Weakly-Supervised Semantic Segmentation via Sub-category Exploration

Yu-Ting Chang[1] Qiaosong Wang[2] Wei-Chih Hung[1] Robinson Piramuthu[2]
Yi-Hsuan Tsai[3] Ming-Hsuan Yang[1,4]
[1]UC Merced [2]eBay Inc. [3]NEC Labs America [4]Google Research

1. Overview

In this supplementary document, we provide additional visualizations and experimental results. Section 2 presents the performance comparison with the baseline. Section 3 shows visual comparisons of clustering results at different rounds of training. Finally, more qualitative results are presented in Section 4, including initial response maps, semantic segmentation results, and few failure cases.

2. Comparison with the Baseline Model

We compare our result with the AffinityNet [1] method which is our baseline model. The AffinityNet approach adopted wide ResNet-38 [3] as their backbone network to train the segmentation model, while we utilize the ResNet-101 architecture. Although these two backbones have been shown to perform competitively, for a fair comparison, we generate pseudo ground truth via the AffinityNet approach and re-train the segmentation network with the ResNet-101 backbone. In Table 1, we show the mean IoU of semantic segmentation results. We present the result of AffinityNet that applies both backbones, including the ones with and without CRF refinement. With the proposed method, the results validate that better semantic segmentation performance is achieved based on our improved initial response maps.

3. Quality of Clustering

Figure 1 to Figure 6 present exemplar results of the clustering for 6 different classes in the PASCAL VOC 2012 dataset [2] (i.e., class of bird, boat, cow, dog, sheep, and tv monitor). For each class, we show the clustering result of the first round and the third round model, which are the beginning and the final clustering result during the iterative training process.

By observing the visual change between Round-1 and Round-3, we can find the quality of clustering results is enhanced. For example, in Figure 1, images of the bird flying in the sky are clustered into different clusters at Round-1, yet such images can be gathered into the same cluster at Round-3. Note that in Figure 1 to Figure 6, we use yellow boxes to mark the images that have a similar visual style but are clustered into different clusters at Round-1. Whereas images with the coherent visual style are clustered into one cluster at Round-3, which is marked by red boxes.

Results of the final round of clustering present the consistency within a cluster, including size, type, context, and the interaction between objects. The visual change between the first and the last round of clustering demonstrates the effectiveness of our iterative training process, which validates that our learned feature representations are enhanced via the sub-category objective in an unsupervised manner.

Table 1: Performance comparison with baseline models in mIoU (%) for evaluating semantic segmentation results on the PASCAL VOC validation set.

Method	Segmentation Backbone	w/o CRF	w/ CRF
AffinityNet [1]	Wide ResNet-38	61.7	–
AffinityNet [1]	ResNet-101	61.9	63.9
Ours	ResNet-101	64.8	66.1

4. Qualitative Comparisons

In this section, we provide additional results for qualitative comparisons, including the visual comparison of initial response prediction and segmentation result. Figure 7 presents both initial response maps and segmentation results of the AffinityNet [1] method and ours. We illustrate intermediate results to demonstrate that a better initial seed can benefit the quality of segmentation result, and a number of qualitative examples of our final model are presented in Figure 8.

4.1. Images with Single-object

We analyze how our method benefits completeness of the response map for single-object images and provide example results. The motivation of our approach is as follows. For one parent class, a single label for all of these variations encourages classification models to pick common discriminative attributes (e.g., facial features) across these variations. Therefore, the network attends to highly discriminative regions (e.g., face), resulting in a sparse response map for localization. In Figure 9, it is likely for the network to attend on cat faces but miss the body, since face is sufficient for classifying and is the *common* area to tell cat apart from other parent classes. Our self-supervised task of classifying sub-categories is richer and can guide the network to additionally look at *other* parts of the objects.

4.2. Failure Case

In addition, we also show some failure segmentation cases in Figure 10. There are two main issues that would affect the quality of segments: 1) the incompleteness on detailed parts and 2) the ambiguity on object boundaries, which are two challenging problems of the segmentation task. Although there are some failure examples, our approach can produce high quality semantic segmentation results.

References

[1] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *CVPR*, 2018. 1, 2, 9

[2] Mark Everingham, Luc J. Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010. 1

[3] Zifeng Wu, Chunhua Shen, and Anton Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90:119–133, 2019. 1

Figure 1: Visual comparison of the different rounds of clustering of *bird* class. Red boxes at later round demonstrate sets of image with visual consistency compared to images marked by yellow boxes at the beginning round.

Figure 2: Visual comparison of the different rounds of clustering of *boat* class. Images of sailboat are in different clusters at Round-1 while such image can be clustered to one cluster (i.e., Cluster-10) at Round-3.

Figure 3: Visual comparison of the different rounds of clustering of *cow* class. Images of cattle in the meadow in distant view are in different clusters at Round-1 while such images can be clustered to one cluster (i.e., Cluster-8) at Round-3.

Figure 4: Visual comparison of the different rounds of clustering of *dog* class. Images of dog in the meadow are in different clusters at Round-1 while such images can be clustered to one cluster (i.e., Cluster-4) at Round-3.

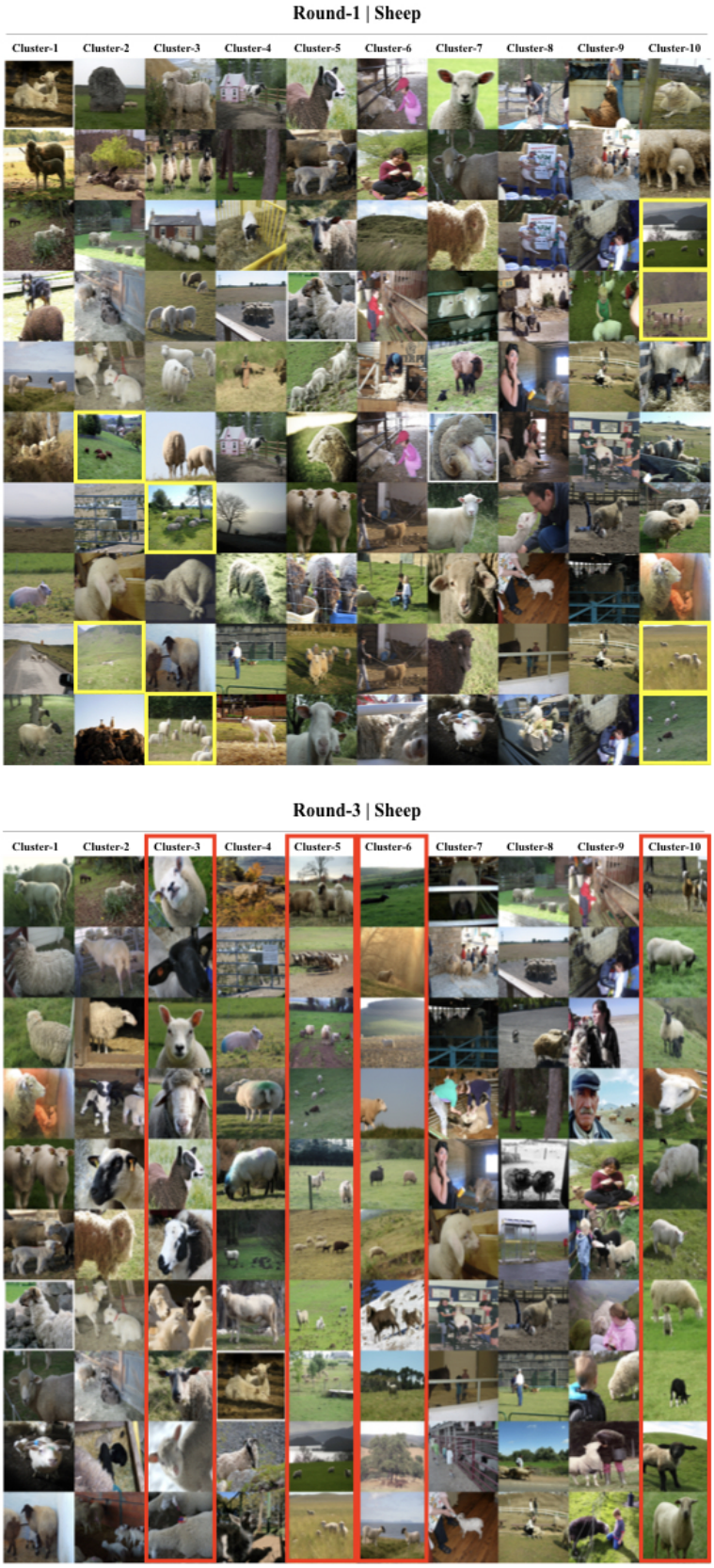

Figure 5: Visual comparison of the different rounds of clustering of *sheep* class. Images of sheep in the meadow in distant view are in different clusters at Round-1 while such images can be clustered to one cluster (i.e., Cluster-5) at Round-3.

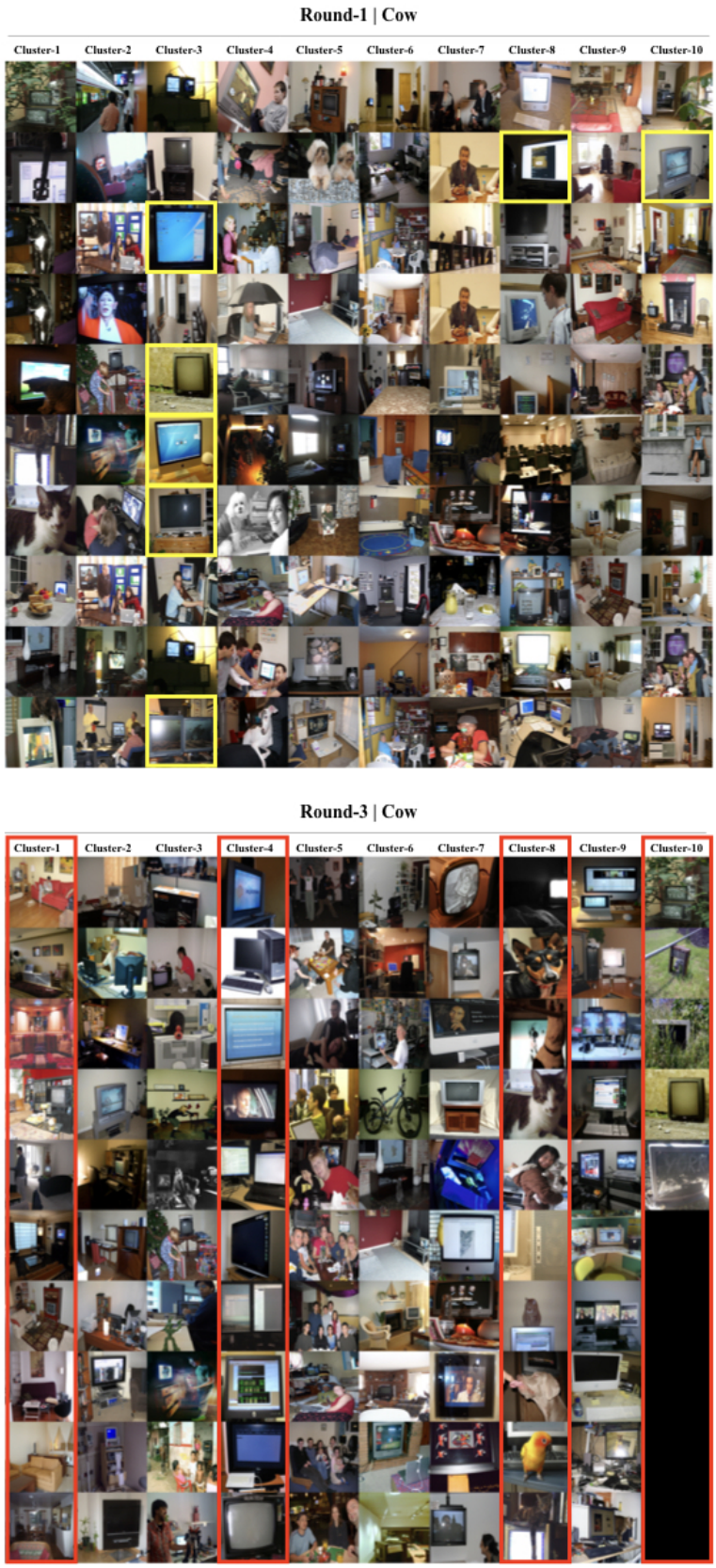

Figure 6: Visual comparison of the different rounds of clustering of *monitor* class. Images of large screen are in different clusters at Round-1 while such images can be clustered to one cluster (i.e., Cluster-4) at Round-3.

Figure 7: Qualitative comparison of the initial response map and semantic segmentation map. We compare our intermediate and final results with the AffinityNet [1] approach.

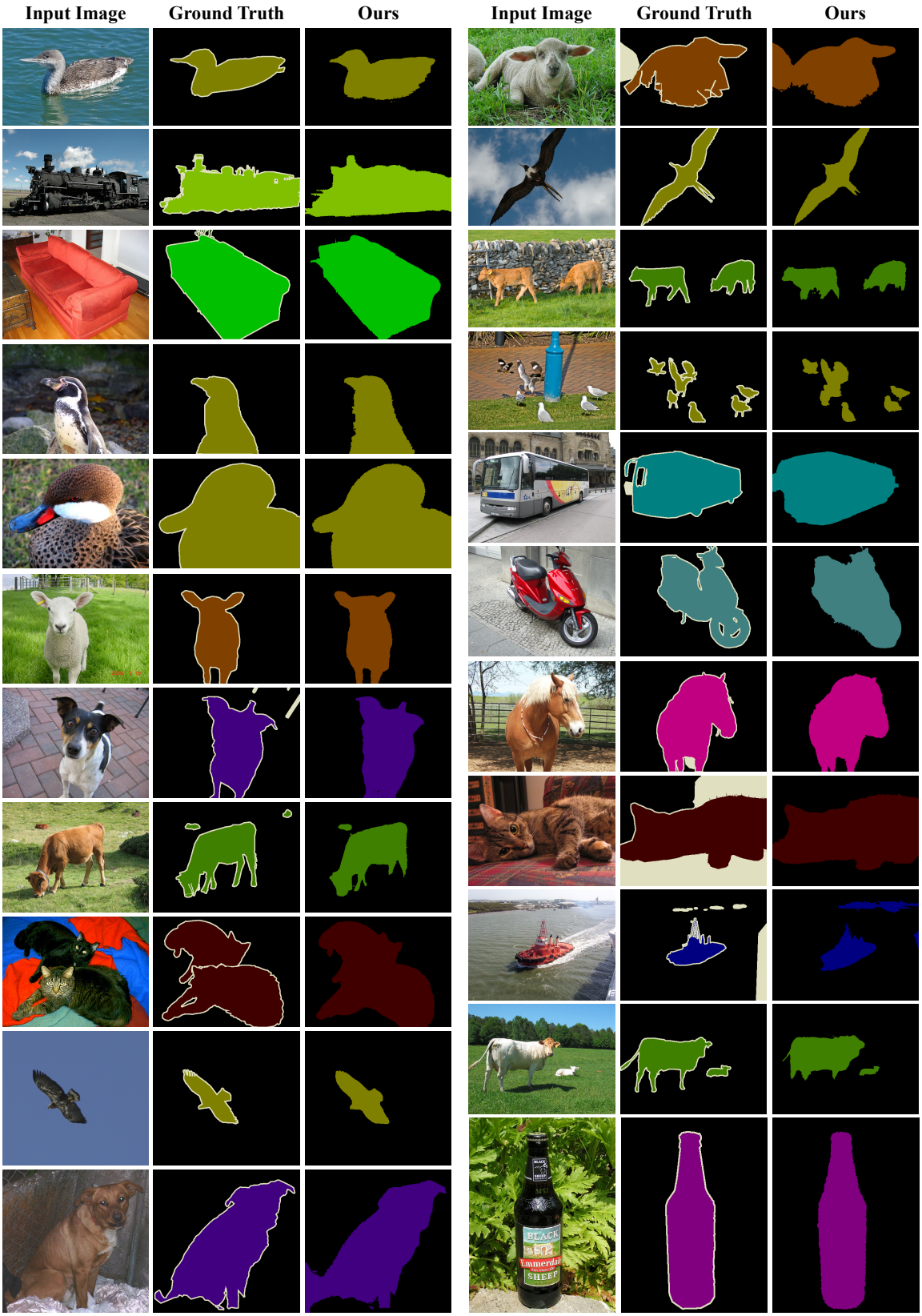

Figure 8: Semantic segmentation results on the PASCAL VOC 2012 val images.

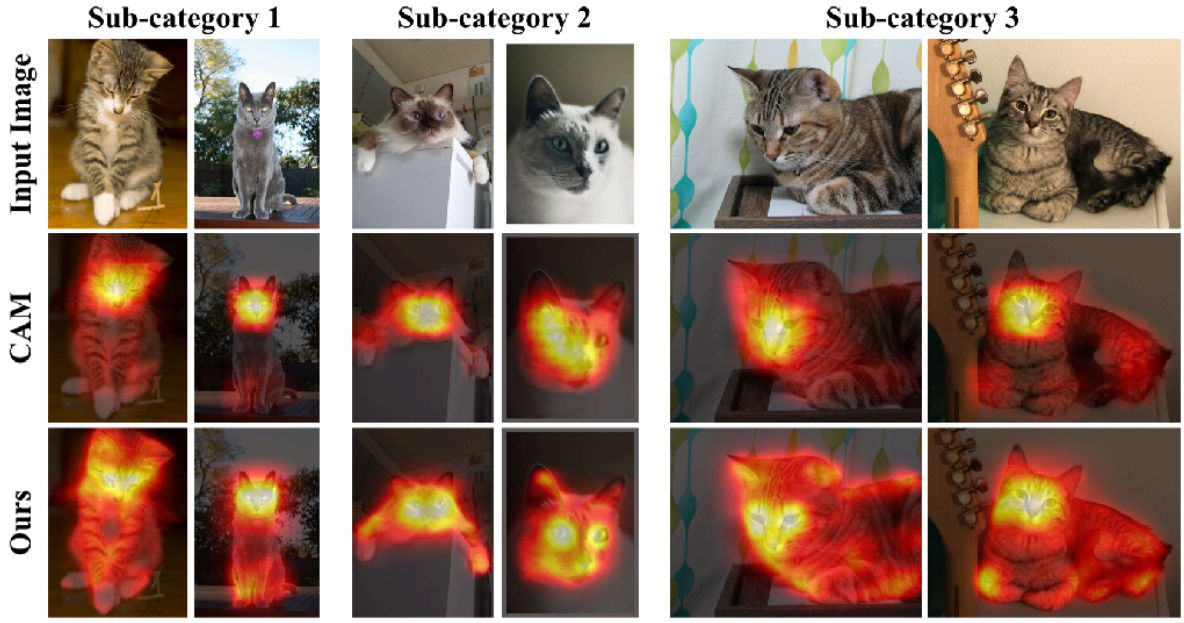

Figure 9: Response maps of three sub-categories of *cat*. All of them contain only one object, but in different variations. Results show that each sub-category is able to capture a particular form of the object. The resulting response maps cover more complete object regions compared to the original CAM approach.

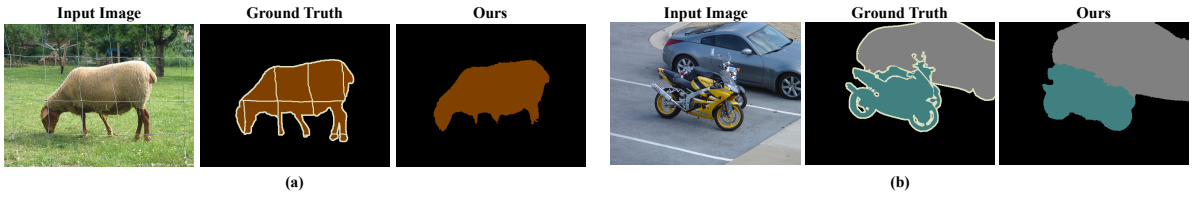

Figure 10: Failure semantic segmentation results. (a) Failure cases of the incompleteness on detailed parts. Legs of the animal are missing in the segment. (b) Failure case of the ambiguity on object boundary. There are errors on the boundary region between two objects.