

Supplementary Materials: Few-shot Semantic Segmentation via Perceptual Attention and Spatial Control

Anonymous Authors

In this supplemental material, we first elaborate on additional descriptions of datasets (i.e., PASCAL-5ⁱ and MS COCO). Then, we explore weight settings of two loss functions we used. Afterwards, we explore the impact of different k -shot fusion solutions on segmentation. Finally, we show detailed results of some experiments, which are shown as overall results in our manuscript due to limited space.

1 DETAILED DESCRIPTION OF DATASETS

PASCAL VOC 2012 [1] has 21 classes (including a background class) and three subsets (i.e., training (train), validation (val), and test with 1464, 1449, and 1456 images, respectively). Following [3], we use PASCAL VOC 2012 and additional annotations from SDS to build an augmented dataset PASCAL-5ⁱ, where 80 categories are divided into 4 splits and each split contains 5 categories. The category division is shown in Table 1. MS COCO 2014 [2] contains 81 classes, including a background class, which is divided as shown in Table 2. This dataset contains 80K training images and 40K validation images. Although the two datasets contain labels for object detection and semantic segmentation, we only used image-level class labels for the FSS task.

Table 1: The class division of the PASCAL-5ⁱ dataset.

Split	Categories
0	aeroplane, bicycle, bird, boat, bottle
1	bus, car, cat, chair, cow
2	diningtable, dog, horse, motobike, person
3	potted plant, sheep, sofa, train, tv/monitor

2 SETTING THE WEIGHTS OF LOSS FUNCTIONS

To explore the optimal weight setting for two loss functions, we compare the results of DiffSeg training with different weights of loss functions, as shown in Table 3.

The image loss focuses on the segmentation results, while the latent loss pays attention to the output of perceptual comparison. It can be seen from Table 3 that the impact of different weights on the results is limited, but when the weights of the two losses are 2 and 1, the best results are achieved.

3 K-SHOT SOLUTIONS

In order to explore the impact of different k -shot fusion solutions on segmentation, we compared the results with different fusion solutions, i.e. Logical OR, Average fusion for masks (Mask-avg), Average fusion for features (Feature-avg) and Attention.

Table 2: The class division of the MS COCO dataset.

Split	Categories
0	airplane, apple, backpack, banana, baseball bat baseball glove, bear, bed, bench, bicycle bird, boat, book, bottle, bowl broccoli, bus, cake, car, carrot
1	cat, cell phone, chair, clock, cow cup, dining table, dog, donut, elephant fire hydrant, fork, frisbee, giraffe, hair drier handbag, horse, hot dog, keyboard, kite
2	knife, laptop, microwave, motorbike, mouse orange, oven, parking meter, person, pizza potted plant, refrigerator, remote, sandwich, scissors sheep, sink, skateboard, skis, snowboard
3	sofa, spoon, sports ball, stop sign, suitcase surfboard, teddy bear, tennis racket, tie, toaster toilet, tooth brush, traffic light, train, truck tv monitor, umbrella, vase, wine glass, zebra

Table 3: Results of DiffSeg training with different weights of loss functions.

Weights	1-shot	5-shot
3,1	69.2	70.8
2,1 (ours)	69.3	72.1
1,1	68.9	69.7
1,2	69.0	71.5
1,3	68.2	70.4

The results are shown in Table 4, where mask-based fusion methods (i.e., Logical OR and Mask-avg) perform fusion operations on segmentation results, which cannot improve the performance. Thanks to the ability of Clip to align image with text and the great prior knowledge of diffusion, the fusion on feature level and attention have equivalent results, where attention is proven to be an effective k -shot fusion method in [4]. However, computational burdens of attention operations are significantly heavier than Feature-avg. Therefore, we select simple but effective average fusion as our k -shot solution.

Table 4: Results of DiffSeg with different 5-shot solutions.

Method	5-shot	increment
1-shot baseline	69.3	0
Logical OR	52.7	-16.6
Mask-avg	69.8	0.5
Feature-avg (ours)	72.1	2.8
Attention	72.1	2.8

Table 5: Results of DiffSeg with different modules in 1-shot and 5-shot segmentation.

UNet	ECM	PCM	1-shot					5-shot				
			split-1	split-2	split-3	split-4	mean	split-1	split-2	split-3	split-4	mean
×	×	√	54.9	59.5	51.0	49.5	53.7	57.8	61.7	50.8	53.6	56.0
√	×	√	65.8	71.1	63.9	62.3	65.8	69.3	74.7	65.7	59.9	67.4
√	√	×	43.8	49.5	38.9	36.5	42.2	45.3	52.0	41.6	39.5	44.6
√	√	√	70.5	75.6	67.9	63.2	69.3	72.8	77.6	68.7	69.3	72.1

Table 6: Results of DiffSeg in 1-shot and 5-shot segmentation when different features are selected as prior knowledge.

Methods	SA	CA	1-shot					5-shot				
			split-1	split-2	split-3	split-4	mean	split-1	split-2	split-3	split-4	mean
Attention	√	×	61.8	66.0	59.4	50.4	59.4	61.7	68.6	57.5	63.0	62.7
	×	√	65.5	73.2	64.6	57.9	65.3	66.1	73.4	63.0	65.1	66.9
	√	√	70.5	75.6	67.9	63.2	69.3	72.8	77.6	68.7	69.3	72.1
FRB	—	—	51.4	57.6	47.4	44.0	50.1	53.5	58.7	47.7	51.3	52.8
FST	—	—	54.8	57.4	52.9	45.3	52.6	55.4	61.0	52.6	47.8	54.2

Table 7: Results of DiffSeg in 1-shot and 5-shot segmentation when certain attention operations are removed or changed in PAM. “×” and “○” present removing attention and changing attention to convolutional operations, respectively.

Att_1	Att_2	Att_3	1-shot					5-shot				
			split-1	split-2	split-3	split-4	mean	split-1	split-2	split-3	split-4	mean
×	×	×	45.2	48.3	41.9	41.8	44.3	47.7	50.7	44.5	43.1	46.5
×	√	√	58.3	59.1	61.3	55.7	58.6	60.4	61.3	63.4	56.6	60.4
√	×	√	51.6	57.1	54.7	45.0	52.1	53.4	58.9	56.6	45.9	53.7
√	√	×	64.0	70.5	64.1	58.2	64.2	67.0	73.4	67.2	59.5	66.8
○	○	○	58.2	65.1	56.8	56.7	59.2	60.0	66.7	59.0	58.8	61.1
○	√	√	63.1	68.5	59.2	58.8	62.4	64.6	69.6	60.6	60.0	63.7
√	○	√	59.9	67.8	55.6	56.3	59.9	61.3	69.2	57.3	58.3	61.5
√	√	○	66.3	71.2	63.8	60.3	65.4	68.4	72.7	65.9	61.8	67.2
√	√	√	70.5	75.6	67.9	63.2	69.3	72.8	77.6	68.7	69.3	72.1

Table 8: Results of DiffSeg in 1-shot and 5-shot segmentation when SCM is inserted into different levels of diffusion UNet.

8×8	16×16	32×32	Params sharing	1-shot					5-shot				
				split-1	split-2	split-3	split-4	mean	split-1	split-2	split-3	split-4	mean
×	√	√	√	63.0	68.5	63.2	55.3	62.5	65.6	71.2	66.9	59.1	65.7
√	×	√	√	61.3	67.1	59.6	53.2	60.3	63.2	68.7	61.1	55.5	62.1
√	√	×	√	67.9	74.5	66.4	62.8	67.9	70.5	76.4	68.4	65.5	70.2
√	√	√	×	70.2	75.1	68.4	63.1	69.2	72.7	77.2	70.9	67.6	72.1
√	√	√	√	70.5	75.6	<u>67.9</u>	63.2	69.3	72.8	77.6	<u>68.7</u>	69.3	72.1

4 DETAILED EXPERIMENTAL RESULTS

Due to limited space in our manuscript, we only show overall results for some experiments. In this supplementary, we present detailed results of these experiments.

The detailed results of our method with different modules, different selection of prior knowledge are shown in Tables 5 and 6, respectively. Results of DiffSeg when certain attention operations are removed or changed in PAM and when SCM is inserted into different levels of diffusion UNet are shown in Tables 7 and 8, respectively. Fig. 1 and 2 show some additional qualitative results of DiffSeg.

REFERENCES

- [1] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. 2010. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* 88, 2 (2010), 303–338.
- [2] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Proceedings of ECCV*. 740–755.
- [3] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. 2017. One-Shot Learning for Semantic Segmentation. In *Proceedings of BMVC*.
- [4] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. 2019. CANet: Class-Agnostic Segmentation Networks With Iterative Refinement and Attentive Few-Shot Learning. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5217–5226.

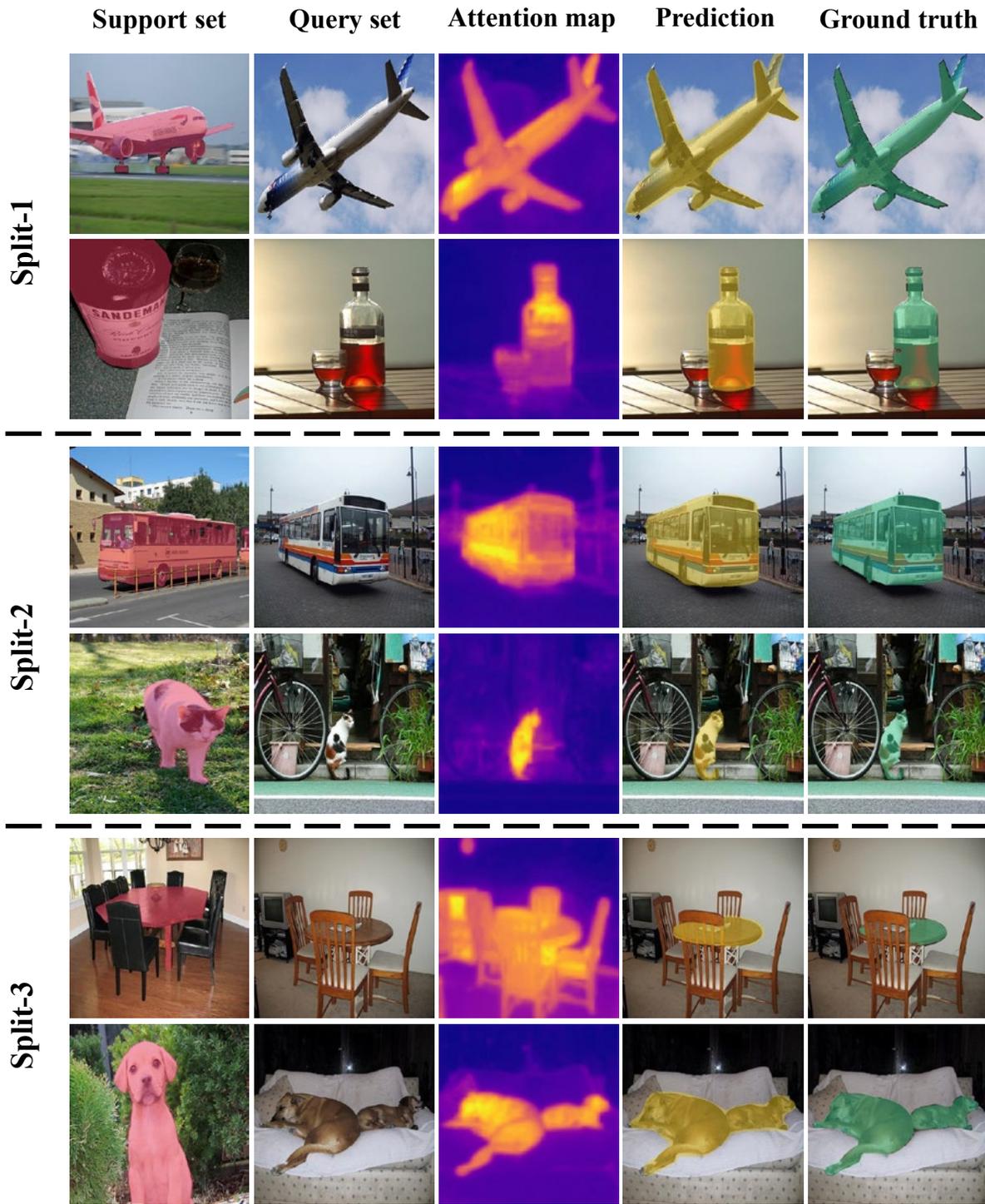


Figure 1: Additional qualitative results of our method. (Part 1)

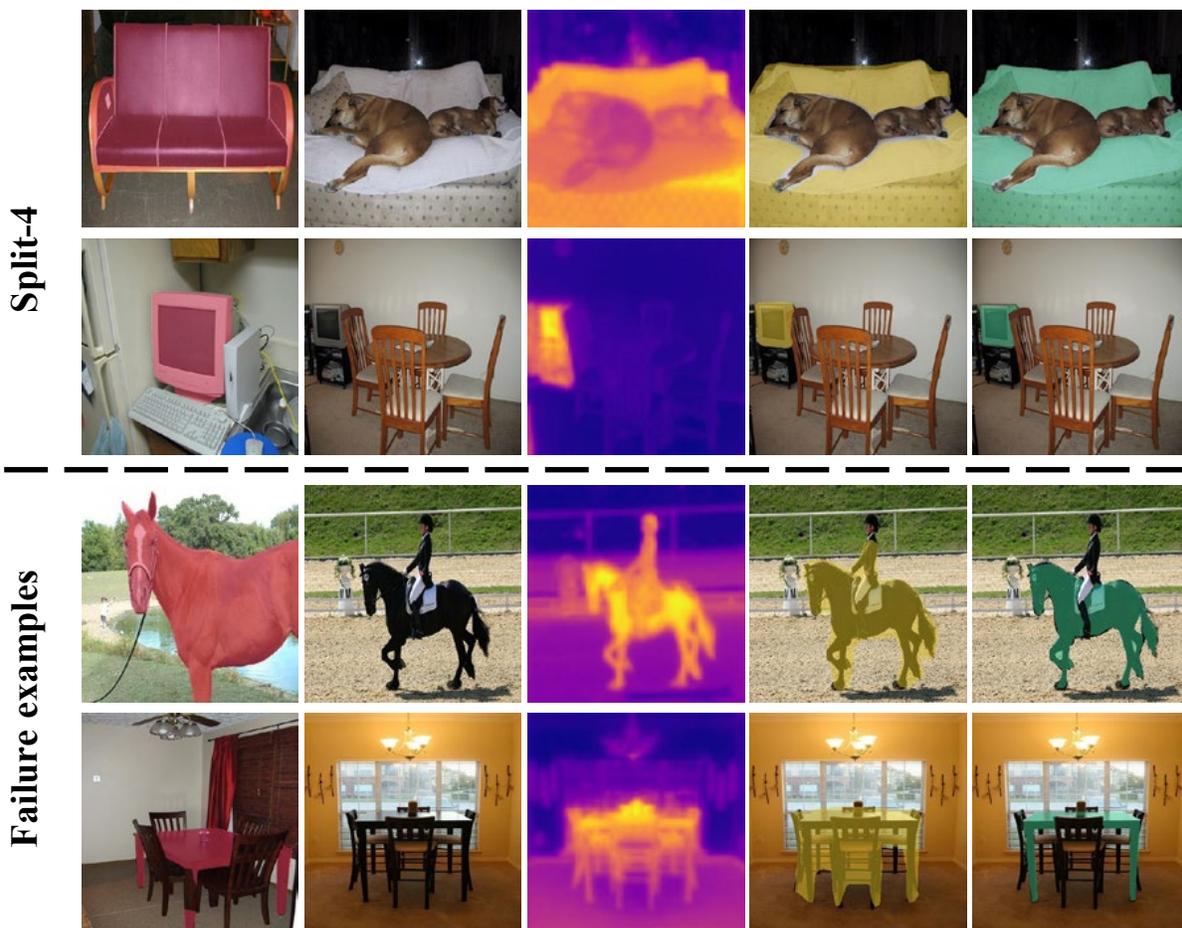


Figure 2: Additional qualitative results of our method. (Part 2)