Delving into Shape-aware Zero-shot Semantic Segmentation
Supplementary Material

Xinyu Liu[1,2], Beiwen Tian[2,4], Zhen Wang[3], Rui Wang[3], Kehua Sheng[3], Bo Zhang[3],
Hao Zhao[2], Guyue Zhou[2]
[1]Xidian University [3]Didi Chuxing
[2]Institute for AI Industry Research (AIR), Tsinghua University
[4]Department of Computer Science and Technology, Tsinghua University
liuxinyu@stu.xidian.edu.cn, zhaohao@air.tsinghua.edu.cn

In this supplementary material, we provide more experiment results to show the effectiveness of **S**hape-**A**ware **Z**ero-**S**hot semantic segmentation framework (**SAZS**). In the following sections, we provide per-category evaluation results. Then, more visualization results on PASCAL-5i and COCO-20i are displayed. Lastly, more scatterplots of compactness (CO) are presented.

Our code and checkpoints can be found at SAZS.

1. Per-Category Evaluation

Tab. 2, Tab. 3 demonstrate our per-category zero-shot semantic mIoU results on COCO-20^i [2] and PASCAL-5^i [1] respectively. The mIoU of SAZS demonstrates the superior performance of our proposed network structure And we observe that some categories often appear as small regions, like tie, or have a complicated internal structure, like person. For these categories, textual feature guidance alone cannot provide sufficient information for semantic parsing. Hence baseline without shape-aware cannot segment under self-supervision effectively. However, when using a SAZS model, the mIoUs of these categories are better aligned with shapes of objects than baseline, which verifies shape awareness does help zero shot learning.

2. Speed and Complexity

We conduct experiments by analyzing the per-episode inference time and floating point operations per second (FLOPs) to demonstrate the complexity of the proposed approach. Tab. 1 summarizes the results on COCO-20^i dataset. Compared with the baseline without fusion module, the inference time of SAZS is slower but the performance of SAZS is much better. Even though losses including L_{shape} in our model do not introduce time cost during inference, there is still room for optimization regarding inference speed and model complexity, which is exactly the direction for our future exploration.

Model	Backbone	mIoU	time(s)	FLOPS(G)
w/o fusion	DRN	26.6	177.43	275.76
w/o fusion	ViT-L	29.1	196.95	345.99
SAZS	DRN	35.2	230.54	275.76
SAZS	ViT-L	35.3	222.52	345.99

Table 1. More quantitative results on COCO-20^i.

3. More Qualitative Results

In this section, we present additional qualitative results on PASCAL-5i and COCO-20i using our model with ViT-L backbone. Specifically, Fig. 2 shows the results on PASCAL-5i. All categories are novel (unseen) in their corresponding fold. Taking into account the variety of images, we display all different categories of visualizations of SAZS. As shown in Fig. 2, SAZS achieves precise semantic parsing in all these scenes. For example, bicycle, diningtable, and tvmonitor in Fig. 2 show the ability of SAZS to discriminate target semantic objects from other objects (distractors), such as person, dog, keyboard. Furthermore, in Fig. 2, train, pottedplant, and tvmonitor, the model segments are precise even if the target instance contains more than one.

The visualization of COCO-20i is shown in Fig. 3, with both seen and unseen categories are displayed. We select 20 various scene and attribute labels with different semantics and multiple objects. Facing a more noisy and complex scene, SAZS is still able to recognize the novel(unseen) categories that are small and complicated, for example, broccoli, pottedplant and skis in Fig. 3. Particularly, in the second image in lines 2 and 3 of Figure 1, where multiple species appear in the scene with multiple objects and complex shapes, SAZS can accurately distinguish broccoli, carrots and hot dogs with sharp object edge segmentation.

Considering the diversity of scenes, we believe SAZS is precise enough for various applications including open scenario understanding and intelligent service robots.

4. More Scatter Analysis

Fig. 1 provides more scatterplots and the corresponding Pearson analysis results on the pascal dataset. The coordinates of the sample points in Fig. 1 represent the IoU result and CO variance of the corresponding model, and they are all negatively correlated. The results show that shape-aware can increase the correlation between the per-category iou results of our approaches and CO. For example, in the third column of the Fig. 1, the Pearson correlation coefficient r of SAZS is 0.13 higher than the baseline.

References

[1] Mark Everingham, SM Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 1

[2] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

Table 2. Per-category zero-shot semantic segmentation results on COCO-20^i.

Method	Backbone	person	Bicycle	Car	motorbike	aeroplane	Bus	train	truck	boat	trafficlight	firehydrant	stopsign	parkingmeter	bench	bird	cat	dog	horse	sheep	cow	mIoU
Baseline	DRN	35.7	55.5	38.2	43.6	69.2	69.9	15.1	27.2	**20.2**	**24.0**	12.2	5.9	57.2	66.5	11.9	43.3	12.3	**19.7**	25.3	20.7	35.2
Ours	DRN	**36.0**	**61.5**	**38.5**	**55.7**	66.7	72.2	**17.9**	**29.4**	16.7	14.4	12.2	5.9	53.8	65.8	11.5	**46.2**	13.5	15.5	**27.6**	**22.5**	35.2
Baseline	ViT-L	35.7	55.1	32.1	47.2	**75.6**	**83.5**	16.2	20.3	16.1	12.4	12.2	5.9	**60.1**	**72.2**	12.0	36.3	11.0	15.8	25.2	20.7	34.7
Ours	ViT-L	35.7	56.5	33.4	48.2	74.7	83.2	16.2	25.0	17.6	13.1	12.1	**7.3**	56.4	71.9	**12.3**	35.3	**13.8**	17.6	25.3	21.1	**35.3**

Method	Backbone	elephant	bear	zebra	giraffe	backpack	umbrella	handbag	tie	suitcase	frisbee	skis	snowboard	sportsball	kite	baseballbat	baseballglove	skateboard	surfboard	tennisracket	bottle
Baseline	DRN	**25.6**	**67.0**	**23.9**	16.5	**64.6**	74.5	35.5	27.5	**55.3**	**49.7**	10.3	21.1	41.0	78.8	**28.7**	33.5	18.8	18.2	12.1	60.5
Ours	DRN	24.3	63.8	23.2	16.5	57.0	74.5	**35.8**	38.2	53.3	40.6	**10.9**	21.1	38.5	63.9	20.9	30.3	20.3	18.2	**15.9**	62.2
Baseline	ViT-L	15.9	61.3	18.8	**16.9**	60.0	**79.0**	35.5	52.1	51.9	47.0	10.4	21.1	**43.7**	**85.6**	20.1	**33.9**	**24.7**	18.9	13.0	**69.9**
Ours	ViT-L	16.2	57.1	19.3	16.5	61.0	78.7	35.5	**53.7**	52.1	43.6	10.3	21.1	40.8	78.5	19.8	32.8	21.4	**18.9**	15.0	69.2

Method	Backbone	wineglass	cup	fork	knife	spoon	bowl	banana	apple	sandwich	orange	broccoli	carrot	hotdog	pizza	donut	cake	chair	sofa	pottedplant	bed
Baseline	DRN	**16.8**	56.2	**60.3**	**47.3**	72.5	73.9	4.7	9.8	9.6	18.8	3.7	**46.7**	**38.1**	62.2	10.2	17.2	28.1	**13.0**	42.0	36.7
Ours	DRN	15.0	57.0	49.6	44.8	73.5	77.1	**5.0**	8.6	9.7	**23.4**	3.9	46.4	37.4	**67.7**	10.5	17.2	37.0	12.0	**49.1**	47.0
Baseline	ViT-L	15.4	**62.5**	52.3	38.8	78.8	79.5	4.8	8.3	**9.8**	16.1	3.7	42.9	37.4	60.0	10.5	**25.5**	**39.6**	11.6	39.1	41.5
Ours	ViT-L	14.5	58.6	58.9	39.0	**79.3**	**80.2**	4.8	**10.2**	9.7	16.0	**4.1**	44.6	37.4	60.6	10.5	18.1	36.7	11.4	46.7	**47.3**

Method	Backbone	diningtable	toilet	tvmonitor	laptop	mouse	remote	keyboard	cellphone	microwave	oven	toaster	sink	refrigerator	book	clock	vase	scissors	teddybear	hairdrier	toothbrush
Baseline	DRN	39.6	**55.9**	44.6	74.3	77.7	70.6	14.0	32.5	13.1	**12.0**	**8.1**	25.9	24.0	34.3	**43.7**	25.7	37.3	11.4	30.9	33.5
Ours	DRN	**54.7**	44.9	47.6	60.5	70.5	64.0	**16.7**	**34.5**	**13.8**	11.2	7.2	26.2	24.8	**42.6**	43.6	26.0	**47.6**	**15.5**	**32.6**	**28.4**
Baseline	ViT-L	39.2	32.6	45.7	70.4	**79.5**	67.0	14.0	20.6	13.4	10.7	6.4	26.1	24.1	37.7	41.7	25.7	32.6	11.9	27.4	26.8
Ours	ViT-L	43.9	38.2	**53.1**	**77.0**	78.9	**71.7**	14.0	16.4	13.1	10.8	6.2	**26.9**	**27.9**	40.8	41.8	**26.8**	41.9	12.1	29.8	28.1

Table 3. Per-category zero-shot semantic segmentation results on PASCAL-5^i.

Method	Backbone	Aeroplane	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	mIoU	FBIoU
Baseline	DRN	58.6	20.9	68.4	50.6	46.2	76.9	47.6	70.1	10.6	75.9	45.5	61.7
Ours	DRN	65.4	26.3	78.6	61.5	54.8	78.1	48.3	78.0	17.9	79.5	55.5	66.4
Baseline	ViT/L	**76.1**	34.6	82.4	**64.3**	**58.2**	73.4	51.7	**84.7**	18.9	83.1	58.4	68.3
Ours	ViT/L	74.8	**34.9**	**83.0**	63.6	56.9	**78.9**	**54.3**	84.0	**20.9**	**83.2**	**59.4**	**69.0**

Method	Backbone	Diningtable	Dog	Horse	Motorbike	Person	Pottedplant	Sheep	Sofa	Train	Tvmonitor
Baseline	DRN	4.8	66.1	68.0	61.0	4.1	18.4	60.5	30.1	63.9	8.0
Ours	DRN	40.0	76.5	73.8	65.2	36.6	**20.7**	66.5	42.9	70.1	29.2
Baseline	ViT/L	40.0	81.5	73.4	63.3	36.7	19.9	80.6	**47.4**	69.0	29.4
Ours	ViT/L	**40.5**	**81.8**	**73.8**	**70.1**	**37.0**	19.3	**81.8**	44.1	**75.8**	**30.1**

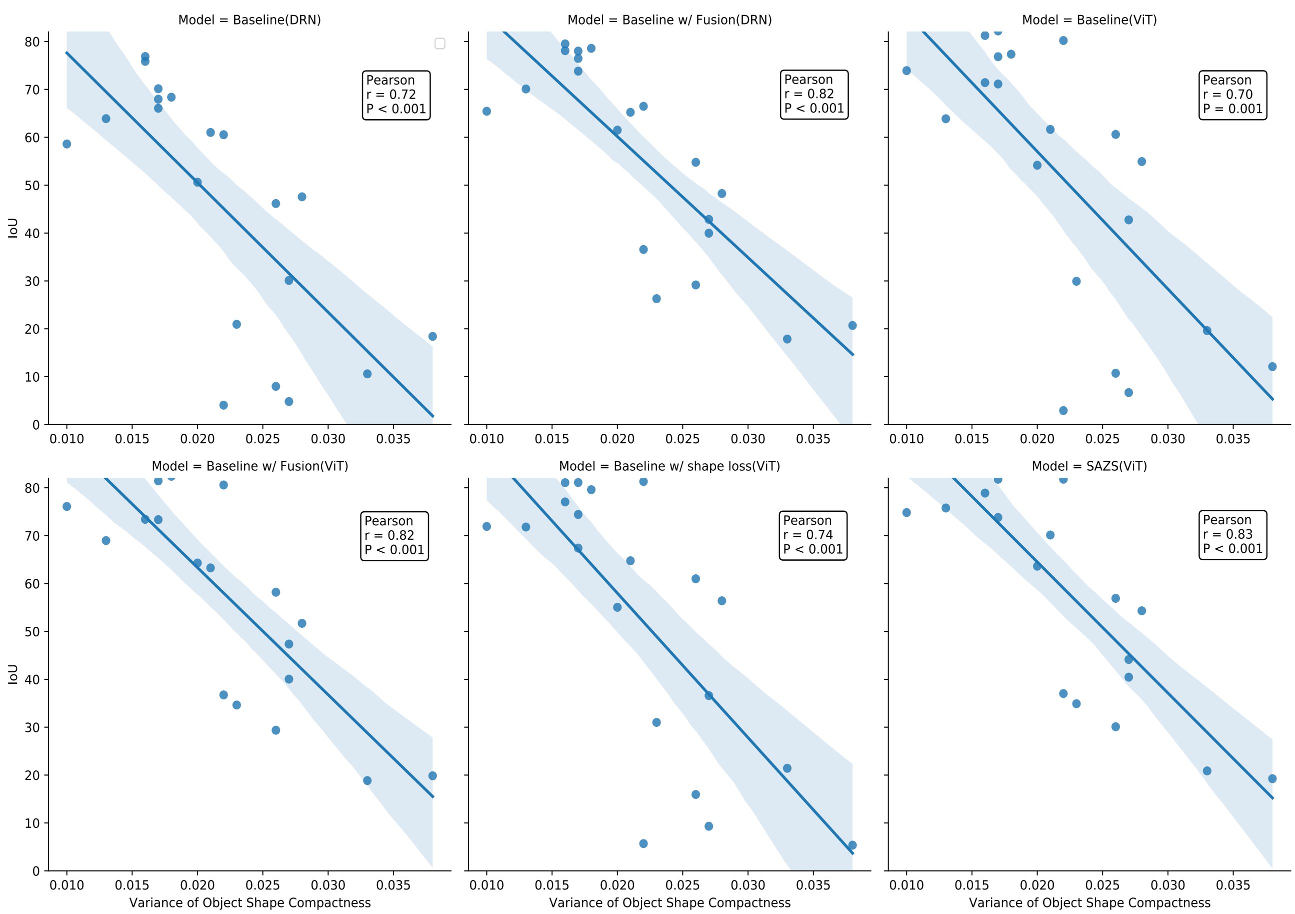

Figure 1. More scatterplots on PASCAL-5^i.

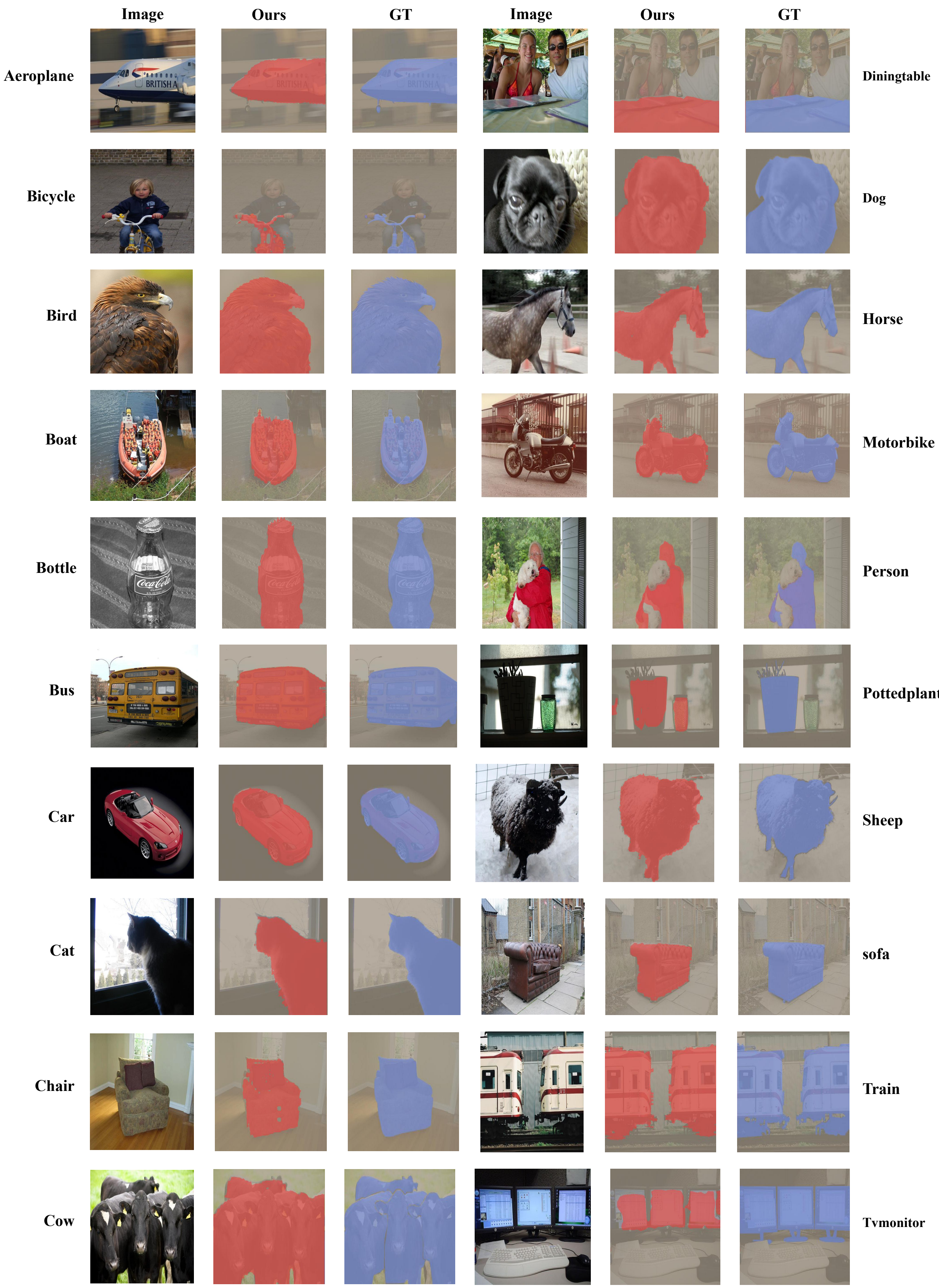

Figure 2. More qualitative results on PASCAL-5^i.

Figure 3. More qualitative results on COCO-20^i.