SAMU: An Efficient and Promptable Foundation Model for Medical Image Segmentation

Joseph Bae[1], Xueqi Guo[1], Halid Yerebakan[1], Yoshihisa Shinagawa[1], and Sepehr Farhand[1]

Siemens Healthineers, Malvern, PA 19355, USA
{joseph.bae, sepehr.farhand}@siemens-healthineers.com

Abstract. Segmentation of 3D medical images is a labor-intensive task with important clinical applications. Recently, foundation models for image segmentation have received significant interest. Specifically, many works have proposed methods for the adaptation of promptable natural image foundation models to medical image segmentation. However, the shift to 3D volumes from 2D natural images has proven difficult, and many approaches have limited real-world clinical applicability due to large model sizes and corresponding heavy computational requirements. Here, we present an original model for generalized, promptable 3D medical image segmentation. Our approach leverages a lightweight convolutional backbone while simultaneously integrating information from single-point prompts at multiple spatial resolutions. Our approach dramatically reduces the computational burden for promptable segmentation while also outperforming similar recent works on a diverse dataset of 98,699 image-mask pairs from CT and MRI datasets.

Keywords: Foundation Models · Segmentation · Prompting

1 Introduction

Medical image segmentation is an important task for the diagnosis, treatment, and management of human diseases. Automated delineation of abnormalities can prevent missed diagnoses and enable early detection of pathology. Interventional treatments including radiotherapy for malignancies rely heavily on accurate healthy organ segmentation in order to minimize damage to non-targeted structures. Finally, longitudinal contouring of pathologies can inform treatment decisions based on changes in disease presentation. However, manual segmentation of 3D medical images is a labor-intensive process in routine clinical workflows. While recent foundation models have been presented for generalized segmentation of both natural and medical images, the massive models proposed are difficult to train, finetune, or use in low-resource settings including for individual hospitals or clinics. This is particularly salient when many foundation models trained on natural images must be finetuned or completely re-trained from scratch on medical imaging datasets in order to overcome the large domain shift from natural to medical imaging. Inspired by the Segment Anything Model

(SAM) [9] created by Meta AI for natural image segmentation, we propose a lightweight, CNN-based architecture for promptable segmentation of 3D medical images. Our approach is adaptable to multiple anatomical sites and imaging modalities while requiring significantly fewer computational resources and outperforming recently proposed Vision Transformer [5] (ViT)-based methods.

SAM is among the most widely used foundation models for segmentation of natural images and can leverage a single-point prompt as input. However, it has been observed that the original implementation of SAM has difficulty generalizing to medical imaging tasks [11]. As a result, multiple strategies have been proposed to adapt SAM to medical images with varying success [4,10,14,13]. Currently, most of these works have focused primarily on 2D image segmentation [4,10], with 3D volumes being segmented on a slice-by-slice basis. This approach is generally ineffective at producing high-fidelity segmentations with spatial consistency across slices, and also requires modeling every individual image slice with associated high-computation costs and a need for per-slice prompts. SAM-Med3D [14] attempted to address these problems by training a SAM model using 3D ViT blocks on a large dataset of 131,000 mask-image pairs, thereby natively accommodating 3D inputs with a 3D architecture. However, the 100 million trainable parameters in the model make it difficult to finetune and result in relatively slow inference times. FastSAM3D [13] was proposed to reduce this computational burden through a knowledge distillation process in which a lightweight TinyViT architecture was trained as a student model with SAM-Med3D as a teacher. Additionally, FastSAM3D incorporated 3D sparse flash attention to further improve model efficiency. While FastSAM3D was able to effectively reduce the number of trainable parameters required for promptable segmentation to 53 million, the model is still large enough to be potentially prohibitive for hospitals and clinics with low computational resources. Further, both SAM-Med3D and FastSAM3D perform relatively poorly in the one-point prompt segmentation setting.

We focus here primarily on the single-point prompt segmentation task for the following reasons. First, in busy clinical settings, point annotation of only one slice in a 3D volume is more efficient and practical than requiring multiple annotations. Second, many previous implementations of multiple-point prompting are ambiguous, and may not be a realistic reflection of presumed real-world usage. For instance, in some implementations, each point beyond the first is chosen only within the false-negative region, implying a gradual and interactive refinement of model outputs in inference. This can prohibitively increase inference times due to repetitive modeling of a single image and requires significantly more human resources compared to single-point prompting.

In this work, we propose a lightweight model inspired by **SAM** and **U**Net (SAMU) for promptable 3D medical image segmentation. Our approach is motivated by a need for a computationally inexpensive model for one-point volumetric segmentation and is trained and evaluated on a large dataset of 98,699 image-mask pairs (N=2,703 subjects) spanning CT and MRI modalities. Our approach is comprised of the following contributions:

1. SAMU enables 3D medical image segmentation with significantly fewer parameters than other 3D approaches.
2. SAMU incorporates multi-scale prompt encoding to more thoroughly exploit positional information from single-point prompts.
3. SAMU outperforms previously published baselines on a diverse dataset of >90,000 mask-image pairs.

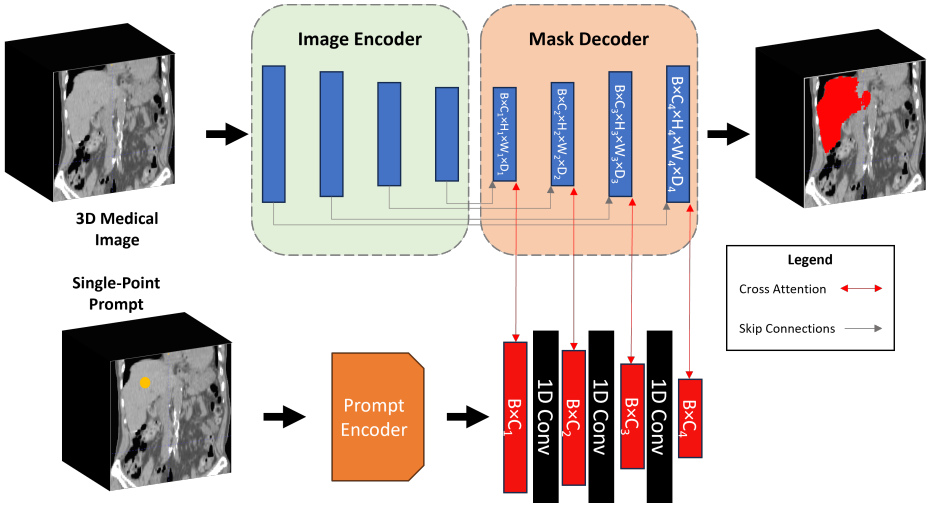

Fig. 1. SAMU Architecture. Shown is the proposed architecture for our promptable segmentation framework. SAMU leverages a UNet backbone for image encoding and mask decoding. The prompt encoding module from SAM is utilized to generate feature representations of single-point prompts. Cross-attention is performed at multiple spatial resolutions for prompt supervision of mask decoding.

2 Methodology

We first describe the promptable segmentation framework popularized by SAM. We then elaborate upon our proposed architecture and highlight key innovations. An overview of our approach is presented in Figure 1.

2.1 SAM Overview

SAM is composed of an image encoder, a prompt encoder, and a mask decoder. The image encoder is a ViT-H model pre-trained via a masked auto-encoder strategy [7,16] that leverages sequential attention and multi-layer perceptron blocks to extract image representations for the segmentation task. The original SAM prompt encoder flexibly allows point, bounding box, mask, and text

prompts. For point prompts, a learned representation is created by concatenating a positional encoding vector with learnable tokens for segmentation. The mask decoder performs cross-attention between prompt and image embeddings before upsampling image embeddings to an output mask prediction.

2.2 Proposed SAMU

SAMU Encoder-Decoder Backbone Unlike SAM, SAMU integrates both image encoder and mask decoder into a single UNet-like architecture. Specifically, for the encoding branch, SAMU utilizes 3D convolutions combined with max-pooling operations for feature downscaling. In the decoder branch, 3D transposed convolutions are used to recover the spatial dimensions of the input. Further, skip connections between corresponding spatial dimensions in the encoder and decoder branches of SAMU are implemented. This architecture allows the network to leverage the multi-scale feature information present at different spatial resolutions, a capability demonstrated to be significant in other medical imaging applications of UNet. Further, this framework is significantly less computationally expensive when compared to ViT encoders.

SAMU One-Point Prompt Encoding For each volumetric 3D medical image studied, a single-point prompt is encoded into a positional embedding with concatenated learnable tokens as described in SAM. We discard the bounding box, mask, and text prompt encoding modules of SAM. Each learned prompt representation is incorporated into the decoder branch of SAMU at multiple resolutions. Specifically, at each spatial resolution in the UNet decoder, cross-attention is performed between the prompt feature vector and the image embedding. To align the dimensions of the prompt feature vector with the channel dimension of each image embedding, 1D convolutions are performed. Therefore, the prompt feature vector p_n is of $B \times C_n$ dimensionality where B is the batch size and C_n is the number of channels of the image representation r_n at the n^{th} level of the UNet backbone. As a result of these steps, important localizing information present in the single-point prompts is provided to the network at multiple spatial resolutions, ensuring that the model can adequately learn which regions to segment during inference.

3 Experiments and Results

3.1 Datasets and Implementation

We studied one-point segmentation of a variety of anatomical and pathological targets of interest. For each volumetric medical image, a single-point prompt was randomly chosen within the entire 3D ground truth of each target volumetric mask. 15 abdominal organs from the AMOS [8] abdominal CT/MRI dataset (N=360), 3 tumor regions from the Brain Tumor Segmentation 2021 (BraTS)

[1,12,2] dataset (N=1,251), and 117 anatomical structures from the Total Segmentator (TotalSeg) [15] dataset (N=1,092) were studied. Data partitioning was performed in the same manner as in SAM-Med3D [14] and FastSAM3D [13] for fair comparisons. The official training split for AMOS was further randomly divided for training and validation while the official validation split was used for testing. The training split of BraTS was randomly divided into training, validation, and testing splits at a 70/10/20 ratio. Individual MRI sequences in BraTS were treated as independent samples, but all images for a given patient could only appear in one of the train, validation, or test splits in order to avoid data leakage. Training, validation, and testing splits were used as described by the Total Segmentator dataset. A single instance of SAMU was trained concurrently on all three training datasets and evaluated on all three test datasets; individual models were not independently trained for each dataset and modality. Prior to image input, each volume was z-score normalized and cropped to a 128x128x128 voxel patch as described in SAM-Med3D [14] and FastSAM3D [13]. Random flipping was used for data augmentation. The Dice score metric was used to compare segmentation quality.

All models were trained on the same single NVIDIA RTX 6000 Ada Generation GPU. SAMU was trained with a batch size of 16 and optimized with AdamW with a learning rate of 0.001. Pre-trained weights for Med-SAM3D and FastSAM3D were obtained from their respective authors.

Table 1. Dice Scores of Segmentation Results

		AMOS	BraTS	TotalSeg
2D	SAM [9]	0.049	0.108	0.202
	SAM-Med2D [4]	0.097	0.013	0.008
	MedSAM [10]	0.004	0.008	0.006
3D	SAM-Med3D [14]	0.453	0.365	0.334
	FastSAM3D [13]	0.307	0.315	0.242
	SAMU (ours)	**0.677**	**0.434**	**0.756**

3.2 Segmentation Results

Table 1 displays Dice scores results for SAMU as well as 2D and 3D baselines across all datasets studied when using a single-point prompt. Note that 2D results are directly reported in [13]; we were unable to replicate each 2D approach due to limited resources. Furthermore, each 2D method required slice-by-slice rather than full volume inference, therefore also requiring a single-point prompt per slice rather than a single-point prompt per image volume. Despite this advantage, SAMU and other 3D approaches still outperform their 2D counterparts on medical imaging datasets. Furthermore, SAMU outperforms SAM-Med3D and FastSAM3D in the single-point prompt evaluation schema across all datasets.

Qualitative results from SAM, SAM-Med3D, FastSAM3D, and SAMU are presented in Figure 2. The original SAM model was frequently unable to achieve satisfactory performance on medical imaging tasks. Additionally, SAM-Med3D and FastSAM3D seem to heavily rely on prompt location and do not always produce quality segmentations conforming to anatomical structures. In contrast, SAMU produces segmentations that more accurately delineate targets of interest. Note that the same single-point prompt was provided to each model studied.

Table 2. Dice Scores for Ablation Studies

	AMOS	BraTS	TotalSeg
Baseline 1	0.586	0.358	0.694
Baseline 2	0.643	0.378	0.733
Baseline 3	0.628	0.400	0.676
SAMU (ours)	**0.677**	**0.434**	**0.756**

3.3 Ablation Experiments

To evaluate the impact of different components of SAMU, we compared model performance to that of a variant with no prompt provided (Baseline 1), a version of SAMU in which prompt cross-attention is only performed at the bottom layer of the UNet architecture (Baseline 2), and a SAMU variant in which cross-attention is applied to both image embeddings and skip-connected residual feature maps (Baseline 3). The results are shown in Table 2. As might be expected, the addition of prompts (Baselines 2 and 3, SAMU) improves model performance. Further, the absence of multi-scale prompt information seems to heavily reduce performance (Baseline 2 vs. SAMU), as does cross attention between prompts and residual features (Baseline 3 vs. SAMU).

3.4 Computational Requirements

Table 3 displays the resource requirements of SAMU when compared with SAM, SAM-Med3D, and FastSAM3D. In addition to using substantially fewer parameters, SAMU is faster in inference. It should be noted that 3D models are significantly faster for 3D medical imaging inference as 2D SAM approaches require per-slice inference along with input of per-slice single-point prompts rather than per-volume prompts.

4 Discussion and Conclusion

Here we present SAMU, a lightweight promptable architecture for medical image segmentation. SAMU builds upon medical SAM models for 3D image segmentation while remaining accessible to users without powerful computational

Table 3. Computational Requirements

	Params	Inference Time (s)
SAM [9]	636M	52.521
SAM-Med3D [14]	100M	1.188
FASTSAM3D [13]	53M	0.280
SAMU (ours)	**12M**	**0.049**

resources. One clinical scenario with clear potential applications for promptable segmentation is in the field of radiation oncology, where physician annotation of anatomical structures is a labor-intensive, but necessary task [3]. One-point prompting for 3D segmentation in this context would save human resources while enabling improved patient care. SAMU outperforms previous methods on multiple diverse 3D medical image datasets. In addition to reducing computational requirements by utilizing a convolution UNet backbone, SAMU more completely leverages single-point prompt information by incorporating prompt features at multiple spatial resolutions in the decoder branch of the network. This may provide a stronger supervision signal than the prompt attention method used in other SAM architectures.

One limitation of our work is an inability to fully replicate the large-scale training datasets used by SAM-Med3D as the full data has not been made publicly available at this time. However, for the three datasets studied in this work, the authors of SAM-Med3D and FastSAM3D have confirmed previously that the same data splits were used for their model training and evaluation, thereby allowing a fair comparison of model performance. Further, while improved when compared to other baselines, SAMU still does not achieve outstanding segmentation performance. This may be due to the great heterogeneity in modalities and structures present in the training and evaluation data. Nevertheless, SAMU's ability to produce reasonable segmentations on diverse CT and MRI images for a wide range of targets reflects a promising generalizability as an efficient medical imaging foundation model. Due to its relatively low computational cost, SAMU is particularly well-suited to future study in meta-learning [6] frameworks which might improve model generalization to the heterogenous datasets prevalent in the medical domain. We believe SAMU represents an important contribution to 3D medical imaging segmentation which may allow for improvements in clinical workflows.

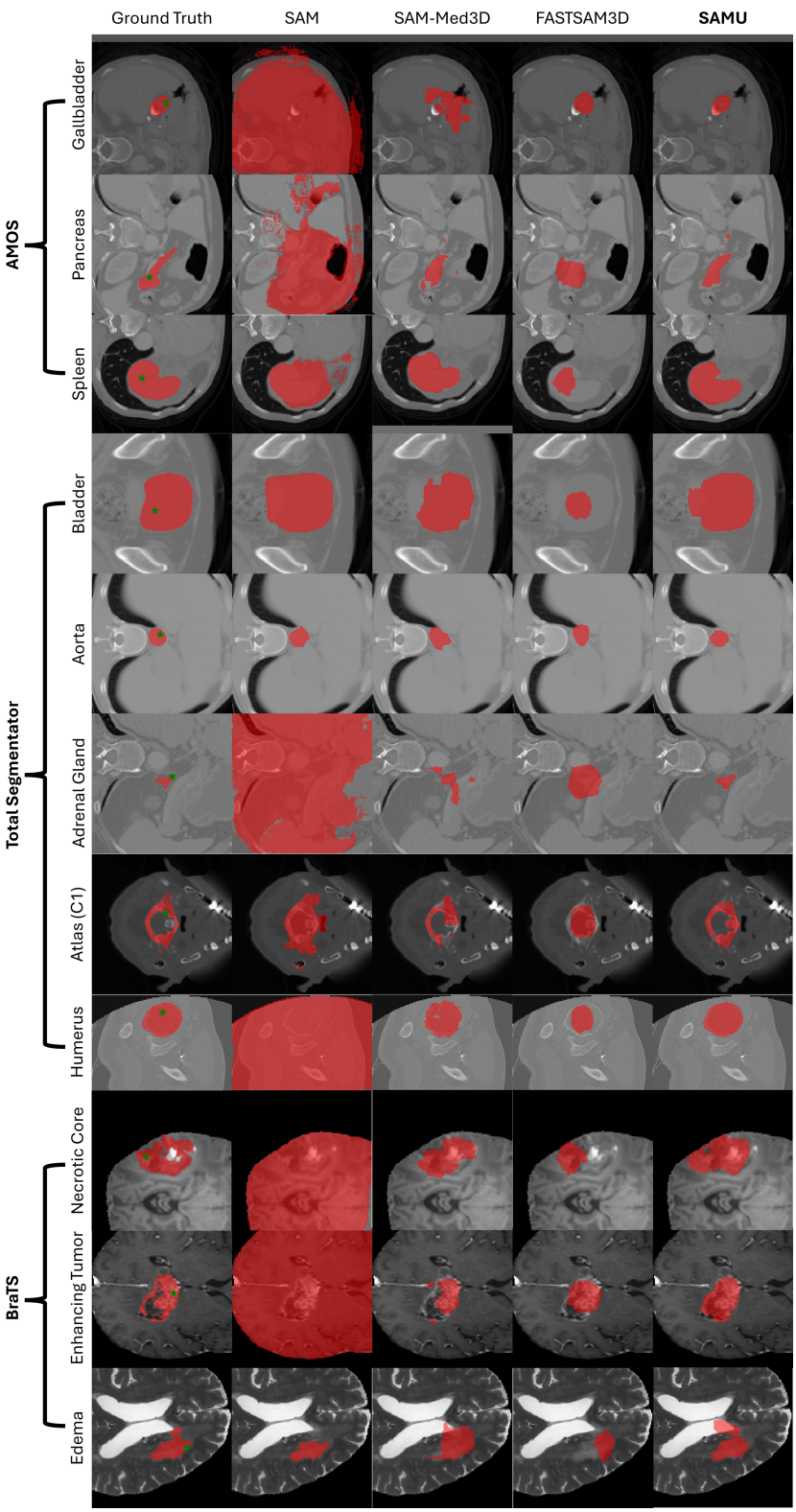

Fig. 2. Qualitative Results. Shown are example segmentations for the studied approaches. Single-point prompts were identically chosen for each method (green star).

References

1. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., et al.: The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021)
2. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific data **4**(1), 1–13 (2017)
3. Cardenas, C.E., Yang, J., Anderson, B.M., Court, L.E., Brock, K.B.: Advances in auto-segmentation. In: Seminars in radiation oncology. vol. 29, pp. 185–197. Elsevier (2019)
4. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., et al.: Sam-med2d. arXiv preprint arXiv:2308.16184 (2023)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: International conference on machine learning. pp. 1126–1135. PMLR (2017)
7. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
8. Ji, Y., Bai, H., Ge, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhanng, L., Ma, W., Wan, X., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. Advances in neural information processing systems **35**, 36722–36732 (2022)
9. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
10. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
11. Mazurowski, M.A., Dong, H., Gu, H., Yang, J., Konz, N., Zhang, Y.: Segment anything model for medical image analysis: an experimental study. Medical Image Analysis **89**, 102918 (2023)
12. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)
13. Shen, Y., Li, J., Shao, X., Romillo, B.I., Jindal, A., Dreizin, D., Unberath, M.: Fastsam3d: An efficient segment anything model for 3d volumetric medical images. arXiv preprint arXiv:2403.09827 (2024)
14. Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., Fu, B., Zhang, S., He, J., Qiao, Y.: Sam-med3d. arXiv preprint arXiv:2310.15161v2 (2024)
15. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al.: Totalsegmentator: robust segmen-

tation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5) (2023)
16. Zhou, L., Liu, H., Bae, J., He, J., Samaras, D., Prasanna, P.: Self pre-training with masked autoencoders for medical image classification and segmentation. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). pp. 1–6. IEEE (2023)