Reliability-Aware Cross-Prompt Aggregation for Propagation-Based Segmentation of Arbitrary 3D Medical Objects

Haoshen Li[1], Zifan Chen[1(✉)], Guangzhengao Yang[2], Guanyu Chen[3], Ziling Lin[2], Xinyu Nan[2], Jie Zhao[4,5], Li Zhang[2(✉)]

[1]Center for Machine Learning Research, Peking University, Beijing, China
[2]Center for Data Science, Peking University, Beijing, China
[3]College of Engineering, Peking University, Beijing, China
[4]National Engineering Laboratory for Big Data Analysis and Applications, Peking University, Beijing, China
[5]Peking University Changsha Institute for Computing and Digital Economy, Changsha, China
czifan@pku.edu.cn, zhangli_pku@pku.edu.cn

Abstract. Segment Anything Models (SAM) enable prompt-driven medical image segmentation with minimal annotation. Propagation-based methods extend this paradigm to volumetric data by iteratively propagating semantic guidance from a single prompted slice. However, existing approaches suffer from two critical limitations: error accumulation, where early prediction errors compound, and fixed propagation strategies that fail to adapt to heterogeneous anatomical structures. To address these issues, we propose a training-free Reliability-Aware Cross-Prompt Aggregation framework. Instead of relying on a single evolving prompt, our method jointly leverages two complementary prompts at each propagation step: the initial prompted slice providing stable global guidance, and the dynamically propagated front slice capturing local appearance. Their predictions are adaptively fused through a reliability-aware mechanism that combines spatial distance weighting with an unsupervised estimation of prompt reliability. This allows the model to intelligently down-weight unreliable predictions and prevent error compounding. Experiments on ten external multi-modal medical imaging datasets demonstrate that our approach consistently outperforms state-of-the-art Medical SAM baselines across all six evaluation metrics, achieving an average Dice score improvement of 4.6% over the best baseline.

Keywords: Propagation-based 3D medical image segmentation · Cross-prompt aggregation · Reliability-aware training-free inference

1 Introduction

Volumetric medical image segmentation is a fundamental task in medical image analysis and underpins a wide range of clinical applications, including diagnosis [6, 13], surgical planning [8, 14], and treatment monitoring [9]. Unlike 2D segmentation, accurate 3D segmentation requires maintaining both spatial and semantic consistency across adjacent slices, while accounting for substantial appearance variations along the anatomical axis. These inter-slice dependencies and structural complexities make volumetric segmentation considerably more challenging than its 2D counterpart.

Despite steady progress in automated methods, manual delineation remains prevalent in clinical practice [4, 20], largely due to the limited generalization of existing models across different anatomies, imaging modalities, and disease patterns. The high cost of expert annotations and the scarcity of labeled 3D data further motivate the development of more general and annotation-efficient segmentation frameworks [16]. Inspired by foundation models, the Segment Anything Model (SAM) [12] demonstrated strong generalization in natural images by enabling prompt-based segmentation with minimal user input. This success has spurred growing interest in adapting SAM-style models to medical imaging scenarios.

Recent SAM-based approaches for medical image segmentation [5, 7, 16, 17] can be broadly categorized into three paradigms. The first paradigm performs slice-wise segmentation by treating each 2D slice independently, as exemplified by MedSAM [16]; while simple and flexible, this strategy often struggles to maintain volumetric consistency without repeated prompting. The second paradigm extends SAM to volumetric inputs by predicting 3D patches, such as SegVol [7], which captures 3D context but incurs higher computational cost and may still face generalization challenges.

The third paradigm, propagation-based methods, addresses the limitations of the first two paradigms by generating complete 3D segmentations from a minimal 2D prompt through iterative slice-wise propagation [5, 17]. PAM [5] exemplifies this approach, leveraging CNN-based intra-slice representations and Transformer cross-attention to propagate semantic guidance across adjacent slices. This design enables strong volumetric coherence and cross-object generalization without task-specific retraining, effectively addressing the volumetric consistency issues of slice-wise methods while avoiding the high computational cost of 3D patch-based approaches. However, propagation-based methods such as PAM inherently suffer from two critical limitations. First, error accumulation occurs as small inaccuracies in early predictions are progressively amplified during iterative propagation, degrading segmentation quality in distant slices, as shown in the *PAM Propagate* part of Figure 1. Second, existing methods typically rely on rigid, distance-based update heuristics (e.g., refreshing prompts strictly every 20mm). These fixed strategies are completely blind to anatomical variability and local prediction reliability, severely limiting their adaptability to heterogeneous organ structures with abrupt morphological changes.

To address these critical limitations, we extend PAM [5] in a training-free manner and propose the Reliability-Aware Cross-Prompt Aggregation framework for arbitrary 3D medical objects segmentation. Unlike existing methods that rely on a single evolving prompt, our framework effectively mitigates error accumulation by jointly leveraging global semantic guidance from the ini-

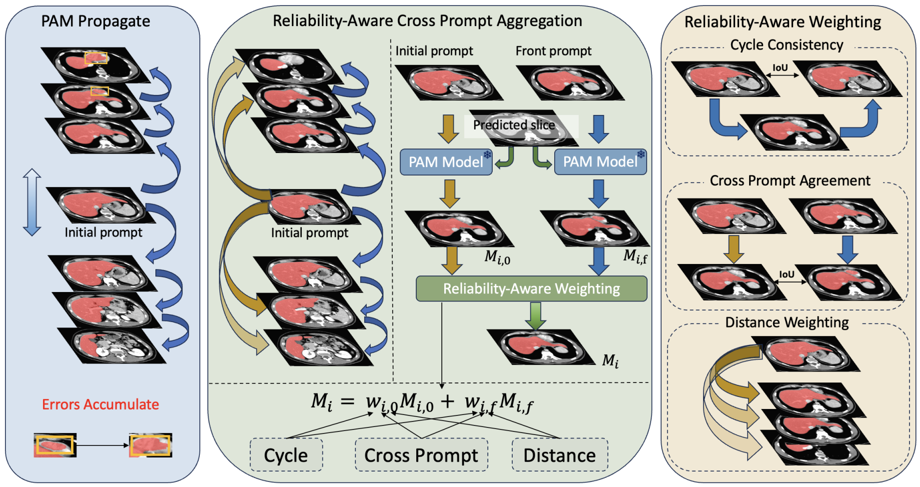

Fig. 1. Overview of the Reliability-Aware Cross-Prompt Aggregation framework. The method leverages two complementary prompts (initial prompted slice and dynamically propagated front slice) and fuses their predictions through distance-aware and confidence-aware weighting mechanisms.

tial prompted slice and local anatomical cues from a dynamically propagated front slice. To seamlessly integrate these complementary signals, we design a reliability-aware dynamic fusion mechanism that eliminates the need for rigid update heuristics. Instead, our method adaptively weights prompt predictions based on their spatial distance and an unsupervised estimation of prompt reliability, which is rigorously evaluated via two unsupervised metrics: cycle consistency and cross-prompt agreement. Ultimately, extensive experiments across ten diverse external medical imaging datasets demonstrate the superior zero-shot generalization of our approach. Without requiring any task-specific retraining, our method consistently outperforms state-of-the-art Medical SAM baselines across all six evaluation metrics, achieving a notable 4.6% improvement in Dice score.

2 Method

2.1 Reliability-Aware Cross-Prompt Aggregation Framework

Figure 1 illustrates the overall workflow of our Reliability-Aware Cross-Prompt Aggregation framework. Given a volumetric medical image, our framework propagates semantic guidance along the anatomical axis by jointly leveraging two complementary prompts at each step: the initial prompted slice, which provides stable semantic guidance throughout propagation, and the dynamically propagated front slice, which captures region-specific appearance near the propagation frontier. For a target slice i, predictions are obtained independently from both prompts and fused through a reliability-aware weighting mechanism:

$$M_i = w_{i,0}\, M_{i,0} + w_{i,f}\, M_{i,f}, \tag{1}$$

where $M_{i,0}$ and $M_{i,f}$ denote the predictions produced by the frozen PAM model given the initial prompt and the front prompt for slice i, while $w_{i,0}$ and $w_{i,f}$ represent the fusion weights of the two input prompts for slice i, which are derived from proposed distance-aware and confidence-aware weighting mechanism (detailed in the following section):

$$w_{i,j} = w_{\text{dist}}(i,j)\, w_{\text{conf}}(j), \tag{2}$$

where $j \in \{0, f\}$ indexes the two prompts, $w_{\text{dist}}(i,j)$ accounts for spatial relevance between slices, and $w_{\text{conf}}(j)$ measures the confidence of prompt j by averaging cycle consistency and cross-prompt prediction agreement:

$$w_{\text{conf}}(j) = \frac{C_{\text{cycle}}(j) + C_{\text{cross-prompt}}(j)}{2}, \tag{3}$$

The two composite reliability scores $w_{i,0}$ and $w_{i,f}$ are subsequently normalized so that their final fusion weights sum to one.

Eq. 1 naturally induces a recursive propagation process. In this way, the proposed Reliability-Aware Cross-Prompt Aggregation framework dynamically balances robustness and adaptivity, mitigating error accumulation while avoiding rigid, distance-agnostic update strategies, thereby better adapting to heterogeneous organs with diverse shapes, sizes, and appearance patterns. As illustrated in Figure 1, compared with the original PAM propagation, our method effectively alleviates error accumulation and achieves more accurate volumetric segmentation.

2.2 Distance-Aware Weighting Scheme

To explicitly encode the spatial relevance of each prompt, we propose a distance-aware weighting scheme to smoothly attenuate the influence of distant prompts while emphasizing the spatially closer front slice:

$$w_{\text{dist}}(i,j) = \exp\left(-\frac{|i-j| \cdot s_z}{\tau}\right), \quad j \in \{0, f\}. \tag{4}$$

where s_z denotes the slice spacing in physical units, and τ is a temperature parameter controlling the decay rate (set to 100.0 in our experiments). The two distance scores $w_{\text{dist}}(i,0)$ and $w_{\text{dist}}(i,f)$ are then normalized to sum to one before being used in the fusion weights.

This formulation assigns higher weights to spatially closer prompts while smoothly attenuating the influence of distant ones. By incorporating physical distance into the fusion process, the proposed scheme enforces locality-aware propagation, enabling adaptive emphasis on recent predictions without discarding the stabilizing guidance provided by the initial prompt.

2.3 Confidence-Aware Weighting Scheme

To mitigate error accumulation caused by propagating unreliable predictions, we introduce two unsupervised reliability metrics—cycle consistency and cross-prompt prediction agreement—to assess the trustworthiness of the propagated front slice and down-weight predictions with low confidence, thereby preventing unreliable slices from dominating subsequent propagation. Given that the new propogated front slice M_i has been obtained by fusing predictions from the initial prompt and the previous front prompt, we then compute cycle-consistency score $C_{\text{cycle}}(i)$ and cross-prompt agreement score $C_{\text{cross-prompt}}(i)$ for M_i, which together serve as slice-level reliability metrics to guide subsequent propagation.

Cycle Consistency. Cycle consistency measures how well a predicted slice can be reversibly propagated back to the original prompt slices that were used to generate it. Specifically, we use M_i as input to propagate back to the two prompt masks M_0 and M_f, and take the average of predicted IoU scores as the cycle consistency of M_i:

$$C_{\text{cycle}}(i) = \frac{1}{2}\Big(\text{IoU}(M_{i\to 0},\, M_0) + \text{IoU}(M_{i\to f},\, M_f)\Big), \tag{5}$$

This metric quantifies the self-consistency of the prediction: high cycle consistency values indicate that the slice prediction accurately preserves semantic information from both prompt sources and can be reliably used for subsequent propagation, while low values signal potential errors that should be down-weighted to prevent error accumulation.

Cross-Prompt Prediction Agreement. While cycle consistency captures self-reliability, it does not reflect agreement across different prompts. To bridge this gap, we introduce a cross-prompt consensus metric by evaluating the IoU overlap between the predictions from the two prompts:

$$C_{\text{cross-prompt}}(i) = \text{IoU}(M_{i,0},\, M_{i,f})\,, \tag{6}$$

Predictions affected by drift or local failure tend to disagree with the other prompt, resulting in lower agreement scores. Conversely, high cross-prompt agreement indicates that both prompts produce consistent predictions, which serves as strong evidence for prediction reliability.

The final confidence score for slice M_i is obtained by averaging these two measures as defined in Eq. 3. By jointly modeling cycle-consistency and cross-prompt agreement, the proposed confidence-aware weighting provides a principled and training-free reliability estimation, effectively attenuating unreliable predictions while preserving stable semantic guidance during propagation.

3 Experiments

Datasets. We evaluate our method on ten external medical imaging datasets that are consistent with those used in PAM [5], covering diverse modalities and anatomical structures. The datasets include: D1: Adrenal-ACC-Ki67-Seg [1]

Table 1. Performance comparison of different models on external datasets. Values are reported as mean ± standard deviation.

Methods	DSC ↑	Sensitivity ↑	Specificity ↑	HD95 ↓	NSD ↑	ASD ↓
SegVol	0.590 ± 0.218	0.587 ± 0.229	0.997 ± 0.004	15.457 ± 19.783	0.482 ± 0.195	4.255 ± 5.377
MedSAM	0.515 ± 0.156	0.567 ± 0.240	0.994 ± 0.007	27.951 ± 39.774	0.526 ± 0.263	6.383 ± 9.007
MedSAM2	0.534 ± 0.200	0.785 ± 0.171	0.988 ± 0.026	12.800 ± 7.013	0.606 ± 0.102	2.753 ± 1.383
PAM	0.785 ± 0.098	0.845 ± 0.100	0.997 ± 0.004	7.669 ± 5.992	0.774 ± 0.101	1.463 ± 0.897
Ours	**0.831** ± 0.101	**0.853** ± 0.099	**0.999** ± 0.001	**6.974** ± 6.361	**0.824** ± 0.102	**1.258** ± 0.947

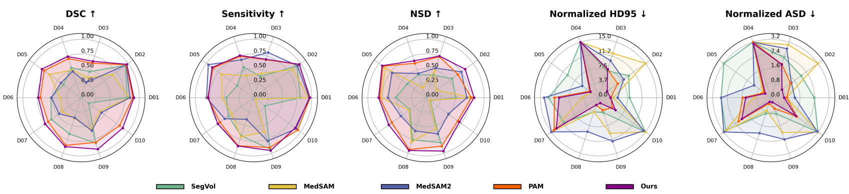

Fig. 2. Radar charts across ten datasets with five evaluation metrics.

(CT), D2: CHAOS-CT [10] (CT), D3: HaN-Seg [19] (CT), D4: HCC-TACE-Seg [18] (CT), D5: LNQ2023 [11] (CT), D6: QUBIQ [2] (CT), D7: WORD [15] (CT), D8: ACDC [3] (MR), D9: CHAOS-MR [10] (MR), and D10: MouseKidney-SRX (SRX). For preprocessing, we use 2D mask prompts as input. Images are cropped using a dynamic crop ratio of 1.25, where the crop region is determined based on the mask bounding box and then resized to 224×224.

Implementation Details. Our method is training-free and built upon the PAM [5]. We use pre-trained PAM model weights and apply our Reliability-Aware Cross-Prompt Aggregation framework during inference. All experiments are conducted on a NVIDIA A800 GPU. Considering that the initial 2D prompt typically exhibits high reliability (annotated by medical experts in practical scenarios), and that there is no reference slice available for computing reliability metrics for the initial prompt, we set its confidence weight to 1. The distance temperature coefficient τ for distance weighting is set to 100.0.

Comparison Methods. We compare our method against state-of-the-art segmentation methods: SegVol [7], MedSAM [16], MedSAM2 [17], and PAM [5]. All methods are evaluated on the ten external datasets without fine-tuning.

Evaluation Metrics. We report six standard segmentation metrics: Dice Similarity Coefficient (DSC), Hausdorff Distance at the 95th percentile (HD95), Average Surface Distance (ASD), Normalized Surface Dice (NSD), Sensitivity, and Specificity. Higher values indicate better performance for DSC, NSD, Sensitivity, and Specificity, while lower values are better for HD95 and ASD.

Comparison with Baseline Methods. Table 1 presents the quantitative results on six metrics averaged across the ten datasets. Our method consistently outperforms all baselines across all metrics, achieving a mean Dice score of 0.831 (4.6% improvement over PAM's 0.785), Sensitivity of 0.853 (0.8% over PAM), and the highest Specificity (0.999) among all methods. For boundary metrics, our method achieves the best HD95 (6.974 mm) and ASD (1.258 mm), outperform-

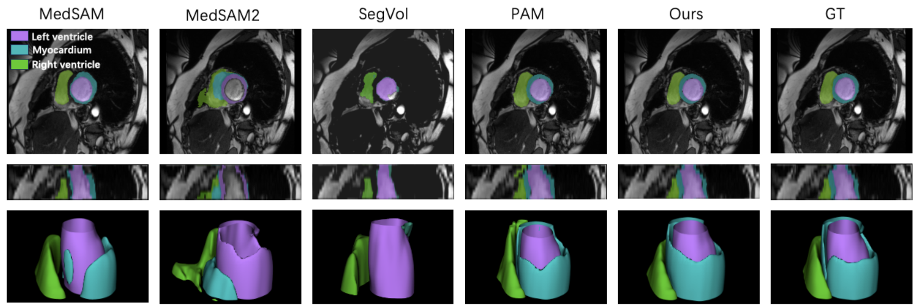

Fig. 3. Qualitative segmentation results comparing different methods.

ing PAM (7.669 mm and 1.463 mm, respectively), with NSD of 0.824 compared to PAM's 0.774. Moreover, statistical analysis confirms that our method significantly outperforms the strongest baseline, PAM, in terms of Dice score, with all per-dataset paired Wilcoxon signed-rank tests yielding $p < 1e-5$. Figure 2 visualizes performance across all ten datasets using radar charts for each metric, revealing consistent superiority across diverse datasets and modalities, demonstrating the effectiveness of our proposed method.

Figure 3 presents qualitative segmentation results across different methods. We can see that PAM and our method demonstrate significantly superior segmentation quality compared to the other methods, highlighting the effectiveness of iterative semantic propagation for volumetric segmentation. Notably, in the Right ventricle case, PAM suffers from error accumulation, causing over-propagation beyond the true boundary. In contrast, our confidence-aware weighting mechanism effectively identifies reliable stopping points during propagation, resulting in more accurate boundary localization and improved segmentation fidelity.

Ablation Studies. Table 2 shows ablation results averaged across all ten datasets. In fixed-distance propagation, performance improves as update distance increases (5mm: Dice 0.745, 20mm: 0.785), and using only the initial prompt achieves the best performance (Dice: 0.803), confirming that excessive prompt updates exacerbate error accumulation. However, the single-prompt approaches have limitations in adapting to region-specific appearance changes for front slices. Cross-prompt fusion strategies (Intersection: Dice 0.815, Average: 0.820) outperform the best fixed strategy (0.803), demonstrating the effectiveness of cross-prompt aggregation. The Average fusion achieves 1.7% Dice improvement over initial prompt only. Moreover, Distance weighting alone achieves Dice 0.825 (0.5% over Average), while confidence weighting alone reaches 0.826 (0.6% over Average). Combining both achieves the best performance (Dice: 0.831, HD95: 6.974 mm, NSD: 0.824), representing 2.8% and 1.1% Dice improvement over single-prompt methods and Average fusion, respectively, demonstrating the complementary benefits of spatial relevance modeling and reliability estimation.

Table 2. Ablation study results averaged across ten datasets.

Method	DSC ↑	Sensitivity ↑	Specificity ↑	HD95 ↓	NSD ↑	ASD ↓
Propagation Distance						
5mm updates	0.745	0.835	0.995	10.541	0.729	1.989
20mm updates	0.785	0.845	0.997	7.669	0.774	1.463
Initial prompt only	0.803	0.834	0.998	7.377	0.801	1.370
Fusion Strategy						
Intersection	0.815	0.799	0.999	9.484	0.770	1.661
Union	0.665	**0.871**	0.987	12.405	0.736	2.131
Average	0.820	0.857	0.998	6.999	0.822	1.282
Weighting Components						
Distance only	0.825	0.860	0.998	7.184	0.822	1.294
Confidence only	0.826	0.854	0.998	7.109	0.822	1.281
Distance + Confidence	**0.831**	0.853	**0.999**	**6.974**	**0.824**	**1.258**

Table 3. Correlation coefficients between reliability metrics and Dice scores.

Type	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	Mean
Cycle Consistency											
Pearson	0.6982	0.7153	0.6156	0.6862	0.6154	0.6630	0.4919	0.6295	0.6735	0.4734	0.6262
Spearman	0.7733	0.6122	0.6767	0.7532	0.6878	0.7028	0.7026	0.8151	0.6805	0.5316	0.6936
Cross-Prompt Agreement											
Pearson	0.7350	0.7760	0.6050	0.5562	0.6173	0.6299	0.4310	0.5909	0.6098	0.4988	0.6050
Spearman	0.7881	0.7367	0.6711	0.5900	0.6758	0.7133	0.6097	0.6815	0.7389	0.5219	0.6727

Reliability Metrics Analysis. We analyze the relationship between our proposed reliability metrics (cycle consistency and cross-prompt prediction agreement) and actual slice segmentation performance (Dice score) across all slices in the ten datasets. Table 3 presents Pearson and Spearman correlation coefficients for each dataset. Cycle consistency shows notable positive correlations with mean Pearson and Spearman coefficients of 0.626 and 0.694, respectively. Cross-prompt prediction agreement also demonstrates consistent positive correlations, with mean coefficients of 0.605 (Pearson) and 0.673 (Spearman). These notable positive correlations validate the effectiveness of our unsupervised reliability assessment mechanism in predicting segmentation quality without requiring ground truth labels, enabling effective identification and prioritization of more reliable predictions during aggregation.

4 Conclusion

We proposed a training-free Reliability-Aware Cross-Prompt Aggregation framework to extend PAM for reducing error accumulation and improving anatomical adaptability in 3D medical image segmentation. Our method jointly leverages two complementary prompts (initial and propagated front slices) at each step and adaptively fuses their predictions through a reliability-aware weighting mechanism that combines distance-aware weighting to model spatial relevance and confidence-aware weighting to estimate prompt reliability using cycle consistency and cross-prompt prediction agreement. Extensive experiments on ten external datasets demonstrate significant improvements over state-of-the-art Medical SAM methods. Furthermore, our proposed two reliability metrics exhibit notable positive correlations with ground-truth slice Dice scores, validating their effectiveness in identifying reliable predictions without requiring supervision.

Acknowledgement Zifan Chen was partly supported by the China Postdoctoral Science Foundation (2025M782893). Li Zhang was partly supported by NSFC 72573036, Clinical Medicine Plus X-Young Scholars Project of Peking University PKU2023LCXQ041 and Capital's Funds for Health Improvement and Research (2026-1Q-1093).

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

References

1. Ahmed, A., Elmohr, M., Fuentes, D., Habra, M., Fisher, S., Perrier, N., Zhang, M., Elsayes, K.: Radiomic mapping model for prediction of ki-67 expression in adrenocortical carcinoma. Clinical Radiology **75**(6), 479–e17 (2020)
2. Becker, A.S., Chaitanya, K., Schawkat, K., Muehlematter, U.J., Hötker, A.M., Konukoglu, E., Donati, O.F.: Variability of manual segmentation of the prostate in axial t2-weighted mri: A multi-reader study. European journal of radiology **121**, 108716 (2019)
3. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al.: Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE transactions on medical imaging **37**(11), 2514–2525 (2018)
4. Chen, H., Chen, Z., Zhao, J., Li, H., Li, J., Liu, Y., Yuan, M., Bao, P., Nan, X., Dong, B., et al.: Msi-unet: A flexible unet-based multi-scale interactive framework for 3d gastric tumor segmentation on ct scans. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2024)
5. Chen, Z., Nan, X., Li, J., Zhao, J., Li, H., Lin, Z., Li, H., Chen, H., Liu, Y., Tang, L., et al.: Pam: a propagation-based model for segmenting any 3d objects across multi-modal medical images. npj Digital Medicine (2025)
6. De Fauw, J., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., et al.: Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature medicine **24**(9), 1342–1350 (2018)
7. 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)
8. Ferrari, V., Carbone, M., Cappelli, C., Boni, L., Melfi, F., Ferrari, M., Mosca, F., Pietrabissa, A.: Value of multidetector computed tomography image segmentation for preoperative planning in general surgery. Surgical endoscopy **26**(3), 616–626 (2012)

9. He, M., Chen, Z.F., Zhang, L., Gao, X., Chong, X., Li, H.s., Shen, L., Ji, J., Zhang, X., Dong, B., et al.: Associations of subcutaneous fat area and systemic immune-inflammation index with survival in patients with advanced gastric cancer receiving dual pd-1 and her2 blockade. Journal for immunotherapy of cancer **11**(6), e007054 (2023)
10. Kavur, A.E., Gezer, N.S., Barış, M., Aslan, S., Conze, P.H., Groza, V., Pham, D.D., Chatterjee, S., Ernst, P., Özkan, S., et al.: Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. Medical image analysis **69**, 101950 (2021)
11. Kikinis, E.Z.R., Pieper, S.: Mediastinal lymph node quantification (lnq): Segmentation of heterogeneous ct data. In: International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI). vol. 5 (2023)
12. 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)
13. Li, H., Ai, T., Wang, Y., Ji, Z., Yu, Q., Lu, L., Dong, B., Zhang, L., Ye, X., Zhao, K., et al.: Lymph node metastasis classification with prototype-guided multiple instance aggregation and heterogeneous feature fusion. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 312–321. Springer (2025)
14. Li, H., Wang, Y., Yu, Q., Zhu, J., Yan, K., Guo, D., Lu, L., Dong, B., Zhang, L., Ye, X., et al.: Metastatic lymph node station classification in esophageal cancer via prior-guided supervision and station-aware mixture-of-experts. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 360–370. Springer (2025)
15. Luo, X., Liao, W., Xiao, J., Chen, J., Song, T., Zhang, X., Li, K., Metaxas, D.N., Wang, G., Zhang, S.: Word: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image. Medical Image Analysis **82**, 102642 (2022)
16. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature communications **15**(1), 654 (2024)
17. 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)
18. Morshid, A., Elsayes, K.M., Khalaf, A.M., Elmohr, M.M., Yu, J., Kaseb, A.O., Hassan, M., Mahvash, A., Wang, Z., Hazle, J.D., et al.: A machine learning model to predict hepatocellular carcinoma response to transcatheter arterial chemoembolization. Radiology: Artificial Intelligence **1**(5), e180021 (2019)
19. Podobnik, G., Ibragimov, B., Tappeiner, E., Lee, C., Kim, J.S., Mesbah, Z., Modzelewski, R., Ma, Y., Yang, F., Rudecki, M., et al.: Han-seg: the head and neck organ-at-risk ct and mr segmentation challenge. Radiotherapy and Oncology **198**, 110410 (2024)
20. Yu, Q., Wang, Y., Yan, K., Li, H., Guo, D., Zhang, L., Shen, N., Wang, Q., Ding, X., Lu, L., et al.: Effective lymph nodes detection in ct scans using location debiased query selection and contrastive query representation in transformer. In: European Conference on Computer Vision. pp. 180–198. Springer (2024)