Anytime, Anywhere, Anyone: Investigating the Feasibility of Segment Anything Model for Crowdsourcing Medical Image Annotations

Pranav Kulkarni*, Adway Kanhere*, Dharmam Savani*, Andrew Chan, Devina Chatterjee, Paul Yi, Vishwa S. Parekh

University of Maryland School of Medicine, Baltimore, MD, USA

** Authors contributed equally to this work*

Can **Segment Anything Model** crowdsource annotations for medical image segmentation?

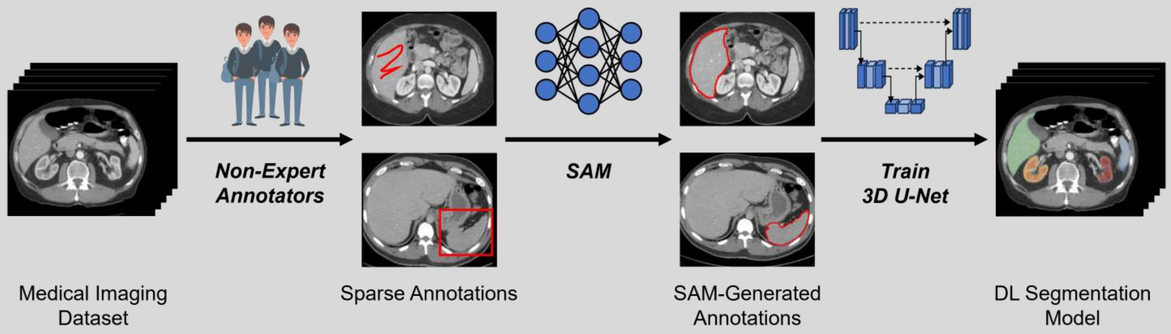

Figure 1. Pipeline for crowdsourcing sparse annotations for objects of interest (e.g., organs, tumors, etc.) from non-expert annotators for the purpose of training 3D DL segmentation models (e.g., U-Net) using SAM-generated annotations.

Introduction

- Medical image annotation for deep learning (DL) is a labor-intensive task[1].
- There is a critical need for an annotation process to enable non-experts to annotate datasets with sparse annotations without the need for an expert in the loop.
- The Segment Anything Model (SAM) has revolutionized segmentation with strong zero-shot generalizability and holds a lot promise for annotating datasets (**Fig. 2**)[2,3].
- We evaluated SAM for crowd-sourcing medical image annotations from non-experts and evaluated SAM-generated annotations for training 3D DL segmentation models.

Methods

- We used the BTCV dataset of $n = 30$ abdominal CT scans with annotations for 13 organs[4]. The dataset was split into train and test sets ($n = 15$, both).
- We included five organs of interest: aorta, left and right kidneys, liver, and spleen.
- We used the OpenLabeling tool to annotate the BTCV train set. Each slice was annotated by four non-experts using bounding boxes.
- We used SAM ViT-Huge to generate masks for the organs. Each slice was passed to SAM with its corresponding boxes. The SAM-generated annotations were converted to NIfTI.
- We measured the mean slice and volume Dice score of SAM-generated annotations on the ground-truth annotations of train set.
- We trained nnU-Net[5] models, a self-configuring SOTA 3D U-Net, on the SAM-generated ("SAM-nnU-Net") and ground-truth ("GT-nnU-Net") annotations.
- We compared the mean volume Dice scores on the ground-truth BTCV test set using Wilcoxon signed-rank tests. Statistical significance was defined as $p < 0.05$.

Results

- The non-experts annotated 651 slices with 1,840 bounding boxes (**Fig. 3**). They took 55.60 ± 8.76 mins (mean 3.29 ± 1.04 secs per slice) to annotate an organ.
- We excluded volumes with missing annotations ($n = 4$).
- SAM-generated annotations have high slice Dice scores but low volume Dice scores (**Table 1**, left). Furthermore, the SAM-nnU-Net model performs significantly worse than the GT-nnU-Net model (**Table 1**, right; **Fig. 4**).

Table 1. Mean slice and volume Dice scores of SAM-generated annotations on the BTCV train set (**left**). Mean volume Dice scores of the GT-nnU-Net and SAM-nnU-Net on the BTCV test set (**right**).

Organ	Slice Dice	Volume Dice	GT-nnU-Net	SAM-nnU-Net	p-value
Mean	0.88 ± 0.02	0.75 ± 0.09	0.90 ± 0.05	0.80 ± 0.05	< 0.001
Aorta	0.88 ± 0.01	0.70 ± 0.09	0.92 ± 0.01	0.78 ± 0.04	< 0.001
Left Kidney	0.88 ± 0.04	0.74 ± 0.11	0.87 ± 0.12	0.78 ± 0.08	0.02
Right Kidney	0.90 ± 0.02	0.78 ± 0.10	0.87 ± 0.11	0.78 ± 0.07	0.06
Liver	0.84 ± 0.02	0.73 ± 0.13	0.94 ± 0.05	0.84 ± 0.04	< 0.001
Spleen	0.91 ± 0.03	0.80 ± 0.14	0.88 ± 0.11	0.80 ± 0.11	< 0.001

Discussion

- SAM lacks spacial relationships since it is designed for 2D segmentation. This can be addressed by adapting SAM for 3D medical image segmentation[6].
- There is potential for unreliable annotations from non-experts and quality assessment is critical for filtering them out without manual intervention from an expert.
- While we may not be ready for non-expert annotations yet, they have a potential for streamlining medical image annotation.

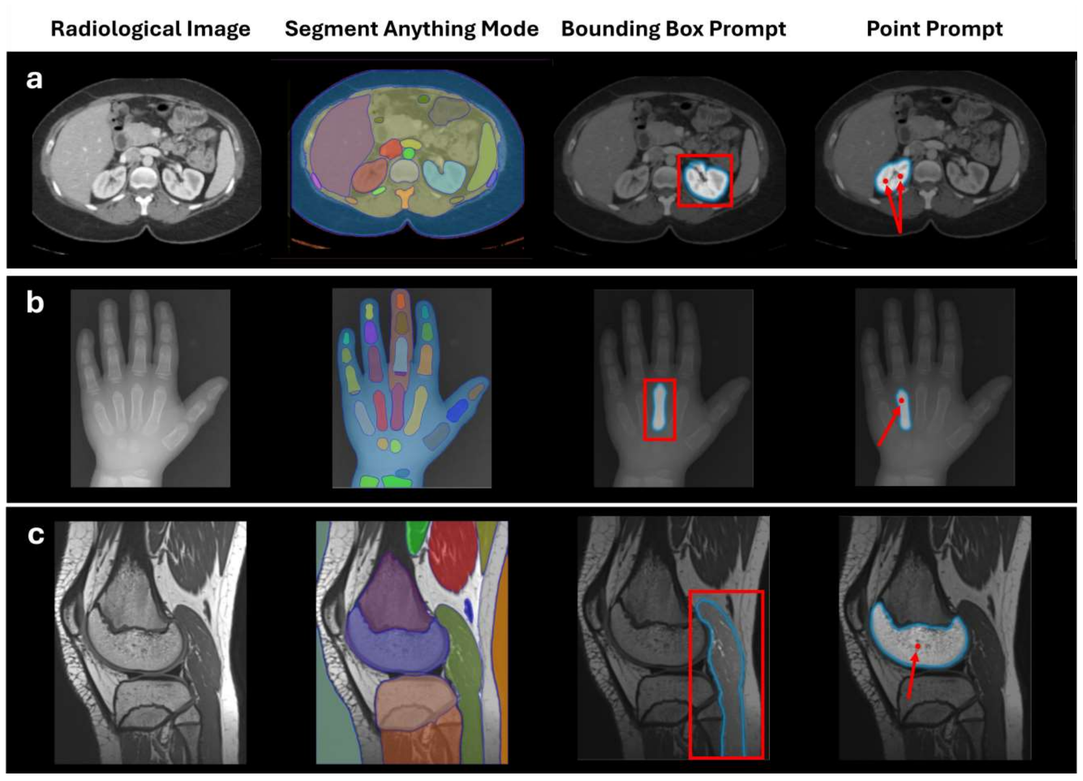

Figure 2. Example of SAM on **(a)** an abdominal CT, **(b)** a hand x-ray, and **(c)** a knee MRI. SAM can operate in either "segment anything" mode (column 2) or "prompting" mode (columns 3,4).

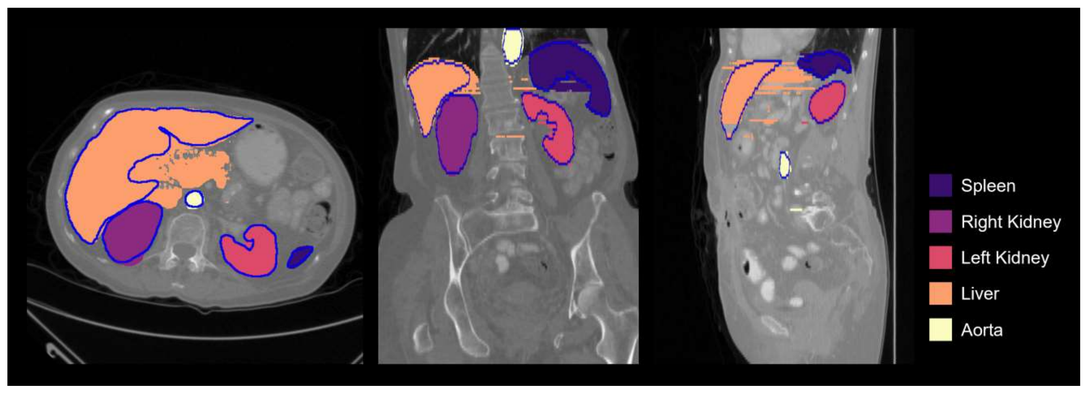

Figure 3. Example of crowdsourced SAM-generated annotations from the BTCV train set in the axial, coronal, and sagittal views. The ground-truth annotations are outlined in blue.

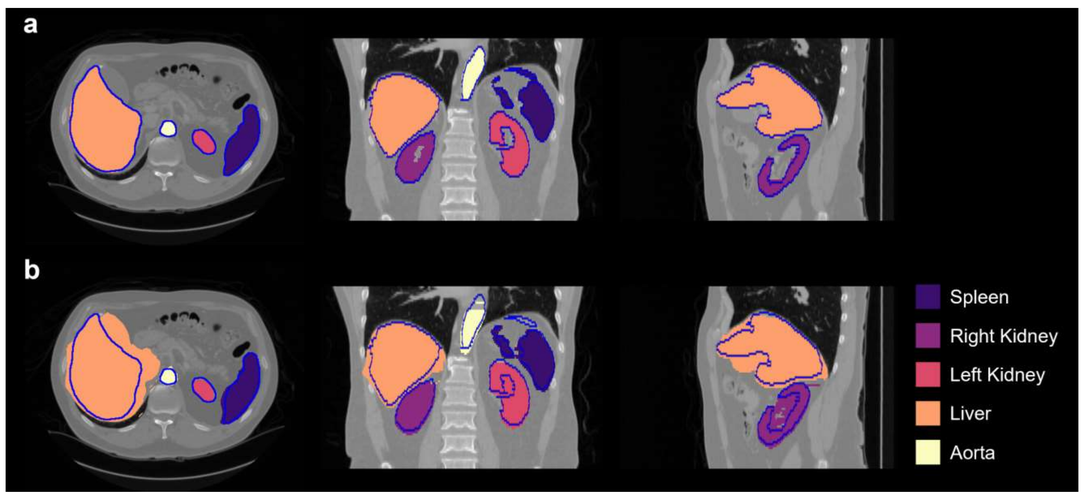

Figure 4. Example of **(a)** GT-nnU-Net and **(b)** SAM-nnU-Net segmentation from the BTCV test set in the axial, coronal, and sagittal views. The ground-truth annotations are outlined in blue.

References

1. Diaz-Pinto, A., ... & Cardoso, M. J. (2024). Monai label: A framework for ai-assisted interactive labeling of 3d medical images. *Medical Image Analysis*, 95:103207.
2. Kirillov, A., ... & Girshick, R. (2023). Segment anything. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015-4026.
3. Ma, J., ... & Wang, B. (2024). Segment anything in medical images. *Nature Communications*, 15(1):654.
4. Landman, B., ... & Klein, A. (2015). Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. *Proceedings of MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, 5:12.
5. Isensee, F., ... (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203-211.
6. Bui, N. T., ... & Le, N. (2023). Sam3d: Segment anything model in volumetric medical images. *arXiv preprint arXiv:2309.03493*.

Pranav Kulkarni
University of Maryland School of Medicine
pkulkarni@som.umaryland.edu
@itspranavk
itspranavk

Read the full paper!