MS COCO 2014 *train* set. From left to right: (a) Image, (b) Our CAMs, (c) Ground-truth.



(a) (b) (c)

Figure A6. Semantic segmentation results on the MS COCO 2014 *val* set. From left to right: (a) Image, (b) Semantic segmentation prediction overlaid with image, (c) Ground-truth.