Unleashing the Potential of SAM for Medical Adaptation via Hierarchical Decoding

Supplementary Material

In this supplementary material we first provide more implementation details as in Sec. 3 and Sec. 4 for training strategies and datasets. Then, we present more visualized results with zoom-in analysis for H-SAM. Finally, we present an organ-by-organ analysis of the result of H-SAM on Synapse CT dataset.

A. Implementation Details

A.1. Training strategy

We provide training strategy and hyper-parameter setting as supplementary for Sec. 3 and Sec. 4. We adopt warmup during training. As shown in Table 7, the initial learning rate is set to 0.0025, and the warmup period is set to 250. The training loss we use is a combination of Dice loss $\mathcal{L}_{dice}$ and MSE loss $\mathcal{L}_{ce}$. To be specific, as shown in Table 8, the weight of each loss is 0.9 for $\mathcal{L}_{dice}$ and 0.1 for $\mathcal{L}_{ce}$. For our 2-stage hierarchical structure, there is also a λ_w for each loss in the 2 stages. The final loss $\mathcal{L}_{total}$ a sum of $\lambda_w \mathcal{L}_{stage1}$ and $(1 - \lambda_w)\mathcal{L}_{stage2}$. The parameter λ_w is set to gradually decrease in the way of exponential decay, from 0.4 to 0 in 300 training epochs. The first decoder output is supervised by 1/4 resolution ground truth, and the second output by full resolution. The final output is ensembled from the two outputs, where we utilize a mean value of results from the two stages.

A.2. Additional datasets information.

In Sec. 4, we present our dataset settings. The Synapse dataset we experiment on is from MICCAI 2015 Multi-Atlas Abdomen Labeling Challenge, containing 3779 axial contrast-enhanced abdominal CT images in total and the training set contains 2212 axial slices. We follow TransUnet to evaluate eight abdominal organs (aorta, gallbladder, spleen, left kidney, right kidney, liver, pancreas, stomach). For fully-supervised training, we strictly follow TransUnet of the division between training and testing cases. For the few-shot setting on the Synapse dataset, we adopt a slice-based dataset selection. We select 10% training data, i.e., 221 slices, *randomly* from different subjects in the training volumes, which contains 2212 axial slices in total.

B. Additional ablation analysis

B.1. Additional organ-by-organ analysis

In Table 4, we provide an ablation study of the Effectiveness of the key contributions in H-SAM: Learnable Mask-Attention, CMAttn, and Hierarchical Pixel Decoder. In Table 11, here we present an additional organ-by-organ analysis to further prove the validity of H-SAM's innovation. The implementation of Learnable Mask-Attention alone brings a 2.1% improvement in terms of mean dice. Learnable Mask-Attention also achieves the highest 94.43% results for the organ Liver. Both its combination with CMAttn and Hierarchical Pixel Decoder achieves promising results on some relatively small-scale organs, such Pancreas (58.18%) and Aorta (84.37%). The combination of all three implementations shows promising results on most organs, reflecting the meticulous design of our hierarchical decoding strategy.

Config	Setting
Optimizer	AdamW
Learning rate	2.5e-3
Batch size	32
Weight decay	0.1
Optimizer momentum	$\beta_1 = 0.9$ $\beta_2 = 0.999$
Warmup period	250

Table 7. Training setting

Hyper parameter	Setting
$\mathcal{L}_{dice}$	0.9
$\mathcal{L}_{ce}$	0.1
$\lambda_{w_{start}}$	0.4
$\lambda_{w_{end}}$	0

Table 8. 2 stage hyper-parameter setting

B.2. Additional analysis under Synapse semi-supervised setting

For the few-shot setting on the Synapse dataset, our H-SAM shows outstanding performance under a slice-based dataset selection. We also validate H-SAM in volume-based dataset selection strictly following the same training split (5 subjects) in MagicNet. As shown in Table 9, similar to our observation for PROMISE12 and LA, H-SAM also outperforms the SOTA semi-supervised method MagicNet [7] without using any unlabeled data.

Methods	Scans used Labeled	 Unlabeled	Mean Dice (%)↑
SS-Net [61]	5(30%)	13(70%)	56.74
UA-MT [62]			61.20
MagicNet [R1]			75.53
H-SAM (ours)	5(30%)	0(0%)	**79.36**

Table 9. Semi-supervised results on Synapse Dataset

Rank size	Mean Dice (%)	Mean HD
1	71.14	27.03
4	80.35	15.54
8	79.15	16.19
16	76.14	16.30

Table 10. Ablation study on rank size of LoRA layers

B.3. Ablation study on the LoRA component

In Table 10, we discuss the effectiveness of the layers of LoRA component. We discover the performance of H-SAM increases to rank=4, but the performance drops gradually when the rank is too large.

C. Visualization

C.1. Zoom-in analysis

As shown in Figure 6 is the zoom-in visualization of H-SAM results against other SAM prompt-free variants. On the Synapse multi-organ CT dataset, H-SAM performs precise segmentation for small-scale organs. The pancreas marked as yellow in the figure is represented in a small region. SAM Adapter and AutoSAM provide a multi-organ segmentation with noise. While SAMed outputs a result with lesser noise, it is also confused by the shape of the organ. H-SAM outputs a perfect result with the correct shape and no noise.

C.2. Visualization on Synapse dataset

As shown in Figure 7, we present the visualization of semantic segmentation predictions on the Synapse dataset. Compared to ground truth, H-SAM performs promising results with both multiple organs (up to 8) and fewer organs.

C.3. Visualization on 2 stages

Here we present the visualization of outputs from different stages. As shown in Figure 8, benefit from the joint training design, both of the 2 stages perform excellent segmentation predictions. In row 5, we present a failure case where stage 2 takes an erroneous prediction from stage 1 as the prior and mistakes a background region to kidney. However, in most cases, the stage-2 prediction takes and corrects stage-1 results as the prior to generating finer segmentation, which especially can be reflected from small organs like Pancreas, as shown in rows 3 and 4.

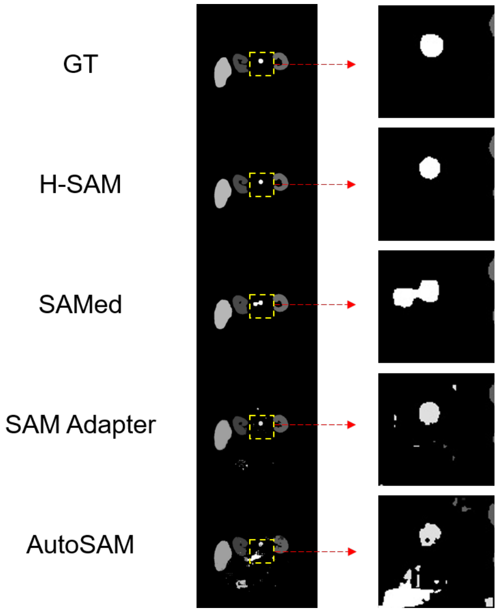

Figure 6. The zoom-in analysis of H-SAM results against other SAM prompt-free variants. H-SAM performs precise segmentation for small-scale organs.

Learnable Mask-Attention	Hierarchical Pixel Decoder	CM Self-Attention	Spleen	Right Kidney	Left Kidney	Gallbladder	Liver	Stomach	Aorta	Pancreas	Mean Dice (%)
✗	✗	✗	85.82	82.26	82.62	63.15	92.71	67.20	78.72	52.12	75.57
✓	✗	✗	89.56	84.18	82.06	62.48	**94.43**	70.97	85.22	52.58	77.68
✓	✓	✗	89.91	83.93	79.70	**70.87**	94.16	70.85	82.47	56.72	78.58
✗	✓	✗	87.32	84.78	79.85	69.39	93.86	68.48	79.43	53.31	77.05
✗	✓	✓	89.11	**85.04**	83.77	69.79	94.00	**77.93**	81.71	50.94	79.03
✗	✗	✓	89.86	84.08	82.38	65.02	94.05	73.82	81.87	50.61	77.71
✓	✗	✓	87.51	83.98	80.95	65.25	94.13	75.66	84.37	**58.18**	78.76
✓	✓	✓	**90.21**	84.16	**85.65**	70.70	94.29	76.10	**85.54**	56.17	**80.35**

Table 11. Additional ablation result of Learnable Mask-Attention, CMAttn, and Hierarchical Pixel Decoder.

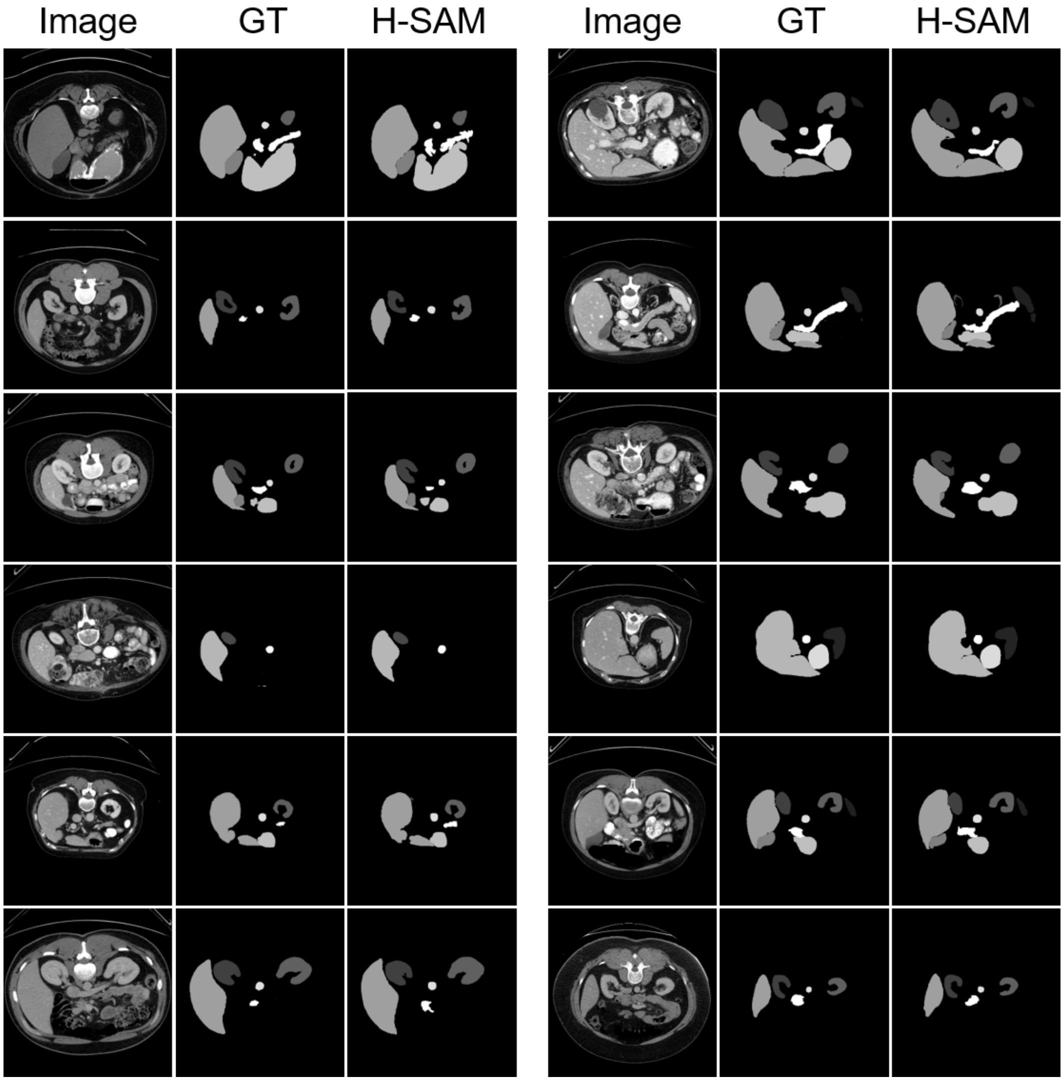

Figure 7. Visualization of semantic segmentation predictions on the Synapse dataset. First and fourth columns: raw image. Second and fifth columns: ground truth. Third and sixth columns: prediction.

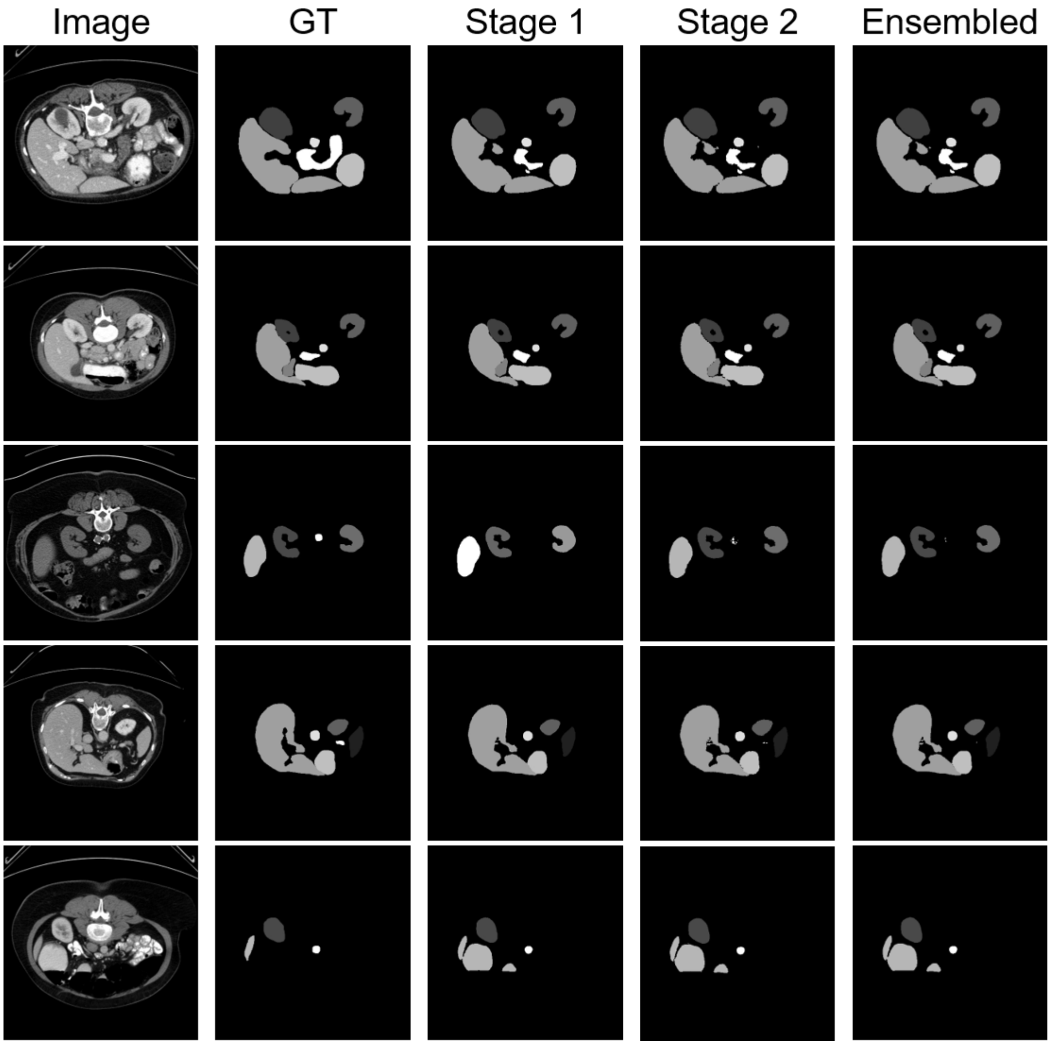

Figure 8. Visualization of the outputs from different stages. First column: raw image. Second column: ground truth. Third column: stage-1 output. Fourth column: stage-2 output. Fifth column: ensembled output from 2-stage outputs. The last row shows failure cases where stage 2 takes an erroneous prediction from stage 1 as the prior.