SIIM CAIMI25
Conference on Artificial Intelligence in Medical Imaging
OCT 20-21 | UNIVERSITY OF CALIFORNIA, IRVINE

Promptable 3D Lesion Segmentation in CT: Converting 2D PACS Annotations into 3D Training Datasets

Samuel D. Church, MS, University of Wisconsin – Madison; Joshua D. Warner, MD, PhD, CIIP; Danyal Maqbool, MS; Xin Tie, PhD; Junjie Hu, PhD; Meghan G. Lubner, MD, FSAR, FACR; Tyler J. Bradshaw, PhD

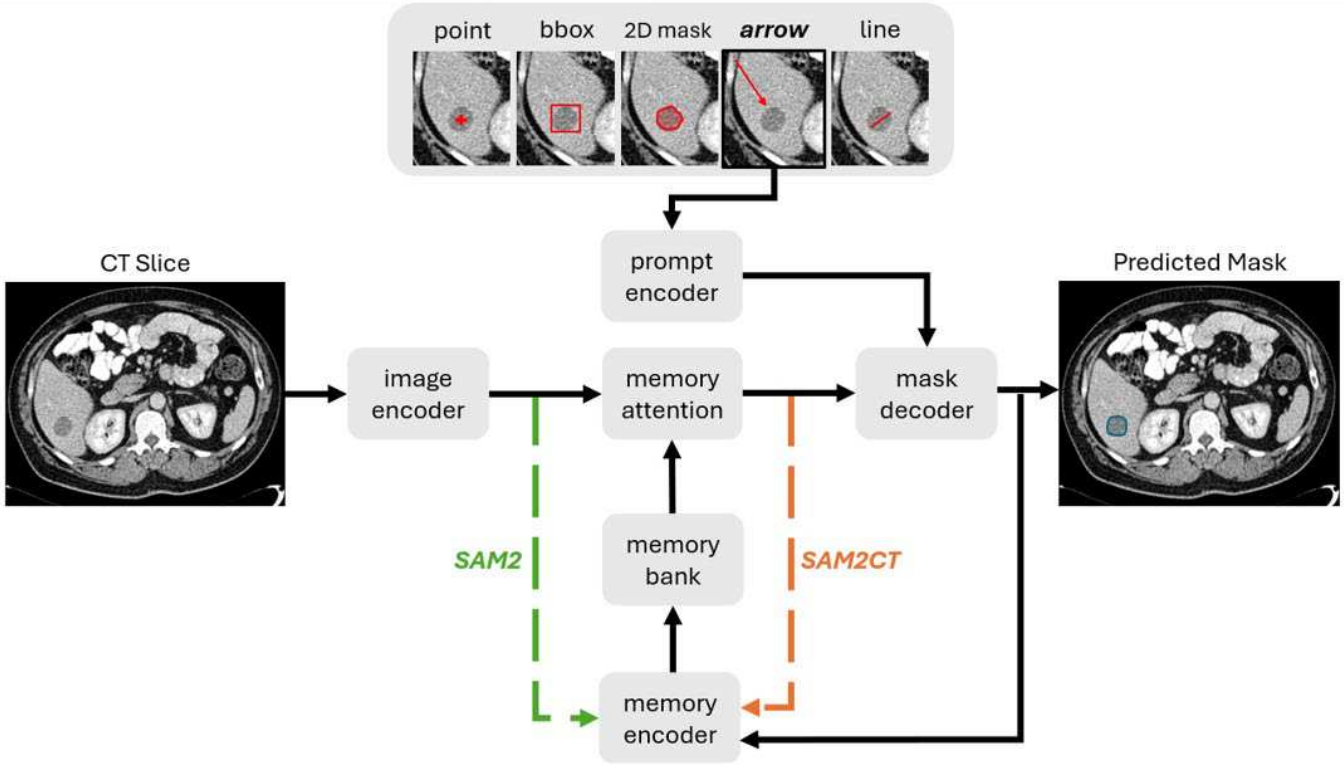
Introduction/Background
Promptable medical segmentation models can generate 3D contours from simple 2D prompts (e.g., bounding boxes). However, current approaches have ignored the most common prompts used by radiologists when annotating clinical images in PACS, such as arrows and lines, limiting retrospective construction of large AI training datasets. We developed and evaluated promptable segmentation models that leverage DICOM GSPS annotations, including arrows and lines, to generate 3D lesion contours from oncologic CT scans.

Methods/Intervention
We developed promptable segmentation models using 2,304 lesion contours from public CT datasets, which included lesions of the kidney, liver, colon, lung, pancreas, and abdominal/mediastinal lymph nodes. Four models were trained to convert 2D prompts (arrows, lines, boxes, points, 2D masks) into 3D lesion segmentations. For training, prompts were synthesized from ground truth segmentations. Swin UNETR and DynUNet were trained from scratch, receiving the prompt as an additional binary input channel. We adapted SAM2, a promptable video segmentation model, to support line and arrow inputs by extending its prompt encoder and fine-tuned it on the CT dataset. We introduced SAM2CT, a variant of SAM2 that creates memory tokens using memory-attended image embeddings instead of unconditioned image embeddings. All models were evaluated on a hold-out test set of 193 lesions. Additionally, two board-certified radiologists scored 60 SAM2CT-generated soft tissue lesion contours from randomly selected real-world PACS annotations (20 arrows, 20 lines, 20 major-minor axis lines).

Results/Outcome
SAM2CT achieved the highest overall Dice score (70.1), outperforming fine-tuned SAM2 (67.5), DynUNet (62.1), and Swin UNETR (59.9) across all lesion and prompt types. SAM2CT achieved average Dice scores of 63.5 for arrow prompts and 70.1 for line prompts. Among the 60 radiologist-reviewed contours generated by SAM2CT, 25% needed minor adjustments, and only 13% needed major corrections.

Conclusion
Promptable medical segmentation models can be adapted to handle prompt types used in clinical workflows, including lines and arrows. Our proposed SAM2CT architecture had the best performance. This work can facilitate the automatic curation of large-scale, annotated 3D datasets from PACS.

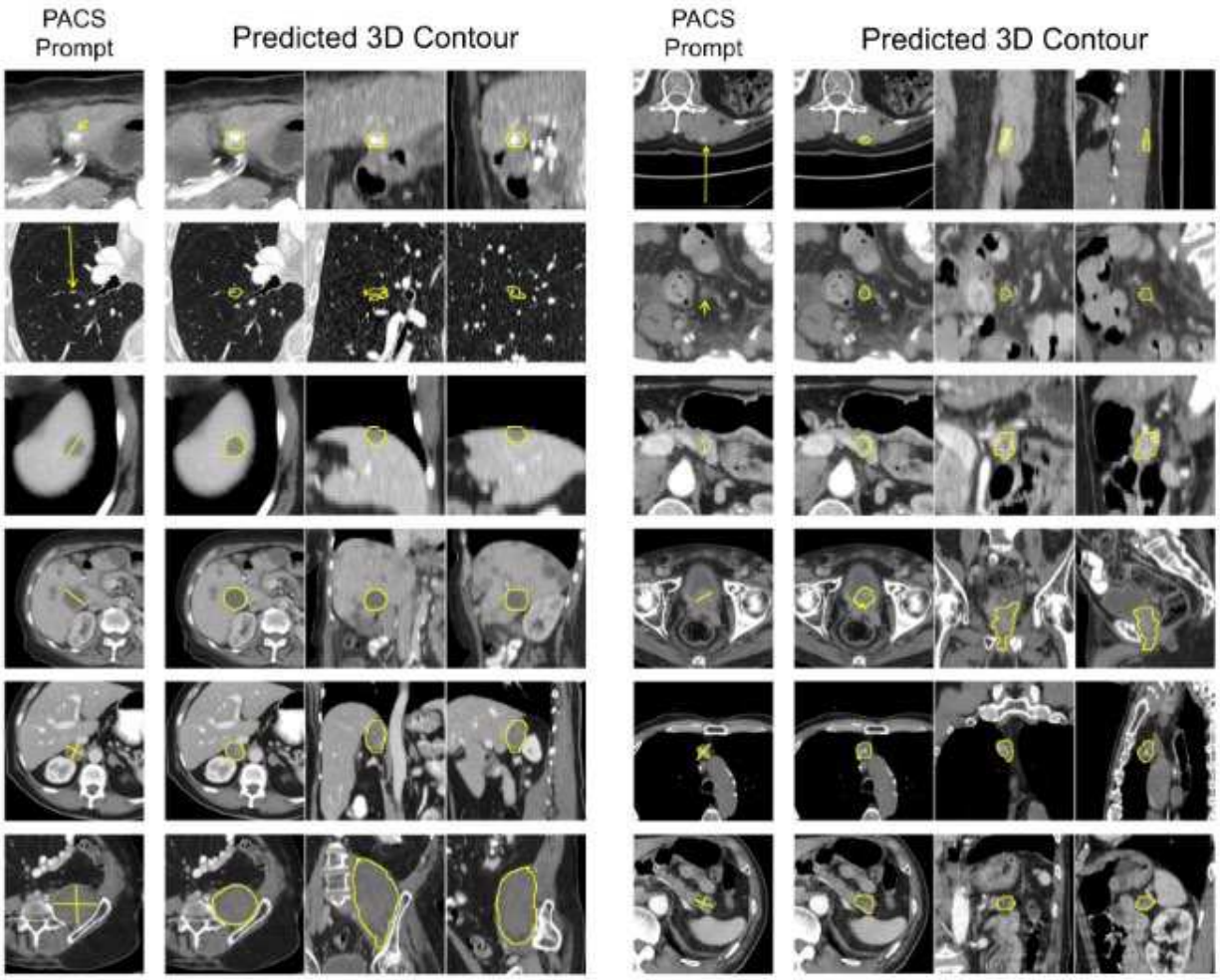
Statement of Impact
Our method enables automatic generation of 3D CT lesion contours directly from existing GSPS annotations within PACS, facilitating large-scale dataset creation without additional physician effort.



Model architecture for SAM2 and SAM2CT, both of which operate on a slice-by-slice basis. Each slice's result is encoded and stored in a memory bank for use in processing subsequent slices. In the illustrated workflow, SAM2 passes the output of the image encoder directly into the memory encoder (green arrow). In contrast, SAM2CT first applies memory attention to the image features and then feeds the attended output into the memory encoder (orange arrow). At the top, the supported prompt types are shown, including standard SAM2 prompts (point, bounding box, and 2D mask) as well as newly added prompts (arrow and line).

Dataset (train/val/test)	Model	Arrow	Line	Point	BBox	Mask	All Prompts
KiTS23 (217/33/24)	DynUNet	60.5	76.4	54.9	78.4	77.3	69.5
	Swin UNETR	41.3	74.7	36.8	75.1	77.9	61.2
	SAM2 (FT)	**72.2**	84.1	**78.4**	85.6	87.1	81.5
	SAM2CT	71.6	**85.0**	77.6	**86.4**	**88.6**	**81.8**
LiTS17 (495/63/54)	DynUNet	47.2	63.4	45.2	68.9	71.1	59.2
	Swin UNETR	41.8	67.5	39.4	69.4	76.4	58.9
	SAM2 (FT)	61.3	66.4	63.2	73.0	79.2	68.6
	SAM2CT	**63.3**	**69.1**	**66.5**	**76.2**	**82.2**	**71.5**
MSD-Colon (79/11/10)	DynUNet	37.1	61.0	37.5	57.4	60.7	50.7
	Swin UNETR	57.9	53.4	54.9	59.8	64.4	58.1
	SAM2 (FT)	59.5	64.8	58.6	65.8	73.5	64.4
	SAM2CT	**65.3**	**63.6**	**61.1**	**71.1**	**76.5**	**67.5**
MSD-Lung (37/16/10)	DynUNet	47.8	69.3	56.8	**71.3**	74.6	64.0
	Swin UNETR	39.0	68.3	41.7	66.9	73.1	57.8
	SAM2 (FT)	49.3	66.2	**61.9**	69.8	73.5	64.1
	SAM2CT	**50.2**	**70.6**	60.7	71.1	**77.4**	**66.0**
MSD-Pancreas (177/28/19)	DynUNet	64.3	75.8	63.4	74.9	80.5	71.8
	Swin UNETR	61.1	71.1	54.0	74.2	79.1	67.9
	SAM2 (FT)	70.6	75.0	66.5	76.4	78.3	73.4
	SAM2CT	**73.7**	**76.6**	**68.2**	**80.2**	**82.2**	**76.2**
NIH-ABD (389/145/43)	DynUNet	60.3	64.6	57.1	66.7	69.6	63.7
	Swin UNETR	58.1	61.3	48.6	65.3	68.0	60.3
	SAM2 (FT)	60.0	62.2	56.9	63.4	64.1	61.3
	SAM2CT	**63.8**	**66.9**	**60.2**	**70.4**	**71.2**	**66.5**
NIH-MED (302/119/33)	DynUNet	47.6	**61.0**	42.8	63.1	63.4	55.6
	Swin UNETR	52.0	56.8	46.5	59.3	61.6	55.2
	SAM2 (FT)	56.4	60.4	51.8	62.0	65.8	59.3
	SAM2CT	**56.9**	59.2	**53.9**	**64.9**	**69.9**	**61.0**
All Lesions	DynUNet	52.1	67.4	51.1	68.7	71.0	62.1
	Swin UNETR	50.2	64.7	46.0	67.1	71.5	59.9
	SAM2 (FT)	61.3	68.4	62.5	70.9	74.5	67.5
	SAM2CT	**63.5**	**70.1**	**64.0**	**74.3**	**78.3**	**70.1**

3D Dice scores on the test set for all four models. Results are reported both at the dataset level and by prompt type. Dataset abbreviations: KiTS23 – Kidney Tumors, LiTS17 – Liver Tumors, NIH-ABD – Abdominal Lymph Nodes, NIH-MED – Mediastinal Lymph Nodes.



Example SAM2CT predictions on real PACS images with GSPS annotations. The two “PACS Prompt” columns show the original physician-provided annotation. The “Predicted 3D Contour” columns display SAM2CT predicted 3D lesion contours across three views: transaxial (left), coronal (middle), and sagittal (right).

Keywords

Deep Learning; Segmentation; CT; Automated Labeling