MIQ-SAM3D: FROM SINGLE-POINT PROMPT TO MULTI-INSTANCE SEGMENTATION VIA COMPETITIVE QUERY REFINEMENT

Jierui Qu

National University of Singapore
College of Design and Engineering
117575, Singapore

*Jianchun Zhao**

Xi'an Jiaotong University
School of Electrical Engineering
Xi'an 710049, China

ABSTRACT

Accurate segmentation of medical images is fundamental to tumor diagnosis and treatment planning. SAM-based interactive segmentation has gained attention for its strong generalization, but most methods follow a single-point-to-single-object paradigm, which limits multi-lesion segmentation. Moreover, ViT backbones capture global context but often miss high-fidelity local details. We propose MIQ-SAM3D, a multi-instance 3D segmentation framework with a competitive query optimization strategy that shifts from single-point-to-single-mask to single-point-to-multi-instance. A prompt-conditioned instance-query generator transforms a single point prompt into multiple specialized queries, enabling retrieval of all semantically similar lesions across the 3D volume from a single exemplar. A hybrid CNN-Transformer encoder injects CNN-derived boundary saliency into ViT self-attention via spatial gating. A competitively optimized query decoder then enables end-to-end, parallel, multi-instance prediction through inter-query competition. On LiTS17 and KiTS21 dataset, MIQ-SAM3D achieved comparable levels and exhibits strong robustness to prompts, providing a practical solution for efficient annotation of clinically relevant multi-lesion cases.

***Index Terms*—** SAM, Volumetric medical images Segmentation, Promptable Segmentation

1. INTRODUCTION

Accurate medical image segmentation is foundational to modern clinical practice. It underpins disease diagnosis, treatment planning, and prognostic assessment [1]. In recent years, deep learning—particularly convolutional neural networks (CNNs) such as U-Net and its variants—has achieved strong performance in three-dimensional (3D) medical image segmentation [2, 3]. However, most of these models remain task-specific and generalize poorly. Moreover, the inherently local receptive fields of CNNs limit their ability to model global context and long-range dependencies [4, 5, 6, 7].

To overcome the limitations of traditional CNNs in understanding global context, the research community has turned to exploring Vision Transformer (ViT) architectures with robust long-range dependency modeling capabilities [8]. Among these, the large-scale pre-trained foundation model based on ViT—the Segment Anything Model (SAM)—has garnered significant attention for its exceptional zero-shot and promptable segmentation capabilities on natural images [9, 10, 11]. This breakthrough has rapidly spurred extensive research aimed at adapting SAM's powerful capabilities to 3D medical image segmentation through strategies like parameter-efficient fine-tuning (PEFT), bridging the significant domain gap between natural and medical images [12]. However, current research is constrained by two core limitations: most frameworks adhere to a "single-point prompt, single-object segmentation" paradigm, struggling to address clinically prevalent diffuse lesion instance segmentation tasks; moreover, their inherent ViT backbone networks sacrifice high-fidelity local details crucial for ambiguous boundaries while capturing global context.

To systematically address the aforementioned limitations, this paper proposes the MIQ-SAM3D segmentation framework, which can simultaneously handle multi-instance objects and perceive high-fidelity details. We introduce a prompt-initiated instance segmentation paradigm inspired by modern detection models [13], which transforms a single-point user prompt into a set of dynamic object queries via a prompt-conditioned instance query generator (PC-IQG). These queries are then fed into a Competitive Query Refinement Decoder, enabling the identification of all similar instances throughout the 3D volume using the user-clicked lesion as an "example." Additionally, we designed a hybrid CNN-Transformer encoder. A parallel 3D CNN branch specializes in capturing local features, while a Spatial Gate Module deeply integrates high-fidelity details into the self-attention computation of the ViT backbone, significantly improving segmentation accuracy for blurred boundaries.

The main contributions of this paper can be summarized as follows:

1. We propose a prompt-initiated MIQ-SAM3D segmentation framework that achieves comparable performance in segmenting multiple disjoint object instances with a single point prompt. This is accomplished through our designed PC-IQG and competitive query optimization decoder.
2. We designed an efficient hybrid CNN-Transformer encoder whose core gated cross-fusion mechanism synergistically leverages CNN's local advantages and ViT's global perspective, achieving outstanding segmentation performance with high-fidelity details.
3. Our approach was validated on multiple public 3D tumor segmentation benchmarks. Experimental results demonstrate comparable performance on multi-instance segmentation tasks.

2. METHOD

The overall architecture of the multi-instance 3D segmentation framework proposed in this study is shown in Fig. 1. For the input 3D medical image $\mathbf{X} \in \mathbb{R}^{D\times H\times W}$, local feature extraction and global feature modeling are performed via a dual-branch CNN-Transformer encoder. User prompts $\mathbf{p}$ are fed into the Prompt-

*Corresponding Author, email: zhao_jianchun@stu.xjtu.edu.cn.

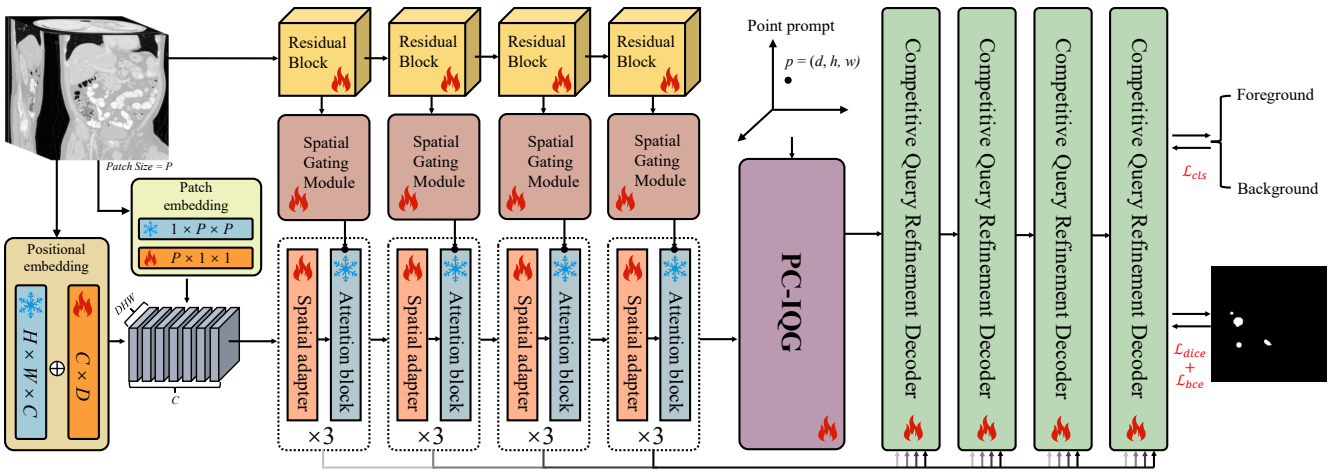

Fig. 1. Overview of our proposed method for MIQ-SAM3D.This framework consists of a hybrid CNN-Transformer encoder, a prompt-conditioned instance query generator (PC-IQG), a competitive query-optimized decoder (CQRD), and an instance prediction head.

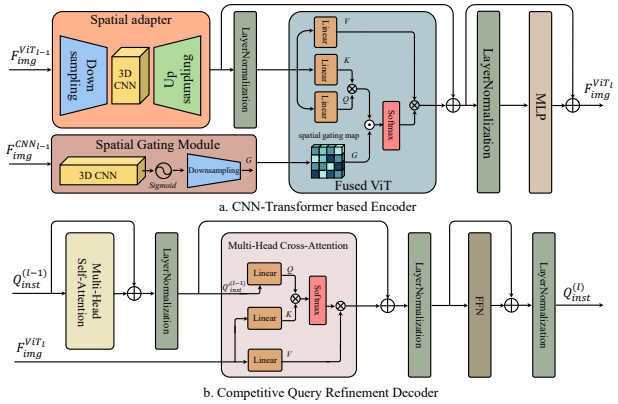

Fig. 2. Structures of the encoder and decoder. (a)Structure diagram of the dual-branch hybrid CNN-Transformer encoder, which uses a spatial gating module to fuse dual-channel features. (b)Structure diagram of CQRD.

Conditioned Instance Query Generato(PC-IQG) module to extract seed prototypes and dynamically generate N instance queries Q_{inst}. Q_{inst} and image features undergo iterative optimization via Competitive Query Refinement Decode(CQRD), enabling competing queries to independently lock onto distinct target instances. Finally, the optimized queries are fed into two parallel prediction heads, enabling end-to-end prediction from single-point prompting to multi-instance segmentation.

2.1. Hybrid CNN-Transformer Encoder

The hybrid CNN-Transformer encoder adopts a dual-branch architecture. Through the Spatial Gating Module, it integrates the local detail features extracted by the Residual Block into the ViT module, guiding the extraction of global features, as shown in Fig. 2(a). The ViT branch inherits the pre-trained weights from the 3DSAM-adapter [14], responsible for capturing long-range dependencies and global semantic information. Most of its parameters remain frozen to preserve pre-trained knowledge. The CNN branch employs lightweight residual blocks trained from scratch to extract high-frequency details like edges and textures across multiple scales. These features are fed into the Spatial Gating Module to generate the spatial gating map G:

$$G = Sigmoid(Conv_{1\times1\times1}(F_{img}^{CNN_{l-1}})) \quad (1)$$

In the Fused ViT block, input features undergo linear transformations to yield queries Q, keys K, and values V, and compute the attention score matrix: $A = \frac{QK^T}{\sqrt{d_k}}$. The gating graph G is downsampled to a dimension compatible with the attention score matrix and applied to the attention scores:

$$Attention_{fued} = (Softmax\,(A \odot Downsample\,(G)))\,V \quad (2)$$

The symbol $\odot$ denotes element-wise multiplication. This gated fusion mechanism generates feature representations for subsequent instance retrieval that combine global context with high-fidelity details.

2.2. Prompt-Conditioned Instance Query Generator(PC-IQG)

This paper proposes the PC-IQG module, which uses the features of a user-specified single point location as a category prototype. Based on this, it generates multiple specialized instance detectors to retrieve and segment all semantically similar objects throughout the entire 3D volume. The user-provided 3D coordinates $p = (d, h, w)$ of the query point are input into the PC-IQG module alongside features from the hybrid encoder. A visual sampler [14] extracts a seed prototype vector $v_{seed} \in R^C$ from the feature map at this location.

The seed prototype v_{seed} encodes the semantic category, morphological features, and local contextual information of the target instance clicked by the user, serving as the key conditional signal for subsequent query generation. Subsequently, v_{seed} is fed into a multi-layer perceptron-based query generation network G_θ, which expands the single seed prototype into N diverse instance queries $Q_{inst} \in R^{N\times C}$ through nonlinear transformations:

$$Q_{inst} = \mathcal{G}_\theta(v_{seed}) \quad (3)$$

The dynamic adaptation mechanism endows the model with strong zero-shot generalization capabilities, enabling it to perform segmentation tasks without retraining for new categories.

2.3. Competitive Query Refinement Decoder(CQRD)

CQRD employs a multi-layer Transformer decoder architecture that progressively optimizes instance representations through competitive-collaborative mechanisms among queries and interactions with image features, as illustrated in Fig. 2(b). This decoder consists of four identically structured decoding layers stacked sequentially. At layer l, the input instance queries $Q_{inst}^{(l-1)}$ undergo an inter-query self-attention module to suppress redundant predictions for the same instance, ensuring different queries lock onto distinct instance targets, yielding $Q'^{(l-1)}_{inst}$. Subsequently, $Q'^{(l-1)}_{inst}$ interacts with the image features output by the encoder through a multi-head cross-attention module. Each query selectively aggregates spatial information relevant to its assigned instance from the global feature map, thereby progressively refining the instance's positional and morphological representations:

$$Q''^{(l-1)}_{inst} = \mathrm{CrossAttention}\left(Q'^{(l-1)}_{inst}, F_{img}^{ViT_l}\right) + Q'^{(l-1)}_{inst} \quad (4)$$

Finally, the query's expressive power is further enhanced through nonlinear transformations in the feedforward network, producing the final query representation of this layer:

$$Q_{inst}^{(l)} = \mathrm{FFN}\left(Q''^{(l-1)}_{inst}\right) + Q''^{(l-1)}_{inst} \quad (5)$$

After L layers of iterative optimization, the final query Q_{inst}^{L} is fed into two parallel prediction heads to generate segmentation results. The classification head predicts the probability distribution of categories corresponding to each query, distinguishing foreground instances from background. The mask head generates instance masks by performing dot product operations between query embeddings and image features.

2.4. Loss Function

Since the number of instances M in the ground truth annotations is typically much smaller than N and instances lack a fixed order, this paper employs a set prediction loss based on bipartite graph matching [15] for end-to-end training.Let the set of true labels be $y = y_1, y_2, \ldots, y_M$, where each $y_j = (c_j, m_j)$ contains the category label c_j and the 3D mask m_j. Let the prediction set of the model be $\hat{y} = \hat{y}_1, \hat{y}_2, \ldots, \hat{y}_M$, where each $\hat{y}_j = (\hat{c}_j, \hat{m}_j)$. Define the matching cost between prediction i and actual instance j as $\mathcal{C}(i, j)$:

$$\begin{aligned}\mathcal{C}(i,j) = -\lambda_{cls} \cdot P_i(c_j) + \lambda_{dice} \cdot \mathcal{L}_{dice}(\hat{m}_i, m_j) \\ + \lambda_{bce} \cdot \mathcal{L}_{bce}(\hat{m}_i, m_j)\end{aligned} \quad (6)$$

Here, $P_i(c_j)$ denotes the probability of predicting that i belongs to category c_j. The Hungarian algorithm is used to find the permutation that minimizes the total matching cost.

$$\hat{\sigma} = \arg\min_{\sigma} \sum_{i=1}^{N} \mathcal{C}\big(\hat{y}_{\sigma(i)},\, y_i\big) \quad (7)$$

Thus establishing an optimal bijection between predictions and ground truth, the total loss function is computed only for successfully matched prediction pairs:

$$\begin{aligned}\mathcal{L}_{total} = \sum_{i=1}^{N} [\lambda_{cls}\mathcal{L}_{cls}\left(\hat{p}_{\hat{\sigma}(i)}, c_i\right) + \lambda_{dice}\mathcal{L}_{dice}\left(\hat{m}_{\hat{\sigma}(i)}, m_i\right) \\ + \lambda_{bce}\mathcal{L}_{bce}\left(\hat{m}_{\hat{\sigma}(i)}, m_i\right)]\end{aligned} \quad (8)$$

3. EXPERIENCE

3.1. Datasets

We employed two publicly available datasets for tumor segmentation: the KiTS21 dataset [16] and the LiTS17 dataset [17], both of which feature multiple tumors. The original datasets included organ and tumor segmentation labels, but we exclusively utilized tumor labels for training and testing. The datasets were randomly partitioned into 70%, 10%, and 20% for training, validation, and testing, respectively. The Dice coefficient (Dice) and Normalized Surface Dice (NSD) are used as evaluation metrics.

3.2. Experimental results and visualization

To validate the effectiveness of the proposed framework in multi-instance segmentation tasks for 3D medical images, we selected five representative methods for comparison: nn-UNet [18], Swin-UNETR [19], 3D UX-Net [20], 3DSAM-adapter(1 pt) [14], and 3D-SAutoMed [21]. Detailed results are shown in Table 1. Our method achieved an optimal Dice coefficient of 60.47% on the Liver Tumor dataset, which was 1.37 percentage points higher than 59.10% of the traditional SOTA method nn-UNet and 4.66 percentage points higher than 55.81% of the similar SAM adaptation method 3DSAM-adapter. It is worth noting that liver tumors usually present the characteristics of multiple occurrence and blurred boundaries. This result fully validates the effectiveness of the hybrid CNN-Transformer encoder in capturing fine boundary details and the performance of the prompt conditional query mechanism in multi-instance scenarios. On the Kidney Tumor dataset, our approach attains state-of-the-art performance with a Dice score of 74.83%. The NSD metric demonstrates comparable levels across both datasets. Fig. 3 demonstrates typical multi-instance segmentation visualizations. Under single-point-input conditions, our method achieves segmentation of multiple tumor regions with clear and complete segmentation boundaries.

3.3. Ablation Studies

To comprehensively evaluate the effectiveness of each component within the proposed framework, we conducted a series of ablation experiments on the Liver Tumor dataset, as shown in Table 2. First, we evaluated the contribution of PC-IQG+CQRD. When these two modules were removed and the model reverted to the traditional single-instance segmentation paradigm, the Dice score dropped significantly from 60.47% to 56.72%. This 3.75 percentage point performance degradation resulted from the inability to simultaneously capture all similar instances in a scene during a single forward pass, leading to overall segmentation quality deterioration. Second, we validated the necessity of the hybrid encoder design by removing the CNN branch. Results showed that the model relying solely on the ViT backbone achieved a Dice coefficient of 58.95%, a 1.52 percentage point decrease compared to the full model. More notably, the NSD metric plummeted from 74.61% to 68.80%, a significant drop of 5.81 percentage points. This outcome conclusively demonstrates the irreplaceable role of the CNN branch in capturing

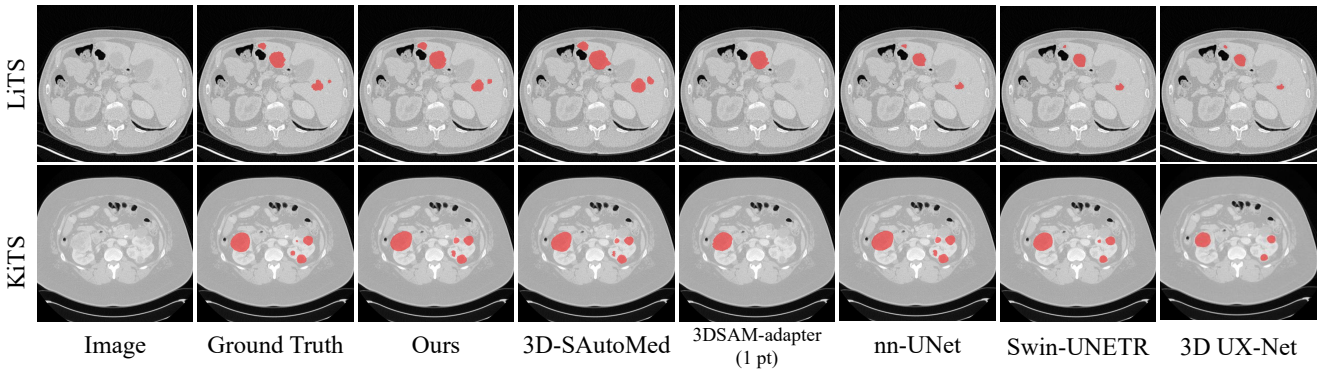

Fig. 3. Comparison with classical medical image segmentation methods on the LiTS17 and KiTS21 datasets. The evaluation metrics for dice scores and normalized surface dice (NSD) were reported.

Table 1. Comparison with classical medical image segmentation methods on the LiTS17 and KiTS21 datasets. The evaluation metrics for dice scores and normalized surface dice (NSD) were reported.

Method	Framework	LiTS17 (Liver Tumor)		KiTS21 (Kidney Tumor)	
		Dice (%) ↑	NSD (%) ↑	Dice (%) ↑	NSD (%) ↑
nn-UNet [18]	U-net	59.10	**77.35**	73.79	81.24
Swin-UNETR [19]	U-net	52.10	68.07	66.04	74.29
3D-UX-Net [20]	U-net	48.27	65.72	60.49	68.23
3DSAM-adapter (1 pt) [14]	SAM	55.81	70.19	73.08	**85.73**
3D-SAutoMed [21]	SAM	58.92	74.83	74.27	82.09
Ours	SAM	**60.47**	74.61	**74.83**	79.57

Table 2. Ablation study on LiTS17 dataset.

Method	Dice (%) ↑	NSD (%) ↑
Ours (Full Model)	60.47	74.61
Ours w/o PC-IQG + CQRD	56.72	75.85
Ours w/o CNN branches	58.95	68.80

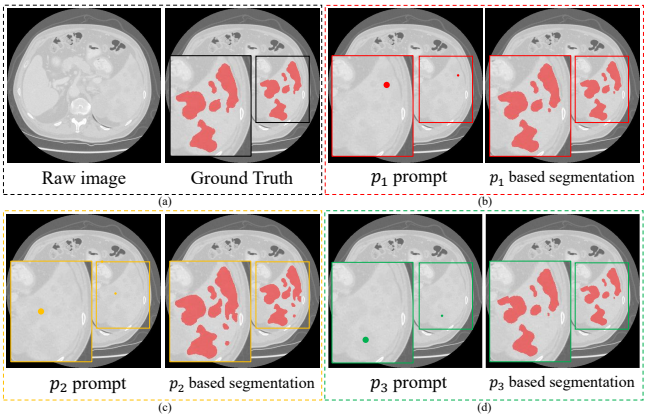

Fig. 4. Segmentation results under different single-point prompts

high-fidelity local details. Finally, we conducted a qualitative assessment of the framework's prompt robustness, as illustrated in Fig. 4. Within the same 3D image containing multiple liver tumors, we provided single-point prompts on different tumor instances (marked by differently colored dots in the figure) and observed the model's segmentation outputs. Experimental results demonstrate that regardless of which specific instance the user selects as the example, the model consistently identifies and segments all tumors of the same type within the scene. The predicted masks exhibit high consistency in both spatial distribution and the number of instances.

4. CONCLUSION

This paper proposes the MIQ-SAM3D multi-instance segmentation framework based on a hybrid CNN-Transformer architecture, achieving a breakthrough from single-point-to-single-mask to single-point-to-multi-instance segmentation. To the best of our knowledge, this is the first work applying prompt-conditioned instance query generation with competitive optimization decoders to 3D medical image segmentation. Our method designs a hybrid encoder that deeply integrates CNN-extracted boundary saliency features into ViT's self-attention computation via a spatial gating module, significantly enhancing the representation of fuzzy boundaries while preserving global semantic understanding. PC-IQG transforms user single-point prompts into dynamically generated object queries, while CQRD enables end-to-end parallel prediction of multiple instances through query-to-query competition. Extensive experiments on Liver Tumor and Kidney Tumor datasets demonstrate that the proposed method surpasses existing state-of-the-art approaches in segmentation accuracy while exhibiting outstanding prompt robustness, providing a practical solution for efficient annotation of clinically prevalent multiple lesions.

5. ACKNOWLEDGMENTS

No funding was received for conducting this study. The authors have no relevant financial or non-financial interests to disclose.

6. REFERENCES

[1] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, and et al., "The medical segmentation decathlon," *Nature communications*, vol. 13, no. 1, pp. 4128, 2022.

[2] et al., "3d u-net: Learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, Sebastien Ourselin, Leo Joskowicz, Mert R. Sabuncu, Gozde Unal, and William Wells, Eds., vol. 9901, pp. 424–432. Springer International Publishing, Cham, 2016.

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, Eds., vol. 9351, pp. 234–241. Springer International Publishing, Cham, 2015.

[4] Omid Nejati Manzari, Javad Mirzapour Kaleybar, Hooman Saadat, and Shahin Maleki, "Befunet: A hybrid cnn-transformer architecture for precise medical image segmentation," Feb. 2024.

[5] Abdelrahman Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan, "Unetr++: Delving into efficient and accurate 3d medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 43, no. 9, pp. 3377–3390, 2024.

[6] Minyoung Park, Seungtaek Oh, Junyoung Park, Taikyeong Jeong, and Sungwook Yu, "Es-unet: Efficient 3d medical image segmentation with enhanced skip connections in 3d unet," *BMC Medical Imaging*, vol. 25, no. 1, pp. 327, Aug. 2025.

[7] Ziyan Huang, Jin Ye, Haoyu Wang, Zhongying Deng, Zhikai Yang, and et al., "Revisiting model scaling with a u-net benchmark for 3d medical image segmentation," *Scientific Reports*, vol. 15, no. 1, pp. 29795, 2025.

[8] Bobby Azad, Reza Azad, Sania Eskandari, Afshin Bozorgpour, Amirhossein Kazerouni, Islem Rekik, and Dorit Merhof, "Foundational models in medical imaging: A comprehensive survey and future vision," Oct. 2023.

[9] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, and et al., "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[10] Mengmeng Zhang, Xingyuan Dai, Yicheng Sun, Jing Wang, Yueyang Yao, and et al., "Hierarchical self-prompting sam: A prompt-free medical image segmentation framework," June 2025.

[11] Jianghao Wu, Yicheng Wu, Yutong Xie, Wenjia Bai, You Zhang, and et al., "Sam-aware test-time adaptation for universal medical image segmentation," June 2025.

[12] Maciej A. Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang, "Segment anything model for medical image analysis: An experimental study," *Medical Image Analysis*, vol. 89, pp. 102918, 2023.

[13] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1290–1299.

[14] Shizhan Gong, Yuan Zhong, Wenao Ma, Jinpeng Li, Zhao Wang, and et al., "3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable medical image segmentation," *arXiv e-prints*, pp. arXiv–2306, 2023.

[15] Nicolas Carion and et al., "End-to-end object detection with transformers," in *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, Eds., vol. 12346, pp. 213–229. Springer International Publishing, Cham, 2020.

[16] Nicholas Heller, Fabian Isensee, Klaus H. Maier-Hein, Xiaoshuai Hou, Chunmei Xie, and et al., "The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge," *Medical image analysis*, vol. 67, pp. 101821, 2021.

[17] Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, and et al., "The liver tumor segmentation benchmark (lits)," *Medical image analysis*, vol. 84, pp. 102680, 2023.

[18] Fabian Isensee, Paul F. Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H. Maier-Hein, "nnu-net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[19] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, and et al., "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Alessandro Crimi and Spyridon Bakas, Eds., vol. 12962, pp. 272–284. Springer International Publishing, Cham, 2022.

[20] Ho Hin Lee, Shunxing Bao, Yuankai Huo, and Bennett A. Landman, "3d ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation," Mar. 2023.

[21] et al., "3d-sautomed: Automatic segment anything model for 3d medical image segmentation from local-global perspective," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, and Ben Glocker, Eds., vol. 15009, pp. 3–12. Springer Nature Switzerland, Cham, 2024.