Dino U-Net: Exploiting High-Fidelity Dense Features from Foundation Models for Medical Image Segmentation

Haoyue Li[1,2†], Yifan Gao[1,3†], Feng Yuan[1,2], Xiaosong Wang[3], Jinjiang Cui[2*], and Xin Gao[2*]

[1] School of Biomedical Engineering (Suzhou), Division of Life Science and Medicine, University of Science and Technology of China, Hefei, China
[2] Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences, Suzhou, China
[3] Shanghai Innovation Institute, Shanghai, China

[†] These authors contributed equally to this work.
[*] Corresponding authors: Xin Gao xingaosam@163.com, Jinjiang Cui cuijj@sibet.ac.cn.

Abstract. The emergence of large-scale pre-trained visual foundation models offers a promising paradigm for medical image segmentation. However, effectively transferring their rich, generic representations to domain-specific clinical tasks remains a significant challenge. This paper presents Dino U-Net, a novel encoder-decoder framework designed to fully leverage the high-fidelity dense features from the DINOv3 foundation model for medical image segmentation. We employ a frozen DINOv3 backbone as the encoder and introduce a dual-branch DINO Adapter to bridge the feature domain gap. To mitigate the loss of fine-grained information during feature dimensionality reduction, we further propose a Fidelity-Aware Projection Module (FAPM). This module utilizes low-rank shared projection coupled with dynamic feature modulation to refine and faithfully propagate features to the decoder. Extensive experiments on seven public medical image datasets demonstrate that Dino U-Net achieves state-of-the-art performance across diverse imaging modalities, excelling in both regional segmentation accuracy and boundary delineation. Moreover, segmentation performance shows consistent improvement as the backbone scales up to 7 billion parameters, confirming the framework's strong scalability. The code is available at https://github.com/yifangao112/DinoUNet.

Keywords: Medical Image Segmentation · Foundation Model · DINOv3 · Feature Adaptation · Transfer Learning.

1 Introduction

Medical image segmentation serves as a cornerstone of computer-aided diagnosis, enabling quantitative analysis for disease screening, treatment planning,

and longitudinal monitoring [3, 5–7, 23]. Consequently, the accuracy and robustness of segmentation algorithms are of paramount importance. Deep learning—particularly the U-Net architecture—has revolutionized this field [22]. With its efficient encoder-decoder structure and skip connections, U-Net and its variants have become the dominant paradigm, demonstrating consistent performance across diverse imaging modalities [12].

Notably, foundation models like the Segment Anything Model (SAM) [11] exhibit impressive zero-shot generalization, with various adaptations for medical tasks [8,27]. However, SAM's reliance on extensive mask-supervision often biases it toward geometric contours, leading to struggles with nuanced semantic relationships or ambiguous anatomical boundaries. In contrast, the self-supervised DINO series [19, 25, 29] learns structural patterns and semantic consistency via masked image modeling. By prioritizing semantic density over geometric contours, DINO offers greater robustness for variable-shaped structures. Specifically, the latest DINOv3 mitigates dense feature degradation during large-scale training, providing high-fidelity representations crucial for localizing fine-grained anatomical structures.

In this paper, we propose Dino U-Net, a novel hybrid architecture designed to fully harness the rich, dense features of DINOv3 for medical image segmentation. We construct an encoder using a frozen DINOv3 backbone and introduce a dual-branch DINO Adapter. Furthermore, to preserve fine-grained information during feature dimensionality reduction, we design a Fidelity-Aware Projection Module (FAPM). This module employs low-rank shared projection and dynamic feature modulation to refine features before projection into the decoder. We hypothesize that by leveraging DINOv3's general high-fidelity representations, U-Net-based segmentation performance can be significantly improved across a wide range of medical imaging tasks.

Our main contributions are as follows: 1) We propose Dino U-Net, a novel architecture that employs a frozen DINOv3 backbone as the encoder integrated with a U-Net decoder. 2) We design a Fidelity-Aware Projection Module (FAPM) to reduce feature dimensionality while maintaining high fidelity. 3) We conduct extensive experiments on seven public datasets, demonstrating state-of-the-art performance and strong generalization across modalities. 4) We show that our method exhibits high scalability, with performance improving as the backbone model scales up.

2 Related work

Medical Image Segmentation The U-Net architecture remains the cornerstone of medical image segmentation. Its encoder-decoder structure and skip connections efficiently capture multi-scale contexts. Numerous variants have optimized this paradigm: U-Net++ [30] and SegResNet [18] improve feature propagation, while the nnU-Net [9] framework establishes a robust automated benchmark. Recent models like U-Mamba [15] and U-KAN [13] further incorporate state-space models and learnable activations to capture long-range dependen-

cies. However, training these models from scratch on limited medical datasets restricts their generalizability across domains. This bottleneck has motivated the transfer of knowledge from large-scale pre-trained foundation models.

Foundation Model-based Medical Image Segmentation Visual foundation models offer powerful, generalizable feature representations. While prompt-based models like SAM and its successors (SAM2 [21], SAM3 [2]) exhibit impressive zero-shot capabilities, their performance on medical images often requires extensive fine-tuning due to significant domain gaps. Parallelly, self-supervised vision transformers, particularly the DINO series [19, 29], learn robust and semantically dense features. The latest iteration, DINOv3 [25], provides unprecedented high-fidelity dense representations that rival specialized networks even without fine-tuning. Recently, hybrid architectures like SAM2-UNet [28] have successfully integrated foundation encoders into U-Net frameworks. However, the potential of utilizing DINOv3's exceptionally dense features for medical segmentation remains underexplored. To address this gap, we propose Dino U-Net to fully harness DINOv3's high-fidelity representations for elevating segmentation performance.

2.1 Preliminary

Vision foundation models, pre-trained on vast datasets of diverse images, have demonstrated a remarkable ability to learn robust and generalizable visual representations. Among these, DINOv3 stands out as a particularly compelling choice for dense prediction tasks such as segmentation. DINOv3 is a self-supervised Vision Transformer, which learns its representations from large-scale, unlabeled natural image data without relying on explicit human annotations. This training paradigm encourages the model to capture a comprehensive understanding of visual scenes, including texture, shape, and context.

The primary advantage of DINOv3 lies in the exceptional quality of its dense features. Unlike its predecessors or other foundation models, DINOv3 was specifically engineered to address the degradation of patch-level feature maps during large-scale training. Through architectural enhancements and a novel training objective named Gram anchoring, it produces feature representations that are remarkably clean and coherent, showing superior performance in tasks that require precise spatial understanding. This proven ability to generate high-fidelity features from general-domain images presents a significant opportunity for medical image segmentation. These rich, dense-pretrained representations can provide a powerful inductive bias for segmenting complex anatomical structures, allowing our model to achieve a new level of accuracy and robustness.

3 Methodology

3.1 Overview of the Dino U-Net

As illustrated in Fig. 1, Dino U-Net employs an encoder-decoder design. The encoder comprises a frozen DINOv3 backbone, a dual-branch DINO Adapter, and

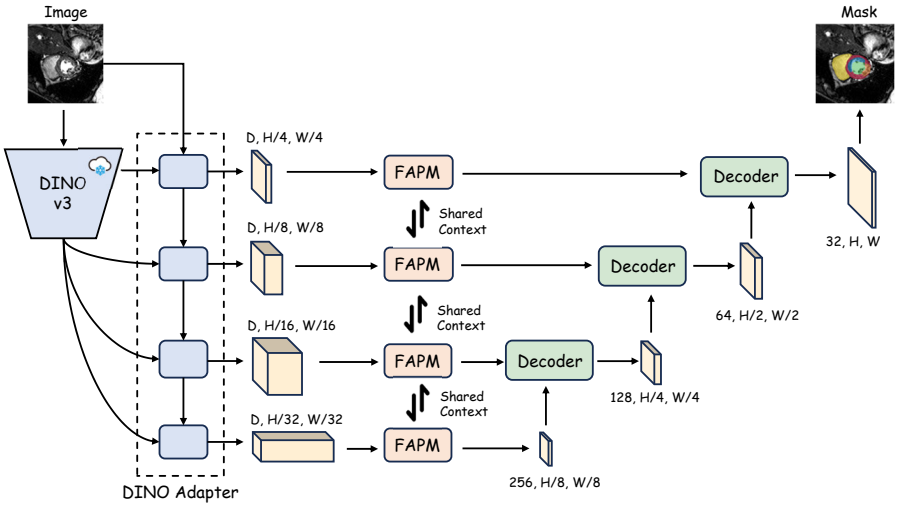

Fig. 1. Architectural overview of the Dino U-Net. The encoder consists of a frozen DINOv3 backbone, a DINO Adapter, and a Fidelity-Aware Projection Module (FAPM), connected to a standard U-Net decoder.

a Fidelity-Aware Projection Module (FAPM). During forward propagation, the adapter efficiently bridges the domain gap by processing the input concurrently through two branches: a spatial prior branch capturing high-resolution geometric textures, and a semantic branch extracting deep representations from the frozen DINOv3. At each interaction stage, a deformable cross-attention mechanism uses geometric cues from the spatial branch to selectively fuse relevant semantic content from DINOv3. After multi-stage fusion, the adapter outputs four hierarchical feature maps with spatial resolutions of $H/4 \times W/4$, $H/8 \times W/8$, $H/16 \times W/16$, and $H/32 \times W/32$, respectively, each with D channels, where D denotes the hidden dimension of the selected DINOv3 backbone. These hierarchical features balance spatial acuity with semantic depth. Subsequently, these maps are refined by the FAPM to produce high-fidelity representations, which are transmitted via skip connections to the U-Net decoder for progressive spatial recovery and final mask generation. Only the DINO Adapter, FAPM, and decoder are updated during training.

3.2 Fidelity-Aware Projection Module

The DINO Adapter produces semantically rich multi-scale features (F) in the high-dimensional DINOv3 space, which are incompatible with the channel dimensions of a standard U-Net decoder. Naïve projection methods (e.g., linear layers or 1×1 convolutions) often sacrifice fine-grained details, compromising feature fidelity. To address this, we propose the Fidelity-Aware Projection Mod-

ule (FAPM) for dimensionality reduction while preserving feature integrity (see Fig. 2).

FAPM first decouples input feature F via a dual-branch structure: a Shared Conv branch with low-rank shared weights (channel dimension 256) extracts cross-scale contextual information F_{share}, while a Specific Conv branch captures scale-specific spatial details F_{spe}. Then, F_{share} is fed into the Modulator Generation unit (a lightweight 1×1 convolution) to produce two spatially adaptive parameters: the scaling factor α and the shifting factor β. These parameters perform an affine transformation on the specific feature F_{spe}. Specifically, F_{spe} is multiplied by α and then added to β. This process enhances local details, yielding the modulated feature F_{mod}:

$$F_{mod} = F_{spe} \odot \alpha + \beta \tag{1}$$

where $\odot$ denotes element-wise multiplication.

F_{mod} is processed through dual paths for integration: a refinement path P_{refine} employing depthwise separable convolutions and a Squeeze-and-Excitation (SE) block to polish local details, and a shortcut path $P_{shortcut}$ that uses a convolution to align channels and preserve the original feature manifold. The final high-fidelity output F' is obtained as:

$$F' = P_{refine}(F_{mod}) + P_{shortcut}(F_{mod}) \tag{2}$$

which is subsequently passed via skip connection to the U-Net decoder to facilitate high-resolution mask reconstruction.

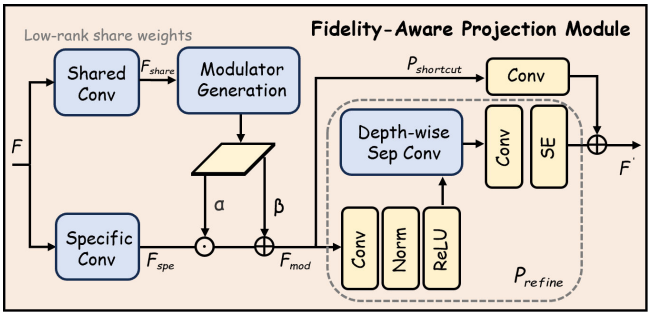

Fig. 2. Architecture of the FAPM. The input feature maps are first decoupled into task-specific and shared branches. A modulator generation unit utilizes the shared features to dynamically calibrate the specific features. The calibrated features subsequently pass through a refinement path (P_{refine}) and are merged with a shortcut path ($P_{shortcut}$) to generate the high-fidelity output.

4 Experiments

Datasets To comprehensively evaluate the effectiveness and generalization capabilities of the proposed Dino U-Net, we conducted extensive experiments on seven public datasets. These datasets were selected to cover diverse challenges in medical image segmentation, encompassing multiple imaging modalities (e.g., MRI, ultrasound, endoscopy), various anatomical targets, and different pathology types. Detailed characteristics of each dataset are listed in Table 1.

Table 1. Summary of the seven public datasets used for evaluation.

Dataset	Category	Modality	Target Classes	Cases
Kvasir-SEG [10]	Polyp	Endoscopy	Polyps	1000
BUSI [1]	Breast Tumor	Ultrasound	Benign & Malignant Tumors	780
CellBinDB [24]	Cell	Microscopy	Cells (Mouse & Human)	1044
PROSTATEx-Seg-Zones [17]	Gland	MRI	4 Prostate Zones	98
Drishti-GS [26]	Optic Nerve Head	Fundus	Optic Disc & Cup	101
MyoPS20 [4, 20, 31]	Myocardial Pathology	CMR	Scar, Edema, Myocardium, LV/RV Pools	25
m2caiSeg [16]	Surgical Instrument	Endoscopy	19 classes (Organs, Instruments, Fluids)	307

Table 2. Comparison of segmentation performance on seven medical datasets. We report the Dice (%) and the HD95. For the Dice, higher is better (↑); for HD95, lower is better (↓). For our Dino U-Net variants, scores outperforming the best baseline are marked in **bold**, and scores outperforming the second-best baseline are underlined.

Method	Kvasir-SEG		Drishti-GS		BUSI		CellBinDB		MyoPS20		PROSTATEx		m2caiseg	
	Dice↑	HD95↓	Dice↑	HD95↓	Dice↑	HD95↓	Dice↑	HD95↓	Dice↑	HD95↓	Dice↑	HD95↓	Dice↑	HD95↓
nnU-Net [9]	80.85	61.48	82.94	1.60	65.51	77.06	87.82	2.71	61.98	22.46	69.52	5.11	51.55	50.48
SegResNet [18]	85.98	35.08	82.91	1.62	67.83	64.95	87.03	3.12	67.60	14.24	71.93	4.62	43.18	53.72
UNet++ [30]	85.69	44.08	82.82	1.65	68.96	66.63	87.43	2.89	67.42	16.33	71.82	4.76	49.80	43.03
U-Mamba [15]	85.24	40.91	82.16	1.78	68.53	56.05	87.74	3.15	62.51	28.00	72.02	4.89	52.49	45.74
U-KAN [13]	85.33	41.32	83.89	1.48	68.29	**44.88**	86.99	3.92	73.59	11.15	72.65	4.54	51.01	45.28
Swin U-Mamba† [14]	81.16	59.59	80.06	2.03	66.43	66.20	86.90	3.15	70.39	11.00	69.81	5.57	44.65	51.94
SAM2-UNet [28]	89.79	27.07	83.43	1.64	66.04	59.20	87.23	3.22	74.07	10.39	71.67	4.39	49.70	52.74
Dino U-Net S	88.87	33.24	84.19	1.43$^{\sim}$	**71.87**‡	58.40	87.38	2.80	74.92‡	10.21‡	72.02	**4.16**$^{\sim}$	51.13	52.08
Dino U-Net B	88.56	28.03	83.90	1.47	70.78‡	57.36	87.56	3.19	**75.41**‡	8.39‡	72.19	4.38	52.39	47.23
Dino U-Net L	89.63	27.81	84.41	**1.40**$^{\sim}$	71.25‡	56.44	87.51	**2.53**$^{\sim}$	74.22$^{\sim}$	10.49	72.50	4.33	52.78$^{\sim}$	45.53
Dino U-Net 7B	**90.77**‡	**25.44**‡	**84.48**$^{\sim}$	1.48	70.63‡	46.85	**87.92**$^{\sim}$	2.64	74.63‡	**8.35**‡	**73.09**$^{\sim}$	4.53	**53.46**‡	**42.01**‡

$\sim$:p-value>0.05, ‡:p-value<0.05, in comparison between the best and best baseline results.

Comparison Methods and Metrics We evaluated four Dino U-Net variants (S, B, L, and 7B, scaled by backbone size) against seven state-of-the-art baselines: CNN-based architectures (nnU-Net, SegResNet, UNet++), recent Mamba/KAN-based models (U-Mamba, U-KAN), and foundation-model adaptations (Swin U-Mamba† [14], SAM2-UNet). Segmentation accuracy was measured using the Dice Similarity Coefficient (Dice) and 95% Hausdorff Distance (HD95). Additionally, we reported the number of active parameters to evaluate model efficiency and scalability. We assessed statistical significance using the Wilcoxon signed-rank test ($p < 0.05$).

Table 3. Overall efficiency and average performance. Efficiency is measured by the number of active parameters (Params). **Bold** denotes the best results; red indicates improvement over the best baseline.

Method	Params ↓	Dice (%) ↑	HD95 ↓
nnU-Net [9]	30.45	71.45	31.56
SegResNet [18]	10.26	72.35	25.34
UNet++ [30]	14.39	73.42	25.62
U-Mamba [15]	63.76	72.96	25.79
U-KAN [13]	9.38	74.54	21.80
Swin U-Mamba† [14]	37.76	71.34	28.50
SAM2-UNet [28]	5.38	74.56	22.66
Dino U-Net S	**5.11**	75.77 (+1.21)	23.19
Dino U-Net B	11.65	75.83 (+1.27)	21.44 (-0.36)
Dino U-Net L	18.14	76.04 (+1.48)	21.22 (-0.58)
Dino U-Net 7B	228.97	**76.43** (+1.87)	**18.76** (-3.04)

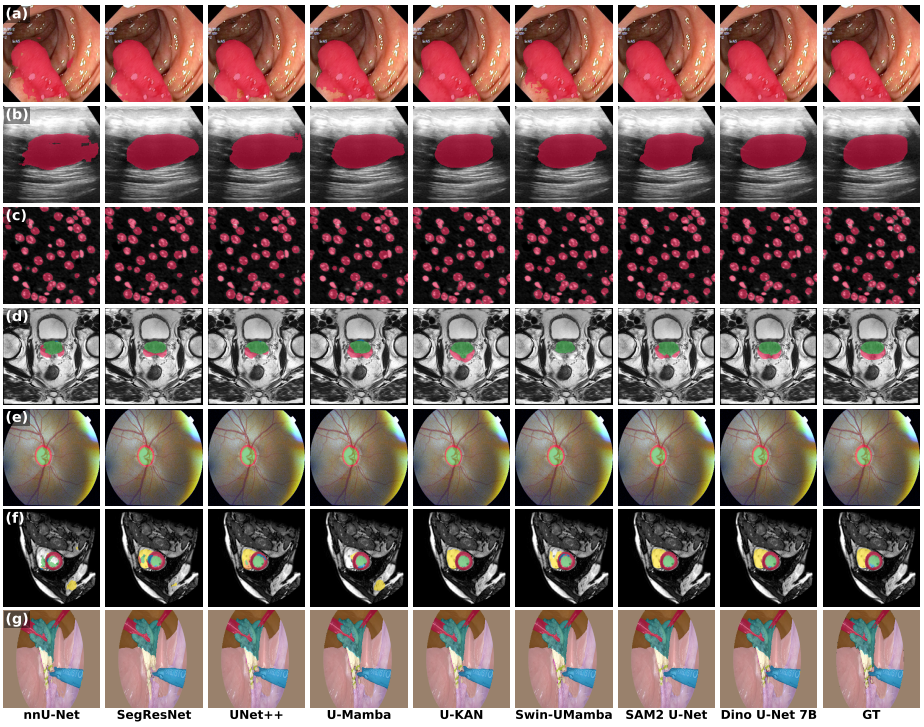

Fig. 3. Qualitative comparison of segmentation results on representative samples from the seven evaluation datasets. Each row corresponds to a different dataset (from top to bottom: Kvasir-SEG, BUSI, CellBinDB, PROSTATEx-Seg-Zones, Drishti-GS, MyoPS20, and m2caiSeg). Columns display the results of different comparison methods, our method (Dino U-Net 7B), and the Ground Truth (GT).

Implementation details Implemented in PyTorch, we strictly followed the default nnU-Net pipeline for preprocessing, data augmentation, and sliding-window

inference. The input patch sizes and batch sizes were determined by the nnU-Net planner according to the image statistics of each dataset. Datasets were randomly split into 8:2 training and testing sets. Models were trained using a combined Dice and Cross-Entropy loss, optimized by Adam (initial learning rate 1×10^{-3} with polynomial decay). Training lasted 200 epochs (250 iterations per epoch) on NVIDIA H100 GPUs.

Table 4. Ablation study on the FAPM. We report the relative changes in parameter counts (Δ Params), Dice score (Δ Dice), and HD95 (Δ HD95) after integrating the FAPM module across various model scales. ↑ indicates higher is better, and ↓ indicates lower is better.

Metrics	**Model Size**			
	S	**B**	**L**	**7B**
Δ Params (M)	+0.5	+0.3	+0.2	−0.8
Δ Dice (%) ↑	+0.79	+0.77	+0.56	+0.75
Δ HD95 (mm) ↓	−1.75	−0.09	−0.37	−1.04

Results Quantitative evaluations (Table 2 and Table 3) demonstrate that Dino U-Net consistently outperforms seven state-of-the-art methods across diverse modalities, with improvements being statistically significant in most cases. Notably, the largest 7B variant sets new records on five of the seven datasets, achieving the best average Dice score of 76.43% (+1.87%) and HD95 of 18.76 (-3.04). The architecture also exhibits exceptional scalability and efficiency: performance improves consistently as the backbone scales, yet even the lightweight small variant surpasses the average performance of all baselines. Qualitatively (Fig. 3), our method achieves more precise boundary delineation than contour-reliant models like SAM2-UNet, showing superior robustness in low-contrast textures.

Ablation Study To validate FAPM, we replaced it with simple 1×1 convolutions across all scales (Table 4). Regarding efficiency, FAPM's low-rank shared basis design demonstrates scalability: while adding marginal parameters to smaller models (S, B, L), it effectively reduces the overall parameter count for the largest 7B model. Furthermore, removing FAPM consistently degrades segmentation accuracy across all scales, with Dice dropping by up to 0.79% and HD95 worsening by up to 1.75mm. This confirms that FAPM is crucial for preserving high-fidelity features and precise boundary delineation.

5 Discussion and Conclusion

In this paper, we propose Dino U-Net, which integrates the DINOv3 foundation model with a dual-branch adapter and a novel FAPM to leverage high-fidelity dense features from large-scale pre-training. Dino U-Net's superior performance

stems from leveraging DINOv3's self-supervised representations and the FAPM's high-fidelity projection, which provide a more robust semantic inductive bias than contour-biased models. The scaling success from small version to 7B variants further confirms that medical segmentation significantly benefits from the massive representational power of billion-parameter foundation models. Despite these strengths, the current framework is limited to 2D tasks, necessitating future extension to 3D volumetric data. Additionally, while the frozen backbone ensures parameter efficiency, the inference overhead of the 7B model remains high, suggesting future research into knowledge distillation to develop more compact yet powerful models.

Extensive experiments on seven public datasets confirm that our method significantly outperforms state-of-the-art models across various imaging modalities. This study demonstrates that the rich, pre-trained representations from next-generation self-supervised models provide a superior inductive bias for medical segmentation. Future work will focus on extending this framework to 3D volumetric segmentation tasks.

Acknowledgments. This work was supported by the National Science Foundation of China (Grant No. 82372052, U24A20757, 82402373), and Taishan Industrial Experts Program (Grant No. tscx202312131).

Disclosure of Interests. The authors declare that they have no conflicts of interest.

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in brief **28**, 104863 (2020)
2. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
3. Dai, Y., Gao, Y., Liu, F.: Transmed: Transformers advance multi-modal medical image classification. Diagnostics **11**(8), 1384 (2021)
4. Gao, S., Zhou, H., Gao, Y., Zhuang, X.: Bayeseg: Bayesian modeling for medical image segmentation with interpretable generalizability. Medical image analysis **89**, 102889 (2023)
5. Gao, Y., Dai, Y., Liu, F., Chen, W., Shi, L.: An anatomy-aware framework for automatic segmentation of parotid tumor from multimodal mri. Computers in Biology and Medicine **161**, 107000 (2023)
6. Gao, Y., Dong, Y., Wu, W., Ge, C., Yuan, F., Sheng, J., Li, H., Gao, X.: Wega: Weakly-supervised global-local affinity learning framework for lymph node metastasis prediction in rectal cancer. arXiv preprint arXiv:2505.10502 (2025)
7. Gao, Y., Rui, S., Su, H., Xiang, J., Wu, L., Wang, X.: A composite alignment-aware framework for myocardial lesion segmentation in multi-sequence cmr images. arXiv preprint arXiv:2507.11886 (2025)
8. Gao, Y., Sheng, J., Wu, W., Li, H., Dong, Y., Ge, C., Yuan, F., Gao, X.: Safeclick: Error-tolerant interactive segmentation of any medical volumes via hierarchical expert consensus. arXiv preprint arXiv:2506.18404 (2025)

9. 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)
10. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., de Lange, T., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26. pp. 451–462. Springer (2020)
11. 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)
12. Krithika Alias AnbuDevi, M., Suganthi, K.: Review of semantic segmentation of medical images using modified architectures of unet. Diagnostics **12**(12), 3064 (2022)
13. Li, C., Liu, X., Li, W., Wang, C., Liu, H., Liu, Y., Chen, Z., Yuan, Y.: U-kan makes strong backbone for medical image segmentation and generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 4652–4660 (2025)
14. Liu, J., Yang, H., Zhou, H.Y., Xi, Y., Yu, L., Li, C., Liang, Y., Shi, G., Yu, Y., Zhang, S., et al.: Swin-umamba: Mamba-based unet with imagenet-based pretraining. In: International conference on medical image computing and computer-assisted intervention. pp. 615–625. Springer (2024)
15. Ma, J., Li, F., Wang, B.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722 (2024)
16. Maqbool, S., Riaz, A., Sajid, H., Hasan, O.: m2caiseg: Semantic segmentation of laparoscopic images using convolutional neural networks. arXiv preprint arXiv:2008.10134 (2020)
17. Meyer, A., Schindele, D., Von Reibnitz, D., Rak, M., Schostak, M., Hansen, C.: Prostatex zone segmentations [data set]. The Cancer Imaging Archive p. 131 (2020)
18. Myronenko, A.: 3d mri brain tumor segmentation using autoencoder regularization. In: International MICCAI brainlesion workshop. pp. 311–320. Springer (2018)
19. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
20. Qiu, J., Li, L., Wang, S., Zhang, K., Chen, Y., Yang, S., Zhuang, X.: Myops-net: Myocardial pathology segmentation with flexible combination of multi-sequence cmr images. Medical image analysis **84**, 102694 (2023)
21. 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)
22. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
23. Shaker, A., Maaz, M., Rasheed, H., Khan, S., Yang, M.H., Khan, F.S.: Unetr++: delving into efficient and accurate 3d medical image segmentation. IEEE Transactions on Medical Imaging **43**(9), 3377–3390 (2024)
24. Shi, C., Fan, J., Deng, Z., Liu, H., Kang, Q., Li, Y., Guo, J., Wang, J., Gong, J., Liao, S., et al.: Cellbindb: a large-scale multimodal annotated dataset for cell segmentation with benchmarking of universal models. GigaScience **14**, giaf069 (2025)
25. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A.,

Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3 (2025), https://arxiv.org/abs/2508.10104
26. Sivaswamy, J., Krishnadas, S.R., Datt Joshi, G., Jain, M., Syed Tabish, A.U.: Drishti-gs: Retinal image dataset for optic nerve head(onh) segmentation. In: 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI). pp. 53–56 (2014). https://doi.org/10.1109/ISBI.2014.6867807
27. Wu, J., Wang, Z., Hong, M., Ji, W., Fu, H., Xu, Y., Xu, M., Jin, Y.: Medical sam adapter: Adapting segment anything model for medical image segmentation. Medical image analysis **102**, 103547 (2025)
28. Xiong, X., Wu, Z., Tan, S., Li, W., Tang, F., Chen, Y., Li, S., Ma, J., Li, G.: Sam2-unet: Segment anything 2 makes strong encoder for natural and medical image segmentation. arXiv preprint arXiv:2408.08870 (2024)
29. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605 (2022)
30. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE transactions on medical imaging **39**(6), 1856–1867 (2019)
31. Zhuang, X.: Multivariate mixture model for myocardial segmentation combining multi-source images. IEEE transactions on pattern analysis and machine intelligence **41**(12), 2933–2946 (2018)