Generative Semantic Segmentation

Jiaqi Chen[1], Jiachen Lu[1], Xiatian Zhu[2] and Li Zhang[1*]
[1]Fudan University, [2]University of Surrey

Code is available

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Unleash the power of generative model for semantic segmentation

- **G**enerative **S**emantic **S**egmentation (GSS) explores a novel route to achieve **visual perception on generative paradigm**.
- According to our GSS, the task of semantic segmentation is perceived as an **image generation** problem.
- New **state-of-the-art** performance on Cross-domain setting (the MSeg dataset). Competitive on the Cityscapes and ADE20K datasets.

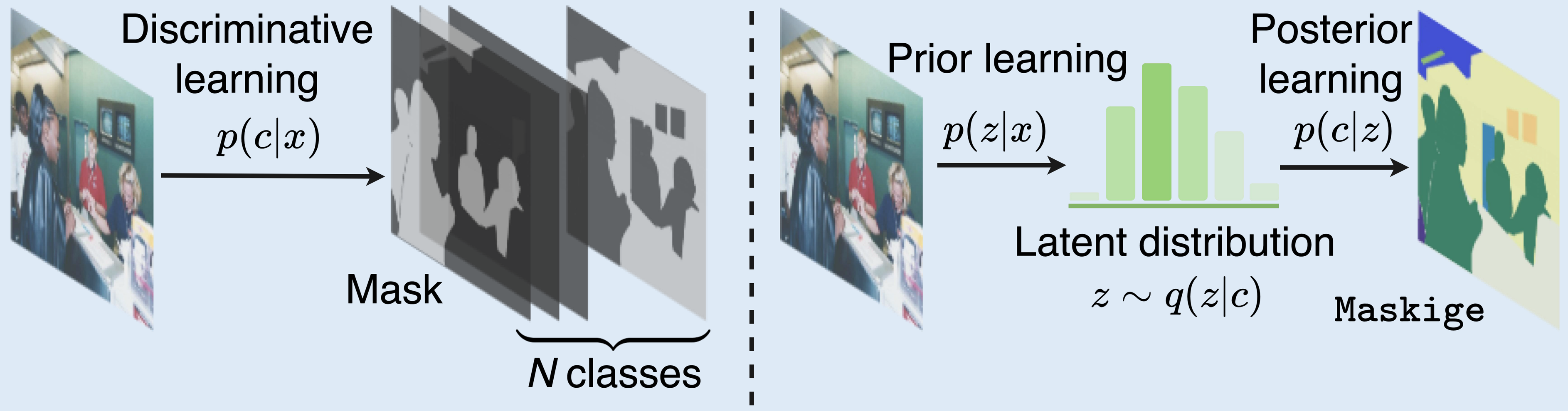
Discriminative learning *v.s.* Generative learning

- Traditional discriminative learning methods directly optimize,

$$\max_{\pi} \log p_{\pi}(c|x)$$

- In generative learning, each category owns a distinctive **color.** We turn segmentation mask into a special image, **maskige.**
- Then, our GSS optimizes its **evidence lower bound (ELBO)**,

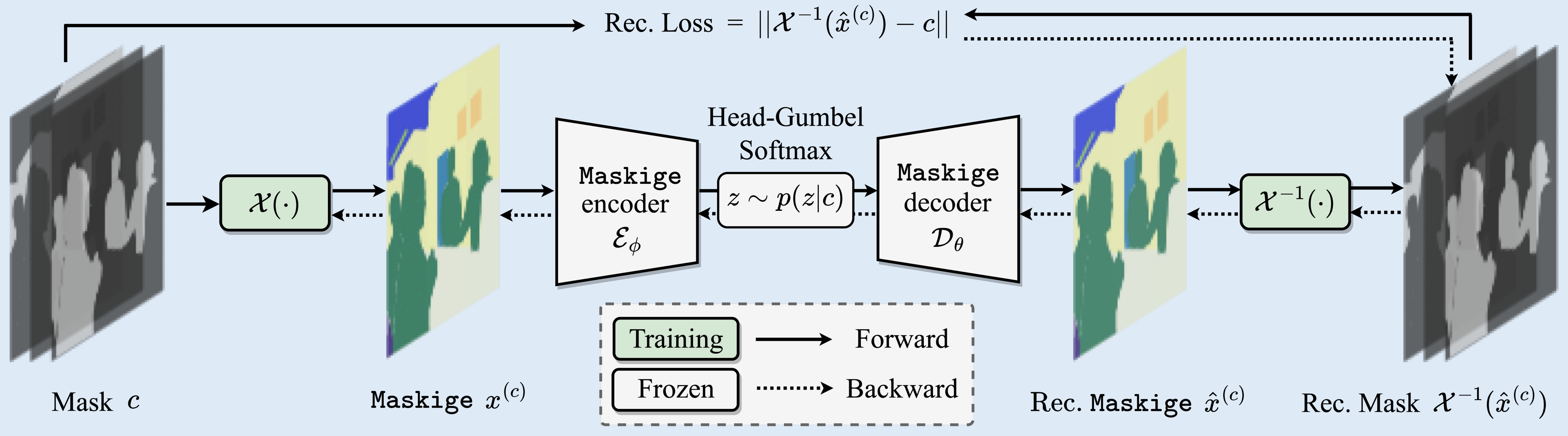
$$\log p(c|x) \geq \mathbb{E}_{q_\phi(z|c)}\left[\log \frac{p(z,c|x)}{q_\phi(z|c)}\right]$$
$$= \mathbb{E}_{q_\phi(z|c)}\left[\log p_\theta(c|z)\right] - D_{KL}\Big(q_\phi(z|c), p_\psi(z|x)\Big)$$



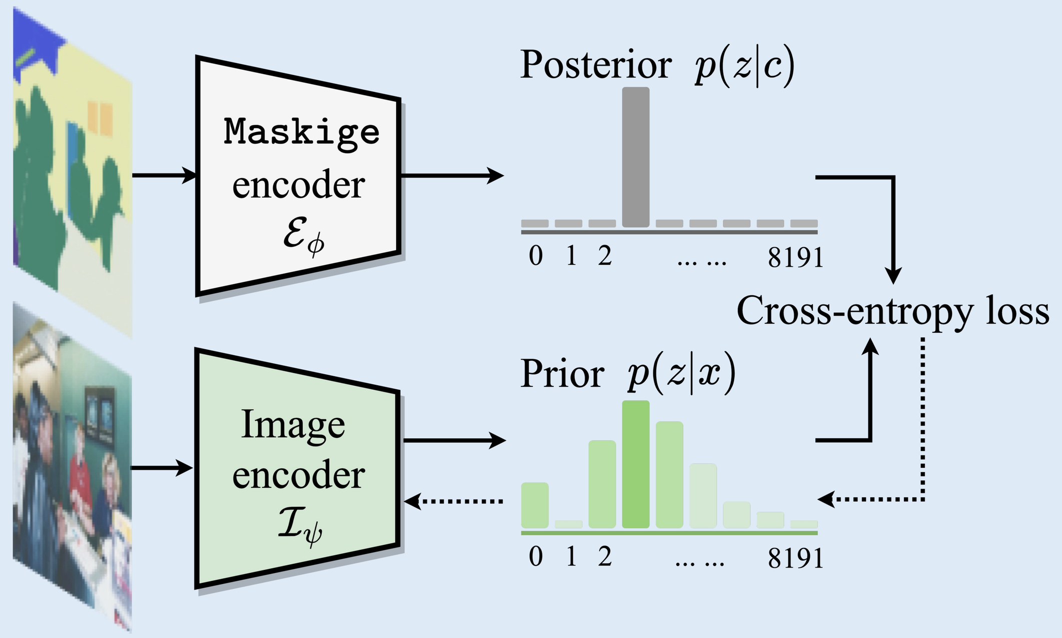
(a) Discriminative semantic segmentation (b) Generative semantic segmentation

How to train GSS in a generative way?

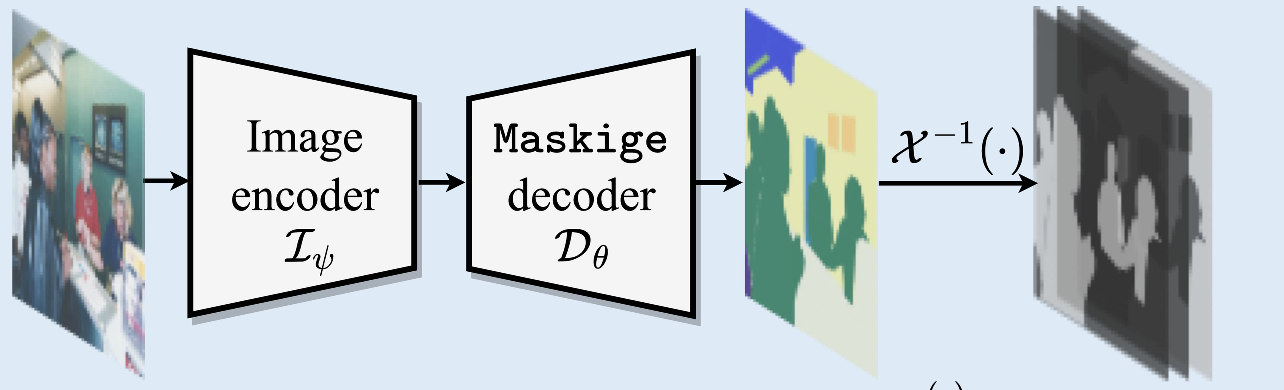
- **Stage I:** Effective posterior learning
  - Learn a mapping between mask and maskige (*i.e.* $\mathcal{X}$ and $\mathcal{X}^{-1}$).
  - Keep the $\mathcal{E}_\phi$ and $\mathcal{D}_\theta$ (DALL-E pretrained VQ-VAE) frozen.



- **Stage II:** Latent pirior learning
  - Using a frozen $\mathcal{E}_\phi$, map the ground truth maskige to the posterior distribution $p(z|c)$.
  - Train the $\mathcal{I}_\psi$ to predict the prior distribution $p(z|x)$ and align with $p(z|c)$ using the Cross-entropy loss.



Generative inference



Experiments

Method	Pretrain	Backbone	Iteration	mIoU
- Discriminative modeling:				
FCN	1k	ResNet-101	80k	77.02
PSPNet	1k	ResNet-101	80k	79.77
DeepLab-v3+	1k	ResNet-101	80k	80.65
NonLocal	1k	ResNet-101	80k	79.40
CCNet	1k	ResNet-101	80k	79.45
Maskformer	1k	ResNet-101	90k	78.50
Mask2former	1k	ResNet-101	90k	80.10
SETR	22k	Swin-Large	80k	78.10
UperNet	22k	Swin-Large	80k	82.89
Mask2former	22k	Swin-Large	90k	**83.30**
Segformer	1k	MiT-B5	160k	82.25
- Generative modeling:				
UViM†	22k	Swin-Large	160k	70.77
GSS-FF	1k	ResNet-101	80k	77.76
GSS-FT-W	22k	ResNet-101	80k	78.46
GSS-FF	22k	Swin-Large	80k	78.90
GSS-FT-W	22k	Swin-Large	80k	**80.05**

Table 1: Results on Cityscapes val split.

Method	Pretrain	Backbone	Iteration	mIoU
- Discriminative modeling:				
FCN	1k	ResNet-101	160k	41.40
CCNet	1k	ResNet-101	160k	43.71
DANet	1k	ResNet-101	160k	44.17
UperNet	1k	ResNet-101	160k	43.82
Deeplab-V3+	1k	ResNet-101	160k	45.47
Maskformer	1k	ResNet-101	160k	45.50
Mask2former	1k	ResNet-101	160k	47.80
OCRNet	1k	ResNet-101	160k	43.25
Segformer	1k	MiT-B5	160k	**50.08**
SETR	22k	ViT-Large	160k	48.28
- Generative modeling:				
UViM†	22k	Swin-Large	160k	43.71
GSS-FF	22k	Swin-Large	160k	46.29
GSS-FT-W	22k	Swin-Large	160k	**48.54**

Table 2: Results on ADE20K val split. UViM† is reproduced by us on PyTorch. GSS-FF is free for train in Stage I, while GSS-FT-W need.

Method	Backbone	Iteration	VOC	Context	CamVid	WildDash	KITTI	ScanNet	h. mean
- Discriminative modeling:									
CCSA	HRNet-W48	500k	48.9	-	52.4	36.0	-	27.0	39.7
MGDA	HRNet-W48	500k	69.4	-	57.5	39.0	-	33.5	46.1
MSeg	HRNet-W48	500k	70.7	42.7	**83.3**	62.0	67.0	48.2	59.2
MSeg	HRNet-W48	160k	63.8	39.6	73.9	60.9	65.1	43.5	54.9
MSeg	Swin-Large	160k	78.7	47.5	75.1	**66.1**	**68.1**	49.0	61.7
- Generative modeling:									
GSS-FF	HRNet-W48	160k	64.1	37.1	72.3	59.3	62.0	40.6	52.6
GSS-FT-W	HRNet-W48	160k	65.2	38.8	75.2	62.5	66.2	43.1	55.2
GSS-FF	Swin-Large	160k	78.7	45.8	74.2	61.8	65.4	46.9	59.5
GSS-FT-W	Swin-Large	160k	**79.5**	**47.7**	75.9	65.3	68.0	**49.7**	**61.9**

Table 3: Results on Cross-domain setting (MSeg test split).

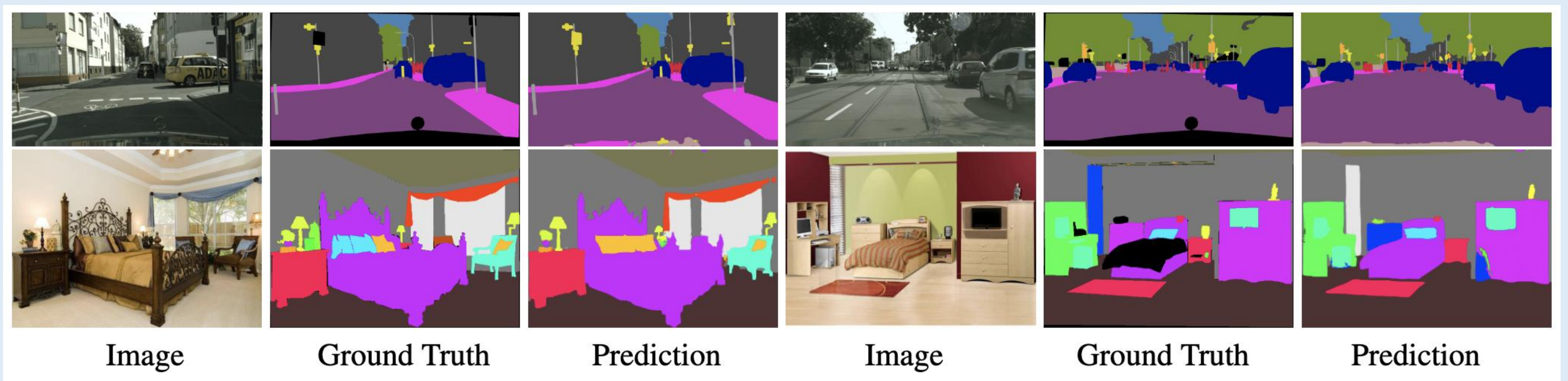


Figure 1: Qutatitive results on Cityscapes and ADE20K