GU-SAM2: A Semi-Supervised Medical Image Segmentation Framework via Gated Feature Fusion of MIM-Pretrained U-Net and SAM 2

Yinjun Liu[1], Mingrui Zhuang[2], Qinhe Zhang[3], Ailian Liu[3], Songyao Zhang[2*], and Hongkai Wang[1*]

[1] Dalian University of Technology, Dalian 116081, China
[2] the Central Hospital of Dalian University of Technology, Dalian 116024, China
[3] the First Affiliated Hospital of Dalian Medical University, Dalian 116000, China
biomedimg@163.com, songyaozhang@dlut.edu.cn

Abstract. Vision foundation models have emerged as a pivotal solution for medical image segmentation, demonstrating remarkable potential in alleviating annotation burdens. However, their direct application to 3D medical data faces significant challenges: the inherent domain gap between natural and medical images, and a mechanistic misalignment where SAM 2's native "object tracking" paradigm struggles with the "semantic matching" required for volumetric anatomy. To address these challenges, we propose GU-SAM2, a novel label-decoupled semi-supervised framework. By strategically separating data utilization, our method effectively bridges the domain gap and resolves mechanistic conflicts: the first stage operates in a label-free manner, employing Masked Image Modeling (MIM) on ubiquitous unlabeled medical images to train a lightweight U-Net for internalizing medical domain knowledge; the second stage operates in a prompt-guided manner, introducing a Gated Feature Fusion mechanism to adaptively inject these medical features into SAM 2. This decoupled fusion strategy enables dynamic, semantic-aware feature selection, correcting the model logic from mere tracking to anatomically consistent segmentation. Extensive experiments across multi-organ and multi-modality datasets demonstrate that GU-SAM2 achieves state-of-the-art performance. Notably, it maintains exceptional stability on small or challenging organs and delivers robust results even in scenarios with extremely scarce annotated data (e.g., 1 or 3 labeled scans).

Keywords: Semi-supervised Learning · SAM 2 · Gated Feature Fusion.

1 Introduction

Accurate segmentation is the cornerstone of clinical workflows [20]. While fully supervised models like nnU-Net [13] excel, they depend heavily on expensive and scarce expert annotations [22]. Semi-Supervised Learning (SSL) attempts to mitigate this by leveraging unlabeled data via pseudo-labeling [24, 29]. However,

in low-data regimes, noisy pseudo-labels often cause the model to reinforce its own erroneous predictions. This "confirmation bias" severely limits the reliability of traditional SSL methods in clinical settings [28, 2].

Foundation models like SAM [14] and SAM 2 [18] offer promising avenues for medical segmentation, yet face two critical limitations. First, a significant "domain gap" prevents them from handling soft-tissue ambiguities due to lacking medical texture priors [15, 4]. Second, treating 3D medical data as pseudo-video sequences induces "mechanistic misalignment": SAM 2's intrinsic "object tracking" paradigm (for spatiotemporal consistency) conflicts with medical segmentation's need for "semantic matching" across slices [27, 5]. This causes "semantic drift"—erroneous background tracking after target organ disappearance (i.e., over-tracking) [11, 16]. Consequently, both constraints make direct SAM 2 application to 3D medical slices suboptimal [27, 11, 16].

To bridge the domain gap, adaptations like MedSAM [15] and SAMed [30] employ fully supervised Parameter-Efficient Fine-Tuning (PEFT), ignoring unlabeled data. Recent SSL attempts, such as VESSA [7] and SSS [31], rely on complex external prompting or pseudo-labeling, which increases complexity and risks "confirmation bias." Most critically, the native "mechanistic misalignment" has not been fully addressed by existing methods, particularly in developing an effective mechanism to inject semantic information into SAM 2.

To address these challenges, we propose GU-SAM2, employing a core strategy of decoupling data utilization. In the first stage, we leverage unlabeled medical images for Masked Image Modeling (MIM) pre-training on a lightweight U-Net to capture medical-specific anatomical consistency and texture priors via pixel-level reconstruction [10, 23]. In the second stage, under supervision, we inject these features into SAM 2 via a Gated Feature Fusion module. This mechanism adaptively regulates injection intensity based on image content, significantly enhancing the injection weights of medical semantic features in the anatomical regions of target organs. This design successfully corrects the inherent "object tracking" tendency to resolve mechanistic misalignment. Furthermore, by operating without pseudo-labels, we effectively avoid confirmation bias.

The main contributions of this paper are summarized as follows:

- We propose GU-SAM2, a semi-supervised framework designed for medical image segmentation. It resolves the native "mechanistic misalignment" of SAM 2 by transforming time-based "object tracking" into anatomy-based "semantic matching" via semantic injection.
- We design a label-decoupled strategy featuring a "MIM pre-training + Gated Feature Fusion" paradigm. This strategy leverages unlabeled data to capture anatomical texture priors, and achieves adaptive feature injection via a gating mechanism under supervision. By abandoning pseudo-labels, it effectively avoids "confirmation bias."
- Extensive experiments demonstrate that GU-SAM2 achieves superior performance, particularly maintaining stability on small or challenging organs. Notably, it delivers exceptional robustness even in low-data regimes (e.g., with only 1-3 labeled samples).

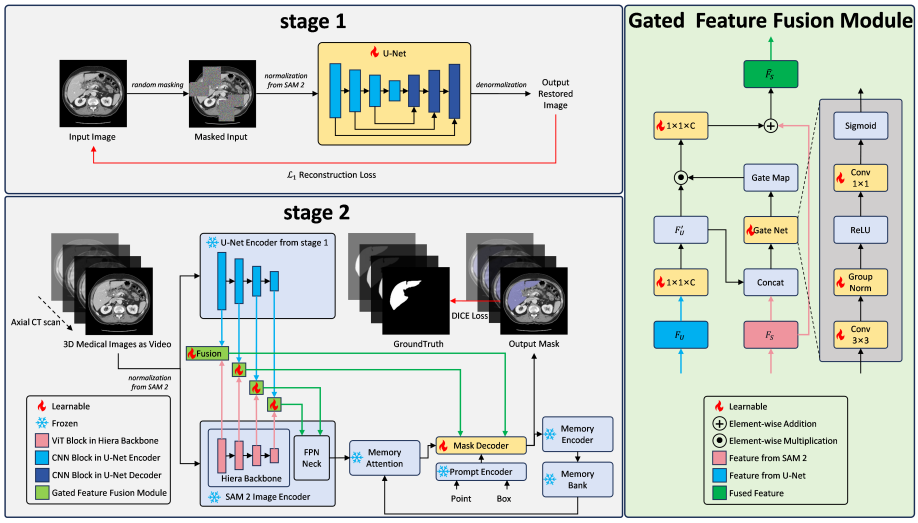

Fig. 1. The architecture of GU-SAM2.

2 Method

2.1 Stage 1: Unsupervised Domain Knowledge Internalization

To bridge the significant domain gap between natural and medical images, we employ Masked Image Modeling (MIM) to pre-train a lightweight U-Net encoder on unlabeled data. We meticulously design the U-Net architecture to consist of four hierarchical convolutional blocks. Crucially, we ensure that the feature maps F_U^i output at each downsampling stage strictly align one-to-one with the feature maps F_S^i of four different scales output by the SAM 2 Hiera (Small) backbone in both spatial resolution and channel dimensions. This structural isomorphism lays the foundation for seamless integration in the second stage, allowing the U-Net to serve as a dedicated medical imaging feature extractor.

The specific pre-training process is as follows: First, we resize each single-channel medical image slice and replicate it along the channel dimension to construct a 3-channel input $I \in \mathbb{R}^{512\times512\times3}$. Subsequently, we apply a random masking strategy to I to generate the masked image I_{mask}. Next, we perform the same normalization processing on I_{mask} as in SAM 2, and input the normalized values into the U-Net for restoration. Finally, the network output undergoes a denormalization operation to revert pixel values. This process prompts the encoder to internalize domain knowledge specific to medical imaging, including tissue textures, boundary geometry, and anatomical consistency. The training objective is to minimize the $\mathcal{L}_1$ reconstruction loss between the restored reconstructed image I_{rec} and the original image I:

$$\mathcal{L}_{rec} = ||I_{rec} - I||_1 \tag{1}$$

This pre-trained U-Net serves as a frozen feature extractor in the second stage, providing domain knowledge specific to medical imaging for SAM 2.

2.2 Stage 2: Supervised Multi-scale Gated Semantic Injection

In the second stage, as illustrated in Fig. 1, we integrate the frozen U-Net encoder into the SAM 2 framework. Crucially, the feature fusion process is executed exclusively after the Hiera image encoder of SAM 2 has completed its forward propagation. This means that we perform semantic injection on the extracted features F_S in a post-processing manner. Let F_S^i and F_U^i denote the frozen features from the i-th stage ($i \in \{1, 2, 3, 4\}$) of SAM 2 and U-Net, respectively. To adaptively inject the texture knowledge learned by the U-Net encoder in Stage 1, we design a Gated Feature Fusion Module (GFFM). First, to align the feature spaces, the U-Net features are projected via a 1×1 convolution:

$$F_U^{'i} = \text{Conv}_{proj}(F_U^i) \tag{2}$$

Subsequently, to achieve dynamic semantic filtering, the aligned features are concatenated with the SAM 2 features and fed into a lightweight gating network. The generated spatial gate map G^i is computed as follows:

$$G^i = \sigma(\text{Conv}_{1\times1}(\delta(\text{GN}(\text{Conv}_{3\times3}([F_S^i; F_U^{'i}]))))) \tag{3}$$

where σ is the Sigmoid function, δ is the ReLU activation, and $[\cdot;\cdot]$ denotes channel concatenation. Specifically, the $\text{Conv}_{3\times3}$ is responsible for neighboring awareness, judging the importance of the current pixel by aggregating information from surrounding pixels; GroupNorm (GN) is employed to address the statistical instability caused by Batch Size=1 in medical image training; the $\text{Conv}_{1\times1}$ compresses the features into a single-channel probability map; and Sigmoid maps these probabilities to the $(0, 1)$ interval. Notably, the obtained spatial gate map G^i maintains the same resolution as F_S^i and $F_U^{'i}$ but with a single channel (i.e., $1 \times H_i \times W_i$), with values ranging in $(0, 1)$. Values approaching 1 indicate the injection of encodings extracted by the U-Net encoder (targeting anatomical regions of the organ), while values approaching 0 indicate suppression (background regions). This injection essentially functions as a dynamic semantic filtering. The final fused feature $\hat{F}_S^i$ is computed as:

$$\hat{F}_S^i = F_S^i + \alpha \cdot (G^i \odot F_U^{'i}) \tag{4}$$

Here, α is a learnable channel scaling factor implemented via a $1\times1\times C$ convolution. It performs scaling along the channel dimension, meaning that one scaling coefficient is shared within the same channel. We adopt a Zero-Initialization strategy ($\alpha = 0$) to preserve the pre-trained representations of SAM 2.

The fused features follow the native routing logic of SAM 2: the high-resolution fused features $(\hat{F}_S^1, \hat{F}_S^2)$ directly enter the Mask Decoder to preserve fine spatial details, while the low-resolution fused features $(\hat{F}_S^3, \hat{F}_S^4)$ are fed into the FPN Neck of the SAM 2 Image Encoder for fusion.

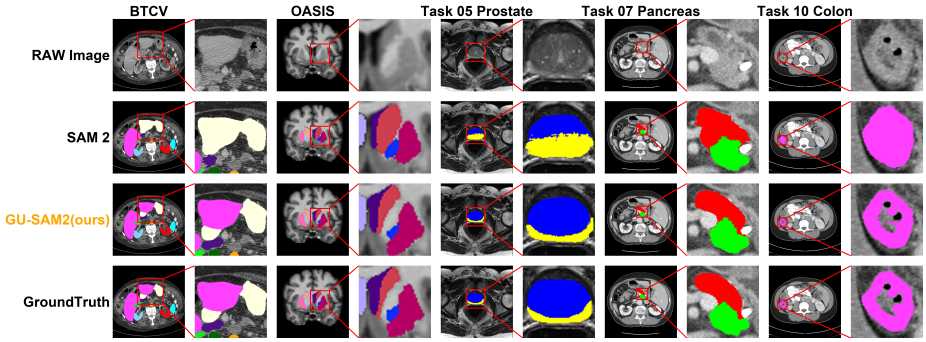

Fig. 2. Qualitative segmentation results comparison on BTCV, OASIS, and MSD datasets.

We reorganize labeled slices into short video sequences. During training, sparse prompts (a box and a point) are provided solely for the first frame, while subsequent frames rely on temporal memory. Adhering to PEFT principles, we freeze most components, fine-tuning only the Mask Decoder and GFFM under Dice Loss supervision. This lightweight design minimizes trainable parameters and overfitting risk, ensuring efficient and stable convergence even with limited data.

3 Experiments

3.1 Datasets and Implementation Details

To evaluate generalization and robustness, we conducted extensive experiments on three benchmark datasets. BTCV [6] (12 abdominal organs, CT) represents standard clinical scenarios; OASIS [17] (14 brain structures, MRI) verifies generalization on MRI modality; and three challenging tasks from MSD [21] (Task 05 Prostate PZ/TZ, Task 07 Pancreas Tumor, Task 10 Colon Tumor) test performance on minute targets and irregular shapes. The Dice Similarity Coefficient (DSC) is used as the metric.

For inference, we implemented tailored strategies to ensure fair comparison. For 2D SAM-based methods, we employ Slice-wise Prompting (providing a Box and Point for every slice). Conversely, for 3D SAM 2-based methods (including GU-SAM2), leveraging their temporal memory, we adopt a First-frame Guidance with Sparse Aid strategy: strict prompts are provided only for the first frame, while subsequent frames receive additional prompts with a probability of 0.25 to simulate efficient clinical interaction, relying otherwise on the Memory module for temporal reasoning.

3.2 Comparison with State-of-the-Arts

To verify performance limits, we conducted evaluations under a fully supervised setting with a strict 80%/20% train-test split, ensuring identical data usage for

Table 1. Quantitative comparison on the BTCV dataset.

Model	Large			Medium				Small/Hard					AVE	
	Spleen	Liver	Stom	R.Kid	L.Kid	Aorta	IVC	Gall	Eso	Veins	Panc	AG	hard	all
nnUNet [13]	0.942	0.948	0.824	0.894	0.910	0.877	0.782	0.704	0.723	0.720	0.680	0.616	0.689	0.802
UNetr [9]	0.968	0.971	0.913	0.924	0.941	0.890	0.847	0.750	0.766	0.788	0.767	0.741	0.762	0.856
Swin-UNetr [8]	0.971	**0.975**	0.921	**0.936**	0.943	0.892	0.853	0.794	0.773	0.812	0.794	0.765	0.788	0.869
TransUNet [3]	0.952	0.969	0.889	0.927	0.929	**0.920**	0.833	0.662	0.757	0.791	0.775	0.637	0.724	0.838
EnsDiff [25]	0.938	0.967	0.910	0.931	0.924	0.869	0.851	0.772	0.771	0.802	0.771	0.745	0.772	0.854
SegDiff [1]	0.954	0.953	**0.927**	0.932	0.926	0.846	0.833	0.738	0.763	0.796	0.782	0.723	0.760	0.847
MedSegDiff [26]	**0.973**	0.973	0.924	0.930	**0.955**	0.907	**0.868**	0.812	0.815	0.825	0.788	0.779	0.804	0.879
SAM [14]	0.504	0.472	0.553	0.662	0.763	0.601	0.410	0.534	0.569	0.535	0.486	0.354	0.496	0.537
SAM 2 [18]	0.528	0.503	0.529	0.705	0.637	0.483	0.394	0.764	0.551	0.618	0.653	0.730	0.663	0.591
MedSAM [15]	0.748	0.887	0.849	0.803	0.871	0.862	0.735	0.756	0.710	0.704	0.756	0.728	0.731	0.784
SAMed [30]	0.841	0.943	0.785	0.698	0.780	0.874	0.782	0.657	0.739	0.771	0.604	0.736	0.701	0.768
SAM-Med2D [4]	0.885	0.947	0.885	0.889	0.930	0.872	0.795	0.813	0.796	0.811	0.784	0.798	0.800	0.851
MedSAM-2 [32]	0.915	0.834	0.872	0.924	0.920	0.894	0.835	0.926	0.908	0.838	0.832	0.900	0.881	0.883
GU-SAM2(ours)	0.921	0.938	0.884	0.931	0.927	0.905	0.850	**0.928**	**0.920**	**0.851**	**0.862**	**0.923**	**0.897**	**0.903**

Table 2. Quantitative comparison on the OASIS dataset.

Model	Basal Ganglia			Large		Small/Hard		AVE	
	Putamen	Caudate	Pallidum	Thalamus	Lat. Ventricle	Hippocampus	Amygdala	hard	all
nnUNet [13]	0.910	**0.903**	0.868	0.918	0.934	0.901	0.842	0.872	0.896
Swin-UNetr [8]	0.894	0.880	0.845	0.905	0.918	0.883	0.821	0.852	0.878
FastSurfer [12]	0.902	0.891	0.854	**0.924**	0.938	0.887	0.835	0.861	0.890
QuickNAT [19]	0.893	0.889	0.866	0.915	**0.942**	0.917	0.859	0.888	0.897
SAM 2 [18]	0.700	0.618	0.707	0.775	0.835	0.715	0.642	0.679	0.713
MedSAM-2 [32]	0.884	0.867	0.869	0.887	0.880	0.848	0.821	0.835	0.865
GU-SAM2(ours)	**0.911**	0.902	**0.921**	0.913	0.933	**0.920**	**0.894**	**0.907**	**0.913**

both Stage 1 pre-training and Stage 2 fine-tuning. As detailed in Tables 1, 2, and 3, GU-SAM2 achieves comprehensive leading performance across BTCV, OASIS, and all challenging MSD tasks (Prostate, Pancreas, Colon Tumor), surpassing existing SOTA methods. A core advantage over traditional models like nnU-Net lies in the robustness to target size, enabled by explicit spatial prompts. We quantified this by calculating the Spearman correlation (r_s) between organ pixel ratios and Dice scores on BTCV. While nnU-Net ($r_s = 0.853$), Swin-UNetr ($r_s = 0.942$), and MedSegDiff ($r_s = 0.893$) exhibit strong positive correlations implying heavy reliance on target size, GU-SAM2 yields a non-significant correlation of $r_s = 0.253$ (P-value $= 0.428$). This indicates that our approach effectively overcomes the long-tailed distribution problem, maintaining stability even for minute or low-contrast structures.

Qualitative results in Fig. 2 further evidence this superiority. Benefiting from the domain knowledge learned by the U-Net encoder, GU-SAM2 successfully delineates targets with blurred boundaries (e.g., Pancreas) or irregular shapes (e.g., Colon Tumor) where the vanilla SAM 2 fails. Moreover, the injected semantic information prevents identity confusion during temporal inference, such as distinguishing the Liver from the Stomach in BTCV cases. This marks that our method successfully elevates the underlying logic from primitive "Object Tracking" to anatomically aware "Semantic Matching."

Table 3. Quantitative comparison on challenging MSD tasks (Task 05, 07, and 10).

Model	Task 05 (Prostate)		Task 07 (Pancreas)		Task 10 (Colon)
	PZ	TZ	Organ	Tumor	Tumor
nnUNet [13]	0.742	0.893	0.824	0.542	0.583
Swin-UNetr [8]	0.729	0.882	0.814	0.531	0.572
SAM 2 [18]	0.613	0.758	0.630	0.676	0.604
MedSAM-2 [32]	0.778	0.861	0.817	0.835	0.826
GU-SAM2(ours)	**0.820**	**0.903**	**0.834**	**0.867**	**0.853**

Table 4. Ablation study on key components of GU-SAM2.

Module Configuration			Dice Score	
U-Net Encoder Init	Fusion Strategy	Decoder Tuning	BTCV	OASIS
None	–	×	0.591	0.713
None	–	✓	0.683	0.768
Random	Add	×	0.307	0.329
Random	GFFM	×	0.582	0.684
Random	Add	✓	0.323	0.356
Random	GFFM	✓	0.656	0.742
MIM	Add	×	0.318	0.335
MIM	GFFM	×	0.585	0.701
MIM	Add	✓	0.814	0.807
MIM	GFFM	✓	**0.903**	**0.913**

3.3 Ablation Study and Interpretability

We conducted strict ablation studies (Table 4) to verify the effectiveness of each module. Results show that the MIM pre-trained U-Net provides crucial texture priors, significantly outperforming random initialization. The GFFM exhibits exceptional "semantic-aware" regulation capabilities: it actively suppresses harmful noise from a Random U-Net (pushing gate values towards 0) to prevent performance collapse, while precisely injecting anatomical features from a MIM U-Net. As visualized in Fig. 3, the gate heatmaps show high responses (Red) exclusively in target anatomical regions while suppressing the background (Blue). This precise, semantically oriented injection serves as the best proof of transforming SAM 2 from low-level "Object Tracking" to high-level "Semantic Matching" in medical image segmentation.

3.4 Data Efficiency Analysis

As shown in Table 5, we investigate data efficiency. First, results confirm a positive correlation between the scale of unlabeled data and final model performance, effectively validating the effectiveness of leveraging large-scale unlabeled data to capture anatomical texture priors. Second, GU-SAM2 demonstrates exceptional robustness with minimal labeled data. Benefiting from domain knowledge

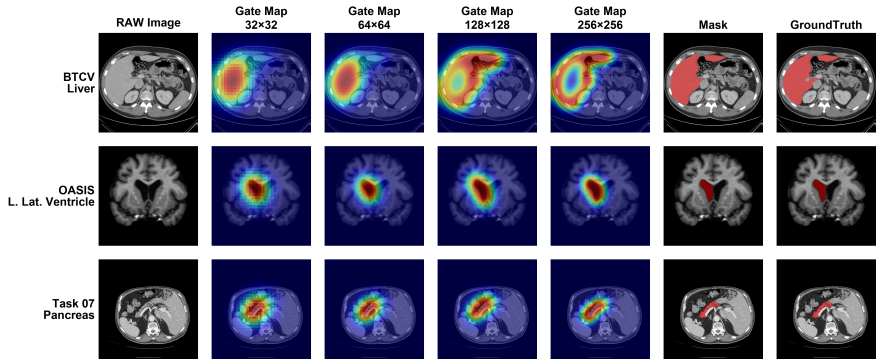

Fig. 3. Visualization of multi-scale gate maps. The $1 \times H \times W$ probability maps (range $[0, 1]$) are displayed using a blue-to-red heatmap, where red and blue indicate values approaching 1 and 0, respectively.

Table 5. Training data efficiency analysis of GU-SAM2 on BTCV and OASIS datasets. Stage 1: Ratio of unlabeled data used for U-Net pre-training. Stage 2: Amount of labeled data used for fine-tuning.

Dataset	Stage 1	Few-shot (Stage 2)			Ratio (Stage 2)		
	Data	1 shot	3 shots	5 shots	10%	50%	100%
OASIS	10%	0.820	0.837	0.843	0.858	0.875	0.881
	50%	0.829	0.856	0.871	0.882	0.894	0.905
	Full	0.846	0.879	0.885	0.898	0.907	0.913
BTCV	10%	0.718	0.727	0.743	0.727	0.772	0.825
	50%	0.782	0.811	0.819	0.811	0.845	0.876
	Full	0.837	0.875	0.882	0.875	0.894	0.903

injected via MIM pre-training, the model adapts rapidly with minimal annotations. This indicates that our "label-decoupled" strategy significantly reduces the reliance on expensive medical annotations, offering high value for clinical deployment.

4 Conclusion

We proposed GU-SAM2 to address the "domain gap" and "mechanistic misalignment" in SAM 2. Central to our approach is a data decoupling strategy: leveraging unlabeled data for MIM pre-training to capture anatomical priors, and injecting semantics via Gated Feature Fusion under supervision. This design transforms "object tracking" into "semantic matching." Extensive experiments across multiple datasets demonstrate that GU-SAM2 achieves SOTA performance in multi-modality scenarios. Notably, it maintains stable segmentation on small or challenging organs and exhibits exceptional data efficiency in extremely low-data regimes.

Acknowledgments. This study was supported in part by the Joint Funds of the National Natural Science Foundation of China (No. U24A20753), the general program of National Natural Science Fund of China (No. 81971693), the Young Scientists Fund of the National Natural Science Foundation of China (No. 62403103), the China Postdoctoral Science Foundation (No. 2025M772947), the Fundamental Research Funds for the Central Universities (DUT24RC(3)050), the funding of Liaoning Key Lab of IC & BME System and Dalian Engineering Research Center for Artificial Intelligence in Medical Imaging.

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

References

1. Amit, T., Shaharbany, T., Nachmani, E., Wolf, L.: Segdiff: Image segmentation with diffusion probabilistic models. arXiv preprint arXiv:2112.00390 (2021)
2. Arazo, E., Ortego, D., Albert, P., O'Connor, N.E., McGuinness, K.: Pseudo-labeling and confirmation bias in deep semi-supervised learning. In: 2020 International joint conference on neural networks (IJCNN). pp. 1–8. IEEE (2020)
3. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
4. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., et al.: Sam-med2d. arXiv e-prints pp. arXiv–2308 (2023)
5. Cuttano, C., Trivigno, G., Averta, G., Masone, C.: Sansa: Unleashing the hidden semantics in sam2 for few-shot segmentation. arXiv preprint arXiv:2505.21795 (2025)
6. Fang, X., Yan, P.: Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction. IEEE Transactions on Medical Imaging **39**(11), 3619–3629 (2020)
7. Guo, J., Li, M., Su, H., López, S., Fan, L., Kim, D., Katsaggelos, A.: Vision-language enhanced foundation model for semi-supervised medical image segmentation. arXiv preprint arXiv:2511.19759 (2025)
8. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI brainlesion workshop. pp. 272–284. Springer (2021)
9. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 574–584 (2022)
10. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
11. He, Y., Guo, P., Tang, Y., Myronenko, A., Nath, V., Xu, Z., Yang, D., Zhao, C., Xu, D., Li, W.: A short review and evaluation of sam2's performance in 3d ct image segmentation. arXiv preprint arXiv:2408.11210 (2024)
12. Henschel, L., Conjeti, S., Estrada, S., Diers, K., Fischl, B., Reuter, M.: Fastsurfer-a fast and accurate deep learning based neuroimaging pipeline. NeuroImage **219**, 117012 (2020)

13. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
14. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
15. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
16. 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)
17. Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L.: Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. Journal of cognitive neuroscience **19**(9), 1498–1507 (2007)
18. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
19. Roy, A.G., Conjeti, S., Navab, N., Wachinger, C., Initiative, A.D.N., et al.: Quicknat: A fully convolutional network for quick and accurate segmentation of neuroanatomy. NeuroImage **186**, 713–727 (2019)
20. Shen, D., Wu, G., Suk, H.I.: Deep learning in medical image analysis. Annual review of biomedical engineering **19**(1), 221–248 (2017)
21. Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063 (2019)
22. Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J.N., Wu, Z., Ding, X.: Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. Medical image analysis **63**, 101693 (2020)
23. Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A.: Self-supervised pre-training of swin transformers for 3d medical image analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20730–20740 (2022)
24. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
25. Wolleb, J., Sandkühler, R., Bieder, F., Valmaggia, P., Cattin, P.C.: Diffusion models for implicit image segmentation ensembles. In: International conference on medical imaging with deep learning. pp. 1336–1348. PMLR (2022)
26. Wu, J., Fu, R., Fang, H., Zhang, Y., Yang, Y., Xiong, H., Liu, H., Xu, Y.: Medsegdiff: Medical image segmentation with diffusion probabilistic model. In: Medical Imaging with Deep Learning. pp. 1623–1639. PMLR (2024)
27. Xu, Q., Zhu, L., Liu, X., Lin, G., Long, C., Li, Z., Zhao, R.: Unlocking the power of sam 2 for few-shot segmentation. arXiv preprint arXiv:2505.14100 (2025)
28. Yang, L., Qi, L., Feng, L., Zhang, W., Shi, Y.: Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7236–7246 (2023)
29. Yu, L., Wang, S., Li, X., Fu, C.W., Heng, P.A.: Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In: International confer-

ence on medical image computing and computer-assisted intervention. pp. 605–613. Springer (2019)
30. Zhang, K., Liu, D.: Customized segment anything model for medical image segmentation. arXiv preprint arXiv:2304.13785 (2023)
31. Zhu, H., Liu, X., Xue, R., Zhang, Z., Xu, Y., Ergu, D., Cai, Y., Zhao, Y.: Sss: Semi-supervised sam-2 with efficient prompting for medical imaging segmentation. arXiv preprint arXiv:2506.08949 (2025)
32. Zhu, J., Hamdi, A., Qi, Y., Jin, Y., Wu, J.: Medical sam 2: Segment medical images as video via segment anything model 2. arXiv preprint arXiv:2408.00874 (2024)