MaskSAM: Auto-prompt SAM with Mask Classification for Volumetric Medical Image Segmentation

Supplementary Material

6. Prompts from Ground Truth.

To fairly compare and demonstrate the effectiveness of our model, we fine-tune it using accurate prompts generated from ground truth (GT) as input. In this version of the model, we remove the prompt generator and instead use the accurate prompts as input for the prompt encoder, while keeping the rest of the components unchanged from our MaskSAM model, including the designed adapters. Using accurate prompts from GT, our model achieves excellent performance, with a 94.70% Dice and 94.30% on Synapse and ACDC datasets, which outperform nnUNet by 8.5% and 2.7%, respectively. Furthermore, these results significantly surpass all existing SAM-based methods. These findings demonstrate that our designed adapters can effectively adapt SAM for medical image segmentation when provided with accurate prompts. Additionally, they show that, under the same conditions with accurate prompts, our model achieves near-perfect performance, significantly outperforming other SAM-based methods.

7. Implementation Details

Our models are based on the codebase of nnUNet for the preprocessing and postprocessing. For the preprocessing, we utilize some data augmentations such as rotation, scaling, Gaussian noise, Gaussian blur, brightness, and contrast adjustment, simulation of low resolution, gamma augmentation, and mirroring. During training, we set the initial learning rate to 0.001 and employ a "poly" decay strategy in Eq. (2).

$$\mathrm{lr}(e) = \mathrm{init\_lr} \times (1 - \frac{e}{\mathrm{MAX\_EPOCH}})^{0.9}, \qquad (2)$$

where e means the number of epochs, MAX_EPOCH means the maximum number of epochs, set it to 500, 1000, and 1000 for ACDC, AMOS2022, and Synapse dataset, respectively, and each epoch includes 250 iterations. For the postprocessing, we crop the whole image into several patches with a half-overlap. For each patch, our model infers 8 times by different three axes (*i.e.* axial, sagittal, and coronal planes) and then averages all outputs to get the final predictions. All experiments are conducted using NVIDIA RTX A6000 GPUs with 48GB memory.

8. Theoretical Comparisons

	AMOS22 [29]	Synapse [31]	ACDC [3]
Median Image Shape	$104 \times 512 \times 512$	$128 \times 512 \times 512$	$9 \times 256 \times 216$
Patch Sizes	$9 \times 512 \times 512$	$26 \times 256 \times 256$	$4 \times 256 \times 256$
SAM Model Types	vit_h	vit_h	vit_b
# Tunable Params	66M	65M	21M

Table 5. Configurations for different datasets.

# Params (M)	nnUNet	MedSAM	SAMed	Med-SA
Tunable Params	29M	91M	19M	13M
Total Params	29M	91M	636M+19M	636M+13M
# Params (M)	SAM3D	3DSAM-Adapter	AutoProSAM	MaskSAM (ours)
Tunable Params	2M	26M	27M	21M
Total Params	91M+2M	91M+26M	91M+27M	91M+21M

Table 6. Comparison of tunable and total parameters on ACDC.

The main contributions of our MaskSAM different from the existing SAM-based models are i) automatic prompt, ii) the classifier to generate semantic labels for each mask, and iii) remain all parameters of the original SAM for zero-shot capabilities. There are several categories of the existing SAM-based models. The first category does not modify SAM, such as MedSAM and Polyp-SAM. These models need manual prompts, such as points or boxes, and cannot classify masks into semantic labels. The second category uses parameter-efficient transfer learning, such as Adapters, into SAM. The popular model, Med-SA, uses the GT to generate prompts during inference, which do not have any practical clinical values. It also includes the non-automatic models of the 3DSAM-Adapter and MA-SAM. These models do not handle the requirements of extra prompts. The third category is that cannot classify masks into semantic labels, such as DeSAM, Med-SA, and MA-SAM. Since SAM only predicts binary masks, these models do not address the lack of classifiers. The fourth category is abandoning the components of SAM, such as Mask Decoder, to handle the inability to classify semantic labels, such as 3DSAM-Adapter. This way inevitably destroys the consistency and zero-shot capabilities of SAM. These models only use the pre-trained ViT encoder, which is not the contribution of SAM.

9. Configurations and Parameters

In Table 5, since our method is built upon the large-scale Segment Anything Model (SAM), it incurs substantial memory overhead during training. Furthermore, SAM is originally designed for 2D natural images, making it less straightforward to directly apply to 3D medical images, where the additional depth dimension (i.e., the z-axis) significantly increases memory consumption. As a result, the usable depth is often constrained to a small value. To enable successful training of SAM under limited GPU memory, we adopt dataset-specific configurations.

For the ACDC dataset, the median depth is only 9 slices, which fits well within the memory constraints and thus does not require major compromises. In this case, we use the SAM base model with 21M tunable parameters and apply a series of CNN-based upsampling operations to expand the spatial resolution in the x and y dimensions to SAM's default input size (1024×1024). This configuration achieves near breakthrough-level performance, demonstrating the strong segmentation capability of SAM when adapted to 3D medical images with shallow depth.

In contrast, the AMOS22 and Synapse datasets contain significantly larger depth (z-axis) values. Reducing the depth too much in these datasets would hinder the model's ability to capture long-range 3D context, leading to weak semantic predictions and degraded performance. Therefore, instead of upsampling x and y dimensions to 1024×1024 via CNNs, we do not change the patch size for SAM's modules, allowing more slices along the z-axis to be processed under the same memory budget. Given the increased difficulty of the two datasets, larger anatomical variation, more target classes, and greater image complexity, we adopt the SAM huge model as the backbone, resulting in 66M and 65M tunable parameters for AMOS22 and Synapse, respectively.