

ϵ -Seg: Sparsely Supervised Semantic Segmentation of Microscopy Data

Supplementary Material

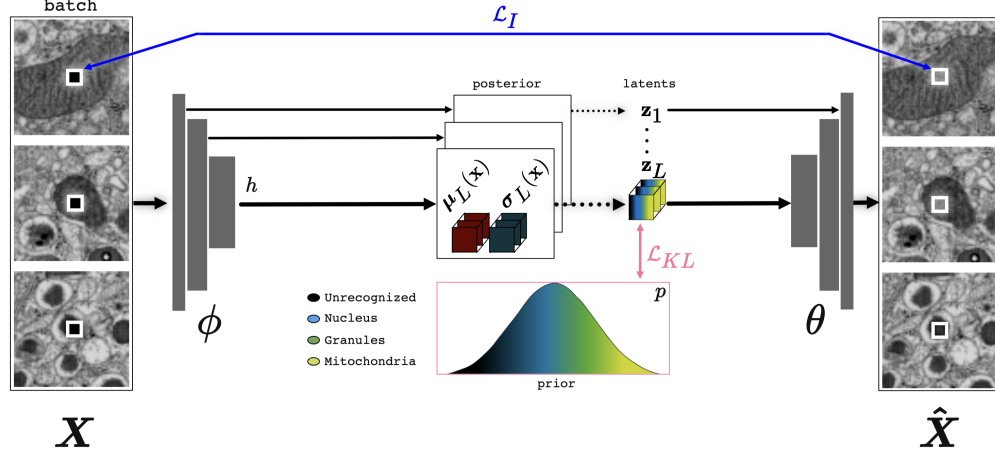


Figure S1: The overall pipeline of Vanilla HVAE in Table S1 (first row in Table 7), which is trained on an inpainting task (of the center-region masked inputs). ϕ and θ are encoder and decoder of the network, respectively. Dotted arrows show sampling from a distribution. h is an intermediate feature embedding of input x coming from the encoder ϕ and it is posterior distribution's parameters which is divided into two chunks shown as μ_L and σ_L . z_L is a sample from $\mathcal{N}(\mu_L(x), \sigma_L^2(x))$. For \mathcal{L}_I inpainting loss and \mathcal{L}_{KL} refer to Equations 1 and 7 respectively.

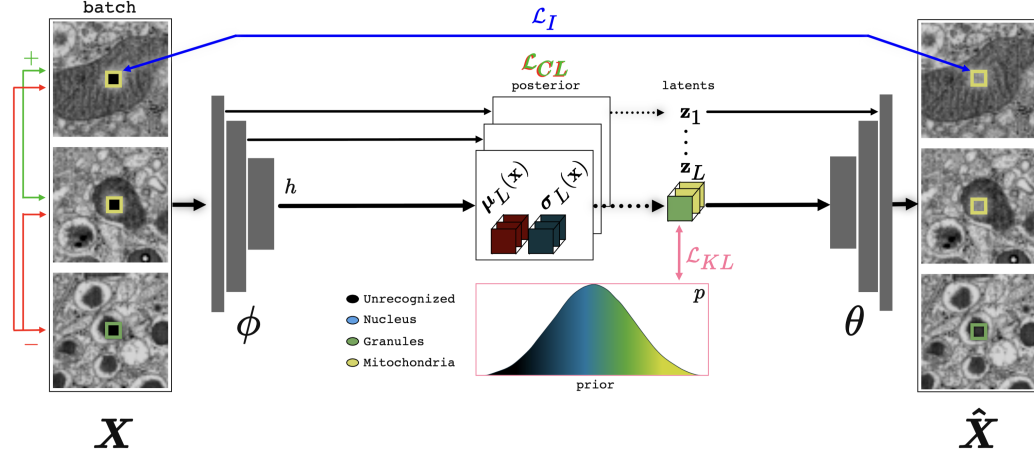


Figure S2: The overall pipeline of Vanilla HVAE with only CL added in the pipeline in the second row in Table 7, which is trained on an inpainting task (of the center-region masked inputs). Green and red arrows are showing positive and negative pair respectively, in a batch. ϕ and θ are encoder and decoder of the network, respectively. Dotted lines show sampling from a distribution. h is an intermediate feature embedding of input x coming from the encoder ϕ and it is posterior distribution's parameters which is divided into two chunks shown as μ_L and σ_L . z_L is a sample from $\mathcal{N}(\mu_L(x), \sigma_L^2(x))$. For \mathcal{L}_I inpainting loss, \mathcal{L}_{CL} contrastive loss and \mathcal{L}_{KL} refer to Equations 1, 16 and 7 respectively.

Model	Learning Paradigm	U	N	G	M	Avg DSC
Vanilla HVAE* [24]	Self-Supervised	0.44	0.55	0.34	0.13	0.37
Labkit [2]	Sparsely Supervised	0.85	0.44	0.68	0.61	0.65
U-net [26]	Fully Supervised	0.94	0.99	0.90	0.87	0.93
U-net	Sparsely Supervised	0.90	0.96	0.78	0.66	0.83
Vanilla ViT [11]	Fully Supervised	0.91	0.98	0.77	0.87	0.88
Segmenter [29]	Fully Supervised	0.91	0.99	0.86	0.90	0.92
MAESTER* [34]	Self-Supervised	0.84	0.95	0.56	0.79	0.79
Han et al* [14]	Self-Supervised	-	-	-	-	0.66
ϵ -Seg (+ \mathcal{L}_H)	Sparsely Supervised	0.89	0.98	0.81	0.83	0.88

Table S1: Dice similarity coefficient per class and average across all classes comparing our model with baselines on the “BetaSeg” dataset [22]. Methods marked with an asterisk use K-Means clustering on latent features to conduct semantic segmentation (more explanation can be found in Section 3). U: Unrecognized, N:Nucleus, G:Granules, M:Mitochondria.

# res. blocks	Per-Class Dice Coefficient				Avg DSC
	U	N	G	M	
5	0.86	0.98	0.80	0.75	0.85
4	0.85	0.97	0.80	0.74	0.84
3	0.88	0.96	0.81	0.80	0.86
2	0.87	0.97	0.81	0.77	0.86
1	0.85	0.97	0.80	0.72	0.84

Table S2: Residual blocks ablation (3 latent variables). U: Unrecognized, N: Nucleus, G: Granules, M: Mitochondria.

Entropy-based Loss. When the sample \mathbf{y}' of the Gumbel-Softmax distribution is uniform, the network is maximally unsure about which class to predict for the current input patch. We noticed that this is commonly the case early during training, where the network has not yet seen a lot of patches for which ground truth labels are available.

To encourage the network not to predict a uniform \mathbf{y}' , we introduced an entropy loss for all patches $\mathbf{x}^{(j)} \in \mathbf{X}$ for which we do not have a ground truth class label.

$$\mathcal{L}_H = - \sum_{\mathbf{x}^{(j)} \in \mathbf{X}} \mathbf{y}'^{(j)} \log(\mathbf{y}'^{(j)}). \quad (18)$$

# latent	Per-Class Dice Coefficient				Avg DSC
	U	N	G	M	
2	0.87	0.98	0.81	0.76	0.86
3	0.86	0.98	0.80	0.75	0.85

Table S3: Latent variables ablation (5 res. blocks/layer). U: Unrecognized, N:Nucleus, G:Granules, M:Mitochondria.

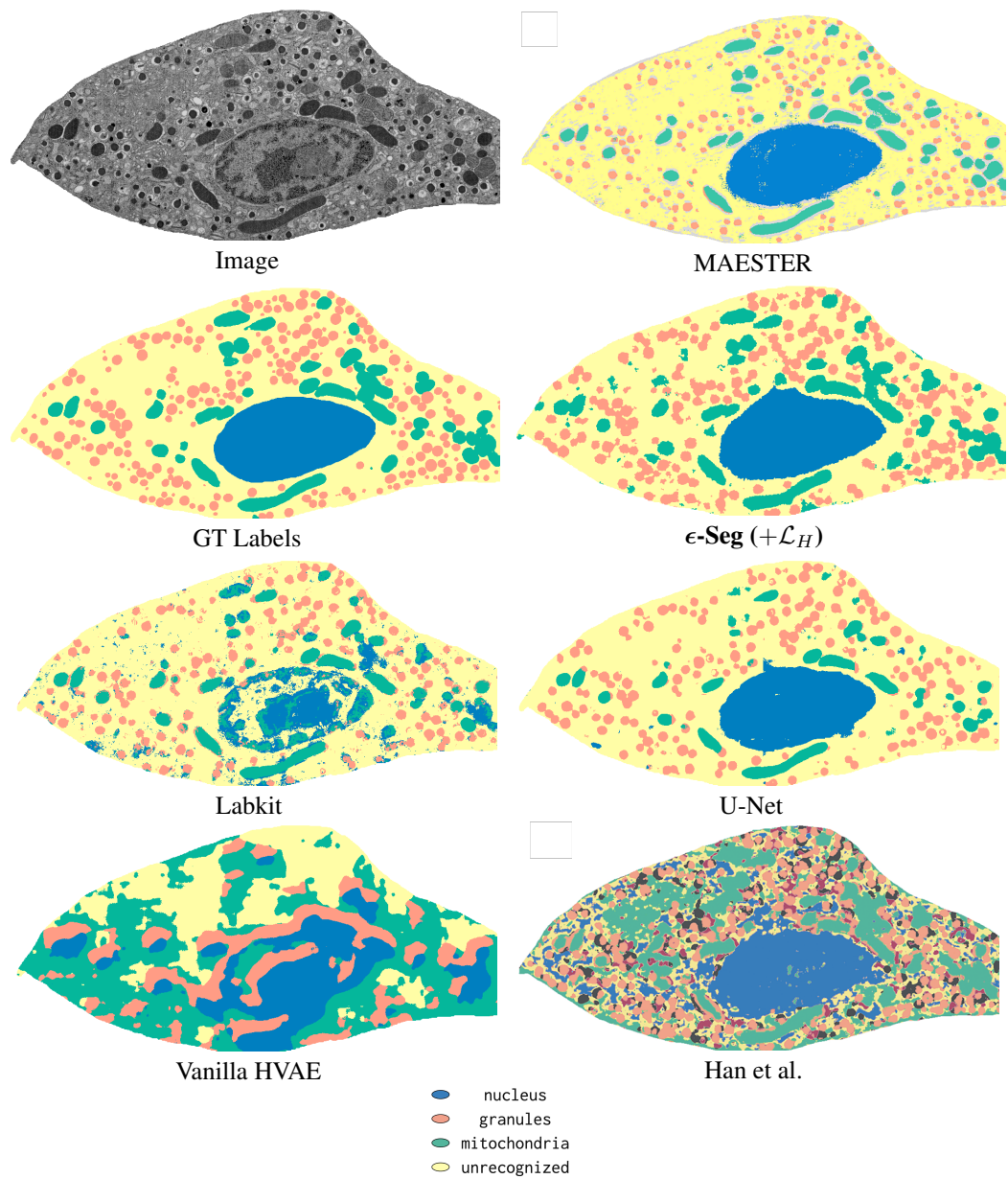


Figure S3: Qualitative segmentation result on part of the test image stack (section 627 of *high_c4* in “BetaSeg” dataset).

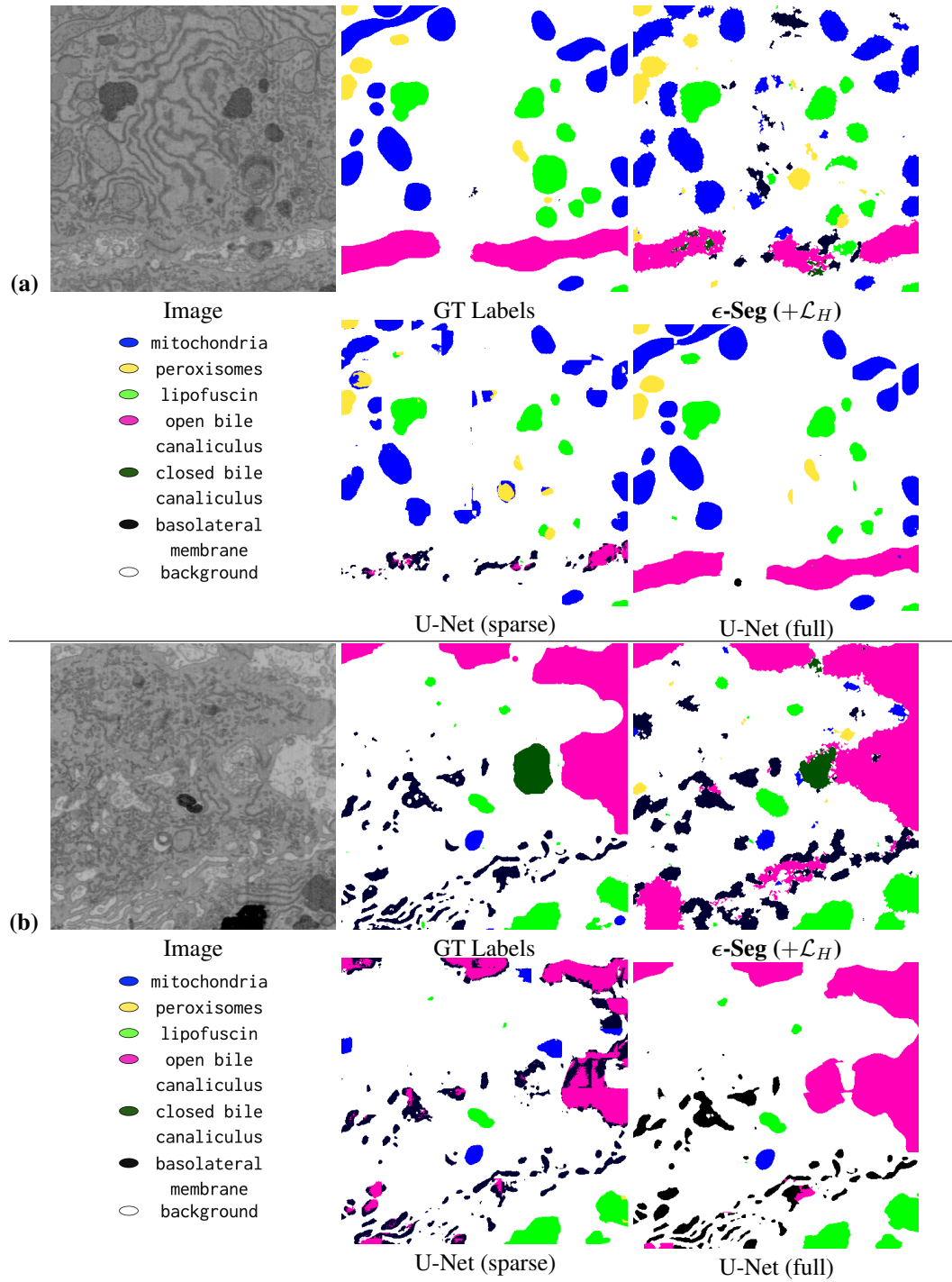


Figure S4: Qualitative segmentation result on two crops of the whole 3D volume. (a) and (b) are section 80 and 26 of crop00 and crop10 in “liver FIBSEM” dataset respectively. U-Net (sparse) and (full) is sparsely-supervised and fully-supervised respectively.

RLF	Model	U	N	G	M	Avg DSC
20	U-net	0.63	0.75	0.51	0.12	0.50
	ϵ -Seg	0.89	0.98	0.81	0.83	0.88
15	U-net	0.53	0.64	0.41	0.14	0.43
	ϵ -Seg	0.88	0.98	0.81	0.78	0.86
10	U-net	0.30	0.20	0.42	0.34	0.31
	ϵ -Seg	0.86	0.98	0.80	0.75	0.85
5	U-net	0.71	0.00	0.00	0.03	0.18
	ϵ -Seg	0.85	0.96	0.77	0.76	0.84
1	U-net	0.17	0.00	0.37	0.02	0.14
	ϵ -Seg	0.79	0.95	0.69	0.69	0.78

Table S4: Comparison between U-Net and ϵ -Seg on the “BetaSeg” dataset under varying label sparsity levels. “RLF” (Relative Labeling Factor) specifies the fraction of available labels, where 20 corresponds to 0.05% and 1 to 0.0025% of total labels. U: Unrecognized, N: Nucleus, G: Granules, M: Mitochondria. Although both models were trained with *balanced supervision*, using patches selected to include all classes, the U-Net still fails to segment the nucleus at very low labeling levels (RLF 1 and 5). This illustrates a key limitation of discriminative models such as U-Net, under extreme supervision sparsity, even balanced examples may not suffice to generalize fine-grained or context-sensitive structures like the nucleus. In contrast, ϵ -Seg benefits from its class-aware latent modeling via the GMM prior, which enables it to extract meaningful representations for different structures and distinguish them semantically. We note that the sparse U-Net reported earlier was trained on slice numbers 800, 600, and 500 of the “high_c1”, “high_c2”, and “high_c3” volumes of the “BetaSeg” dataset. For selecting the same amount of data used in ϵ -Seg, to train the 2D U-Net on, as reported in the table above, we extracted 64x64 patches where except background, different classes are approximately well balanced.