Towards Universal Segmentation for Medical Images with Text Prompts

Ziheng Zhao[1,2], Yao Zhang[2,*], Chaoyi Wu[1,2], Xiaoman Zhang[1,2], Ya Zhang[1,2], Yanfeng Wang[1,2,†], and Weidi Xie[1,2,†]

[1] Shanghai Jiao Tong University, Shanghai, China
[2] Shanghai AI Laboratory, Shanghai, China

Abstract. In this study, we advance the development of medical segmentation by proposing the **first knowledge-enhanced universal segmentation model for 3D medical images driven by text prompts**, termed **SAT** (**S**egment **A**nything in radiology scans, driven by **T**ext prompts). Build on abundant anatomical knowledge and unprecedented segmentation data, it can handle numerous segmentation tasks efficiently with text prompts, in just one model, holding significant potential as a foundation segmentation model, and an agent for large language models.

1 Introduction

Medical image segmentation plays critical roles in clinical scenarios, providing supports for procedures such as diagnosis, therapy planning, disease monitoring. Despite the success achieved by deep learning based segmentation methods, they typically require training and deploying specialist models on each dataset, failing to handle heterogeneous segmentation tasks effectively. Meanwhile, the recent success of large language models (LLM) in medicine necessitates a tailored segmentation tool that can bridge language and visual grounding capabilities.

In this work, we propose SAT, and make the following contributions: *First,* we construct the first multimodal knowledge tree on human anatomy, including over 6K anatomical concepts; And build the most comprehensive, large-scale dataset for 3D medical segmentation from 72 public datasets, termed as **SAT-DS**. It contains over 22K image scans, 302K segmentation annotations, 497 categories from 8 human body regions in 3 modalities (CT, MR and PET). We invest significant efforts in standardizing datasets and unifying annotation labels; *Second*, we propose to inject knowledge into a text encoder via contrastive learning, and then develop a universal segmentation model prompted by medical terminologies in text form; *Third,* we train two model variants: SAT-Pro (447M parameters) and SAT-Nano (110M parameters), conduct comprehensive evaluation to demonstrate the significance of SAT from many aspects.

2 Results

SAT is comparable to specialists. We compare SAT with 72 nnU-Net models trained on each dataset, and demonstrate comparable performance, shown

Table 1. Comparison of nnU-Nets, SAT and SAT-Pro-FT. Results are merged and presented by human body regions and lesions. H&N: head and neck, LL: lower limb, UL: upper limb, WB: whole body, All: average over all the 497 classes.

Metric	Method	Brain	H&N	UL	Thorax	Spine	Abdomen	LL	Pelvis	WB	Lesion	All
DSC↑	nnU-Nets	**81.93**	72.45	89.20	85.60	81.62	**86.96**	82.89	84.29	84.20	**57.52**	80.54
	SAT-Nano	73.90	74.08	89.58	82.09	75.55	77.12	81.79	79.50	77.80	46.04	75.51
	SAT-Pro	77.70	77.29	91.34	86.38	74.73	81.47	84.01	83.01	82.19	51.39	78.81
	SAT-Pro-Ft	78.80	**78.36**	**93.78**	**89.43**	**82.48**	84.11	**87.12**	**86.84**	**84.24**	53.71	**81.41**
NSD↑	nnU-Nets	**81.96**	74.51	86.04	82.66	77.47	**79.54**	80.20	76.41	79.30	**49.89**	78.15
	SAT-Nano	72.95	79.15	90.94	79.68	73.68	67.83	78.42	74.89	75.36	38.17	73.76
	SAT-Pro	77.77	82.45	93.56	85.15	72.87	72.94	82.50	78.66	79.61	44.87	77.82
	SAT-Pro-Ft	80.43	**84.01**	**95.64**	**86.38**	**82.37**	78.42	**87.25**	**82.92**	**82.48**	47.74	**81.60**

in Tab. 1. Meanwhile on model parameters, SAT-Pro (447M) and SAT-Nano (110M) is significantly smaller than 72 nnU-Nets (around 2.2B). This suggest SAT is capable of handling a wide range of tasks efficiently with just one model.

SAT is a foundation segmentation model. We fine-tune SAT-Pro on each dataset, termed as **SAT-Pro-Ft**, and observe significant performance improvement over SAT-Pro, even surpassing the counterpart nnU-Nets. This reveals SAT as a foundation segmentation model pre-trained on large-scale dataset, that can be further adapted to specific datasets for optimized performance.

Text prompt V.S. spatial prompt. On 32 out of 72 datasets, we compare SAT with MedSAM, which is prompted under the minimal rectangle covering the ground-truth segmentation (Oracle Box). As shown in Tab. 2, SAT-Pro consistently outperforms MedSAM on almost all regions, achieving superior segmentation performance with text prompts. While MedSAM exceeds on lesions, the result of oracle box baseline on lesions implies this is primarily due to the strong prior provided by the tight box prompt.

Table 2. Comparison of SAT and MedSAM. Results are merged and presented by human body regions and lesions. H&N: head and neck, All: average over all the classes. Note that SAT is fully automatic methods, while MedSAM is interactive.

Metric	Method	Brain	H&N	Thorax	Spine	Abdomen	Limb	Pelvis	Lesion	All
DSC↑	SAT-Pro	**73.23**	**84.16**	**85.72**	**85.52**	**82.23**	**82.56**	**86.35**	55.69	**82.47**
	SAT-Nano	69.05	77.93	79.06	81.88	78.43	77.6	78.31	47.7	76.66
	Oracle Box	55.84	85.32	72.85	78.11	71.92	78.96	81.3	**67.94**	63.57
	MedSAM	54.35	78.48	73.01	79.09	77.35	80.53	84.42	65.85	75.39
NSD↑	SAT-Pro	**80.46**	**86.12**	**85.87**	**87.01**	**79.98**	**83.53**	**82.39**	49.35	**81.79**
	SAT-Nano	75.62	79.44	77.74	82.7	74.96	77.58	72.98	40.38	74.82
	Oracle Box	60.55	86.83	66.16	69.86	59.51	70.67	71.59	61.59	48.93
	MedSAM	68.62	82.37	68.26	73.29	73.39	73.84	74.9	**62.28**	70.99

A segmentation agent for LLMs. With text as bridge, SAT can be a powerful out-of-the-box agent for any large language models, seamlessly providing visual grounding ability in critical scenarios, *e.g.,* grounded report generation.