Delving into Shape-aware Zero-shot Semantic Segmentation

Xinyu Liu[1,2], Beiwen Tian[2,4], Zhen Wang[3], Rui Wanf[3], Kehua Sheng[3], Bo Zhang[3], Hao Zhao[2], Guyue Zhou[2]

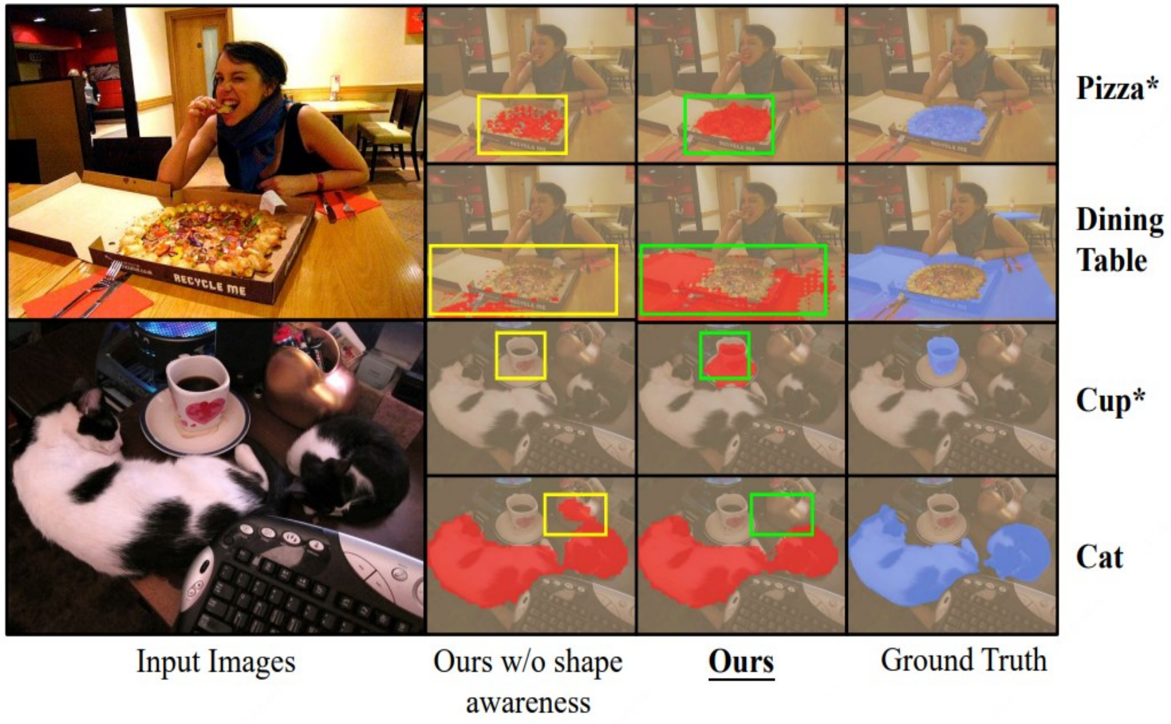
[1]Xidian University, China; [2] Institute for AI Industry Research (AIR), Tsinghua University, China; [3]Didi Chuxing, China; [4]Department of Computer Science and Technology, Tsinghua University, China

liuxinyu@stu.xidian.edu.cn, zhaohao@air.tsinghua.edu.cn

Content

Introduction

Problems:

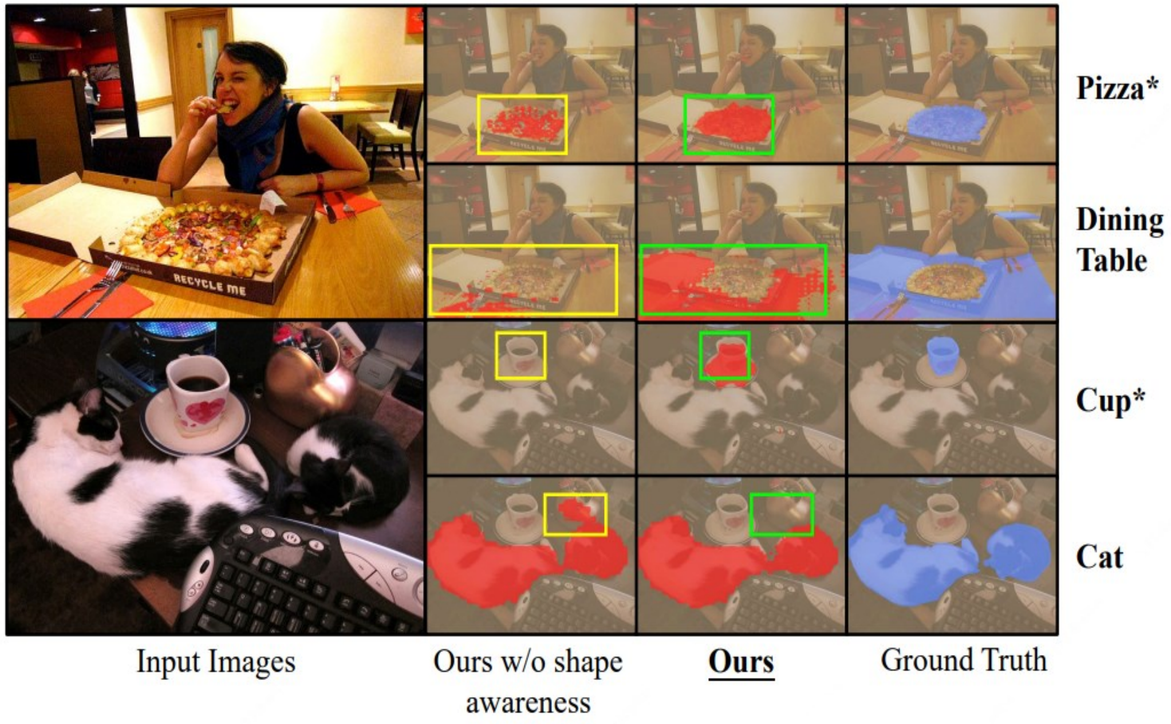


- In real-world, unknown objects may appear, hence there is a need to learn and achieve dense predictions for unknown object categories from limited samples.
- Dense prediction task requires not only accurate semantic understanding but also fine shape delineation.
- However, existing vision-language models are trained with image-level language descriptions.

Introduction

清華大学智能产业研究院
Institute for AI Industry Research, Tsinghua University

Solutions:

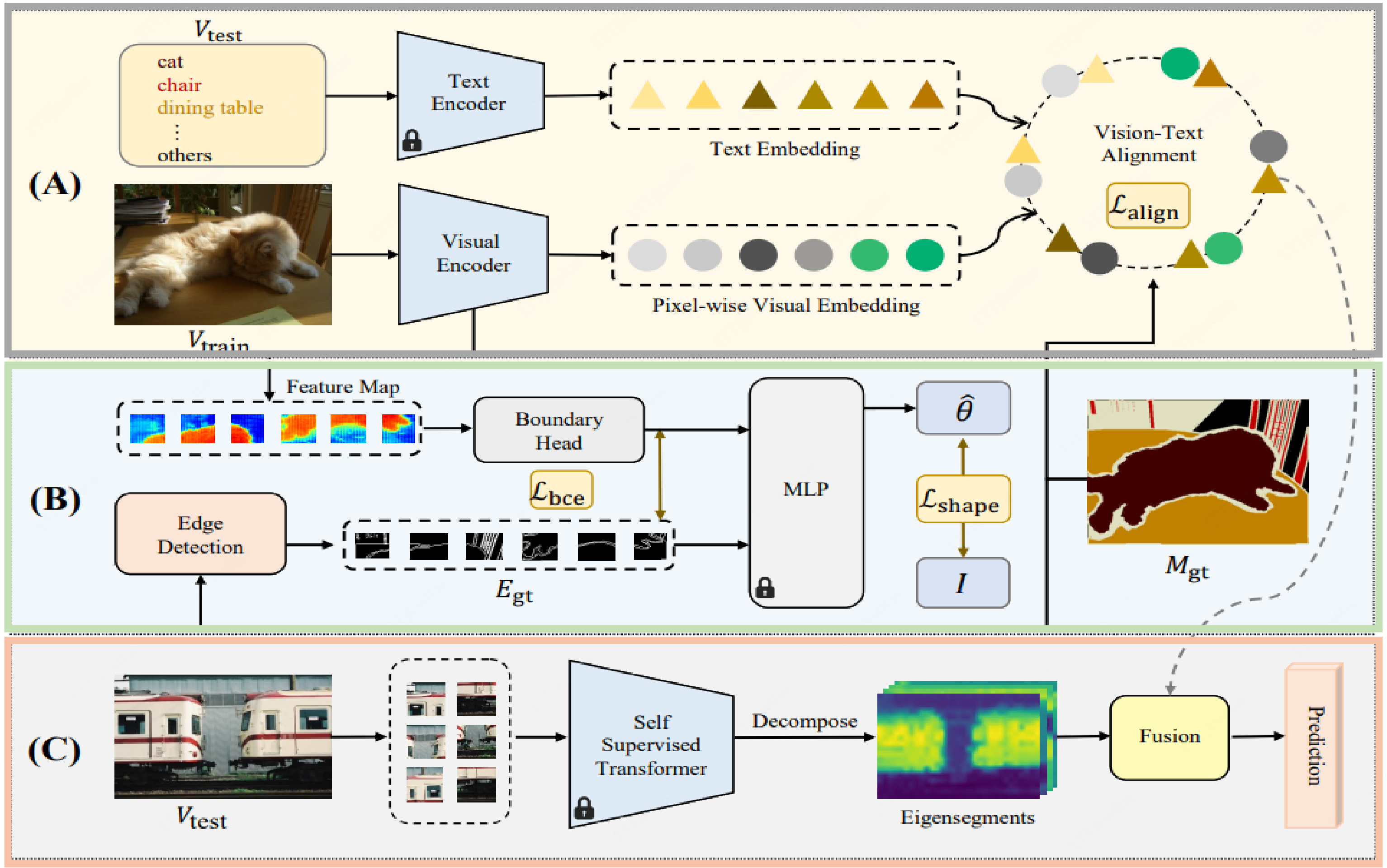


- Leverage the eigen vectors of Laplacian matrices constructed with self-supervised pixelwise features to promote shape-awareness.

- Propose to jointly optimizing a boundary segment constraint that aligns both the boundary of predicted semantic regions and the ground truth regions.

Method

清華大學 智能产业研究院
Institute for AI Industry Research, Tsinghua University



Overall architecture

Experiment

Our results on PASCAL-5^i and COCO-20^i datasets

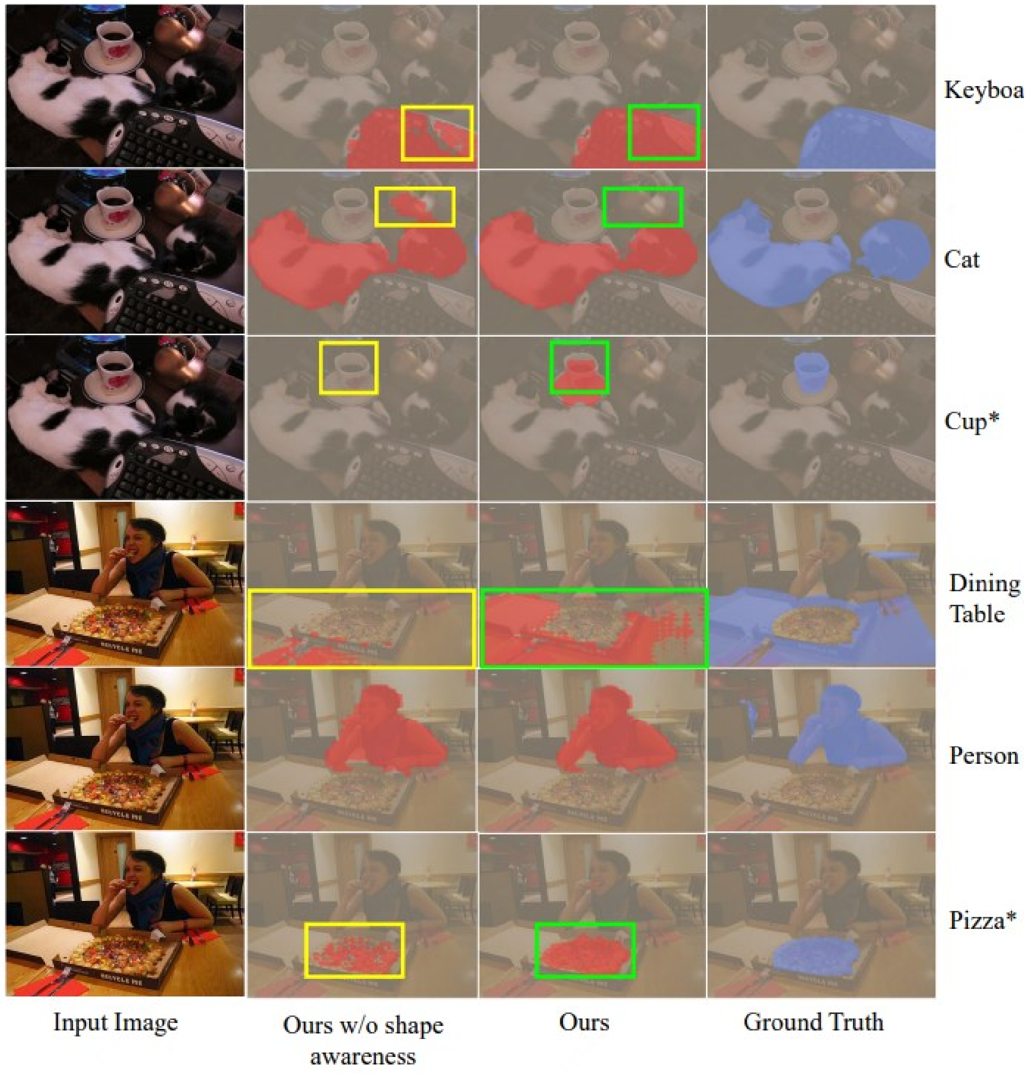
Method	Backbone	Setting	PASCAL-5^i 5^0	5^1	5^2	5^3	mIoU	FBIoU	COCO-20^i 20^0	20^1	20^2	20^3	mIoU	FBIoU
FWB [38]	ResNet	1-shot	51.3	64.5	56.7	52.2	56.2	–	17.0	18.0	21.0	28.9	21.2	–
DAN [52]	ResNet	1-shot	54.7	68.6	57.8	51.6	58.2	71.9	–	–	–	–	24.4	62.3
PFENet [51]	ResNet	1-shot	60.5	69.4	54.4	55.9	60.1	72.9	36.8	41.8	38.7	36.7	38.5	63.0
HSNet [37]	ResNet	1-shot	67.3	72.3	62.0	63.1	66.2	77.6	37.2	44.1	42.4	41.3	41.2	69.1
SPNet [54]	ResNet	zero-shot	23.8	17.0	14.1	18.3	18.3	44.3	–	–	–	–	–	–
ZS3Net [4]	ResNet	zero-shot	40.8	39.4	39.3	33.6	38.3	57.7	18.8	20.1	24.8	20.5	21.1	55.1
LSeg [25]	ResNet	zero-shot	52.8	53.8	44.4	38.5	47.4	64.1	22.1	25.1	24.9	21.6	23.4	57.9
Ours	DRN	zero-shot	**57.3**	**60.3**	**58.4**	**45.9**	**55.5**	**66.4**	**34.2**	**36.5**	**34.6**	**35.6**	**35.2**	**58.4**
LSeg [25]	ViT-L	zero-shot	61.3	63.6	43.1	41.0	52.3	67.6	28.1	27.5	30.0	23.2	27.2	**59.9**
Ours	ViT-L	zero-shot	**62.7**	**64.3**	**60.6**	**50.2**	**59.4**	**69.0**	**33.8**	**38.1**	**34.4**	**35.0**	**35.3**	58.2

The cross dataset results

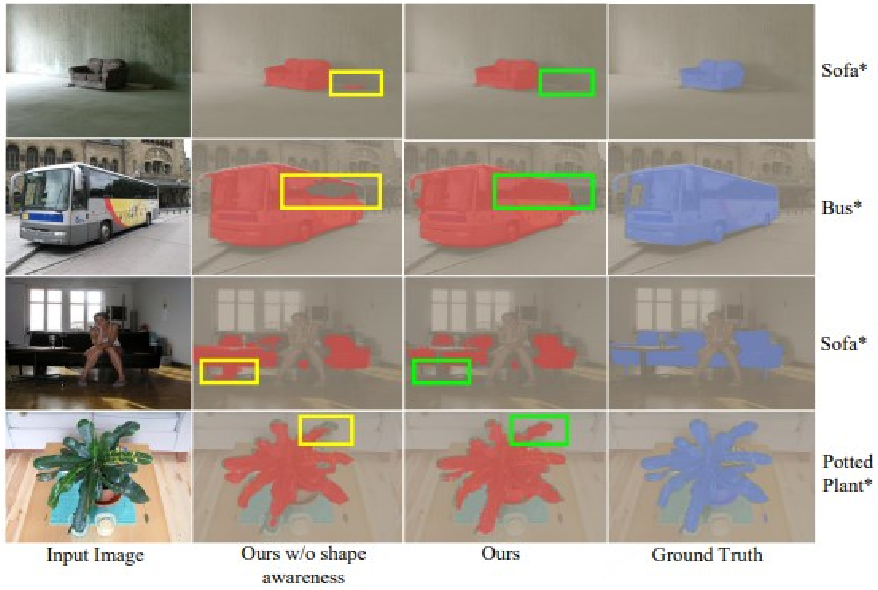
Model	Backbone	external dataset	target dataset	PASCAL-5^i
LSeg	ViT-L	✗	✓(seen classes)	52.3
SPNet	ResNet	✗	✓(seen classes)	18.3
ZS3Net	ResNet	✗	✓(seen classes)	38.3
LSeg	ResNet	✗	✓(seen classes)	47.4
LSeg+	ResNet	COCO	✗	59.0
OpenSeg [14]	ResNet	COCO	✗	60.0
Ours	DRN	COCO	✗	**62.7**

Experiment

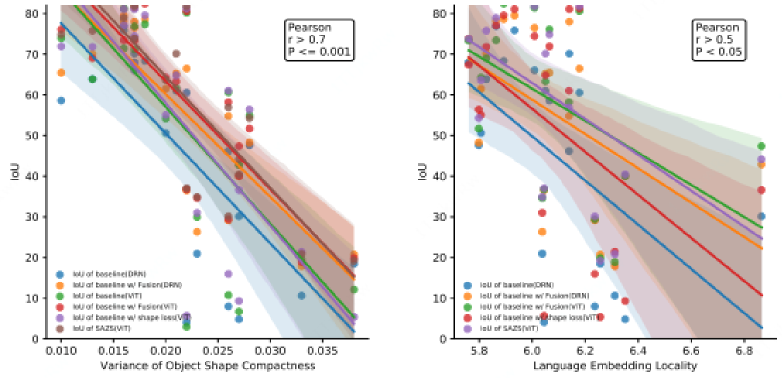
清華大学 智能产业研究院
Institute for AI Industry Research, Tsinghua University



Our results on COCO-20^i dataset



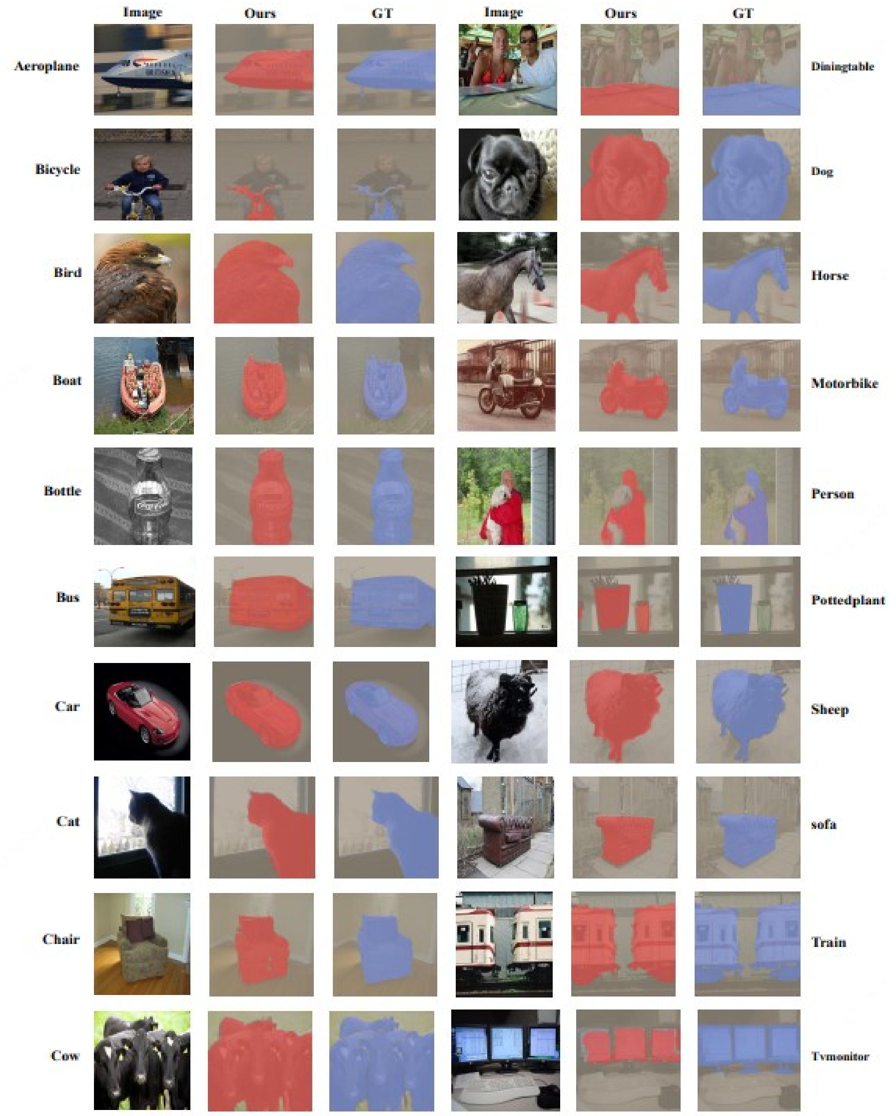
Our results on PASCAL-5^i dataset



Correlation of CO variance and mean language embedding locality on IoU

Experiment

Our results on COCO-20^{i} dataset



Our results on PASCAL-5^{i} dataset

Experiment

Ablation study on COCO-20^i (ViT backbone)

Model	Fusion	$\mathcal{L}_{shape}$	20^0	20^1	20^2	20^3	mIoU
SAZS	✓	✓	**34.2**	36.5	**34.6**	**35.6**	**35.2**
SAZS	✓		33.7	**38.2**	33.4	35.5	35.2
SAZS		✓	28.4	27.6	25.4	25.1	26.6
SAZS			24.2	28.5	24.4	23.3	25.1
LSeg [25]			22.1	25.1	24.9	21.6	23.4

Ablation study on PASCAL-5^i (ViT backbone)

Model	Fusion	$\mathcal{L}_{shape}$	5^0	5^1	5^2	5^3	mIoU
SAZS	✓	✓	62.7	**64.3**	**60.6**	**50.2**	**59.4**
SAZS	✓		**63.1**	62.4	59.0	49.2	58.4
SAZS		✓	59.7	63.4	44.3	42.2	52.4
SAZS			59.2	61.9	43.8	41.9	51.7
LSeg [25]			61.3	63.6	43.1	41.0	52.3

Ablation study on COCO-20^i (DRN backbone)

Model	Fusion	$\mathcal{L}_{shape}$	20^0	20^1	20^2	20^3	mIoU
SAZS	✓	✓	**33.8**	38.1	**34.4**	**35.0**	**35.3**
SAZS	✓		33.3	**39.0**	33.9	32.7	34.7
SAZS		✓	30.0	30.4	27.5	28.5	29.1
SAZS			26.3	32.0	26.2	26.2	27.7
LSeg [25]			28.1	27.5	30.0	23.2	27.2

Impact of Z_{shape} and Z_{sem} of fusion module

Model	external dataset	Z_{shape}	Z_{sem}	PASCAL-5^i
SAZS	COCO			58.4
SAZS	COCO	✓		58.6
SAZS	COCO		✓	62.7

Conslusion

Our contributions:

- We propose a novel methodology for shape-aware zero-shot semantic segmentation models from language supervision, named SAZS, and set new state-of-the-art performance on PASCAL and COCO datasets.
- We propose to leverage the eigen vectors of Laplacian matrices constructed with self-supervised pixelwise features to promote shape-awareness.
- We propose to jointly optimizing a boundary segment constraint that aligns both the boundary of predicted semantic regions and the ground truth regions while narrowing the visual-language embedding space.
- We draw several interesting and conclusive observations: the benefits of promoting shape-awareness highly relates to objects' compactness and language embedding locality.