DINO-Med3D: Bridging Dimension and Domain Gaps in Volumetric Segmentation via Progressive Adaptation

Haoyu Hu[1,2], Xiyao Ma[1,2], Shiqi Liu[2*], Linsen Zhang[2], Xiaoliang Xie[2], Xiaohu Zhou[2], and Zeng-Guang Hou[2⋆]

[1] University of Chinese Academy of Sciences, Beijing, China
[2] Institute of Automation, Chinese Academy of Sciences, Beijing, China
{liushiqi2016, zengguang.hou}@ia.ac.cn

Abstract. Although DINOv3 has demonstrated remarkable semantic discrimination in natural imagery, its direct application to volumetric medical segmentation is hindered by inherent dimension and domain disparities. To resolve these issues, we propose DINO-Med3D, a two-stage progressive framework that repurpose the pre-trained DINOv3 encoder for 3D medical tasks. In the first stage, we mitigate the dimension gap by introducing a multi-slice embedding module that incorporates pseudo-3D context, while simultaneously employing a segmentation proxy task to adapt representations learned from natural scenes to the medical domain. Subsequently, we further enhance volumetric understanding by adding lightweight 3D adapters into the frozen backbone to enforce global inter-slice continuity. Finally, to compensate for the spatial information loss inherent in the embedding process, we design a parallel Detail-Recovery stream to explicitly preserve high-frequency boundary cues. Extensive experiments on five public datasets demonstrate that our approach successfully adapts DINOv3 to the medical domain and significantly outperforms state-of-the-art baselines. The source code is available at https://github.com/HaoyuHu1/DINO-Med3D/.

Keywords: DINOv3 · Medical Image Segmentation · Foundation Models.

1 Introduction

Medical image segmentation plays a pivotal role in computer-aided diagnosis by facilitating quantitative clinical assessment. Despite the architectural evolution from CNNs [6,7,11,20,27] to recent Transformers and Mamba-based models [9,17], performance remains constrained by the scarcity of high-quality annotations. Models trained on isolated, small-scale datasets often struggle with fully leveraging their potentials.

Foundation models offer a promising avenue to circumvent this data bottleneck. While the recent Segment Anything paradigm [13] and its medical adaptations

⋆ Corresponding author

[12, 15] demonstrate impressive zero-shot capabilities, they typically rely on interactive prompts or heuristic auto-prompting mechanisms. In contrast, self-supervised discriminative models, exemplified by DINOv3 [5, 19, 22], learn robust semantic representations directly from data without requiring user interaction. Following this paradigm, a surge of recent studies has adapted these general-purpose encoders to the medical domain [8, 21, 25]. However, directly repurposing 2D natural image encoders for 3D medical volumes is non-trivial. As highlighted by recent benchmarks [16], DINOv3 performance degrades significantly in tasks involving substantial domain shifts or dense 3D predictions, often lagging behind specialized frameworks like nnU-Net [11]. This suggests that naive fine-tuning or simple projection is insufficient to bridge the feature gap between natural imagery and complex volumetric medical data.

We attribute this performance degradation to two primary disparities: the dimension gap and the domain gap. First, adapting 2D encoders to 3D volumes forces slice-wise processing, inevitably discarding spatial correlations along the z-axis. Second, DINOv3 representations are biased toward the distinct shapes and high contrast of natural scenes, generalizing poorly to medical tasks that rely on subtle texture cues to define low-contrast boundaries.

To systematically bridge these gaps, we propose DINO-Med3D, a progressive framework that evolves the 2D foundation model into a 3D medical segmentor through a two-stage paradigm. In the first stage, we address the initial disparities by designing a specialized multi-slice embedding module, which aggregates contextual information from adjacent slices. Simultaneously, we mitigate the domain gap by employing a segmentation proxy task to adapt fundamental representations learned from natural scenes to medical domain. In the second stage, we freeze the aligned backbone and introduce lightweight 3D adapters alongside Low-Rank Adaptation (LoRA) [10] to enforce global inter-slice continuity, thereby further bridging the dimension gap. Concurrently, a parallel Detail-Recovery stream is designed to explicitly capture high-frequency boundary cues, facilitating precise adaptation to the textural properties of medical images. Our contributions are summarized as follows:

- We propose DINO-Med3D, a novel architecture that effectively repurposes the 2D DINOv3 for 3D medical segmentation by progressively bridging dimension and domain gaps.
- We introduce a decoupled two-stage transfer strategy: the first stage aligns feature distributions and input dimensions, while the second stage fosters volumetric reasoning and texture refinement via 3D adapters and the proposed Detail-Recovery stream.
- We conduct extensive experiments on five public datasets spanning multiple modalities, anatomical structures, and pathological textures. The results demonstrate that our method consistently bridges the aforementioned gaps and significantly outperforms SOTA baselines.

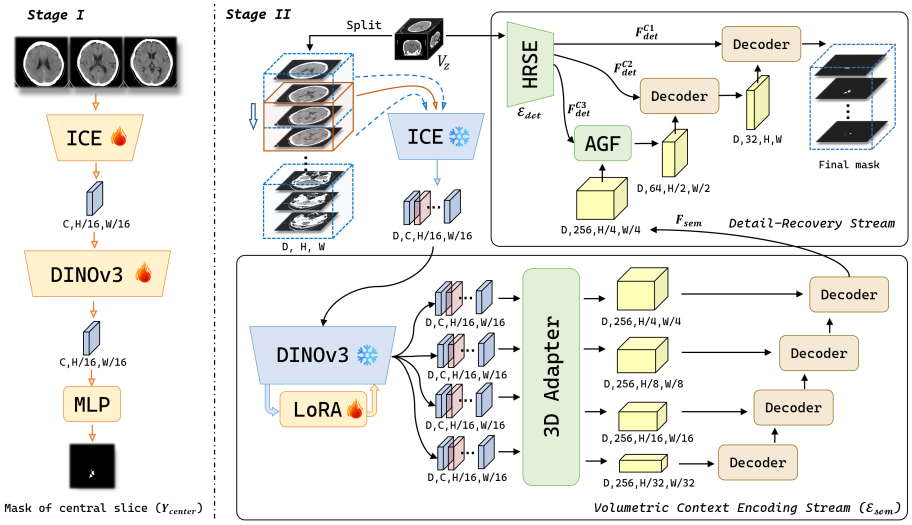

Fig. 1. Schematic illustration of DINO-Med3D framework. Stage I aligns the pre-trained 2D DINOv3 backbone with volumetric medical data via the Inter-slice Context Embedding (ICE) module and a pseudo-3D proxy task. Stage II freezes the aligned backbone and introduces a dual-stream architecture: a Volumetric Context Encoding Stream leveraging LoRA and lightweight 3D adapters for semantic reasoning, and a parallel Detail-Recovery Stream for capturing high-frequency boundary information. Finally, features are integrated via Adaptive Gated Fusion (AGF) for dense prediction.

2 Methods

2.1 Overall Architecture

As illustrated in Fig. 1, the proposed framework bridges 2D pre-training and 3D volumetric segmentation through a two-stage approach. In Stage I, we adapt the standard 2D DINOv3 encoder to the medical domain by employing a dimension-aware adaptation mechanism alongside a segmentation proxy task. Subsequently, in Stage II, we construct a dual-stream architecture comprising (1) a Volumetric Context Encoding Stream initialized with Stage I weights for high-level semantic extraction; and (2) a lightweight Detail-Recovery Stream to capture high-frequency spatial information. These features are aggregated via a fusion module and decoded to generate the final segmentation masks.

2.2 Stage I: Volumetric Adaptation and Alignment

To bridge the gap between 2D pre-training and volumetric medical data, we introduce a pseudo-3D adaptation strategy facilitated by a depth-aware embedding module and a proxy task.

Inter-slice Context Embedding(ICE) Direct 3D input is incompatible with DINOv3's pre-trained 2D Rotary Positional Embeddings, while naive slice-wise processing discards critical through-plane continuity. To resolve this, we propose ICE to replace the original 2D patch embedding. ICE processes a sub-volume slab $V \in \mathbb{R}^{K\times H\times W}$ to yield a pseudo-3D patch embedding of size $C \times \frac{H}{16} \times \frac{W}{16}$ via depth-aware 3D convolution. Here, C denotes the embedding dimension of DINOv3 and K is a tunable hyperparameter representing the input slices of ICE. In this way, while maintaining dimensional compatibility with the 2D backbone, ICE effectively fuses inter-slice context and learns volumetric structure.

Proxy Task: Domain Adaptation We further align the feature space with medical semantics via a proxy segmentation task. A lightweight MLP head projects the encoder's output from $\mathbb{R}^{C\times \frac{H}{16}\times \frac{W}{16}}$ to $\mathbb{R}^{H\times W}$ to predict the segmentation map $\hat{Y}_{center}$ of the central slice. The ICE module and backbone are fully fine-tuned to minimize the discrepancy between $\hat{Y}_{center}$ and the ground truth, compelling the model to adapt feature representations from natural scenes to medical images.

2.3 Stage II: Dual-Stream Segmentation Network

To synergize semantic abstraction with spatial precision, we propose a dual-stream architecture. Leveraging the Stage I initialization, the framework processes input $X \in \mathbb{R}^{D\times H\times W}$ to generate a spatially corresponding dense prediction map M, where D represents the input depth. Two parallel paths are constructed: a context stream $\mathcal{E}_{sem}$ for long-range semantic features (F_{sem}) and a lightweight detail stream $\mathcal{E}_{det}$ for high-frequency features (F_{det}). These hierarchical features are integrated via a fusion module $\mathcal{F}$ and reconstructed by decoder $\mathcal{D}$ to yield M:

$$M = \mathcal{D}\left(\mathcal{F}(F_{sem}, F_{det}), F_{det}\right) = \mathcal{D}\left(\mathcal{F}(\mathcal{E}_{sem}(X), \mathcal{E}_{det}(X)), \mathcal{E}_{det}(X)\right) \quad (1)$$

Volumetric Context Encoding Stream This stream($\mathcal{E}_{sem}$) leverages the representations from Stage I and extends these capabilities to full volume. Given an input volume $X \in \mathbb{R}^{D\times H\times W}$, we construct a sequence of pseudo-3D slabs V. For depth z=1,2,...,D, adjacent K slices are stacked to form a multi-channel slab $V_z = [x_{z-\frac{K-1}{2}}, \ldots, x_z, \ldots, x_{z+\frac{K-1}{2}}] \in \mathbb{R}^{K\times H\times W}$, utilizing zero-padding at boundaries. These D slabs are processed by the ICE and backbone sequentially and then stacked to restore volumetric geometry. To preserve the 3D medical representation acquired in stage I, we freeze the pre-trained backbone and inject trainable LoRA matrices into the Transformer blocks. A lightweight 3D Adapter aligns hierarchical features via 3D deconvolutions to transition from the stacked pseudo-3D features to genuine volumetric representation, followed by a UperNet [24] decoder to yield the coarse semantic map $F_{sem} \in \mathbb{R}^{D\times C_{sem}\times \frac{H}{4}\times \frac{W}{4}}$.

Detail-Recovery Stream Complementing the semantic stream, the High-Resolution Spatial Encode (HRSE, $\mathcal{E}_{det}$) compensates for the loss of high-frequency spatial details inherent in patch-based embedding. Designed as a

lightweight 3D CNN with residual blocks, HRSE prioritizes spatial precision by avoiding aggressive downsampling, extracting a feature hierarchy $F_{det} = \{C_1, C_2, C_3\}$ at original, 1/2, and 1/4 scales. This preserves fine-grained boundary information essential for segmentation. To bridge the semantic-spatial gap between two streams, we introduce the Adaptive Gated Fusion (AGF, $\mathcal{F}$) module, which acts as a content-aware residual correction mechanism. Given the semantic features F_{sem} and local details F_{det}^{C3}, AGF first generates a spatial attention map G to highlight regions requiring refinement:

$$G = \sigma\left(\mathcal{K}_{1\times1\times1}\left([F_{sem}; F_{det}^{C3}]\right)\right) \tag{2}$$

where $[\,;\,]$ denotes channel-wise concatenation, $\mathcal{K}_{1\times1\times1}$ is a $1\times1\times1$ convolution and σ is the sigmoid function. The gate $G \in [0,1]$ then modulates the injection of high-frequency details into the semantic stream:

$$\tilde{F}_{det}^{C3} = F_{sem} + G \odot F_{det}^{C3} \tag{3}$$

Finally, the refined $\tilde{F}_{det}^{C3}$ is progressively aggregated with lower-level features ($F_{det}^{C1,C2}$) via a UperNet decoder to yield the final probability map.

3 Experiments and Results

3.1 Dataset

We evaluated our method on five public datasets across CT and MRI modalities. CT datasets include: (1) AISD [14]: 397 brain scans with infarct lesions; (2) MSD-Pancreas and (3) MSD-Colon [1]: 281 and 126 abdominal scans, cropped to 50 and 30 slices along the z-axis, respectively. MRI datasets include: (4) BraTS 2020 [2,3,18]: 368 brain scans (FLAIR, Whole Tumor), downsampled to 4 mm spacing; and (5) ACDC [4]: 100 cardiac scans. All images were resized to 512×512 and randomly split (8:1:1). Specific windowing (WW, WL) was applied for CTs: AISD (100, 40), Pancreas (255, 50), and Colon (400, 40). MRIs were clipped to the [0.5, 99.5] percentile range and normalized.

3.2 Implementation Details and Evaluation Metrics

Standard data augmentation (rotation, scaling, intensity jitter) was employed. We optimized a compound loss $\mathcal{L} = 0.2\mathcal{L}_{CE} + 0.8\mathcal{L}_{Dice}$ for both stage (50 epochs each). Stage 1: batch size (bs) 24, learning rate (lr) 2×10^{-5}. Stage 2: initialized from the best Stage 1 checkpoint, fine-tuned with LoRA ($r = 16, \alpha = 16, p = 0.1$), bs 1, and lr 1×10^{-4}. The ICE input slices K is set to 3. Our proposed method variants are denoted as DINO-Med3D-S/B/L. All baseline methods were implemented following their default configurations. All models were trained for 100 epochs. Evaluation metrics include Dice Similarity Coefficient (DSC) and Hausdorff Distance (HD95).

Table 1. Quantitative comparison. The best baselines are in gray. Scores exceeding the best and second-best baseline are **bolded** and underlined respectively.

Method	AISD		MSD-Pan.		MSD-Colon		BraTS		ACDC		Params (M)
	DSC↑	HD95↓	DSC↑	HD95↓	DSC↑	HD95↓	DSC↑	HD95↓	DSC↑	HD95↓	
nnU-Net	47.44	57.09	56.93	6.93	38.13	34.44	88.73	3.62	88.63	1.57	31
nnFormer	34.87	107.75	54.42	10.83	50.68	68.91	87.31	24.40	75.77	6.03	150
SwinUNETR	42.73	45.91	51.85	20.51	20.84	79.12	88.76	7.38	85.96	3.64	62
Umamba	41.92	62.82	53.00	14.62	32.33	64.20	89.61	3.30	87.70	1.32	43
VoCo(Swin-L)	43.99	39.26	51.37	33.72	33.51	60.68	89.19	9.26	85.57	6.56	248
Dino U-Net	48.19	56.79	48.25	14.01	31.26	91.51	89.09	5.38	91.19	1.17	318
our-Small	**51.86**	51.76	**58.58**	9.36	47.08	50.36	**89.96**	**2.18**	88.81	**1.12**	47
our-Base	**52.72**	49.75	**59.50**	7.08	**53.28**	34.61	**90.44**	**2.34**	90.82	**0.91**	112
our-Large	**53.85**	44.23	**61.20**	**6.40**	**62.28**	**27.70**	**90.27**	**2.06**	**91.52**	**0.86**	334

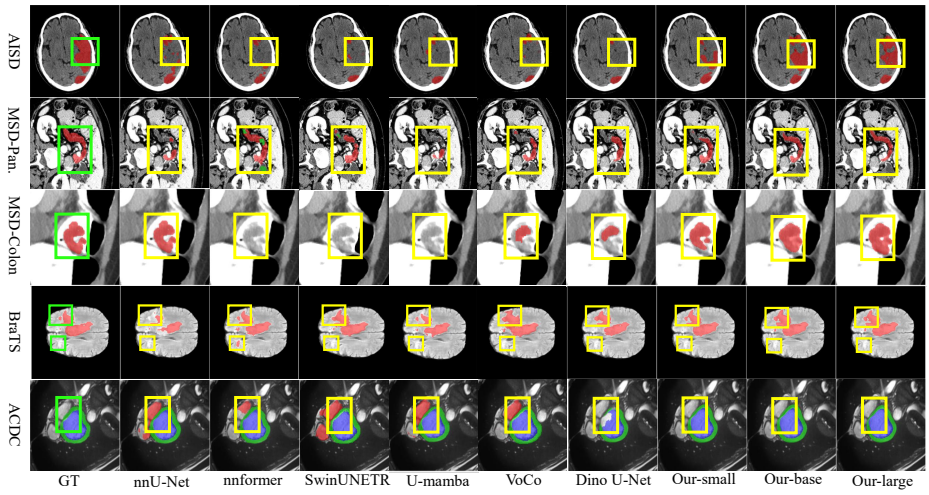

Fig. 2. Qualitative comparison on representative samples. Images are cropped and zoomed for better visualization. Prominent regions are highlighted with boxes.

3.3 Result

Table 1 presents the quantitative comparisons across different benchmarks and Figure 2 provides a qualitative comparison. Our proposed framework consistently outperforms SOTA baselines, including the CNN-based nnU-Net [11], the Transformer-based nnFormer [26] and SwinUNETR [9], and the Mamba-based U-Mamba [17]. Furthermore, we benchmark our method against VoCo [23], a representative self-supervised pre-training framework. We also include Dino U-Net [8], a representative DINOv3-based adaptation method for medical imaging. Notably, the DINO-Med3D-L variant achieves the best overall performance in terms of both DSC and HD95. This demonstrates the framework's robustness

and potent segmentation capabilities across multiple modalities (CT and MRI), diverse anatomical structures, and varying segmentation targets.

Our method demonstrates superior robustness in challenging scenarios characterized by high anatomical variability, such as the MSD-Colon and AISD datasets. Notably, on MSD-Colon, DINO-Med3D-L achieves a DSC of 62.28%, outperforming SOTA method nnFormer (50.68%) by a substantial margin of 11.60%. In contrast, competing approaches suffered from severe under-segmentation, further highlighting the efficacy of our framework in handling complex cases.

Direct comparison with Dino U-Net highlights the efficacy of our architectural adaptations. To ensure a fair evaluation, we benchmark our model against the Large variant model of Dino U-Net which is built upon the DINOv3 architecture with large backbones. Despite both methods leveraging DINOv3 encoders, DINO-Med3D yields consistently higher accuracy, particularly on texture-rich datasets (MSD-Panc. and Colon). Furthermore, we observe a positive scaling law: performance improves monotonically from Small to Large variants (Params: 47M to 334M). It is worth noting that even our efficiency-focused Small variant (47M) outperforms the much larger Dino U-Net (318M) on AISD, MSD and BraTS, striking a favorable trade-off between computational cost and segmentation accuracy.

3.4 Ablation Study

To validate the efficacy of the proposed components, we conducted ablation studies on three datasets—MSD-Pancreas, MSD-Colon, and BraTS—covering diverse organs and modalities. All implementations were built upon the DINO-Med3D-B.

Table 2 analyzes the impact of ICE input slices K. As presented in Table 2, performance improves as the number of layers increases, peaking at three layers across all datasets, before degrading with further additions. Specifically, on MSD-Pancreas, the DSC rises from 54.97% to 59.50% and subsequently drops to 58.72%. This suggests that while multi-layer encoding facilitates better feature extraction, excessive depth may compromise the model's representation capability.

Table 2. Impact of the ICE input slices K on segmentation performance. Results are reported in DSC. Best scores are bolded.

Dataset	ICE input slices K				
	1	2	3	4	5
MSD-Pan.	54.97	55.97	**59.50**	58.79	58.72
MSD-Colon	44.26	51.99	**53.28**	47.53	33.10
BraTS	89.66	90.01	**90.44**	89.69	90.20

Table 3 dissects the efficacy of the proposed multi-stage framework. A frozen DINOv3 backbone with a direct MLP decoder yields suboptimal results (Row 1).

Table 3. Ablation study results. The bolded row represents the full implementation of our proposed method.

Stage I		Stage II		Dice Similarity Coefficient ↑		
UperNet	MLP	Stream 1	Stream 2	MSD-Pan.	MSD-Colon	BraTS
				28.60 (-30.90)	17.38 (-35.90)	78.46 (-11.98)
	✓			54.63 (-4.87)	43.11 (-10.17)	89.16 (-1.28)
	✓	✓		57.00 (-2.50)	44.93 (-8.35)	89.91 (-0.53)
	✓	✓	✓	**59.50**	**53.28**	**90.44**
		✓	✓	54.78 (-4.72)	45.96 (-7.32)	90.19 (-0.25)
✓		✓	✓	57.42 (-2.08)	47.95 (-5.33)	90.39 (-0.05)

Conversely, activating Stage I—unfreezing the backbone via the proxy segmentation task with ICE module—yields significant gains. In Stage II, introducing Stream 1 (Volumetric Context Encoding) boosts MSD-Pancreas DSC by 2.37%, confirming the necessity of transforming pseudo-3D features into intrinsic 3D representations. Integrating Stream 2 (Detail-Recovery) further refines boundaries, with qualitative improvements shown in Fig. 3.

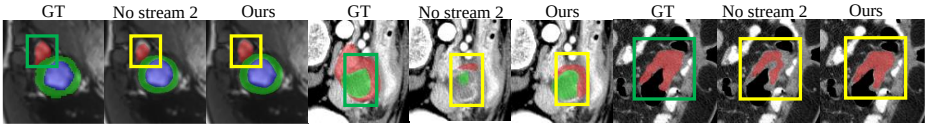

Fig. 3. Visual ablation study of the Detail-Recovery Stream. Cases from ACDC and MSD (Panc. & Colon) are presented. Each triplet shows: (a) Ground Truth, (b) w/o Stream 2, and (c) Full Method. Boxes indicate major improvements in detail preservation.

Finally, we verified the role of the proxy task and the decoder architecture. Removing Stage I leads to a 4.72% DSC drop on MSD-Pancreas. Notably, this drop occurs despite the application of LoRA in Stage II. This demonstrates that Stage I's full fine-tuning is a prerequisite for robust transfer learning, as the Parameter-Efficient LoRA adaptation in Stage II cannot fully bridge for the dimension and domain gaps. While retaining Stage I and substituting the MLP decoder with a UperNet decoder improves performance compared to the no-Stage-I setting (+2.64% DSC), it fails to match the optimal results, thereby demonstrating the effectiveness of our lightweight decoder in proxy task.

4 Conclusion and Discussion

In this work, we proposed DINO-Med3D, a progressive adaptation framework designed to repurpose the 2D DINOv3 foundation model for volumetric medical segmentation. Comprehensive evaluations across five datasets demonstrate that

DINO-Med3D consistently outperforms state-of-the-art baselines, validating the efficacy of our dual-stream architecture.

Despite these advancements, certain limitations warrant further investigation. First, the performance variability of the Detail-Recovery Stream across datasets suggests a dependency on texture complexity and lesion size that warrants deeper investigation. Second, the first stage necessitates full fine-tuning of the backbone to align feature distributions, imposing a substantial computational burden. Future work will focus on developing more parameter-efficient alignment strategies.

Acknowledgements. This work was supported in part by the National Key Research and Development Program of China under Grant 2024YFF1206902 and 2025YFSH0038; in part by the Development Project of National Major Scientific Research Instrument under Grant 82327801; in part by the National Natural Science Foundation of China under Grant 62303463; in part by the Beijing Natural Science Foundation under Grant L251071, L246047, L232137, and Z241100009024031.

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

References

1. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., et al.: The medical segmentation decathlon. Nature communications **13**(1), 4128 (2022)
2. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific Data **4**(1), 1–13 (2017)
3. Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv preprint arXiv:1811.02629 (2018)
4. 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)
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
6. 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)
7. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. pp. 424–432. Springer (2016)

8. Gao, Y., Li, H., Yuan, F., Wang, X., Gao, X.: Dino u-net: Exploiting high-fidelity dense features from foundation models for medical image segmentation. arXiv preprint arXiv:2508.20909 (2025)
9. 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)
10. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
11. 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)
12. Jiang, C., Ding, T., Song, C., Tu, J., Yan, Z., Shao, Y., Wang, Z., Shang, Y., Han, T., Tian, Y.: Medical sam3: A foundation model for universal prompt-driven medical image segmentation. arXiv preprint arXiv:2601.10880 (2026)
13. 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)
14. Liang, K., Han, K., Li, X., Cheng, X., Li, Y., Wang, Y., Yu, Y.: Symmetry-enhanced attention network for acute ischemic infarct segmentation with non-contrast ct images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 432–441. Springer (2021)
15. Liu, A., Xue, R., Cao, X.R., Shen, Y., Lu, Y., Li, X., Chen, Q., Chen, J.: Medsam3: Delving into segment anything with medical concepts. arXiv preprint arXiv:2511.19046 (2025)
16. Liu, C., Chen, Y., Shi, H., Lu, J., Jian, B., Pan, J., Cai, L., Wang, J., Zhang, Y., Li, J., et al.: Does dinov3 set a new medical vision standard? arXiv e-prints pp. arXiv–2509 (2025)
17. Ma, J., Li, F., Wang, B.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722 (2024)
18. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE Transactions on Medical Imaging **34**(10), 1993–2024 (2014)
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. 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)
21. Scholz, D., Erdur, A.C., Ehm, V., Meyer-Baese, A., Peeken, J.C., Rueckert, D., Wiestler, B.: Mm-dinov2: Adapting foundation models for multi-modal medical image analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 320–330. Springer (2025)
22. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
23. Wu, L., Zhuang, J., Chen, H.: Large-scale 3d medical image pre-training with geometric context priors. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)

24. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Proceedings of the European conference on computer vision (ECCV). pp. 418–434 (2018)
25. Yang, S., Wang, H., Xing, Z., Chen, S., Zhu, L.: Segdino: An efficient design for medical and natural image segmentation with dino-v3. arXiv preprint arXiv:2509.00833 (2025)
26. Zhou, H.Y., Guo, J., Zhang, Y., Han, X., Yu, L., Wang, L., Yu, Y.: nnformer: Volumetric medical image segmentation via a 3d transformer. IEEE transactions on image processing **32**, 4036–4045 (2023)
27. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: International workshop on deep learning in medical image analysis. pp. 3–11. Springer (2018)