Glasses: Adapter-Enhanced DINOv3 with Boundary-Aware Multi-Scale Representations for Medical Image Segmentation

Jiahao Zhu[1], Hengfei Cui[1,2(✉)], Hanlin Yin[1], Geng Chen[1], and Yong Xia[1]

[1] National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China
hfcui@nwpu.edu.cn

[2] Chongqing Innovation Center, Northwestern Polytechnical University, Chongqing 401135, China

Abstract. Accurate medical image segmentation requires simultaneous modeling of robust global semantics and precise anatomical boundaries. While Vision Transformers (ViTs) exhibit strong global representation capability, they often struggle to capture fine-grained local structures and boundary details, particularly in scenarios with low contrast and blurred boundaries. To address this limitation, we propose Glasses, a novel framework that equips pre-trained Transformers with a structure-enhanced visual adaptation mechanism for medical image segmentation. Specifically, Glasses employs a frozen DINOv3 as the backbone encoder, with only the terminal layers fine-tuned. We introduce a Multi-Scale Deformable Attention Adapter that performs cross-scale interaction between lightweight convolutional structural priors and Transformer tokens extracted from specific intermediate layers of DINOv3, thereby transforming single-scale token embeddings into hierarchical representations enriched with both local structural cues and global context. This design preserves the stability of pre-trained semantic representations while improving the modeling of local anatomical structures and scale variations. Furthermore, we introduce an independent Edge-Aware Convolutional Encoder that explicitly learns boundary-enriched multi-scale representations under nested boundary supervision, providing a strong source of fine-grained local structural information that complements the adapter-enhanced Transformer features during decoding. A U-Net-style decoder then progressively fuses the adapter-enhanced Transformer features and the boundary-enriched convolutional representations to recover high-resolution segmentation outputs. Extensive experiments on the Synapse and ISIC datasets demonstrate that Glasses achieves Dice coefficients of 83.52% and 91.60%, respectively, significantly outperforming state-of-the-art (SOTA) methods. Source code is available at https://github.com/friay/Glasses.

Keywords: Segmentation · DINOv3 · Boundary supervision.

1 Introduction

Medical image segmentation is a fundamental task in computer-assisted diagnosis and treatment planning [16]. Medical images often exhibit low contrast, ambiguous boundaries, and subtle inter-class variations, making accurate delineation of anatomical structures particularly challenging [19,13,21]. Effective segmentation models must therefore jointly capture robust global semantic consistency while preserving fine-grained structural details.

Convolutional neural networks (CNNs) have dominated medical image segmentation owing to their strong spatial inductive bias and capability in modeling local textures and boundaries [19,13,21,7]. However, their inherently local receptive fields limit global context modeling. Vision Transformers (ViTs) [9], empowered by self-attention mechanisms and large-scale self-supervised pre-training, offer superior global representation capabilities. In particular, foundation models such as DINOv3 [20] provide powerful and transferable semantic features that generalize across domains. Nevertheless, adapting such pre-trained Transformers to medical image segmentation introduces a critical challenge: while their global semantic representations are robust, their single-scale token embeddings lack explicit hierarchical structure and boundary-aware inductive bias. This limitation becomes especially pronounced when the backbone is largely frozen to preserve pre-trained semantics and avoid overfitting to limited medical data. Existing hybrid CNN-Transformer architectures typically rely on parallel feature extraction or late-stage feature fusion [6,5,28,29]. However, such strategies often fail to effectively inject structural priors into frozen Transformer representations. Simply fusing convolutional features does not fundamentally transform token embeddings into structure-aware hierarchical representations. Therefore, a principled adaptation mechanism is required to enhance structural perception in frozen foundation Transformers without disrupting their semantic stability.

To address this challenge, we propose **Glasses**, a structure-enhanced adaptation framework built upon a largely frozen DINOv3 backbone. Specifically, we design a Multi-Scale Deformable Attention Adapter that performs cross-scale interaction between lightweight convolutional structural priors and intermediate Transformer tokens, transforming single-scale embeddings into hierarchical multi-scale representations enriched with both global semantics and local structural cues while preserving semantic stability [31]. In parallel, an independent Edge-Aware Convolutional Encoder explicitly learns boundary-enriched multi-scale convolutional features under nested boundary supervision, providing high-resolution structural information complementary to the adapter-enhanced features [23,26]. During decoding, the adapter-enhanced Transformer features and boundary-enriched convolutional representations are progressively fused within a U-Net-style decoder to produce high-resolution and structurally consistent segmentation outputs [19].

The main contributions are three-fold: (1) We propose a structure-enhanced adaptation framework with largely frozen DINOv3 backbone for medical image segmentation. (2) We design a Multi-Scale Deformable Attention Adapter that enables cross-scale interaction between lightweight convolutional structural pri-

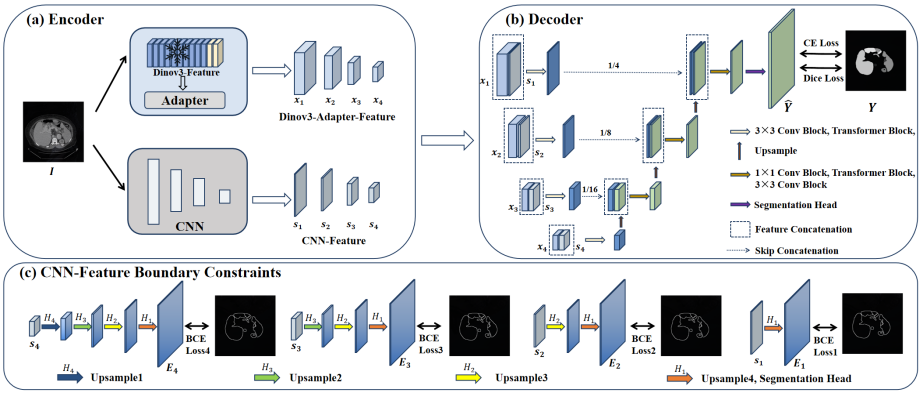

Fig. 1. Overview of the Glasses framework, including a largely frozen DINOv3 backbone with a Multi-Scale Deformable Attention Adapter, an Edge-Aware Convolutional Encoder with nested boundary supervision, and a feature fusion decoder for final segmentation prediction.

ors and intermediate Transformer tokens, providing structure-aware adaptation without modifying the frozen backbone. (3) We introduce an Edge-Aware Convolutional Encoder with nested boundary supervision to provide complementary boundary-enriched multi-scale representations.

2 Methodology

2.1 Overall Framework

An overview of the proposed Glasses framework is illustrated in Fig. 1. Glasses adopts a dual-encoder architecture built upon a largely frozen DINOv3 backbone, enhancing structural perception while preserving pre-trained semantic representations. Given an input image $I \in \mathbb{R}^{H\times W\times C}$, the Transformer-based branch first extracts intermediate token embeddings from a largely frozen DINOv3, with only the final layers fine-tuned. Instead of directly using single-scale token outputs, we introduce a Multi-Scale Deformable Attention Adapter to perform structure-aware representation adaptation [31]. Through cross-scale interaction between lightweight convolutional structural priors and intermediate Transformer tokens, the Adapter transforms single-scale embeddings into hierarchical multi-scale semantic features $\{x_1, x_2, x_3, x_4\}$, where lower-index features correspond to higher spatial resolutions. These representations primarily preserve global semantic consistency inherited from DINOv3, while incorporating moderate structural cues to facilitate spatially aware adaptation. In parallel, an independent Edge-Aware Convolutional Encoder processes the same input image to produce boundary-enriched multi-scale structural representations $\{s_1, s_2, s_3, s_4\}$. Unlike the lightweight structural priors used inside the adapter, this branch is explicitly optimized for high-resolution structural modeling. Nested boundary

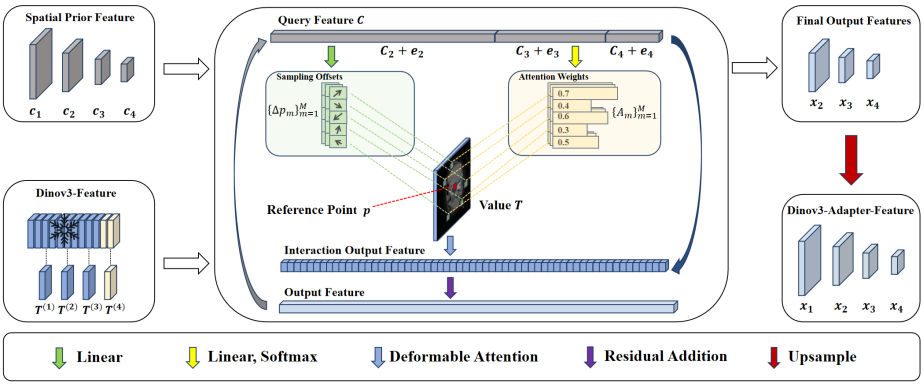

Fig. 2. Architecture of the Multi-Scale Deformable Attention Adapter.

supervision is applied at multiple scales to encourage precise contour localization and spatial continuity. This design ensures that structural representations are learned in a dedicated manner rather than implicitly inferred from semantic features. The two sets of features, $\{x_i\}$ and $\{s_i\}$, are progressively fused within a U-Net-style hierarchical decoder. At each decoding stage, the scale-aligned semantic feature x_i and structural feature s_i are concatenated and refined to recover higher spatial resolution.

2.2 Multi-Scale Deformable Attention Adapter

As illustrated in Fig. 2, the proposed Multi-Scale Deformable Attention Adapter serves as a representation transformation mechanism that restructures flat Transformer token embeddings into hierarchical structure-aware features under a largely frozen backbone constraint [31,20]. Instead of directly modifying the internal layers of DINOv3, the adapter introduces an external structural token space that progressively aligns with semantic representations across multiple Transformer depths.

Given an input image I, a lightweight Spatial Prior Module (SPM) first produces hierarchical convolutional features $\{c_1, c_2, c_3, c_4\}$. Features c_2, c_3, c_4, corresponding to resolutions of 1/8, 1/16 and 1/32, are projected to the Transformer embedding dimension and flattened into token sequences. Learnable level embeddings e_i are added to encode scale identity, and the resulting tokens are concatenated to form the initial structural representation $C^{(0)}$, maintaining explicit multi-scale awareness during subsequent interactions.

Intermediate Transformer tokens $\{T^{(1)}, T^{(2)}, T^{(3)}, T^{(4)}\}$ are extracted from selected DINOv3 layers according to predefined interaction indices. At the k-th interaction stage, the structural tokens $C^{(k-1)}$ serve as queries, while the Transformer tokens $T^{(k)}$ provide semantic key-value representations. For each query location q and attention head h, M sampling offsets $\Delta p_{k,h,m}$ and attention weights $A_{k,h,m}$ are predicted relative to a normalized reference point p_q. The

deformable attention response is computed as:

$$\mathrm{MSDeformAttn}(q, p_q, T^{(k)}) = \mathrm{Proj}\Big(\mathrm{Concat}_h\big(\sum_{m=1}^{M} A_{k,h,m} \cdot T^{(k)}(p_q + \Delta p_{k,h,m})\big)\Big). \tag{1}$$

The structural tokens are then updated in a residual manner:

$$C^{(k)} = C^{(k-1)} + \mathrm{MSDeformAttn}\big(C^{(k-1)}, T^{(k)}\big). \tag{2}$$

Through iterative deformable alignment, structural tokens progressively absorb semantic information from Transformer layers of increasing abstraction. The originally convolutional structural tokens thus evolve into semantically enriched yet spatially grounded hierarchical representations. After the final stage, $C^{(k)}$ is reshaped back into multi-scale feature maps $\{x_2, x_3, x_4\}$. A higher-resolution feature x_1 is obtained by upsampling x_2 and integrating it with c_1. The resulting hierarchical representations $\{x_1, x_2, x_3, x_4\}$ are forwarded to the decoder, providing scale-aware semantic features while preserving the stability of the largely frozen DINOv3 backbone.

2.3 Edge-Aware Convolutional Encoder

Although the adapter branch leverages lightweight structural priors via the SPM, it primarily serves representation interaction rather than explicit structural modeling. To provide complementary boundary-enriched multi-scale representations, we introduce an Edge-Aware Convolutional Encoder with nested boundary supervision, as illustrated in Fig. 1(c).

The encoder comprises four downsampling stages that produce hierarchical feature maps $\{s_1, s_2, s_3, s_4\}$. Due to convolutional locality and dense spatial connectivity, these features preserve spatial continuity and fine-grained texture information, forming a complementary structural representation stream alongside the adapter-enhanced semantic features $\{x_i\}$.

Boundary prediction is formulated as a nested multi-scale projection governed by a shared predictor $H_1(\cdot)$. The highest-resolution boundary response is $E_1 = H_1(s_1)$. For deeper features, boundary maps are obtained through progressive alignment followed by projection into the same boundary representation space: $E_2 = H_1(H_2(s_2))$, $E_3 = H_1(H_2(H_3(s_3)))$, $E_4 = H_1(H_2(H_3(H_4(s_4))))$. Here, H_2, H_3, and H_4 denote stage-specific alignment operators that progressively restore spatial resolution and channel compatibility. Together with the shared predictor H_1, these operators form a recursive projection pathway applied consistently across hierarchical levels. This nested design enforces multi-scale structural features to reside in a shared boundary representation manifold. Consequently, structural representations across different depths are regularized toward a consistent contour space, promoting cross-scale structural coherence and stabilizing boundary learning.

The encoder outputs both structural features $\{s_i\}$ and boundary responses $\{E_i\}$. The former are fused with semantic features during decoding, while the latter provide explicit supervision for contour refinement.

2.4 Loss Function

Glasses is optimized under joint semantic and boundary supervision. Given an input image I, the network produces a segmentation prediction $\hat{Y}$ and multi-scale boundary responses $\{E_i\}_{i=1}^{4}$. For semantic segmentation, we adopt a hybrid loss combining cross-entropy (CE) and Dice losses:

$$L_{\text{seg}} = \alpha L_{\text{CE}} + \beta L_{\text{Dice}}, \tag{3}$$

where $\alpha = 0.5$ and $\beta = 0.8$. To regularize structural learning, boundary supervision is imposed on all hierarchical edge predictions. The ground-truth boundary map is derived from the segmentation mask using gradient operators. Each boundary response E_i is optimized with a BCE loss, with progressively smaller weights assigned to deeper predictions to account for resolution differences:

$$L_{\text{edge}} = \sum_{i=1}^{4} w_i L_{\text{BCE}}(E_i). \tag{4}$$

The **overall loss** is defined as:

$$L = L_{\text{seg}} + \lambda(t)\, L_{\text{edge}}, \tag{5}$$

where $\lambda(t)$ gradually decays during training, placing greater emphasis on boundary learning at early stages while allowing semantic optimization to dominate later. During inference, only the final segmentation prediction is used.

3 Experiments and Results

3.1 Datasets and Implementation Details

We evaluate Glasses on two public benchmarks: the **Synapse** multi-organ CT dataset [6] and the **ISIC 2018** skin lesion segmentation dataset [8]. For Synapse, abdominal CT volumes are sliced into 2D axial images following the standard protocol. The official train-test split is adopted to ensure fair comparison with existing SOTA methods. All slices are resized to 224×224 with random rotation and flipping applied for augmentation. For ISIC 2018, the task is binary lesion segmentation. Images are resized to 256×256 and normalized with ImageNet statistics. All experiments are implemented in PyTorch. DINOv3 ViT-B/16 is initialized with publicly released self-supervised weights and kept frozen except for the last two Transformer blocks and normalization layers. Optimization is performed with AdamW. The fine-tuned backbone layers use an initial learning rate of 1×10^{-5}, while newly introduced modules use 1×10^{-4}. A polynomial decay schedule is applied. The batch size is 20 for Synapse and 16 for ISIC 2018. Training is conducted for 400 and 200 epochs, respectively. **Evaluation metrics** include Dice and Hausdorff Distance (HD) for Synapse, and Dice, Sensitivity, Specificity, and Accuracy for ISIC 2018.

Table 1. Quantitative comparison of Dice (%) and HD (mm) on the Synapse dataset. The best and second-best results are highlighted in **bold** and underline.

Method	Dice↑	HD↓	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
U-Net [19]	76.85	39.70	89.07	69.72	77.77	68.60	93.43	53.98	86.67	75.58
nnUNet [13]	79.40	14.57	**90.24**	44.23	87.30	**87.94**	90.10	70.10	85.45	79.82
TransUNet [6]	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
Swin-Unet [5]	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
Mamba-Unet [24]	80.58	21.95	87.23	68.25	84.66	80.41	94.03	58.92	90.12	81.05
HiFormer [11]	80.39	14.70	86.21	65.69	85.23	79.77	94.61	59.52	90.99	81.08
MISSFormer [12]	81.96	18.20	86.99	68.65	85.21	82.00	94.41	65.67	**91.92**	80.81
Effformer [14]	80.79	17.00	85.81	66.89	84.10	81.81	94.80	62.25	91.05	79.58
DAE-Former [3]	82.63	16.39	87.84	71.65	87.66	82.39	95.08	63.93	91.82	80.77
GLoG-GSUnet [10]	83.36	23.02	88.43	**73.20**	84.50	80.98	95.14	**71.02**	91.51	82.10
CASTformer [27]	82.55	22.73	89.05	67.48	86.05	82.17	**95.61**	67.49	91.00	81.55
PVT-CASCADE [18]	81.06	20.23	83.01	70.59	82.23	80.37	94.08	64.43	90.10	83.69
Rskd [15]	82.33	17.21	87.11	68.56	86.68	84.54	93.51	65.40	90.06	82.81
Ours	**83.52**	**12.28**	88.34	70.59	**88.38**	86.24	95.12	63.25	91.66	**84.60**

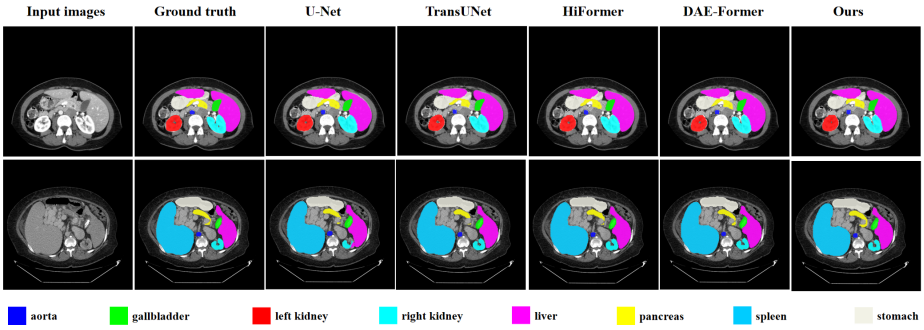

Fig. 3. Visual comparison of segmentation results on the Synapse dataset. Glasses generates spatially coherent boundaries, particularly for anatomically complex regions.

3.2 Comparison with SOTA Methods

We compare Glasses with representative convolution-based, Transformer-based, and hybrid methods on the Synapse and ISIC 2018 datasets. All results follow standard evaluation protocols.

On Synapse, quantitative results are summarized in Table 1. Glasses achieves a mean Dice of 83.52% and a Hausdorff Distance of 12.28 mm, outperforming all SOTA methods. At the organ level, Glasses attains the highest Dice on multiple structures, including Kidney (L) and Stomach, and remains competitive on other organs. Qualitative comparisons are shown in Fig. 3. Methods including U-Net [19], TransUNet [6], HiFormer [11], and DAE-Former [3] occasionally produce fragmented predictions or irregular contours, especially in low-contrast regions and at organ interfaces. In contrast, Glasses generates continuous and anatomically coherent boundaries with reduced leakage and improved structural completeness. On ISIC 2018, results are presented in Table 2. Glasses achieves the highest Dice of 91.60% and Sensitivity of 92.17%. Meanwhile, Specificity

Table 2. Quantitative comparison of Dice (%), Sensitivity (%), Specificity (%) and Accuracy (%) on the ISIC 2018 dataset. The best and second-best results are highlighted in **bold** and underline.

Method	Dice↑	Sensitivity↑	Specificity↑	Accuracy↑
U-Net [19]	85.45	88.00	96.97	94.04
nnUNet [13]	90.16	91.02	97.55	96.40
TransUNet [6]	84.99	85.78	96.53	94.52
Swin-Unet [5]	89.46	90.56	97.98	**96.45**
MedT [22]	83.89	82.52	96.37	93.58
FAT-Net [25]	89.03	91.00	96.99	95.78
TMU-Net [4]	90.59	90.38	97.46	96.03
VM-UNet [17]	89.71	91.12	96.13	94.91
TransNorm [2]	89.51	87.50	97.90	95.80
DAE-Former [3]	91.47	91.20	97.80	96.41
Effformer [14]	89.04	88.61	96.98	95.19
CCA [30]	89.16	90.98	95.94	94.77
MCGU-Net [1]	89.50	84.80	**98.60**	95.50
Ours	**91.60**	**92.17**	96.74	95.39

Table 3. Ablation study on the Synapse dataset. We progressively add the Adapter, backbone fine-tuning, CNN branch, and two boundary supervision strategies (Independent vs. Nested). Dice (%) and HD (mm) are reported.

No.	DINOv3	Adapter	Finetune	CNN	Boundary	Dice↑	HD↓
1	✓				×	75.25	29.32
2	✓	✓			×	81.53	25.71
3	✓	✓	✓		×	81.83	25.11
4	✓	✓	✓	✓	×	82.49	23.32
5	✓	✓	✓	✓	Independent	82.99	19.49
6	✓	✓	✓	✓	Nested	**83.52**	**12.28**

and Accuracy remain competitive, demonstrating that our model improves recall without sacrificing precision.

3.3 Ablation Studies

To systematically evaluate the contribution of each component, we conduct ablation experiments on the Synapse dataset, with results summarized in Table 3. Starting from the frozen DINOv3 backbone (group 1), the model achieves limited segmentation accuracy. After introducing the Multi-Scale Deformable Adapter (group 2), performance improves substantially, indicating the importance of structural feature interaction for medical segmentation. Allowing limited fine-tuning (group 3) of the final Transformer blocks further enhances performance, suggesting improved alignment between pre-trained representations and the target domain. Notably, introducing the parallel convolutional branch (group 4) leads to consistent improvements in both Dice and HD, indicating enhanced spatial modeling capability. Adding independent boundary supervision (group 5) further increases Dice and reduces HD. Replacing independent supervision with the proposed nested multi-scale boundary modeling (group 6) achieves the best overall performance, reaching 83.52% for Dice and 12.28 mm for HD.

4 Conclusion

In this work, we present Glasses, a structure-enhanced adaptation framework for medical image segmentation built upon a largely frozen DINOv3 backbone. By introducing a Multi-Scale Deformable Attention Adapter, Glasses augments pre-trained Transformer tokens with convolutional structural priors, enabling improved modeling of anatomical details while preserving global semantic representations. In parallel, an independent Edge-Aware Convolutional Encoder provides boundary-enriched multi-scale features that complement the adapter-enhanced Transformer representations during decoding, leading to more precise and spatially consistent segmentation results. Extensive experimental results on the Synapse and ISIC 2018 datasets demonstrate the strong and consistent performances of Glasses across different segmentation tasks. Ablation studies further validate the effectiveness and necessity of the proposed components.

Acknowledgments. The study was supported in part by the National Natural Science Foundation of China under Grant 62271405, in part by the Key Research and Development Program of Shaanxi under Grant 2025CY-YBXM-039, and in part by the Natural Science Foundation of Chongqing, China, under Grant CSTB2023NSCQ-MSX0286.

Disclosure of Interests. The authors declare that there are no conflicts of interest.

References

1. Asadi-Aghbolaghi, M., Azad, R., Fathy, M., Escalera, S.: Multi-level context gating of embedded collective knowledge for medical image segmentation. arXiv preprint arXiv:2003.05056 (2020)
2. Azad, R., Al-Antary, M.T., Heidari, M., Merhof, D.: Transnorm: Transformer provides a strong spatial normalization mechanism for a deep segmentation model. IEEE access **10**, 108205–108215 (2022)
3. Azad, R., Arimond, R., Aghdam, E.K., Kazerouni, A., Merhof, D.: Dae-former: Dual attention-guided efficient transformer for medical image segmentation. In: International workshop on predictive intelligence in medicine. pp. 83–95. Springer (2023)
4. Azad, R., Heidari, M., Wu, Y., Merhof, D.: Contextual attention network: Transformer meets u-net. In: International workshop on machine learning in medical imaging. pp. 377–386. Springer (2022)
5. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision. pp. 205–218. Springer (2022)
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. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368 (2019)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
10. Eghbali, N., Bagher-Ebadian, H., Alhanai, T., Ghassemi, M.M.: Glog-csunet: Enhancing vision transformers with adaptable radiomic features for medical image segmentation. arXiv preprint arXiv:2501.02788 (2025)
11. Heidari, M., Kazerouni, A., Soltany, M., Azad, R., Aghdam, E.K., Cohen-Adad, J., Merhof, D.: Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 6202–6212 (2023)
12. Huang, X., Deng, Z., Li, D., Yuan, X.: Missformer: An effective medical image segmentation transformer. arXiv preprint arXiv:2109.07162 (2021)
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. Li, Y., Yuan, G., Wen, Y., Hu, J., Evangelidis, G., Tulyakov, S., Wang, Y., Ren, J.: Efficientformer: Vision transformers at mobilenet speed. Advances in Neural Information Processing Systems **35**, 12934–12949 (2022)
15. Liang, P., Chen, J., Chang, Q., Yao, L.: Rskd: Enhanced medical image segmentation via multi-layer, rank-sensitive knowledge distillation in vision transformer models. Knowledge-Based Systems **293**, 111664 (2024)
16. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (2017)
17. Liu, J., Yang, H., Zhou, H., Xi, Y., Yu, L., Yu, Y., Liang, Y., Shi, G., Zhang, S., Zheng, H., et al.: Vm-unet: Vision mamba unet for medical image segmentation. arXiv preprint arXiv:2402.02491 (2024)
18. Rahman, M.M., Marculescu, R.: Medical image segmentation via cascaded attention decoding. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 6222–6231 (2023)
19. 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)
20. 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)
21. Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J.: Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE transactions on medical imaging **35**(5), 1299–1312 (2016)
22. Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M.: Medical transformer: Gated axial-attention for medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 36–46. Springer (2021)
23. Wang, R., Chen, S., Ji, C., Fan, J., Li, Y.: Boundary-aware context neural network for medical image segmentation. Medical image analysis **78**, 102395 (2022)

24. Wang, Z., Zheng, J.Q., Zhang, Y., Cui, G., Li, L.: Mamba-unet: Unet-like pure visual mamba for medical image segmentation. arXiv preprint arXiv:2402.05079 (2024)
25. Wu, H., Chen, S., Chen, G., Wang, W., Lei, B., Wen, Z.: Fat-net: Feature adaptive transformers for automated skin lesion segmentation. Medical image analysis **76**, 102327 (2022)
26. Wu, Z., Leng, L.: Edge-aware lightweight network for medical image segmentation. In: International Conference on Multimedia Information Technology and Applications. pp. 79–85. Springer (2025)
27. You, C., Zhao, R., Liu, F., Dong, S., Chinchali, S., Topcu, U., Staib, L., Duncan, J.: Class-aware adversarial transformers for medical image segmentation. Advances in neural information processing systems **35**, 29582–29596 (2022)
28. Zhang, Y., Liu, H., Hu, Q.: Transfuse: Fusing transformers and cnns for medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 14–24. Springer (2021)
29. 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)
30. Zhu, J., Huang, C., Xi, H., Cui, H.: Cca: Contrastive cluster assignment for supervised and semi-supervised medical image segmentation. Neural Networks **188**, 107415 (2025)
31. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)