SSA-Seg: Semantic and Spatial Adaptive Pixel-level Classifier for Semantic Segmentation

NEURAL INFORMATION PROCESSING SYSTEMS

Huawei Noah’s Ark Lab

Xiaowen Ma*, Zhenliang Ni*, Xinghao Chen†

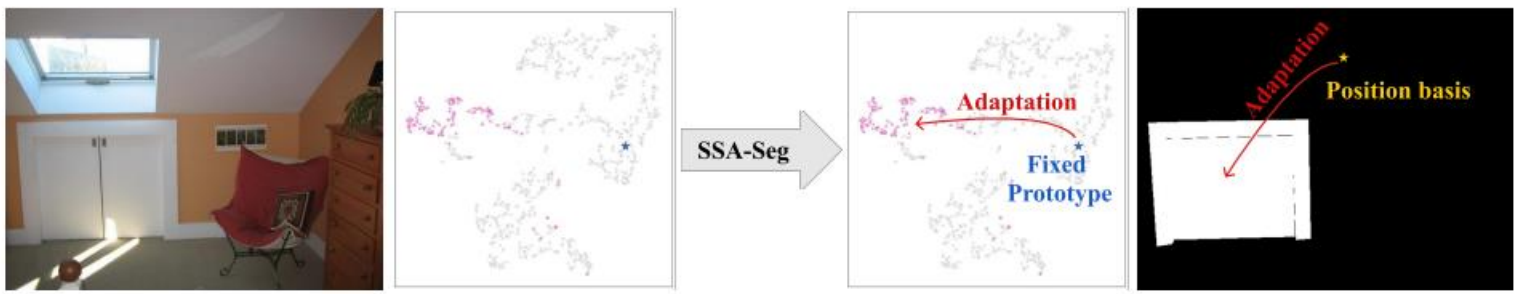
Huawei Noah’s Ark Lab

Paper

Code

Motivation

- Vanilla pixel-level classifiers for semantic segmentation are based on a certain paradigm, involving the inner product of fixed prototypes obtained from the training set and pixel features in the test image
- Limitation: feature deviation in the semantic domain and information loss in the spatial domain
- SSA-Seg: employ the coarse masks obtained from the fixed prototypes as a guide to adjust the fixed prototype towards the center of the semantic and spatial domains in the test image



(a) Test image (b) Feature distribution (c) Semantic prototype adaptation (d) Spatial prototype adaptation

Preliminary

➢ Vanilla pixel-level classifier

$$\mathcal{M}_c = \mathcal{S}_f \otimes \mathcal{S}^T$$

feature fixed prototype

➢ Semantic and Spatial Adaptive Classifier

$$\mathcal{O}_c = (\mathcal{S}_f \oplus \mathcal{P}_f) \otimes (\mathcal{S}_p \oplus \mathcal{P}_p)^T$$

feature

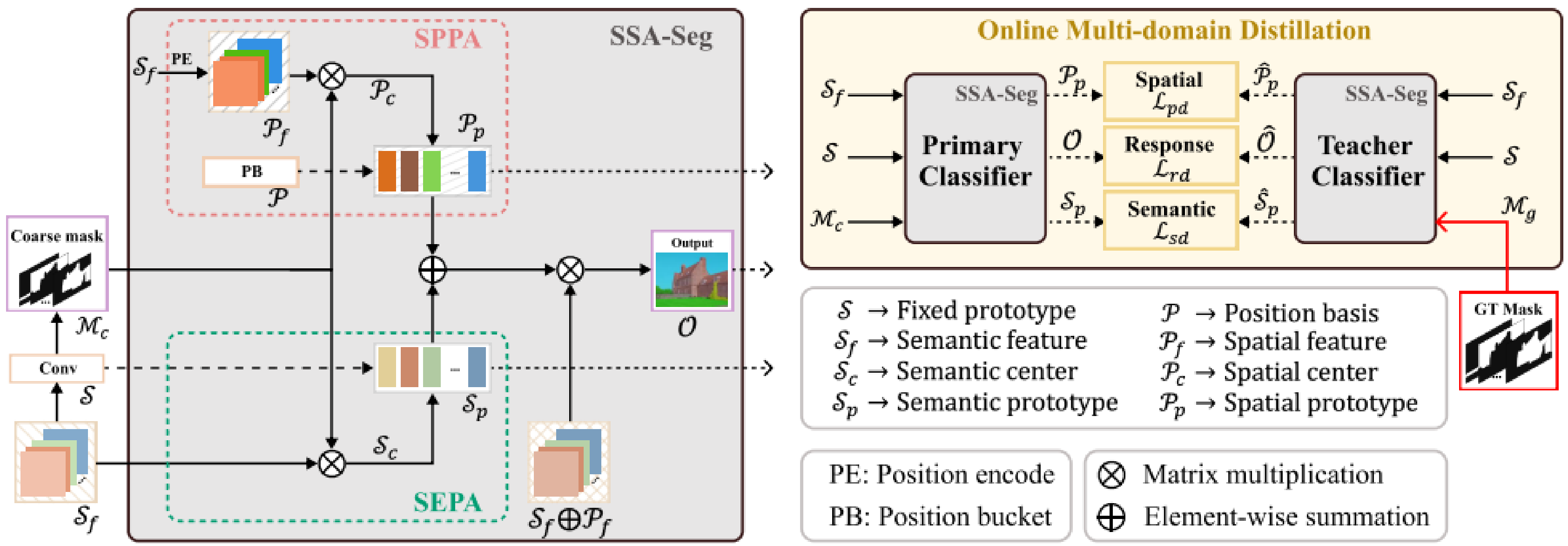
spatial feature

semantic prototype

spatial prototype

Method

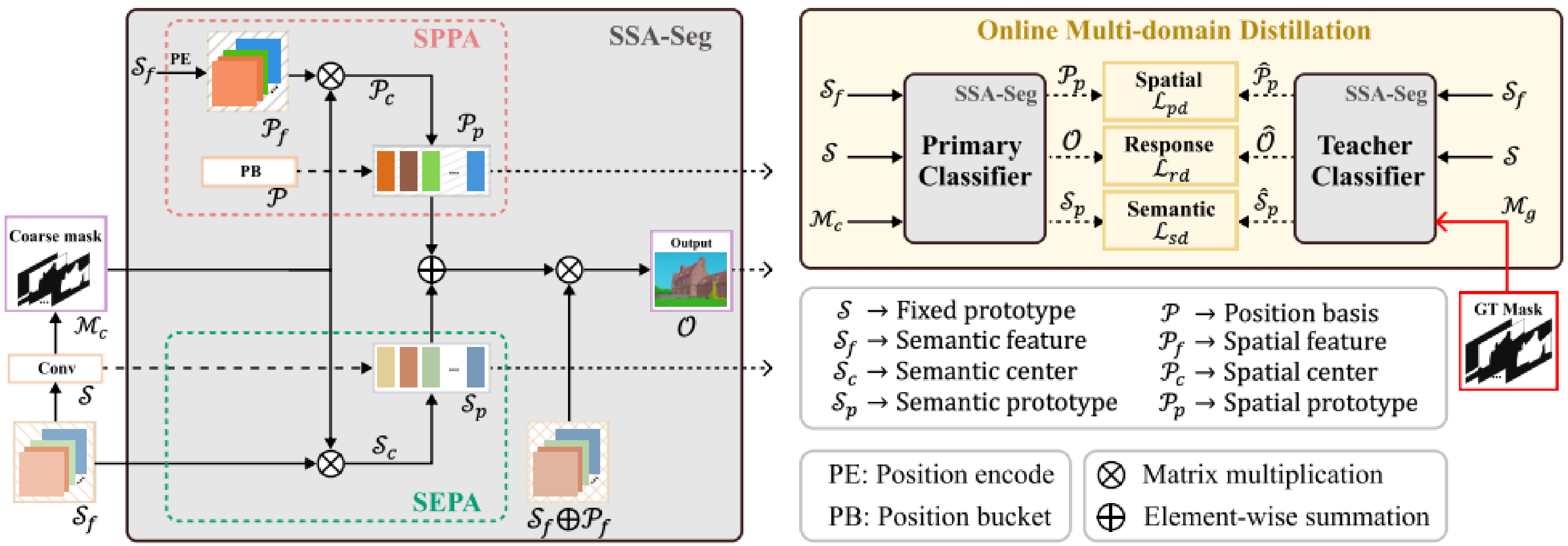
SSA-Seg consists of three parts: semantic prototype adaptation (SEPA), spatial prototype adaptation (SPPA), and online multi-domain distillation



Method

- SEPA offsets fixed semantic prototypes based on coarse mask-guided semantic feature distributions
- It can adapt to the semantic feature distributions of different images, alleviating the feature deviation in the semantic domain

$$\mathcal{S}_c = \text{Softmax}_K(\mathcal{M}_c) \otimes \mathcal{S}_f$$
$$\mathcal{S}_p = \phi_s(\mathcal{S}_c \odot \mathcal{S})$$

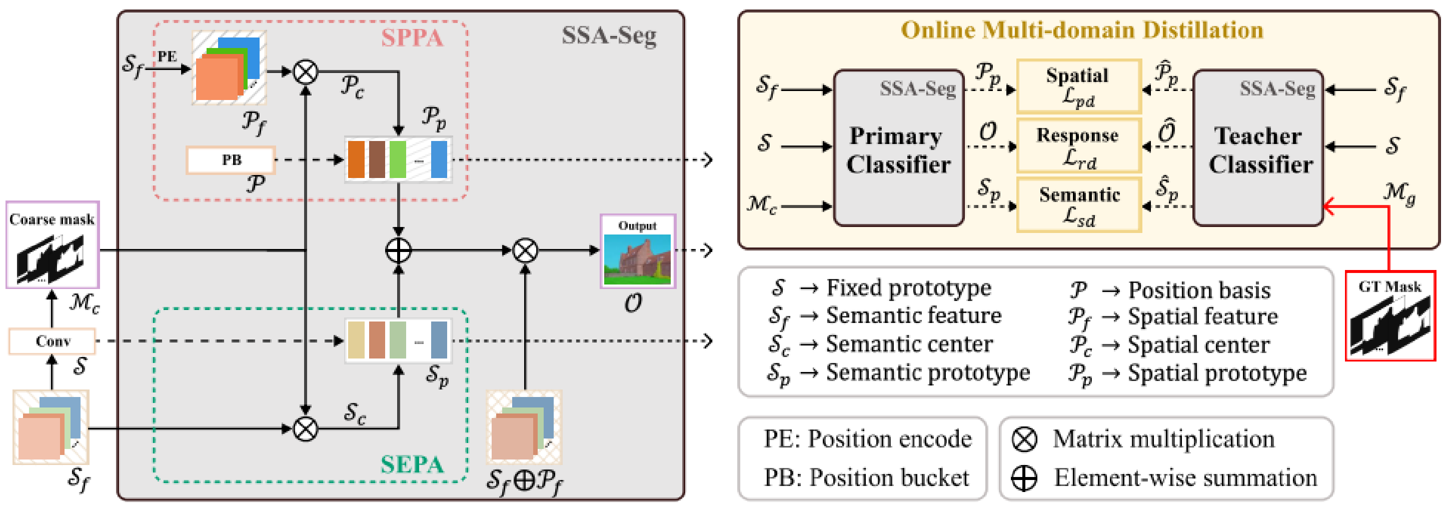


Method

- SPPA aims to make classification decisions with additional consideration of the spatial relation between pixel features and prototypes.
- Modeling the spatial relations of pixel and prototype can introduce structured information about the target objects, thus improving the segmentation performance for boundary regions and small targets

$$\mathcal{P}_c = \mathrm{Softmax}_{HW}(\mathcal{M}_c) \otimes \mathcal{P}_f$$

$$\mathcal{P}_p = \phi_p(\mathcal{P}_c \odot \mathcal{P})$$



Method

- Online multi-domain distillation learning is proposed to optimize the process of feature generation and constrain the adaptation of the semantic and spatial prototype
- It consists of three parts: Response Domain Distillation, Semantic Domain Distillation, Spatial Domain Distillation

Method

➢ Response Domain Distillation

$$\mathcal{L}_{rd}^{i} = -\sum_{j=1}^{K} \psi(\hat{\mathcal{O}})^{i,j} \cdot log(\psi(\mathcal{O}^{i,j})), \quad \mathcal{L}_{rd} = \frac{-1}{HW} \sum_{i=1}^{HW} \mathcal{L}_{rd}^{i}$$

$$\mathcal{L}_{rd} = \frac{-1}{2K} \sum_{k=1}^{K} \left(\frac{\sum_{i=1}^{HW} \mathcal{L}_{rd}^{i} \mathcal{B}_k^i \mathcal{H}^i}{\sum_{i=1}^{HW} \mathcal{B}_k^i \mathcal{H}^i} + \frac{\sum_{i=1}^{HW} \mathcal{L}_{rd}^{i} \cdot \bar{\mathcal{B}}_k^i \mathcal{H}^i}{\sum_{i=1}^{HW} \bar{\mathcal{B}}_k^i \mathcal{H}^i} \right), \quad \mathcal{H}^i = -\sum_{j=1}^{K} \psi(\hat{\mathcal{O}})^{i,j} \cdot log(\psi(\hat{\mathcal{O}}^{i,j}))$$

provide more information to the Primary Classifier

Method

➢ Semantic Domain Distillation

$$\mathcal{M} = \psi(\mathcal{S}_p \mathcal{S}_p^T), \quad \hat{\mathcal{M}} = \psi(\hat{\mathcal{S}}_p \hat{\mathcal{S}}_p^T)$$
$$\mathcal{M}_d = \Lambda(\mathcal{M} - \hat{\mathcal{M}})$$
$$\mathcal{L}_{sd} = \frac{1}{K} \sum_{i=1}^{K} \sum_{j=1}^{K} \mathcal{M}_d^{i,j}$$

guide the offset process of semantic prototypes to exhibit better inter-class separation

Method

- Spatial Domain Distillation

$$\mathcal{L}_{pd} = \frac{-1}{K}\sum_{i=1}^{K}\sum_{j=1}^{D}\psi(\mathcal{P}_p)^{i,j} \cdot log(\psi(\hat{\mathcal{P}}_p^{i,j}))$$

constrain the spatial prototypes guided by the rough mask to be equal to the spatial prototypes guided by the ground-truth mask

Experiment

Method	Backbone	Latency	Params	ADE20K		COCO-Stuff-10K		PASCAL-Context	
				FLOPs	mIoU	FLOPs	mIoU	FLOPs	mIoU
OCRNet [54]	HRNet-W48	67.2	8.6	164.8	43.30	164.8	36.16	143.2	48.22
+SSA-Seg		69.3	8.7	165.0	**47.47** ↑4.17	165.0	**37.94** ↑1.78	143.3	**50.21** ↑1.99
UperNet [48]	Swin-T	52.8	60.0	236.1	44.14	236.1	38.93	207.5	51.93
+SSA-Seg		54.3	61.1	236.3	**47.56** ↑3.42	236.3	**42.30** ↑3.37	207.7	**54.91** ↑2.98
SegFormer [49]	MiT-B5	69.0	82.0	52.5	49.13	52.5	44.07	45.8	58.39
+SSA-Seg		70.1	82.3	52.6	**50.74** ↑1.61	52.6	**45.55** ↑1.48	45.8	**59.14** ↑0.75
UperNet [48]	Swin-L	105.5	233.8	404.9	51.68	404.9	46.85	362.9	60.50
+SSA-Seg		107.3	234.9	405.2	**52.69** ↑1.01	405.2	**48.94** ↑2.09	363.2	**61.83** ↑1.33
ViT-Adapter [8]	ViT-Adapter-L	283.3	363.8	616.1	54.40	616.1	50.16	541.5	65.77
+SSA-Seg		284.9	364.9	616.3	**55.39** ↑0.99	616.3	**51.18** ↑1.02	541.7	**66.05** ↑0.28
AFFormer-B [14]	AFFormer-B	25.1	3.0	4.3	39.94	4.3	33.22	3.7	48.57
+SSA-Seg		26.0	3.3	4.4	**41.92** ↑1.98	4.4	**36.40** ↑3.18	3.7	**49.72** ↑1.15
SeaFormer-B [45]	SeaFormer-B	26.8	8.6	1.8	40.05	1.8	33.29	1.6	45.75
+SSA-Seg		27.3	8.8	1.8	**42.46** ↑2.41	1.8	**35.92** ↑2.63	1.6	**47.00** ↑1.25
SegNeXt-T [17]	MSCAN-T	22.8	4.3	6.2	41.04	6.2	36.39	5.4	50.35
+SSA-Seg		23.3	4.6	6.3	**43.90** ↑2.86	6.3	**38.91** ↑2.52	5.4	**52.58** ↑2.23
SeaFormer-L [45]	SeaFormer-L	29.4	14.0	6.4	42.36	6.4	35.99	5.6	49.14
+SSA-Seg		29.9	14.2	6.4	**45.36** ↑3.00	6.4	**38.48** ↑2.44	5.6	**49.66** ↑0.52

- SSA-Seg significantly improves the segmentation performance of the baseline models with only a minimal increase in computational cost
- By applying SSA-Seg, we achieve the state-of-the-art lightweight segmentation performance

Experiment

NEURAL INFORMATION PROCESSING SYSTEMS

- SSA-Seg outperforms previous classifiers

Method	Backbone	FLOPs	ADE20K	COCO.
FCN [34]	ResNet101 [19]	275.7	39.9	32.5
+ProtoSeg [62]		278.5	41.1 ↑1.2	34.0 ↑1.5
+DNC [46]		278.5	41.1 ↑1.2	33.0 ↑0.5
SSA-Seg		275.9	**44.3** ↑4.4	**36.6** ↑4.1
UperNet	Swin-B [32]	297.2	48.0	42.8
+GMMSeg [28]		302.3	49.0 ↑1.0	44.3 ↑1.5
+DNC [46]		308.6	48.6 ↑0.6	43.1 ↑0.3
+SSA-Seg		297.5	**49.2** ↑1.2	**45.2** ↑2.4
OCRNet [54]	HRNetV2-W48 [43]	164.8	43.3	36.2
+GMMSeg [28]		169.8	44.8 ↑1.5	-
+CAC [44]		164.9	45.7 ↑2.4	-
+SSA-Seg		165.0	**47.5** ↑4.2	**37.9** ↑1.7
SegNeXt-T [17]	MSCAN-T [17]	6.2	41.0	36.4
+CAC [44]		6.2	43.0 ↑2.0	37.5 ↑1.1
+SSA-Seg		6.3	**43.9** ↑2.9	**38.9** ↑2.5
SeaFormer-B [45]	SeaFormer-B [45]	1.8	40.0	33.3
+CAC [44]		1.8	40.1 ↑0.1	35.5 ↑2.2
+SSA-Seg		1.8	**42.5** ↑2.5	**35.9** ↑2.6

- By combining SSA-Seg, the existing pixel-level segmentation baselines achieve a better balance between efficiency and performance compared to mask classification methods

Method	Params	FLOPs	Latency	mIoU
MaskFormer [10]	41.3	55.1	31.0	44.5
Mask2Former [9]	44.0	70.1	55.2	47.2
YOSO [20]	42.0	37.3	28.3	44.7
PEM [4]	35.6	46.9	26.8	45.5
CGRSeg-B [40]	19.1	7.7	36.0	45.5
+SSA-Seg	19.3	7.6	36.0	47.1
CGRSeg-L [40]	35.7	14.9	43.3	48.3
+SSA-Seg	35.8	14.8	42.6	49.0

Experiment

- SSA-Seg has high accuracy for confusing classes
- SSA-Seg can present stronger activation values in the center region of the mask and does not show too much activation in irrelevant regions