Supplementary Material

Vu Minh Hieu Phan[1][0000−0003−3861−0296], Yutong Xie[1], Bowen Zhang[1], Yuankai Qi[2], Zhibin Liao[1], Antonios Perperidis[1], Son Lam Phung[3], Johan W. Verjans[1], and Minh-Son To[4]

[1] Australian Institute for Machine Learning, University of Adelaide, Australia
[2] Macquarie University, Australia
[3] University of Wollongong, Australia
[4] Flinders University, Australia

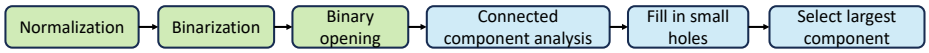

Fig. 1. Traditional image processing for mask extraction consists of two stages: image pre-processing (green blocks) and connected component analysis (blue blocks).

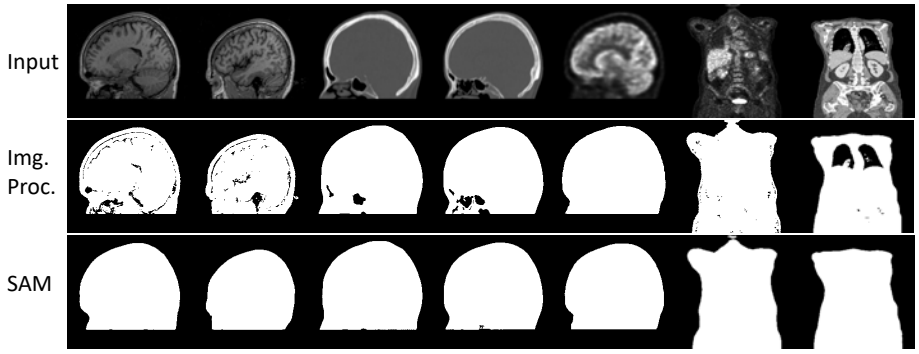

Fig. 2. Extracted masks using standard image processing algorithms (Row 2) and using deep-learning-based Segment-Anything-Model (Row 3).

Table 1. Accuracies on three tasks using different attention mechanisms for foreground (FG) and background (BG). Reversing the attention strategies with structural attention on BG reduces the performance significantly.

FG	BG	PET-CT		MR-CT		MR-PET	
		MAE↓	SSIM↑	MAE↓	SSIM↑	MAE↓	SSIM↑
Global		10.91	75.88	7.84	81.23	7.33	81.96
Local		10.82	75.97	8.12	81.06	7.64	81.53
Local	Struct	10.51	75.91	7.75	81.48	7.21	82.15
Struct	Local	**9.57**	**78.47**	**6.18**	**85.48**	**6.32**	**82.97**

Table 2. Hyper-parameter analysis. Evaluating the effect of mask loss weight (λ_{mask}), local attention window size (win. size), and classification threshold (σ) on PET-CT performance. The model performance is insensitive when classification threshold, $\sigma > 0.5$. Using low $\lambda_{\text{mask}} = 0.1$ can lead to lower performance, showing the effects of learning structural information via the mask loss. A high window size with a value of 8 leads to optimal performance. Using a large window size for local attention on background allows a more effective information exchange between foreground and background features.

λ_{mask}	win. size	σ	MAE	PSNR	SSIM
0.5	2	0.5	9.84	33.89	77.70
0.5	4	0.5	9.71	33.96	78.12
0.5	8	0.05	10.02	33.78	76.12
0.5	8	0.25	9.72	33.98	78.20
0.5	8	0.6	9.65	34.01	78.31
0.5	8	0.8	9.66	34.03	78.38
0.1	8	0.5	9.92	33.86	77.48
0.5	8	0.5	**9.57**	**34.05**	**78.47**
1.0	8	0.5	9.75	33.93	78.06

Table 3. Ablation study on learning masks via a patch classifier versus using ground-truth (GT). GT masks are extracted via image processing or SAM. Plus, stop-gradient (SG) from a mask loss to a patch classifier is ablated. Inference time on 224×224 images is reported. Performance of using GT masks from SAM is similar to using the patch classifier, while increasing computational overhead (from 0.014s to 1.924s), showing effectiveness of our design for *distilling structural knowledge* from a powerful SAM to a lightweight patch classifier. Stopping gradients to patch classifier reduces performance, showing the benefits of training it with mask loss.

Patch cls.	SG	GT mask gen.	Overhead (s)	PET-CT			MR-PET			MR-CT		
				MAE	PSNR	SSIM	MAE	PSNR	SSIM	MAE	PSNR	SSIM
✓	✓	✗	0.014	9.57	34.05	78.47	6.32	34.24	82.97	6.18	35.76	85.48
✓	✗	✗	0.014	9.82	33.82	77.34	6.62	34.11	81.88	6.41	35.23	84.72
✗		Img proc.	0.016 (0.002)	10.51	33.82	77.02	6.87	34.03	81.73	7.25	34.85	83.41
✗		SAM	1.924 (1.91)	9.71	33.97	78.05	6.54	34.15	82.26	6.30	35.55	84.93