SegDINO: Introducing Multi-Scale Structure into DINO for Efficient Medical Image Segmentation

Sicheng Yang[1], Hongqiu Wang[1], Zhaohu Xing[1], Sixiang Chen[1], Qiuxia Yang[2], Yize Mao[3], Guang Yang[4], and Lei Zhu[1]✉

[1] The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
[2] Department of Radiology, State Key Laboratory of Oncology in South China, Guangdong Provincial Clinical Research Center for Cancer, Sun Yat-sen University Cancer Center, Guangzhou 510060, China
[3] Department of Pancreatobiliary Surgery, State Key Laboratory of Oncology in South China, Sun Yat-sen University Cancer Center, Guangzhou, Guangdong, China
[4] Imperial College London, London, United Kingdom

Abstract. Self-supervised DINO models provide strong transferable visual representations, yet applying them directly to image segmentation remains challenging. Existing approaches commonly rely on heavy decoders with complex upsampling, introducing substantial parameter and computational overhead. We observe that introducing scale into DINO features is far more critical than increasing decoder capacity. In this work, we present SegDINO, an efficient segmentation framework that integrates a DINOv3 backbone with lightweight scale modeling. SegDINO introduces Token Pyramid Adaptation (TPA) to reorganize intermediate DINO features into a pseudo multi-scale hierarchy, and Scale-Aware Decoding (SAD) for efficient intra-scale refinement and top-down multi-scale propagation. We further curate PanCT, a new CT dataset containing 284 patients with expert-annotated pancreatic tumors, to assess SegDINO's ability to handle difficult small-lesion cases. Extensive experiments on PanCT and three public benchmarks demonstrate that SegDINO achieves state-of-the-art results with high efficiency. The code is available at https://github.com/script-Yang/segdino_v2.

Keywords: Medical Image Segmentation · Multi-Scale Representation

1 Introduction

Medical image segmentation plays a central role in medical image analysis, serving as a foundation for downstream clinical applications, such as treatment planning and computer-aided diagnosis [3]. Despite remarkable progress achieved by convolutional networks [22], transformer-based models [14], diffusion-based architectures [26], and Mamba-based frameworks [17], these approaches often

✉ Lei Zhu (leizhu@hkust-gz.edu.cn) is the corresponding author.

struggle to achieve strong generalization in medical imaging scenarios, particularly when training data are limited or acquired under varying imaging protocols [29]. Recent SAM-based segmentation models [12,27] exhibit strong zero-shot abilities, but their heavy architectural designs result in high fine-tuning costs when adapted to downstream tasks, limiting their practical efficiency.

Rather than training segmentation networks entirely from scratch, a growing line of work instead leverages pretrained visual representations to obtain richer semantic and structural priors [30]. Within this paradigm, self-supervised foundation models from the DINO family have emerged as especially powerful due to their strong cross-domain transferability [25]. DINO [5] and DINOv2 [20] have proven effective in general-purpose representation learning [31], enabling robust performance on downstream detection [8] and segmentation tasks [2,9,15]. Most recently, DINOv3 [23] introduced substantial improvements in self-supervised pretraining, providing even stronger invariance and scalability.

However, effectively adapting DINO-based representations to segmentation tasks remains a non-trivial challenge. In practice, we observe that features extracted by DINO lack an explicitly defined multi-scale hierarchy, which makes it difficult for overly simple decoders to capture fine-grained semantic structures, particularly for subtle targets (e.g., small lesion segmentation). Several prior works adopt relatively heavy decoder designs, such as stacking multiple multi-scale fusion modules [9] or employing complex upsampling pipelines [28]. These designs largely inherit architectural principles from semantic segmentation networks originally developed for natural images, and while effective, they introduce substantial parameter overhead and computational cost. We argue that, given the strong representational capacity of DINO features, the key challenge lies in carefully designing scale modeling mechanisms, rather than increasing decoder capacity. With an appropriately designed decoding strategy, competitive segmentation performance can be achieved with significantly fewer parameters.

In this paper, we introduce SegDINO, an efficient segmentation framework that adapts DINO-based representations for medical image segmentation using scale modeling. We design Token Pyramid Adaptation (TPA) to reorganize representations extracted from different DINO depths into a pseudo pyramid, injecting scale diversity into patch-level features. We further introduce Scale-Aware Decoding (SAD), a lightweight decoding strategy that combines intra-scale refinement with top-down inter-scale propagation. Finally, we propose a new dataset for small-lesion medical lesion segmentation, named PanCT, which includes 284 patients collected from the Radiology Department, with primary lesions annotated by two experienced radiologists. Extensive experiments on four datasets demonstrate that SegDINO consistently achieves superior segmentation performance while maintaining high efficiency.

2 Method

SegDINO mainly consists of three components: (1) a DINO-based feature extraction module that collects intermediate representations from multiple depths to

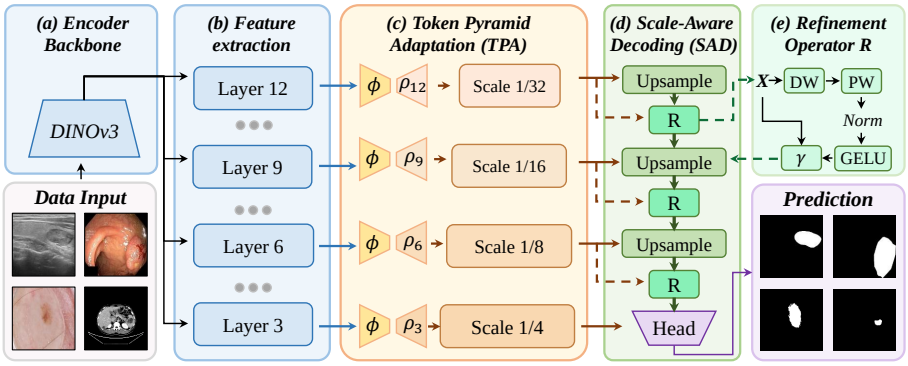

Fig. 1: An overview of the proposed SegDINO framework. A frozen DINOv3 encoder extracts intermediate features from multiple layers. These features are first projected and reorganized by the Token Pyramid Adaptation (TPA) module to construct a pseudo multi-scale pyramid, introducing scale-aware representations. The resulting pyramid is then processed by the Scale-Aware Decoder (SAD), which performs intra-scale refinement and inter-scale information propagation using a residual refinement operator R. Finally, a lightweight prediction head produces the segmentation output.

capture both low-level structures and high-level semantics, (2) a Token Pyramid Adaptation (TPA) module that reorganizes these representations into a pseudo pyramid by introducing scale diversity, and (3) a Scale-Aware Decoding (SAD) module that performs lightweight intra-scale refinement and inter-scale propagation to generate the final segmentation prediction. Fig. 1 illustrates the overall architecture of SegDINO. We describe the details in the following subsections.

2.1 Encoder backbone

We adopt a pretrained DINOv3 [23] as the encoder. As illustrated in Fig. 1 (a), given an input image $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$, it is partitioned into $N = (H/p) \times (W/p)$ non-overlapping patches and embedded into an initial token sequence $\mathbf{Z}^{(0)} \in \mathbb{R}^{N\times d}$. The tokens are then iteratively updated by a stack of L Transformer blocks $\mathcal{B}_\ell$ as

$$\mathbf{Z}^{(\ell)} = \mathcal{B}_\ell\left(\mathbf{Z}^{(\ell-1)}\right), \qquad \ell = 1, \ldots, L. \tag{1}$$

As shown in Fig. 1 (b), to harvest both low-level structure and high-level semantics, we collect intermediate token matrices from a subset of layers

$$\mathcal{L} = \{\ell_1, \ell_2, \ldots, \ell_K\} \subseteq \{1, \ldots, L\}. \tag{2}$$

For each $\ell_k \in \mathcal{L}$, we directly take the patch tokens $\mathbf{Z}_p^{(\ell_k)} \in \mathbb{R}^{N\times d}$ from the ViT output and discard any non-patch tokens (e.g., class or register tokens). The encoder output is the multi-level token set $\left\{\mathbf{Z}_p^{(\ell_k)}\right\}_{k=1}^{K}$.

2.2 Token Pyramid Adaptation

The intermediate token set $\{\mathbf{Z}_p^{(\ell_k)}\}_{k=1}^K$ extracted from the DINO encoder captures representations of increasing semantic abstraction across transformer depth. However, all token matrices are defined on the same patch grid and therefore share an identical spatial resolution. Such a uniform-scale representation is suboptimal for image segmentation, where hierarchical feature scales are essential for integrating global semantic context with fine-grained boundary details.

As illustrated in Fig. 1(c), we introduce a simple yet effective *Token Pyramid Adaptation (TPA)* strategy to explicitly inject scale diversity into DINO features. Inspired by the feature pyramid design in FPN [16], TPA reorganizes intermediate DINO tokens into a hierarchy of spatial feature maps. Specifically, for each intermediate token matrix $\mathbf{Z}_p^{(\ell_k)} \in \mathbb{R}^{N\times d}$, we reshape the patch tokens back to a 2D feature map according to the original patch layout, thereby restoring their spatial arrangement.

$$\mathbf{F}_k = \text{reshape}\left(\mathbf{Z}_p^{(\ell_k)}\right) \in \mathbb{R}^{d\times H_p\times W_p}, \tag{3}$$

where $H_p = H/p$ and $W_p = W/p$ denote the height and width of the patch grid.

Since different transformer layers may produce features with heterogeneous channel statistics, we apply a 1×1 convolution to project all feature maps into a unified embedding space:

$$\tilde{\mathbf{F}}_k = \phi(\mathbf{F}_k), \qquad \phi : \mathbb{R}^d \to \mathbb{R}^{d'}. \tag{4}$$

This projection preserves spatial correspondence while enabling subsequent cross-level interactions. Next, to introduce spatial hierarchy for dense decoding, we design scale-specific spatial transformations to the projected features:

$$\mathbf{P}_k = \rho_k(\tilde{\mathbf{F}}_k), \tag{5}$$

where $\rho_k(\cdot)$ denotes a lightweight resizing operator, implemented using strided convolution. As a result, TPA constructs a pseudo feature pyramid with progressively varying spatial resolutions (e.g., $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$ of the input resolution), while preserving the hierarchical semantic representations extracted by DINO.

2.3 Scale-Aware Decoding

With the pseudo pyramid $\mathcal{P} = \{\mathbf{P}_k\}_{k=1}^K$ constructed by the TPA, as shown in Fig. 1 (d), we introduce a *Scale-Aware Decoding (SAD)* strategy to generate the final segmentation prediction in a lightweight manner. The decoding process is decomposed into two complementary components: *intra-scale updating*, which performs preliminary refinement within each individual scale, and *inter-scale propagation*, which enables information transfer across different spatial resolutions. This design allows the decoder to progressively integrate coarse semantic information with fine-grained spatial details while avoiding heavy modules.

Lightweight residual refinement. As illustrated in Fig. 1 (e), we first define a lightweight residual refinement operator $\mathcal{R}(\cdot)$ to update features:

$$\mathcal{R}(\mathbf{X}) = \mathbf{X} + \gamma \cdot \sigma(\mathrm{Norm}(\mathrm{PW}(\mathrm{DW}(\mathbf{X})))), \tag{6}$$

where $\mathrm{DW}(\cdot)$ and $\mathrm{PW}(\cdot)$ denote depthwise and pointwise convolutions, respectively, $\mathrm{Norm}(\cdot)$ is GroupNorm, $\sigma(\cdot)$ is the GELU activation, and γ is a learnable residual scaling parameter initialized to 0 for stable optimization. This operator introduces only minimal computational overhead.

Intra-scale updating. After spatial resizing and channel unification in TPA, each pyramid level is first pre-conditioned independently:

$$\mathbf{L}_k = \mathcal{R}(\mathbf{P}_k), \qquad k = 1, \ldots, K. \tag{7}$$

This step constructs scale-specific representations by locally refining features within each resolution, preparing them for subsequent multi-scale interaction.

Inter-scale propagation. To fuse representations across different spatial scales, we employ a top-down inter-scale pathway that progressively integrates coarse- and fine-scale features. Let $\mathrm{Up}(\cdot; \mathbf{L}_k)$ denote bilinear interpolation that resizes its input to match the spatial resolution of $\mathbf{L}_k$. Starting from the coarsest level, inter-scale propagation is performed recursively:

$$\mathbf{X}_K = \mathcal{R}(\mathbf{L}_K), \tag{8}$$

$$\mathbf{X}_k = \mathcal{R}(\mathrm{Up}(\mathbf{X}_{k+1}; \mathbf{L}_k) + \mathbf{L}_k), \qquad k = K-1, \ldots, 1. \tag{9}$$

Finally, the segmentation prediction is obtained from the finest-scale representation using a linear prediction head:

$$\mathbf{Y} = \mathcal{H}(\mathbf{X}_1), \tag{10}$$

where $\mathcal{H}(\cdot)$ is implemented as a 1×1 convolution. Through the proposed Scale-Aware Decoding, multi-scale features extracted by TPA are efficiently decoded into the final segmentation output.

3 Experiments

3.1 Datasets

PanCT Dataset. We collected PanCT from the Radiology Department, comprising 284 patients with confirmed pancreatic cancer, including 243 patients for training and 41 for internal testing. The training cohort has a median age of 59 years with 94 female and 149 male patients, while the test cohort has a median age of 56 years with 20 female and 21 male patients. All scans were independently annotated by two experienced radiologists, and all patient data were de-identified under institutional ethical approval. Each 3D CT volume is converted into 2D axial slices for efficient slice-level modeling and fair comparison with 2D baselines; we acknowledge the loss of inter-slice context and leave

2.5D/3D extension for future work. The datasets generated and/or analyzed during the current study are not publicly available, but are available from the corresponding author on reasonable request.

Public benchmarks. We evaluate SegDINO on three public medical image segmentation benchmarks. TN3K [10] is a large-scale thyroid nodule segmentation dataset, containing 3,493 ultrasound images with pixel-level annotations collected from multiple hospitals. Kvasir-SEG [11] is a polyp segmentation dataset derived from colonoscopy examinations, consisting of 1,000 images with high-quality expert annotations. ISIC [7] is a skin lesion segmentation benchmark, providing 2,750 dermoscopic images annotated for lesion boundaries and covering a wide range of lesion types and acquisition conditions.

3.2 Implementation Details

Experimental Settings. For each public dataset, we follow the official training–testing split provided by the organizers to ensure fair comparison. All images are resized to 256×256 for consistent input resolution across models, and normalized with the same mean and standard deviation parameters as in DINOv3 [23]. We implement all experiments using the PyTorch framework [21]. The models are optimized using AdamW [18] with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-4}. Cross-entropy loss is employed as the training objective. Training is conducted for 50 epochs with a batch size of 4. In this work, we adopt the DINOv3-S backbone, from which intermediate features from the 3rd, 6th, 9th, and 12th layers of DINOv3 are extracted. All experiments are run on a cloud platform equipped with four NVIDIA RTX A6000 GPUs.

Evaluation Metrics. For all datasets, we employ the Dice similarity coefficient (DSC, higher is better) to measure the overlap between predictions and the ground truth, together with the 95th percentile Hausdorff Distance (HD95, lower is better) to evaluate boundary localization accuracy.

3.3 Comparison with SOTA Methods

We compare SegDINO with a diverse set of state-of-the-art segmentation models, including U-Net [22], SegNet [4], R2U-Net [1], Attention U-Net [19], TransUNet [6], U-NeXt [24], and U-KAN [13].

Quantitative Comparisons. Table 1 shows that SegDINO consistently outperforms all competing methods across all four datasets. On TN3K, SegDINO improves DSC by 3.64% over the strongest baseline TransUNet, while reducing HD95 by 7.50. On Kvasir-SEG and ISIC, SegDINO achieves DSC gains of 4.64% and 2.41% over SegNet and U-KAN, respectively, accompanied by HD95 reductions of 8.04 and 6.64. On PanCT, a small lesion segmentation task, SegDINO maintains its advantage by achieving the best DSC and HD95 scores, indicating superior accuracy on small-scale targets.

Visual Comparisons. We further present qualitative segmentation results for SegDINO and several competing methods on representative samples from the

Table 1: Comparison with state-of-the-art models.

Methods	TN3K		Kvasir-SEG		ISIC		PanCT	
	DSC↑	HD95↓	DSC↑	HD95↓	DSC↑	HD95↓	DSC↑	HD95↓
U-Net [22]	0.7945	24.59	0.7916	41.58	0.8187	25.12	0.8416	5.76
SegNet [4]	0.7924	22.74	0.8415	25.89	0.8327	21.41	0.7966	6.97
R2U-Net [1]	0.6886	27.89	0.7367	45.64	0.8102	25.36	0.7987	5.58
Att-UNet [19]	0.8015	24.64	0.8016	31.92	0.8275	26.12	0.8373	3.36
TransUNet [6]	0.8027	23.95	0.8054	37.67	0.8186	24.76	0.8465	4.93
U-NeXt [24]	0.7285	31.40	0.6271	58.32	0.8230	23.63	0.7885	7.82
U-KAN [13]	0.7866	24.49	0.7217	36.92	0.8341	23.57	0.8111	3.82
SegDINO	**0.8391**	**16.45**	**0.8879**	**17.85**	**0.8582**	**16.93**	**0.8657**	**2.61**

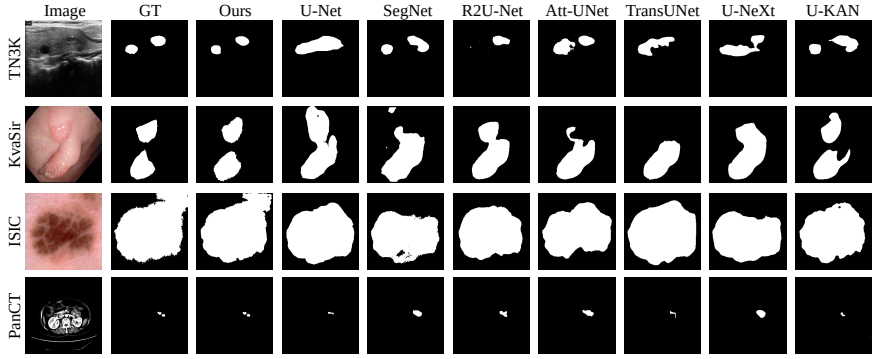

Fig. 2: Visual comparisons of proposed SegDINO and other methods.

four datasets. As shown in Fig. 2, SegDINO produces more accurate object boundaries and cleaner region predictions. On the PanCT dataset for small-lesion segmentation, SegDINO better captures fine-grained structures.

Efficiency Comparisons. As illustrated in Fig. 3, SegDINO achieves strong segmentation performance with low computational overhead. The model contains 27.68M parameters in total, with 21.60M coming from the DINO-S backbone and 6.08M from the remaining components. Even with this parameter scale, SegDINO already outperforms most transformer-based counterparts. Furthermore, SegDINO reaches an inference speed of 51 FPS, surpassing the majority of transformer architectures while remaining competitive with lightweight convolutional models. These results show that SegDINO maintains a well-balanced combination of accuracy, model size, and inference speed, confirming its effectiveness as an efficient solution for medical image segmentation.

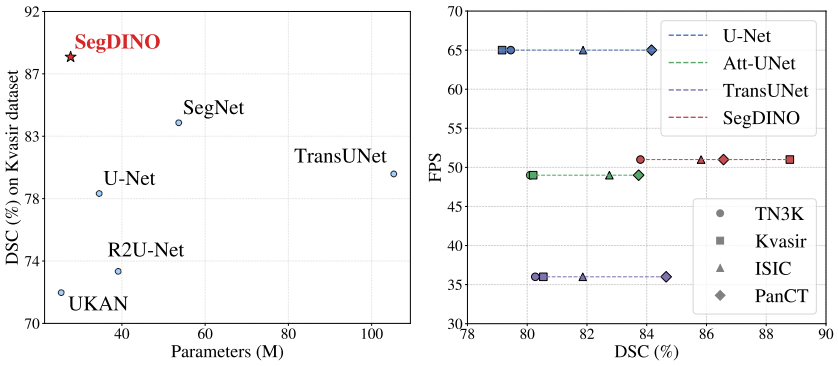

Fig. 3: Overall performance and efficiency comparisons across different datasets.

Table 2: Ablation study of different modules.

Methods	TPA	SAD	TN3K		PanCT	
			DSC ↑	HD95 ↓	DSC ↑	HD95 ↓
Basic			0.8318	18.62	0.7758	7.97
M1	✓		0.8379	17.79	0.8525	3.74
M2		✓	0.8324	19.01	0.7906	7.51
Ours	✓	✓	**0.8391**	**16.45**	**0.8657**	**2.61**

3.4 Ablation Study

We conduct an ablation study to evaluate the contribution of each SegDINO component. The baseline (*"Basic"*) directly projects multi-level encoder features to a unified channel space and concatenates them at a single resolution for segmentation. Using TPA alone (*"M1"*) reorganizes encoder features into a pseudo token pyramid. Using SAD alone (*"M2"*) keeps features at one resolution and performs scale-aware refinement without constructing a pyramid. The full model combines TPA and SAD. As shown in Table 2, TPA consistently boosts performance, and adding SAD further improves results. Their effects vary by dataset: on TN3K, where targets are large and continuous, both modules yield only marginal gains. On PanCT, which contains small and fine-grained lesions, TPA brings substantial improvement (e.g., +7.7% DSC), while SAD alone offers limited benefit. Overall, TPA is the primary contributor to effective multi-scale representation learning, whereas SAD provides lightweight yet useful refinement. Given strong DINO-based features, a simple multi-scale hierarchy plus efficient scale-aware decoding is sufficient to exploit the DINO's inherent information.

4 Conclusion

In this work, we presented SegDINO, a lightweight segmentation framework built upon DINOv3 representations. Experiments on four datasets show consistent improvements in segmentation accuracy and boundary localization, especially on small-lesion tasks. Future work will include broader comparisons and ablations, evaluate generalizability across more modalities and clinical scenarios, and explore extensions to 3D volumetric segmentation.

Acknowledgments. This work was supported by Guangdong Science and Technology Department (2024ZDZX2004) and the National Natural Science Foundation of China (Project No.82572383).

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

References

1. Alom, M.Z., Hasan, M., Yakopcic, C., Taha, T.M., Asari, V.K.: Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. arXiv preprint arXiv:1802.06955 (2018)
2. Ayzenberg, L., Giryes, R., Greenspan, H.: Dinov2 based self supervised learning for few shot medical image segmentation. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2024)
3. Azad, R., Aghdam, E.K., Rauland, A., Jia, Y., Avval, A.H., Bozorgpour, A., Karimijafarbigloo, S., Cohen, J.P., Adeli, E., Merhof, D.: Medical image segmentation review: The success of u-net. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
4. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence **39**(12), 2481–2495 (2017)
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. Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). pp. 168–172. IEEE (2018)
8. Damm, S., Laszkiewicz, M., Lederer, J., Fischer, A.: Anomalydino: Boosting patch-based few-shot anomaly detection with dinov2. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1319–1329. IEEE (2025)
9. 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)

10. Gong, H., Chen, J., Chen, G., Li, H., Li, G., Chen, F.: Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules. Computers in biology and medicine **155**, 106389 (2023)
11. 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: International conference on multimedia modeling. pp. 451–462. Springer (2019)
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, 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. Li, X., Ding, H., Yuan, H., Zhang, W., Pang, J., Cheng, G., Chen, K., Liu, Z., Loy, C.C.: Transformer-based visual segmentation: A survey. IEEE transactions on pattern analysis and machine intelligence (2024)
15. Li, Y., Wu, Y., Lai, Y., Hu, M., Yang, X.: Meddinov3: How to adapt vision foundation models for medical image segmentation? arXiv preprint arXiv:2509.02379 (2025)
16. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
17. 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)
18. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
19. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
20. 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)
21. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
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. 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)
24. Valanarasu, J.M.J., Patel, V.M.: Unext: Mlp-based rapid medical image segmentation network. In: International conference on medical image computing and computer-assisted intervention. pp. 23–33. Springer (2022)
25. Wang, S., Safari, M., Hu, M., Li, Q., Chang, C.W., Qiu, R.L., Yang, X.: Dinov3 with test-time training for medical image registration. arXiv preprint arXiv:2508.14809 (2025)

26. Wu, J., Fu, R., Fang, H., Zhang, Y., Yang, Y., Xiong, H., Liu, H., Xu, Y.: Medsegdiff: Medical image segmentation with diffusion probabilistic model. In: Medical Imaging with Deep Learning. pp. 1623–1639. PMLR (2024)
27. 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. Visual Intelligence **4**(1), 2 (2026)
28. Yang, L., Zhao, Z., Zhao, H.: Unimatch v2: Pushing the limit of semi-supervised semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
29. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning (still) requires rethinking generalization. Communications of the ACM **64**(3), 107–115 (2021)
30. Zhou, T., Xia, W., Zhang, F., Chang, B., Wang, W., Yuan, Y., Konukoglu, E., Cremers, D.: Image segmentation in foundation model era: A survey. arXiv preprint arXiv:2408.12957 (2024)
31. Zhu, L., Wei, F., Lu, Y., Chen, D.: Scaling the codebook size of vq-gan to 100,000 with a utilization rate of 99%. Advances in Neural Information Processing Systems **37**, 12612–12635 (2024)