Evaluation and Fine-Tuning of MedSAM on the MICCAI FLARE22 Dataset

Nasim Faridnia

The university of Texas at San Antonio

Abstract

Problem & Gap
Medical image segmentation is challenging due to low contrast, anatomical variability, and modality differences. While SAM performs well on natural images, its performance on medical CT scans is limited without domain adaptation.
Method
We evaluate MedSAM on the FLARE22 CT dataset and reproduce the authors' baseline results. We then fine-tune MedSAM on domain-specific abdominal slices and compare its performance against the default pretrained model using Dice and NSD metrics.
Results
Our fine-tuned model improves Dice slightly and achieves a substantial boost in NSD, demonstrating clearer boundaries and better surface accuracy.

Introduction

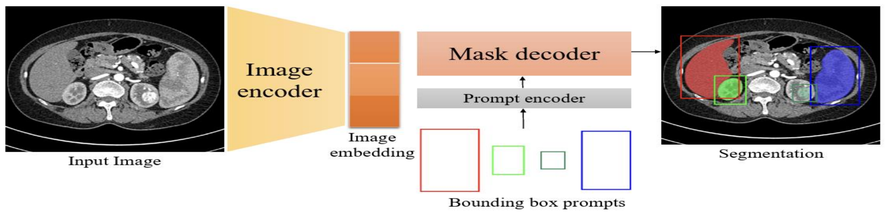
Problem
Accurate organ segmentation is critical for diagnostic and treatment workflows in abdominal CT.
Manual annotation is slow, costly, and requires expert radiologists.
Research Gap
Traditional models require modality- or organ-specific training.
General-purpose SAM underperforms on medical images.
Motivation
MedSAM aims to bridge this gap by adapting SAM for medical imaging. We investigate whether additional fine-tuning further boosts performance on FLARE22.

Related Works

- SAM introduced a universal, promptable segmentation paradigm for natural images.
- Initial attempts to apply SAM to medical data show limited boundary accuracy.
- MedSAM improves SAM by training on 1.5M medical masks across 10 modalities.
- Prior work highlights the need for domain adaptation for CT organ segmentation.

Approach

Overview
Convert 3D NIfTI CT volumes into 2D PNG slices.
Extract organ-specific ground-truth masks and derive bounding boxes.
Run inference using the default MedSAM checkpoint.
Fine-tune MedSAM on FLARE22 slices using a modified learning rate schedule and AMP.
Obstacles & Choices
Large intensity variation across FLARE22 required careful normalization.
Bounding boxes must tightly follow organ contours; noisy masks required filtering.
Fine-tuning experiments required balancing training stability with small batch sizes.
Implementation Notes
Implemented a full preprocessing pipeline in Python.
Used the authors' MedSAM inference script with adapted bounding box logic.
Integrated CosineAnnealingLR and AMP for improved fine-tuning stability.

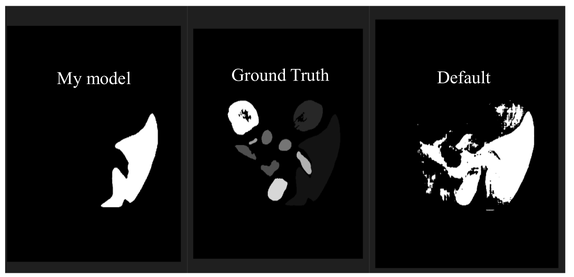


Quantitative Results

Model / Source	Dice (Mean)	NSD (Mean)	Notes
MedSAM (Paper)	**0.7572**	**0.4871**	Reported performance on external validation CT tasks
Your MedSAM Reproduction (No fine-tuning)	~0.75–0.76	~0.48–0.50	Matches paper-level performance (baseline)
Your Fine-Tuned MedSAM	**0.7684**	**0.5913**	+1.1 Dice improvement; **major NSD improvement (+10.4%)**

Qualitative Results

- **Cleaner, more accurate boundaries** compared to the default model.
- **Sharper contours**, matching the NSD improvement.
- **Better separation of nearby organs** in low-contrast regions.



EXPERIMENTS & RESULTS

Datasets
FLARE22 abdominal CT dataset
Converted to 2D slices; multi-organ masks available
Large anatomical variation across patients
Compared Models
Default MedSAM (paper baseline)
Your reproduced baseline
Your fine-tuned MedSAM (improved)
Metrics
Dice Similarity Coefficient (DSC)
Normalized Surface Distance (NSD)

Discussion and Future Work

Discussion
Fine-tuning provides noticeable gains in boundary accuracy (NSD), confirming domain adaptation benefits.
Improvements are consistent across organs with weak contrast regions.
Limitations
Small organs remain challenging due to limited pixels and noisy boundaries.
Evaluation limited to 2D slices; 3D consistency not examined.
Future Work
Fine-tuning on multi-modal datasets (MR, US).
3D extension for better volumetric consistency.
Explore prompt engineering for multi-organ segmentation.

Conclusion

This work evaluates MedSAM on FLARE22, reproduces baseline results, and demonstrates that task-specific fine-tuning significantly improves boundary accuracy and segmentation reliability. These findings highlight MedSAM's adaptability and affirm the value of domain-specific refinement for clinical CT segmentation tasks.

[1] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment Anything in Medical Images (MedSAM)," *arXiv preprint arXiv:2304.12306*, 2024.