MobileSeg3D: A Lightweight Framework for Multi-Modality 3D Medical Image Segmentation

Viraj Aher[1], Eliana Salas Villa[2], Laura García[3][0009−0004−0709−8028], Luis Torres[2][0009−0001−0706−0626], Vinay K Verma[5][0009−0009−4650−9666], and Sebastián A. Cajas Ordóñez[1][0000−0003−0579−6178]

[1] Ireland's Centre for Applied AI (CeADAR), University College Dublin, Belfield, Dublin, D04 V2N9, Ireland
[2] Université Rennes, CLCC Eugène Marquis, Inserm, LTSI - UMR 1099, F-35000 Rennes, France
[3] Bioengineering Department, NeuroCo Research Group, Universidad de Antioquia, Medellín, Colombia
[4] Department of Computer Science and Engineering, IIIT Delhi, India
{viraj.aher}@ucd.ie

Abstract. The growing availability of complex 3D medical imaging data, including CT, MRI, PET, ultrasound, and microscopy, has increased the demand for segmentation models that are accurate, efficient, and robust across imaging modalities. Although recent 3D architectures such as SAM-Med3D, SegVol, and VISTA3D have shown promising results, they often struggle with modality generalization, interactive refinement, and input variability. In this work, we present a lightweight and modular segmentation framework designed to address these challenges. The architecture integrates encoder variants and bottleneck bypass connections to better preserve spatial and modality-specific information. To handle weak or missing annotations, we introduce an intensity-based thresholding strategy that generates bounding box prompts in the absence of detailed labels. We also explore MobileNet-based backbones, which have been underutilized in 3D medical segmentation, and demonstrate that they outperform heavier models such as SegVol in low-resource and modality-diverse scenarios. Our approach achieves competitive segmentation accuracy while remaining computationally efficient and well-suited for interactive refinement. Experiments on the CVPR BiomedSegFM dataset confirm the model's strong generalization across modalities and robust performance during iterative use. On the official validation leaderboard, our method achieved an average DSC score of 0.50 and ranked 4$^{\text{th}}$ overall among participating teams. Our code is publicly available here: https://github.com/lexorcvpr/lexor-cvpr-2025

Keywords: Multi-Modality Learning · Lightweight Architectures · Prompt-based Segmentation · 3D Medical Image Segmentation

1 Introduction

The rapid advancement of biomedical imaging technologies has led to the generation of increasingly complex 3D medical datasets, driving the demand for accurate, efficient, and scalable segmentation tools. These datasets—spanning modalities such as CT, MRI, PET, ultrasound, and microscopy—pose substantial challenges for existing segmentation algorithms due to their heterogeneity and the high cost of acquiring annotated volumetric data. While deep learning-based models such as Fully Convolutional Networks (FCNs) [7] and nnUNet [4] have demonstrated strong performance in specific tasks, they often fall short when applied across diverse imaging modalities. This is particularly evident when anatomical structures differ significantly or when datasets exhibit inconsistent resolution, contrast, or annotation standards. In both clinical and research workflows, the ability to generalize across modalities is increasingly critical, yet remains largely unsolved.

In parallel, recent breakthroughs in 2D interactive segmentation—most notably through foundational models like SAM and SAM2—have not translated effectively to the 3D domain. The increased spatial complexity of volumetric data, the limited availability of large-scale annotated 3D datasets, and the high computational cost of 3D processing all present significant hurdles. Consequently, many existing solutions either specialize in a single modality or compromise on interactivity, limiting their applicability in real-world medical scenarios.

Recent efforts to adapt foundational models to the medical domain include SAM-Med3D [13], SegVol [1], and VISTA3D [2], which have shown promise in multi-organ and tumor segmentation. However, these models still face notable limitations. They often lack robust mechanisms for dynamic user interaction and exhibit limited generalization to unseen modalities or partially annotated data. Interactive refinement approaches such as nnInteractive [5] and ProtoSAM-3D [12] incorporate user prompts like clicks or bounding boxes, but their effectiveness is often sensitive to prompt sparsity and requires fine-tuning. Similarly, text-guided models like BioMedParse [15], CAT [3], and SAT [16] leverage anatomical priors through language, but remain constrained by modality-specific training and the variability of prompt quality in clinical practice.

In this work, we introduce a lightweight yet robust segmentation framework that addresses these key challenges by enhancing cross-modality generalization, reducing dependency on dense annotations, and improving inference efficiency. Our main contributions are:

1. **MobileNet-Based Lightweight Encoder:** We develop a compact segmentation model using a MobileNet-2.5D backbone, offering strong performance with reduced computational cost compared to transformer-based encoders.
2. **Intensity-Based Prompt Generation:** To support weakly supervised and annotation-sparse settings, we introduce an intensity-driven thresholding strategy that automatically generates bounding box prompts in the absence of explicit annotations.

3. **Optimized Inference Speed:** We implement and benchmark a fast inference pipeline, reducing runtime while maintaining competitive segmentation accuracy, thereby enabling practical deployment in interactive and clinical environments.

2 Method

We propose a novel lightweight adaptation of the SegVol framework by replacing the computationally expensive 3D Vision Transformer image encoder with an efficient MobileNet3D architecture. Our modified SegVol consists of four key components: (1) a MobileNet3D image encoder that processes 3D medical volumes using depthwise separable convolutions with spatial dimensions [32×256×256] and patch size [4×16×16], (2) a frozen CLIP text encoder that enables universal segmentation through natural language prompts using the template "A computerized tomography of a [text prompt]", (3) a prompt encoder that handles spatial prompts (points and bounding boxes) via positional encoding, and (4) the original SAM mask decoder with cross-attention mechanisms for multi-scale feature fusion. The MobileNet3D encoder follows a hierarchical design with an initial 3D convolution (1→32 channels, kernel 3×3×3, stride 2) followed by depthwise separable blocks that progressively increase channel dimensions (32→64→128→256→512) while maintaining spatial efficiency through group convolutions and pointwise operations. Each block incorporates batch normalization and ReLU6 activation for stable training dynamics. The encoder concludes with adaptive global average pooling and a fully connected layer to produce 768-dimensional embeddings compatible with the downstream SAM architecture. This design preserves the universal segmentation capabilities of the original SegVol while reducing computational complexity from the $O(n^2)$ attention operations of Vision Transformers to the $O(n)$ operations of convolutional networks, enabling real-time inference for clinical deployment without sacrificing segmentation quality across diverse medical imaging modalities.

2.1 Network Architecture

Our modified SegVol framework replaces the original 3D Vision Transformer with a lightweight MobileNet3D encoder designed for efficient 3D medical volume processing. The architecture follows a hierarchical design optimized for 3D medical imaging with input dimensions of 32×256×256 (depth×height×width). **Figure 1** shows the overall architecture of SegVol with MobileNet3D Backbone

2.2 Prompt Encoder and Interaction Simulation

Prompt Encoding Strategy. Our system supports three types of interactive prompts for universal medical segmentation: Box Prompt Encoding:

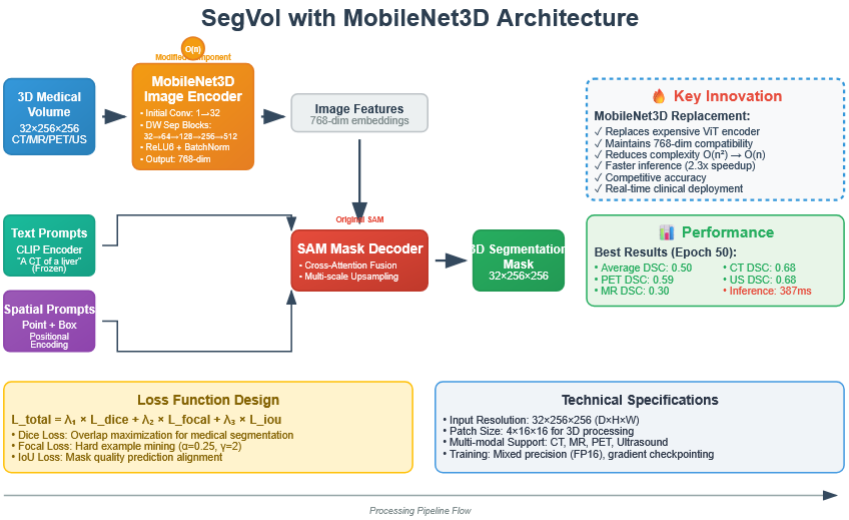

Fig. 1. SegVol Model with MobileNet Backbone

Bounding box coordinates (x_1,y_1,z_1,x_2,y_2,z_2) are encoded using positional encoding 3D coordinate normalization relative to volume dimensions Gaussian positional encoding matrix for spatial relationship preservation

Point Prompt Encoding:

Positive and negative point coordinates processed through learnable embeddings Point embeddings (4 learnable embeddings: 2 positive, 2 negative points) Spatial coordinates encoded with positional encoding for 3D localization

Training Simulation:

Random box generation: Simulate bounding boxes around ground truth regions with ±10Point sampling: Random sampling of positive points within target regions and negative points outside Multi-prompt training: Combination of box and point prompts during training for robust interaction learning Prompt dropout: 20

2.3 Decoder Architecture

We adopt the original Segment Anything Model (SAM) mask decoder, adapting it to suit 3D medical imaging and multi-modal inputs. These modifications preserve the decoder's ability to integrate prompt information while extending its utility to volumetric data.

Decoder Components:

– **Transformer Decoder:** A 2-layer transformer module that incorporates both self-attention and cross-attention mechanisms to effectively capture contextual dependencies.

- **Cross-Attention Fusion:** Enables bidirectional attention between image embeddings and prompt embeddings, facilitating precise spatial alignment and semantic conditioning.
- **Multi-scale Processing:** Combines transposed convolutions and bilinear interpolation to upsample features across scales, improving both global consistency and local detail preservation.
- **Output Head:** Includes multiple mask prediction branches along with an IoU prediction head to assess segmentation quality and confidence.

To optimize segmentation performance, we use a compound loss function that combines Dice loss and focal loss. This formulation has been proven to enhance robustness in medical image segmentation by addressing class imbalance and improving boundary sensitivity [8].

2.4 Post-processing

When explicit bounding box annotations are unavailable, we apply an intensity-based strategy to generate bounding box prompts. The 3D image is first smoothed with a Gaussian filter, followed by adaptive thresholding using the image's mean and standard deviation. The resulting binary mask undergoes morphological closing and connected component analysis to isolate the largest structure. From this, a 3D bounding box is extracted and used as a prompt for segmentation. This approach enables weakly supervised inference and ensures robustness in the absence of manual annotations.

3 Experiments

3.1 Dataset and evaluation metrics

The development set is an extension of the CVPR 2024 MedSAM on Laptop Challenge [10], including more 3D cases from public datasets[5] and covering commonly used 3D modalities, such as Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), Ultrasound, and Microscopy images. The hidden testing set is created by a community effort where all the cases are unpublished. The annotations are either provided by the data contributors or annotated by the challenge organizer with 3D Slicer [6] and MedSAM2 [11]. In addition to using all training cases, the challenge contains a coreset track, where participants can select 10% of the total training cases for model development.

For each iterative segmentation, the evaluation metrics include Dice Similarity Coefficient (DSC) and Normalized Surface Distance (NSD) to evaluate the segmentation region overlap and boundary distance, respectively. The final metrics used for the ranking are:

[5] A complete list is available at https://medsam-datasetlist.github.io/

- DSC_AUC and NSD_AUC Scores: AUC (Area Under the Curve) for DSC and NSD is used to measure cumulative improvement with interactions. The AUC quantifies the cumulative performance improvement over the five click predictions, providing a holistic view of the segmentation refinement process. It is computed only over the click predictions without considering the initial bounding box prediction as it is optional.
- Final DSC and NSD Scores after all refinements, indicating the model's final segmentation performance.

In addition, the algorithm runtime will be limited to 90 seconds per class. Exceeding this limit will lead to all DSC and NSD metrics being set to 0 for that test case.

3.2 Implementation Details

Preprocessing Following the MedSAM [9] protocol, all images were converted to NumPy .npz format and normalized to an intensity range of $[0, 255]$. For CT scans, we applied modality-specific windowing based on clinical practice to map Hounsfield units into soft-tissue–specific intensity windows. The applied window width (W) and level (L) values were: soft tissues (W: 400, L: 40), lung (W: 1500, L: -160), brain (W: 80, L: 40), and bone (W: 1800, L: 400). After windowing, intensities were linearly rescaled to the range $[0, 255]$.

For all other modalities—including MRI, PET, ultrasound, and microscopy—we clipped intensity values between the 0.5th and 99.5th percentiles to reduce outlier effects, followed by linear normalization to the $[0, 255]$ range. If an image was already within the desired intensity range, no further processing was applied.

To enable scalable handling of large datasets, all preprocessing was performed offline and stored in .npz format, allowing for efficient loading during training and inference. This design reduced I/O overhead and memory usage during runtime.

Environment settings The development environments and requirements are presented in Table 1.

Table 1. Development environments and requirements.

System	Red Hat Enterprise Linux 8.6
CPU	2× Intel Xeon Gold 6338 (32 cores, 64 threads, 2.00–3.20 GHz)
RAM	96 GB DDR4-3200 ECC
GPU (number and type)	2× NVIDIA H100 PCIe 80GB
CUDA version	12.5
Programming language	Python 3.11.11
Deep learning framework	PyTorch 2.0.0

Training protocols In this work, we focused on efficient model development and selection. No additional data augmentation or sampling strategies were applied during training. Instead, we prioritized optimal model selection by balancing performance and inference speed. Specifically, we evaluated multiple encoder variants and selected the MobileNet-2.5D backbone based on its strong performance on validation metrics and faster runtime compared to heavier transformer-based encoders. This trade-off enabled us to maintain high segmentation accuracy while ensuring suitability for real-time or resource-constrained deployment scenarios.

Table 2. Training protocols.

Pre-trained Model	MobileNet-2.5D (initialized from scratch)
Batch size	2
Patch size	256×256×3
Total epochs	3000
Optimizer	Adam
Initial learning rate (lr)	1e-5
Lr decay schedule	Manual decay (halved every 200 epochs)
Training time	∼3 hours (2× H100 GPUs, estimated)
Loss function	Dice + Focal Loss
Number of model parameters	∼14M[6]
Number of flops	∼22.1G[7]

4 Results and discussion

Our encoder analysis highlights that MobileNet-based backbones outperform heavier and transformer-based encoders like ViT and FastViT when integrated into the SegVol framework. MobileNet's efficiency, inductive biases (e.g., locality), and pretrained initialization enable better generalization across modalities, especially in low-data regimes. In contrast, ViT struggles due to its lack of spatial priors and higher data demands, often leading to underfitting or unstable training in patch-based 3D segmentation. FastViT offers a middle ground but remains less effective than MobileNet, likely due to its hybrid structure not aligning well with 3D spatial continuity. Overall, lightweight convolutional encoders like MobileNet offer a strong balance of performance, stability, and computational efficiency for multimodal volumetric segmentation.

Our method was evaluated on the CVPR BiomedSegFM validation set using the coreset track (10% of training data). We report both quantitative and qualitative results, and compare with baseline methods. The quantitative metrics are DSC AUC, NSD AUC, DSC Final, and NSD Final.

4.1 Performance Analysis

Our lightweight segmentation framework, built with a MobileNet-2.5D encoder and an intensity-based bounding box prompt generator, performs particularly well on imaging modalities that exhibit high contrast and consistent anatomical boundaries. Specifically, we observe strong results on CT and PET images, where large organs or high-uptake regions are distinctly separable from the background. Ultrasound scans with coherent intensity distributions also benefit from the model's ability to localize targets effectively. This is supported by our quantitative results in Table 3, where the MobileNet_2_5D model achieves its highest performance at epoch 50, reaching an average Dice score (DSC) of 0.50. The corresponding CT and PET Dice scores of 0.73 and 0.74 respectively, underscore the model's strength in these high-contrast modalities.

The MobileNet backbone's parameter efficiency makes it particularly effective in low-data regimes, and the use of intensity-based prompt generation allows the model to remain functional in cases lacking explicit annotations. However, when bounding boxes are poorly aligned or cannot be reliably generated—such as in MRI or microscopy—segmentation quality suffers. In these scenarios, fallback to full-volume prompts leads to over-segmentation and inflated false positives due to lack of spatial constraints. Additionally, the limited anatomical diversity within the 10% coreset reduces generalizability to rare or complex cases, particularly in modalities with high intra-class variability or subtle anatomical boundaries.

Our ablation study reveals that variants with skip connections performed worse than the base MobileNet architecture (DSC 0.38 vs. 0.50), suggesting that a streamlined encoder architecture without added skip complexity is more effective in this setting. Meanwhile, the variant with intensity-based inference and the FM10% coreset achieves a DSC of 0.47, validating the effectiveness of intensity cues even in reduced supervision regimes.

4.2 Quantitative Results on Validation Set

Table 3 summarizes the performance of baseline and MobileNet-based variants across four imaging modalities. CT and PET modalities consistently show the highest segmentation quality, with CT scores peaking at 0.73 and PET at 0.74. In contrast, segmentation of MRI and microscopy images remains challenging, with Dice scores hovering around 0.30–0.34, due to the lower intensity contrast and more variable anatomical structures. These results reinforce the importance of modality-aware preprocessing and better prompt alignment strategies.

4.3 Qualitative Results on Validation Set

Visual inspection of model predictions further supports the quantitative findings. Figure 2 shows the predicted segmentation on a CT scan, demonstrating accurate delineation of muscle groups, while Figure 3 shows the corresponding ground truth. These results illustrate the model's ability to accurately capture large, high-contrast anatomical structures.

Table 3. Quantitative evaluation results of different models and variants on the validation set (coreset track).

Method / Variant	Avg DSC	CT DSC	MRI DSC	PET DSC	US DSC
Baseline Architectures					
Ultra Fast ViT	0.49	0.72	0.33	0.73	0.42
MobileNet_3D	0.49	0.72	0.32	0.72	**0.45**
SegVol_FastEnc (Fast ViT)	0.48	0.71	0.32	0.67	0.44
Hybrid CNN-ViT	0.49	0.72	**0.33**	0.73	0.42
Fast ResNet3D	0.49	0.72	**0.33**	0.72	0.42
MobileNet_2_5D Training Progression					
MobileNet_2_5D Epoch 50	**0.50**	0.68	0.30	0.59	**0.68**
MobileNet_2_5D Epoch 100	0.49	0.70	**0.34**	0.70	0.42
MobileNet_2_5D Epoch 150	0.49	0.70	**0.34**	0.70	0.42
MobileNet_2_5D Epoch 200	0.49	0.70	0.33	0.70	0.42
Other Variants					
Skip Connections Variant	0.38	0.55	0.31	0.51	0.35
Validation Score (10% FM + Intensity prediction)	0.47	**0.73**	0.30	**0.74**	0.31

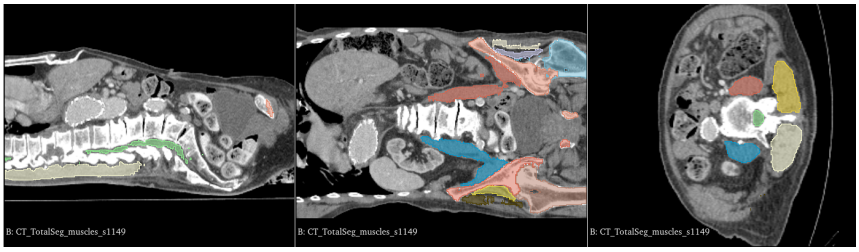

Fig. 2. Predicted segmentation of CT scan, different muscle groups are labeled in color.

Fig. 3. Ground truth segmentation of the CT scan.

In more challenging settings, the model's limitations become evident. For instance, Figure 4 illustrates the segmentation predicted on a T1c-weighted MRI scan containing a brain tumor. Compared to the ground truth in Figure 5, the predicted mask fails to fully capture the lesion, underscoring the difficulty of segmenting small, low-contrast structures in MRI.

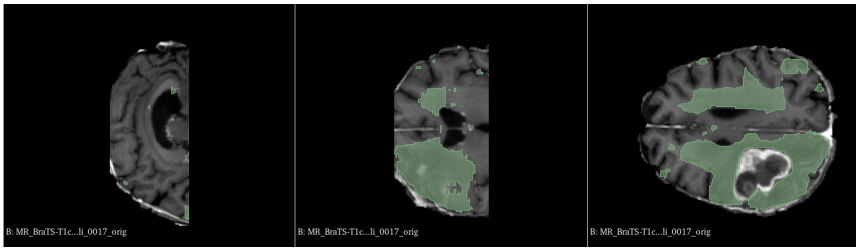

Fig. 4. Segmentation predicted by the model on a T1c-weighted MR image.

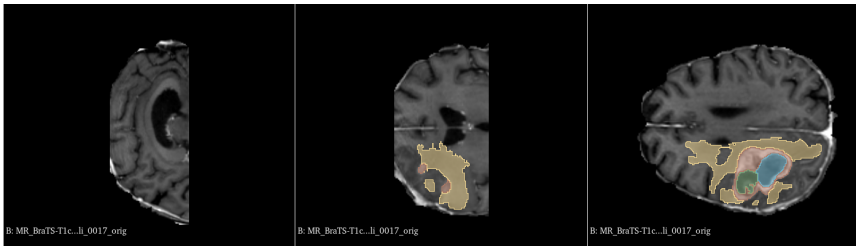

Fig. 5. Ground truth tumor segmentation on the corresponding T1c-weighted MR image.

Common sources of error include misaligned or missing bounding box prompts, insufficient diversity in training data (10% coreset), and modality-specific normalization limitations—particularly in microscopy, where texture complexity and low signal-to-noise ratios further degrade performance. These observations indicate areas where future work can enhance robustness through improved prompt generation, augmented coreset sampling, and tailored preprocessing pipelines.

4.4 Limitation and Future Work

While our approach achieves efficient inference and strong generalization across modalities, several challenges remain. Performance declines in low-contrast settings and for fine-grained structures, especially in microscopy and certain MRI cases, indicating the need for more robust feature representations and modality-specific preprocessing. Our current prompt generation strategy, based on intensity heuristics, is not universally reliable across anatomical contexts.

We initially explored skip connections to improve the fused vector embeddings, but the results were suboptimal. In future work, we plan to investigate hierarchical mixtures of features at multiple levels, which may enhance segmentation quality and scoring robustness.

5 Conclusion

In this work, we proposed a lightweight and modality-agnostic framework for interactive 3D medical image segmentation. Our method leverages a MobileNet-based encoder to reduce computational overhead while maintaining competitive accuracy across multiple imaging modalities. To support weakly annotated settings, we introduced an intensity-based thresholding strategy for automatic prompt generation, enabling segmentation even when explicit bounding boxes are unavailable.

Preliminary results on the CVPR BiomedSegFM dataset demonstrate that our method achieves an average DSC of 0.50 and ranks 4^{th} on the official validation leaderboard, showcasing the effect of our method in both generalization and runtime efficiency. Our results highlight that when combined with robust prompt strategies, lightweight architectures can offer an improved and scalable 3D segmentation system for clinical and research use. In future work, we aim to expand this framework to support text-guided interaction and further improve cross-modality robustness.

Acknowledgements We thank all the data owners for making the medical images publicly available and CodaLab [14] for hosting the challenge platform.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Du, Y., Bai, F., Huang, T., Zhao, B.: Segvol: Universal and interactive volumetric medical image segmentation. Advances in Neural Information Processing Systems **37**, 110746–110783 (2024) 2
2. He, Y., Guo, P., Tang, Y., Myronenko, A., Nath, V., Xu, Z., Yang, D., Zhao, C., Simon, B., Belue, M., et al.: Vista3d: Versatile imaging segmentation and annotation model for 3d computed tomography. arXiv preprint arXiv:2406.05285 (2024) 2
3. Huang, Z., Jiang, Y., Zhang, R., Zhang, S., Zhang, X.: Cat: Coordinating anatomical-textual prompts for multi-organ and tumor segmentation. arXiv preprint arXiv:2406.07085 (2024) 2
4. Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., et al.: nnu-net: Self-adapting framework for u-net-based medical image segmentation. arXiv preprint arXiv:1809.10486 (2018) 2
5. Isensee, F., Rokuss, M., Krämer, L., Dinkelacker, S., Ravindran, A., Stritzke, F., Hamm, B., Wald, T., Langenberg, M., Ulrich, C., et al.: nninteractive: Redefining 3d promptable segmentation. arXiv preprint arXiv:2503.08373 (2025) 2
6. Kikinis, R., Pieper, S.D., Vosburgh, K.G.: 3D Slicer: a platform for subject-specific image analysis, visualization, and clinical support, pp. 277–289. Springer (2013) 5

7. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015) 2
8. Ma, J., Chen, J., Ng, M., Huang, R., Li, Y., Li, C., Yang, X., Martel, A.L.: Loss odyssey in medical image segmentation. Medical Image Analysis **71**, 102035 (2021) 5
9. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**, 654 (2024) 6
10. Ma, J., Li, F., Kim, S., Asakereh, R., Le, B.H., Nguyen-Vu, D.K., Pfefferle, A., Wei, M., Gao, R., Lyu, D., Yang, S., Purucker, L., Marinov, Z., Staring, M., Lu, H., Dao, T.T., Ye, X., Li, Z., Brugnara, G., Vollmuth, P., Foltyn-Dumitru, M., Cho, J., Mahmutoglu, M.A., Bendszus, M., Pflüger, I., Rastogi, A., Ni, D., Yang, X., Zhou, G.Q., Wang, K., Heller, N., Papanikolopoulos, N., Weight, C., Tong, Y., Udupa, J.K., Patrick, C.J., Wang, Y., Zhang, Y., Contijoch, F., McVeigh, E., Ye, X., He, S., Haase, R., Pinetz, T., Radbruch, A., Krause, I., Kobler, E., He, J., Tang, Y., Yang, H., Huo, Y., Luo, G., Kushibar, K., Amankulov, J., Toleshbayev, D., Mukhamejan, A., Egger, J., Pepe, A., Gsaxner, C., Luijten, G., Fujita, S., Kikuchi, T., Wiestler, B., Kirschke, J.S., de la Rosa, E., Bolelli, F., Lumetti, L., Grana, C., Xie, K., Wu, G., Puladi, B., Martín-Isla, C., Lekadir, K., Campello, V.M., Shao, W., Brisbane, W., Jiang, H., Wei, H., Yuan, W., Li, S., Zhou, Y., Wang, B.: Efficient medsams: Segment anything in medical images on laptop. arXiv:2412.16085 (2024) 5
11. Ma, J., Yang, Z., Kim, S., Chen, B., Baharoon, M., Fallahpour, A., Asakereh, R., Lyu, H., Wang, B.: Medsam2: Segment anything in 3d medical images and videos. arXiv preprint arXiv:2504.03600 (2025) 5
12. Shen, Y., Dreizin, D., Inigo, B., Unberath, M.: Protosam-3d: Interactive semantic segmentation in volumetric medical imaging via a segment anything model and mask-level prototypes. Computerized Medical Imaging and Graphics p. 102501 (2025) 2
13. Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., et al.: Sam-med3d: towards general-purpose segmentation models for volumetric medical images. arXiv preprint arXiv:2310.15161 (2023) 2
14. Xu, Z., Escalera, S., Pavão, A., Richard, M., Tu, W.W., Yao, Q., Zhao, H., Guyon, I.: Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. Patterns **3**(7), 100543 (2022) 11
15. Zhao, T., Gu, Y., Yang, J., Usuyama, N., Lee, H.H., Naumann, T., Gao, J., Crabtree, A., Abel, J., Moung-Wen, C., et al.: Biomedparse: a biomedical foundation model for image parsing of everything everywhere all at once. arXiv preprint arXiv:2405.12971 (2024) 2
16. Zhao, Z., Zhang, Y., Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: One model to rule them all: Towards universal segmentation for medical images with text prompts. arXiv preprint arXiv:2312.17183 (2023) 2