2024 IEEE International Symposium on Biomedical Imaging (ISBI) | 979-8-3503-1333-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISBI56570.2024.10635207

PROMISE: PROMPT-DRIVEN 3D MEDICAL IMAGE SEGMENTATION USING PRETRAINED IMAGE FOUNDATION MODELS

Hao Li, Han Liu, Dewei Hu, Jiacheng Wang, Ipek Oguz

Vanderbilt University

ABSTRACT

To address prevalent issues in medical imaging, such as data acquisition challenges and label availability, transfer learning from natural to medical image domains serves as a viable strategy to produce reliable segmentation results. However, several existing barriers between domains need to be broken down, including addressing contrast discrepancies, managing anatomical variability, and adapting 2D pretrained models for 3D segmentation tasks. In this paper, we propose ProMISe, a prompt-driven 3D medical image segmentation model using only a single point prompt to leverage knowledge from a pretrained 2D image foundation model. In particular, we use the pretrained vision transformer from the Segment Anything Model (SAM) and integrate lightweight adapters to extract depth-related (3D) spatial context without updating the pretrained weights. For robust results, a hybrid network with complementary encoders is designed, and a boundary-aware loss is proposed to achieve precise boundaries. We evaluate our model on two public datasets for colon and pancreas tumor segmentations, respectively. Compared to the state-of-the-art segmentation methods with and without prompt engineering, our proposed method achieves superior performance. The code is publicly available at https://github.com/MedICL-VU/ProMISe

***Index Terms*—** Medical image segmentation, lightweight adapter, transfer learning, prompt engineering, pretrained Segment Anything Model (SAM)

1. INTRODUCTION

Recently, image segmentation foundation models [1, 2] have revolutionized the field of image segmentation, demonstrating wide generalizability and impressive performance by training on massive amounts of data to learn general representations. Prompt engineering further improves the segmentation capability of these models. Given proper prompts as additional inputs, these models can handle various zero-shot tasks across domains and produce reliable segmentations during inference. Unlike these broad successes, medical image segmentation is often limited by issues such as expensive data acquisition and time-consuming annotation processing, resulting in a lack of massive public datasets available for training. Thus it is desirable to leverage transfer learning from the natural image domain for robust medical image segmentation [3].

However, directly leveraging pretrained 2D natural image foundation models for 3D medical image segmentation often leads to sub-optimal results [4]. This is primarily because: (1) medical images have their own unique contrast and texture characteristics; (2) anatomical differences among individuals make medical image segmentation challenging; and (3) slice-wise (2D) segmentation with transfer learning discards important depth-related spatial context in 3D medical data. Given these challenges, can we effectively adapt the pretrained models to achieve robust 3D medical segmentations?

In this paper, we propose ProMISe, **pro**mpt-driven 3D **m**edical **i**mage **se**gmentation using pretrained image foundation models (see Fig. 1). Specifically, ProMISe takes a 3D input image and a single point prompt as inputs, and uses image and prompt encoders to produce segmentation. Unlike most promptable models, a shallow convolutional neural network (CNN) is used as complementary path alongside the pretrained transformer image encoder [1], with adapters employed within the transformer to capture 3D depth context. During training, most weights of the adapted transformer encoder remain static; the other components in the proposed method are designed in a lightweight manner for efficiency and trained from scratch. We use a structural loss and a novel boundary-aware loss for precise decisions. The main novel contributions are:

- We propose a method for 3D medical segmentation that adapts pretrained image foundation models. Plug-and-play lightweight adapters are used to better optimize knowledge transfer across domains and more effectively capture fine-grained features. Our method is compatible with various pretrained image models, easy to implement, and cost-effective to train.
- We present a simple yet efficient boundary-aware loss for ambiguous edges. This ready-to-use loss can be seamlessly integrated into any training process without the need for offline edge map generation from ground truth.
- We validate the performance on two public datasets for challenging tumor segmentations. Our method outperforms state-of-the-art segmentation methods consistently.

Related works. Fully fine-tuning image foundation models for a task requires a large amount computational resources

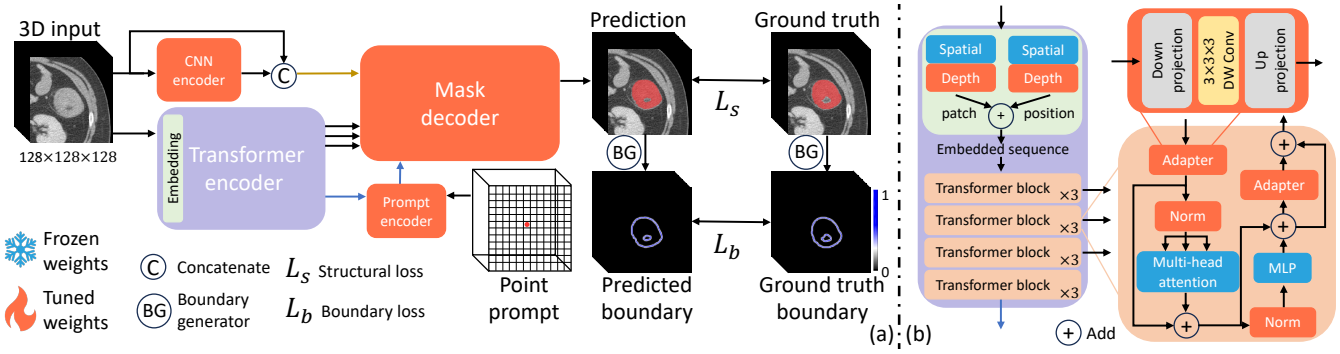

Fig. 1. The proposed framework (ProMISe) and details of transformer encoder are shown as (a) and (b), respectively.

and is not training-efficient. In contrast, partially fine-tuning [5] or introducing and training new shallow layers, such as lightweight adapters [6, 7, 8, 9, 10] and the Low-Rank Adaptation (LoRA) module [11, 12], have demonstrated robust performance as parameter-efficient fine-tuning methods. Recent works use SAM [1] for 3D medical image segmentation in a 2D slice-wise manner, which discard important depth-wise (3D) information and may require additional efforts to create prompts [5, 12]. Other models use adapters; this approach has proven effective for adapting a pretrained model from 2D images to 3D (2D+time) videos [6, 9], and it has subsequently been utilized in 3D medical image segmentation [10] with the use of adapters in the pretrained transformer block [7]. Although these models can segment 3D medical images, the image encoder still operates in a slice-wise (2D) manner with an additional branch for depth information. The weights for this branch that are replicated from the spatial branch demand more computational resources. In contrast, a holistic adaptation of SAM for 3D medical segmentation was proposed in [8], which avoids a depth branch by including an adaptor with depth-wise convolution [6]. However, a single adapter in each transformer block may not fully achieve accurate adaptation due to the notable discrepancies between natural and medical images. Moreover, this method struggles to adequately capture details and can lead to sub-optimal results, especially for tumor segmentation. These challenges and the critical importance of precise segmentation in medical applications motivate our proposed model as a more robust solution.

2. METHODS

Fig. 1 illustrates ProMISe, our proposed framework for 3D medical image segmentation, which employs prompt engineering and a pretrained image foundation model. Specifically, a 3D patch is taken as input and is fed through complementary CNN and transformer encoders. The prompt encoder utilizes the deepest feature from the transformer encoder (blue arrow in Fig. 1) as input together with the point prompt. Subsequently, all features, including the original input, are used to predict the segmentation mask via a lightweight CNN decoder. During training, the transformer encoder is partially tuned, while the rest are trained from scratch.

Image encoders. Our model is designed to effectively capture both global and local information using complementary transformer and CNN encoders, respectively.

For the transformer encoder (Fig. 1(b)), the input 3D image patch first passes through an embedding layer to create tokens with their positional information. Specifically, the pretrained weights from SAM [1] are employed for spatial patch embedding, and we introduced a trainable depth embedding layer for 3D data. The same approach is applied for positional encoding. Furthermore, we adapted the pretrained weights from SAM and fine-tuned the normalization layer in every transformer block. Unlike other works that employ a single lightweight adapter at the beginning of the transformer block [6, 8], an additional adapter is used before the output to optimize knowledge transfer across domains and further refine the image features. Notably, the two adapters employ depth-wise convolution to handle 3D images, and they are identical.

Inspired by the hybrid network design [13], a CNN encoder is used to capture detailed information to complement the transformer. This is particularly desirable for tumor segmentation, as the boundaries are often ambiguous. It is designed as a shallow network for efficiency (Fig. 2(a)).

Prompt encoder. We adapt the visual prompt encoder based on [2] (Fig. 3). Unlike the prompt encoder proposed in SAM [1], we incorporate image embeddings from the transformer encoder as an additional input. Point embeddings are derived from the given point prompt and image embedding using visual sampling (e.g., grid sampling) to ensure that their semantic features are aligned with image embeddings. Subsequently, the self-attention layer is applied to the point embeddings and learnable global queries. Afterwards, the image embeddings are applied to these queries via cross-attention. The output of the prompt encoder is fed to the mask decoder.

During training, 10 random points from background are provided for each input patch to increase the generlizability to noisy prompts. In contrast to previous work that utilized 40 points from target region as prompts [8], we randomly select

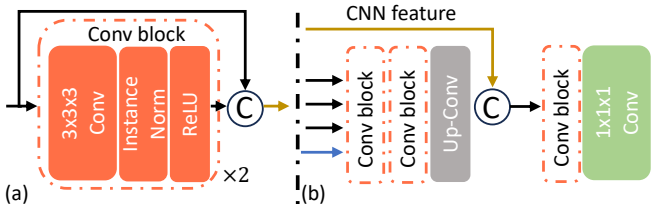

Fig. 2. The details of (a) CNN encoder, and (b) mask decoder.

10 point prompts during each iteration if the input patch contains foreground. For prompt engineering [14], our goal is a single click with minimal prior knowledge, but more prompts are supported if desired during inference.

Mask decoder. Instead of directly adapting the mask decoder from the foundation model in a 2D manner, we designed a shallow network to efficiently capture features in 3D and trained it from scratch (Fig. 2(b)). The multi-level features from the transformer encoder (Fig. 1(b)) are refined by two successive convolutional blocks. These are followed by a transposed convolution to ensure the features remain the same size. The fused features are processed through another convolutional block and a segmentation head for final results.

Boundary-aware loss. In medical image segmentation, accurately delineating the boundaries of objects is important, especially for irregularly shaped objects such as tumors [15]. Besides popular structural segmentation losses, such as the combined Dice loss and cross-entropy loss (denoted as $L_{structural}$), we further propose a boundary loss ($L_{boundary}$) to preserve fine details and produce robust segmentations. Moreover, by emphasizing edge accuracy, the model might generalize better to unseen data for tumor segmentation. As shown in Fig. 1, we extract a smooth boundary map rather than a binary boundary for a more robust representation, and because learning from a binary boundary is a challenging task. Specifically, we use average-pooling operation P_{ave} with kernel size 5 as boundary generator. Given either a logits map or a binary mask M, the smooth boundary is derived: $B(M) = |M - P_{ave}(M)|$. The total objective function is:

$$L(S, G) = \lambda_1 L_{structural}(S, G) + \lambda_2 L_{boundary}(B(S), B(G))$$

where S and G represent segmentation and ground truth. $L_{structural} = L_{Dice} + L_{CE}$ is used to capture the structural information and $L_{boundary} = L_{MSE}$ recovers the detailed contours. Unlike other methods [16] that require complicated offline computation of edge or distance maps to avoid iterative generation, our proposed ready-to-use boundary loss is computationally efficient and can be easily adapted to any segmentation task, and is independent of any augmentation.

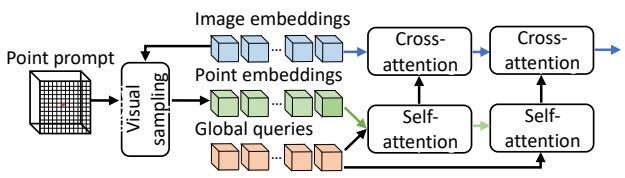

Fig. 3. The details of the proposed prompt encoder.

3. EXPERIMENTS

3.1. Experimental settings

Datasets. We evaluated our proposed method on two public datasets from the Medical Segmentation Decathlon (http://medicaldecathlon.com/) for challenging tumor tasks from pancreas and colon applications, where ambiguous edges are present. These consist of 281 ($0.61 \times 0.61 \times 0.7$ to $0.98 \times 0.98 \times 7.5mm^3$) and 126 ($0.54 \times 0.54 \times 1.25$ to $0.98 \times 0.98 \times 7.5mm^3$) 3D CT volumes, respectively. Following the setup from the prior study [8], we used the same data split for each task with a training/validation/testing split of 0.7/0.1/0.2 and only use tumor labels to focus on binary segmentation.

Preprocessing. We resample to $1mm$ isotropic resolution, intensity clip based on foreground 0.5 and 99.5 percentiles, and Z-score normalize based on all foreground voxels. Four data augmentations were used: random flip, rotation, zoom, and intensity shift. During training, an input patch of size $128 \times 128 \times 128$ was randomly selected such that its center pixel is equally likely to be foreground or background. Subsequently, each dimension was upsampled to 512.

Implementation details. We utilized pretrained ViT-B from SAM [1] as transformer encoder, and set $\lambda_1 : \lambda_2 = 1 : 10$ during training. The batch size was 1, and initial learning rate was 0.0004 with decreased amount $2e^{-}6$ every epoch. The AdamW optimizer was used with a maximum of 200 epochs. We used PyTorch, MONAI and an NVIDIA A6000 GPU for our experiments. The Dice score and normalized surface Dice (NSD) are used for evaluation. Compared state-of-the-art methods include: CNN (nnU-Net [17]), CNN with large kernel (3D UX-Net [18]), Swin-encoder with CNN decoder (Swin-unetr [20]), pure transformer (nnFormer [19]), and adaptation method with adapters (3DSAM-adapter [8]). We retrained using their official codes, and the pretrained weights are also employed if publicly available.

3.2. Results

Quantitative results. Tab. 1 presents a detailed comparison of results for colon and pancreas tumor segmentation. Notably, while CNN-based networks segment these tumors more effectively than transformers, prompt-driven methods outperform others when provided with only a single point in the entire volume. Our proposed method consistently outperforms all in terms of both Dice and boundary (NSD) metrics.

Ablation study. We also investigated the efficiency variations of the proposed ProMISe (Tab. 2). The use of two adapters and the boundary-aware loss mostly improved the

Table 1. Dice and normalized surface Dice (NSD) for colon and pancreas tumor. Bold indicates best performance. Significant improvements (2-tailed paired t-test, $p < 0.05$) are denoted via *. The promptable models use 1 point prompt per 3D volume.

Dataset	Metric	nnU-Net [17]	3D UX-Net [18]	nnFormer [19]	Swin-UNETR [20]	3DSAM-adapter [8]	ProMISe
Colon	Dice	45.60	23.07	21.36	37.23	57.32	**66.81***
	NSD	53.01	32.84	32.05	51.16	73.65	**81.24***
Pancreas	Dice	39.12	37.57	35.98	37.98	54.41	**57.46***
	NSD	57.66	55.25	53.45	56.42	77.88	**79.76***

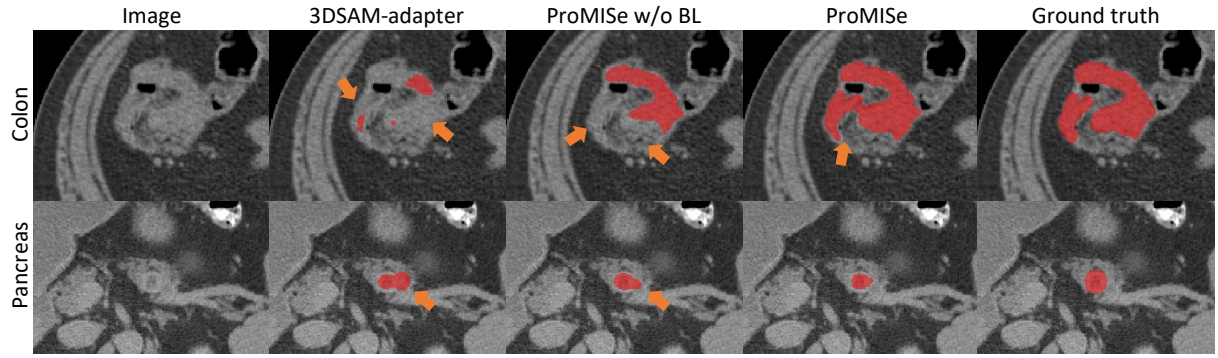

Fig. 4. Qualitative results. BL denotes boundary-aware loss. The major differences are highlighted by orange arrows.

Table 2. Quantitative results of ablation study with single point prompt unless noted. R and C represent residual and concatenate fusions, and B indicates boundary loss. + shows the cumulative variants. Best viewed by individual sections.

	Colon		Pancreas	
Method	Dice	NSD	Dice	NSD
baseline [8]	57.32	73.65	54.41	77.88
+ two adapters	61.61	73.88	56.08	77.89
+ up-Conv	62.92	77.62	55.37	77.38
ProMISe-R	63.67	79.96	55.15	79.02
ProMISe-R-B	64.75	79.77	56.57	79.46
ProMISe-C	64.76	77.59	56.35	78.01
ProMISe-C-B (**proposed**)	66.81	81.24	57.46	79.76
baseline [8] (10 prompts)	63.09	79.97	55.94	79.18
ProMISe-C-B (10 prompts)	67.28	81.63	58.05	80.36

results. Interestingly, switching from trilinear upsampling to up-convolution improved the performance for the colon, but showed a decline for the pancreas. This implies that trilinear upsampling may be more appropriate for pancreas tumors, which are typically round in shape. Using concatenation (-C) in the CNN encoder offers better Dice scores than residual connections (-R), though the latter improves surface quality more. While the performance of ProMISe improves with 10 prompts, the improvement is limited over a single prompt. Furthermore, it is challenging to identify the tumor area due to ambiguous boundaries, making the use of a single click preferable in practice, as it requires less expert knowledge.

Qualitative results. Fig. 4 shows qualitative visualizations from top-performing promptable methods. ProMISe yields results that closely align with the ground truth. 3DSAM-adapter [8] fails to detect certain regions that ProMISe captures, even without the boundary-aware loss. This indicates the improved generalizability of the model through our proposed modifications. Moreover, the use of the boundary-aware loss yields robust segmentations, alleviating issues of both under-segmentation for colon and over-segmentation for pancreas tumors, respectively. Notably, the boundary-aware loss improves segmentation not just for the irregularly shaped colon tumors but also for the pancreas tumors, which typically have a more regular, rounded shape. However, slight under-segmented areas are found in pancreas segmentation.

4. CONCLUSION

In this paper, we propose a promptable network, named ProMISe, designed for robust 3D tumor segmentation using pretrained weights from image foundation models. We evaluate on two public datasets, where our model consistently outperforms state-of-the-art methods across all tasks. Moreover, the critical role of the two adapters and boundary-aware loss techniques are demonstrated. Future work will aim to improve the efficiency through knowledge distillation and further explore different point sampling strategies.

Acknowledgments. This work was supported, in part, by NIH U01-NS106845, and NSF grant 2220401.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by MSD. Ethical approval was not required as confirmed by the license attached with the open access data.

6. REFERENCES

[1] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al., "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[2] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee, "Segment everything everywhere all at once," *arXiv preprint arXiv:2304.06718*, 2023.

[3] Christos Matsoukas, Johan Fredin Haslum, Moein Sorkhei, Magnus Söderberg, and Kevin Smith, "What makes transfer learning work for medical images: Feature reuse & other factors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9225–9234.

[4] Sheng He, Rina Bao, Jingpeng Li, Jeffrey Stout, Atle Bjornerud, P. Ellen Grant, and Yangming Ou, "Computer-vision benchmark segment-anything model (sam) in medical images: Accuracy in 12 datasets," 2023.

[5] Jun Ma and Bo Wang, "Segment anything in medical images," *arXiv preprint arXiv:2304.12306*, 2023.

[6] Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li, "St-adapter: Parameter-efficient image-to-video transfer learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 26462–26477, 2022.

[7] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo, "Adaptformer: Adapting vision transformers for scalable visual recognition," *Advances in Neural Information Processing Systems*, vol. 35, pp. 16664–16678, 2022.

[8] Shizhan Gong, Yuan Zhong, Wenao Ma, Jinpeng Li, Zhao Wang, Jingyang Zhang, Pheng-Ann Heng, and Qi Dou, "3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable medical image segmentation," *arXiv preprint arXiv:2306.13465*, 2023.

[9] Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li, "Aim: Adapting image models for efficient video action recognition," *arXiv preprint arXiv:2302.03024*, 2023.

[10] Junde Wu, Rao Fu, Huihui Fang, Yuanpei Liu, Zhaowei Wang, Yanwu Xu, Yueming Jin, and Tal Arbel, "Medical sam adapter: Adapting segment anything model for medical image segmentation," *arXiv preprint arXiv:2304.12620*, 2023.

[11] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[12] Kaidong Zhang and Dong Liu, "Customized segment anything model for medical image segmentation," *arXiv preprint arXiv:2304.13785*, 2023.

[13] Hao Li, Dewei Hu, Han Liu, Jiacheng Wang, and Ipek Oguz, "Cats: Complementary cnn and transformer encoders for segmentation," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2022, pp. 1–5.

[14] Hao Li, Han Liu, Dewei Hu, Jiacheng Wang, and Ipek Oguz, "Assessing test-time variability for interactive 3d medical image segmentation with diverse point prompts," 2023.

[15] Han Liu, Dewei Hu, Hao Li, and Ipek Oguz, "Medical image segmentation using deep learning," in *Machine Learning for Brain Disorders*, pp. 391–434. Springer, 2023.

[16] Hoel Kervadec, Jihene Bouchtiba, Christian Desrosiers, Eric Granger, Jose Dolz, and Ismail Ben Ayed, "Boundary loss for highly unbalanced segmentation," in *International conference on medical imaging with deep learning*. PMLR, 2019, pp. 285–296.

[17] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[18] Ho Hin Lee, Shunxing Bao, Yuankai Huo, and Bennett A. Landman, "3d UX-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation," in *The Eleventh International Conference on Learning Representations*, 2023.

[19] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu, "nnformer: Interleaved transformer for volumetric segmentation," 2022.

[20] Yucheng Tang, Dong Yang, Wenqi Li, Holger Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh, "Self-supervised pre-training of swin transformers for 3d medical image analysis," 2022.