S-GRPO: Structural Group Relative Policy Optimization for Medical Image Segmentation

Yiru Duan, Minxin Nie, Xiao Ma, Xiaohui Li, Wenbo Zhang, and Qiang Chen

School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
chen2qiang@njust.edu.cn

Abstract. Accurate medical image segmentation requires both pixel-level precision and global anatomical consistency. However, the standard supervised paradigm, focused on pixel-wise distribution fitting, often traps models in sub-optimal local minima and fails to directly optimize non-differentiable structural metrics. In this paper, we propose Structural Group Relative Policy Optimization (S-GRPO), a novel paradigm that reformulates temporal sampling into structural parallel sampling. Departing from standard GRPO, S-GRPO leverages a multi-head architecture to generate a group of diverse segmentation candidates in a single forward pass, markedly reducing training latency. An adaptive Selector Network is designed to identify the optimal "expert" head and estimate relative advantages. Furthermore, we incorporate Advantage-Aware Distillation and a conditional Exponential Moving Average (EMA) update strategy to anchor the policy to a stable reference model, ensuring balanced exploration and exploitation. S-GRPO is designed as a plug-and-play optimization paradigm, compatible with diverse backbones including CNNs, Transformers, and State Space Models. Extensive experiments demonstrate that S-GRPO consistently improves performance across eight representative architectures, achieving superior results compared to conventional supervised baselines. The source code is available at: https://github.com/Duan-yiru/S-GRPO-main.

Keywords: Medical Image Segmentation · Reinforcement Learning · Group Relative Policy Optimization · Structural Diversity · Advantage-Aware Distillation.

1 Introduction

Medical image segmentation is a cornerstone of modern computer-aided diagnosis and treatment, playing a crucial role in downstream clinical tasks such as quantitative lesion analysis, surgical planning, and navigation [8, 12]. High-quality segmentation results require not only high-precision pixel-level predictions but also adherence to global geometric constraints and anatomical plausibility [20]. In recent years, deep learning-based approaches have dominated this field. Convolutional Neural Networks (CNNs), exemplified by U-Net [17] and its variants [1, 14, 24], have established strong baselines through hierarchical

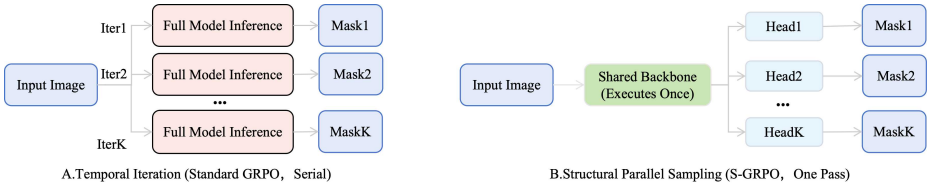

Fig. 1. Comparison of Sampling Strategies. (A) Temporal Sampling used in standard GRPO. (B) Structural Parallel Sampling proposed in S-GRPO.

feature encoding. More recently, Transformer-based architectures like SwinUNet [3] and BATFormer [10], as well as the latest Mamba-based models [4, 11, 23], have further pushed the state-of-the-art by capturing long-range dependencies and global contexts. These methods are typically trained using supervised learning objectives, such as Cross-Entropy or Dice Loss, which serve as differentiable proxies for segmentation quality.

Despite these advancements, a fundamental limitation persists in the standard supervised paradigm: supervised learning, essentially performing pixel-wise distribution fitting, often traps models in sub-optimal local minima [5–7, 22]. Furthermore, the inability to directly optimize non-differentiable metrics (e.g., Hausdorff Distance, topological connectivity) often compromises the structural integrity of segmentation results. Conversely, Reinforcement Learning (RL) [13, 21] offers a compelling alternative by transforming these non-differentiable metrics into direct reward signals. Generally, RL algorithms can be categorized into Actor-Critic methods, which rely on value function estimation, and Critic-free approaches that optimize policies directly. However, adapting RL to medical segmentation faces dual challenges. First, traditional Actor-Critic frameworks (e.g., PPO [18]) suffer from training instability and excessive memory overhead when fitting a high-dimensional value function for dense predictions. Second, emerging Critic-free paradigms like Group Relative Policy Optimization (GRPO) [19], while successful in Large Language Models (LLMs), rely on iterative sampling to estimate baselines. Since standard GRPO necessitates executing the full model repeatedly for a single update, it incurs prohibitive training latency for computationally intensive segmentation models, rendering it impractical for high-resolution medical imaging.

To resolve the conflict between the temporal iterative sampling of standard GRPO and the efficiency requirements of medical segmentation, we propose a paradigm shift towards structural parallel sampling. We present Structural Group Relative Policy Optimization (S-GRPO), representing one of the first attempts to generalize the GRPO paradigm from LLMs to dense medical image segmentation. As illustrated in Fig. 1, instead of executing the full model repeatedly, we reconfigure the segmentation head to generate a diverse group of candidates in a single forward pass. The main contributions of this work are summarized as follows: 1) We propose S-GRPO, a novel framework that adapts the generative GRPO paradigm for dense prediction by replacing temporal sam-

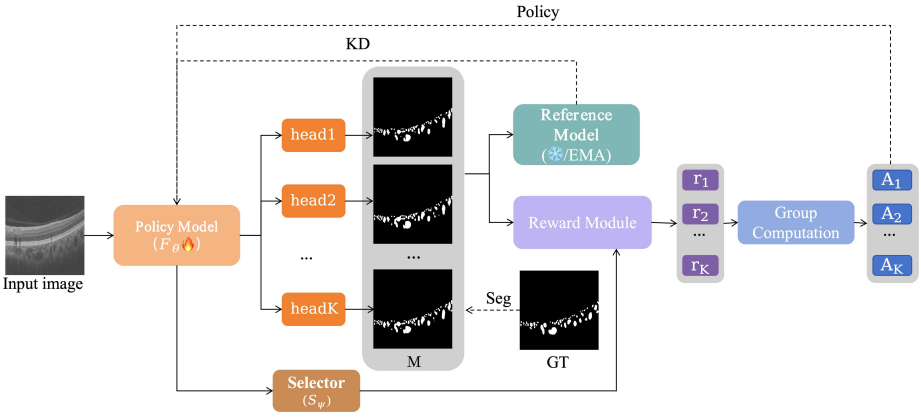

Fig. 2. The S-GRPO framework. The Policy Model generates K diverse hypotheses via structural parallel sampling, with the optimal head identified by the Selector. The Reward Module computes group relative advantages, while the Reference Model provides stable guidance via conditional EMA updates.

pling with structural parallel sampling. 2) We introduce a lightweight "expert" head mechanism coupled with dynamic selection. This allows for the efficient estimation of relative advantages without auxiliary Critic networks. 3) As a versatile "plug-and-play" booster, S-GRPO consistently elevates the performance of eight representative architectures (CNNs, Transformers, and Mamba models), achieving superior Dice scores and boundary accuracy.

2 Method

Given a medical image dataset $\mathcal{D} = \{(\mathbf{X}, \mathbf{Y})\}$, where $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$ denotes the input image and $\mathbf{Y} \in \{0,1\}^{H\times W}$ represents the ground truth. As illustrated in Fig. 2, S-GRPO formulates the segmentation task as a structural reinforcement learning process. The framework is organized into three key stages corresponding to the data flow: (1) *Structural Parallel Sampling and Selection* (Sec. 2.1); (2) *Reward Modeling and Group Relative Advantage Estimation* (Sec. 2.2); and (3) *Adaptive Distillation and Updating Strategy* (Sec. 2.3). In the following, we detail the implementation of these stages.

2.1 Structural Parallel Sampling and Selection

Structural Parallel Sampling. Let $F_\theta(\cdot)$ denote the shared backbone of the Policy Model, where the backbone refers to the base segmentation network used to extract dense feature representations. To achieve structural sampling, we fork the network at the final decoding stage into K parallel segmentation heads. Each head is composed of a Dropout layer followed by a lightweight convolution block. The Dropout layer injects independent stochastic perturbations, enforcing

structural diversity and simulating action space exploration. The output group of probability maps $\mathcal{M} = \{\mathbf{M}_1, \ldots, \mathbf{M}_K\}$ is defined as:

$$\mathbf{M}_k = \text{Conv}_k(\text{Dropout}(F_\theta(\mathbf{X}))), \quad k \in \{1, \ldots, K\} \tag{1}$$

where $\mathbf{M}_k \in [0,1]^{H \times W}$ and Conv_k denotes the lightweight convolutional prediction block of the k-th head. This mechanism allows the model to explore K distinct segmentation strategies simultaneously within a single forward pass.

Selector Network. We introduce a lightweight Selector $S_\psi(\cdot)$ to predict the probability distribution over the K heads, where p_k represents the confidence that the k-th head is the "expert". Given features $\mathbf{F} = F_\theta(\mathbf{X}) \in \mathbb{R}^{C \times H \times W}$ extracted from the shared backbone, the categorical probability distribution $\mathbf{P} = [p_1, \ldots, p_K]$ is computed as:

$$\mathbf{P} = \text{Softmax}(\mathbf{W}_s \cdot \text{GAP}(\mathbf{F}) + \mathbf{b}_s) \tag{2}$$

where $\mathbf{W}_s$ and $\mathbf{b}_s$ are learnable weights and biases, and GAP denotes Global Average Pooling. During training, we sample an index $k^* \sim \mathbf{P}$ using a "Hard Selection" strategy, back-propagating gradients only through the selected head $\mathbf{M}_{k^*}$ to optimize memory efficiency. Rather than directly comparing candidate masks, the selector performs sample-level routing based on global fused representations, while the reward module provides structure-aware feedback to guide its routing policy.

2.2 Reward Modeling and Group Relative Advantage Estimation

Hybrid Reward Modeling. To ensure both high pixel-level accuracy and precise boundary delineation, we define a composite reward r_k for each hypothesis $\mathbf{M}_k$ as a weighted sum of regional overlap and boundary fidelity:

$$r_k = \alpha R_{seg}(\mathbf{M}_k, \mathbf{Y}) + \beta R_{bound}(\mathbf{M}_k, \mathbf{Y}) \tag{3}$$

where α and β are scaling factors. Specifically, the segmentation reward R_{seg} combines the Dice Similarity Coefficient (DSC) and Intersection over Union (IoU):

$$R_{seg}(\mathbf{M}_k, \mathbf{Y}) = \text{Dice}(\mathbf{M}_k, \mathbf{Y}) + \text{IoU}(\mathbf{M}_k, \mathbf{Y}) \tag{4}$$

Their combination provides a more robust regional reward in practice, especially when foreground structures vary in size. To encourage smooth and anatomically plausible contours, we introduce the boundary reward R_{bound} based on the Average Symmetric Surface Distance (ASSD):

$$R_{bound}(\mathbf{M}_k, \mathbf{Y}) = \frac{1}{1 + \text{ASSD}(\mathbf{M}_k, \mathbf{Y})} \tag{5}$$

Group Relative Advantage Estimation. We use the Reference Model's reward r_{ref} as an external baseline to compute raw advantages $\text{Adv}_k = r_k - r_{ref}$. To emphasize relative performance within the structural group, we calculate the normalized advantage:

$$A_k = \text{clip}\left(\frac{\text{Adv}_k - \mu_{group}}{\sigma_{group} + \xi}, -\delta, \delta\right) \tag{6}$$

where μ_{group} and σ_{group} are the mean and standard deviation of rewards within the group, $\xi = 10^{-8}$ is a stability constant, and $\delta = 5$ is the clipping threshold.

2.3 Adaptive Distillation and Updating Strategy

Advantage-Aware Distillation. To prevent policy drift while allowing beneficial exploration, we employ an advantage-aware distillation mechanism. We define the distillation loss $\mathcal{L}_{KD}$ as the discrepancy between the selected head $\mathbf{M}_{k^*}$ and the reference prediction $\mathbf{M}_{ref}$:

$$\mathcal{L}_{KD} = w_{kd} \cdot ||\mathbf{M}_{k^*} - \mathbf{M}_{ref}||_2^2 \tag{7}$$

The distillation weight w_{kd} is dynamically modulated by the advantage A_{k^*}: $w_{kd} = \sigma(-\gamma \cdot A_{k^*})$, where $\sigma(\cdot)$ denotes the Sigmoid function and γ is a hyperparameter controlling the sensitivity of the modulation. When the policy outperforms the reference ($A_{k^*} > 0$), w_{kd} decreases to encourage exploration; otherwise, it increases to enforce stability.

Conditional EMA Update. The Reference Model parameters θ_{ref} are updated via a conditional EMA strategy, gated by performance improvement:

$$\theta_{ref} \leftarrow \begin{cases} \eta\theta_{ref} + (1-\eta)\theta_{policy}, & \text{if } r_{k^*} > r_{ref} \\ \theta_{ref}, & \text{otherwise} \end{cases} \tag{8}$$

where θ_{policy} denotes the current parameters of the Policy Model, and $\eta = 0.995$ is the EMA momentum coefficient controlling the update rate. This ensures the reference anchor only evolves when the policy makes positive progress.

2.4 Optimization Objectives

To ensure pixel-level precision, policy stability, and structural exploration, the training of S-GRPO is guided by several specialized objectives.

First, we apply a supervised loss $\mathcal{L}_{seg}$ to the selected expert head as a combination of Cross-Entropy ($\mathcal{L}_{ce}$) and Dice loss ($\mathcal{L}_{dice}$):

$$\mathcal{L}_{seg} = \mathcal{L}_{ce}(\mathbf{M}_{k^*}, \mathbf{Y}) + \mathcal{L}_{dice}(\mathbf{M}_{k^*}, \mathbf{Y}). \tag{9}$$

Simultaneously, to update the Selector and backbone via relative feedback, we adopt a clipped objective:

$$\mathcal{L}_{policy} = -\mathbb{E}\left[\min(\rho_{k^*} A_{k^*}, \text{clip}(\rho_{k^*}, 1-\epsilon, 1+\epsilon) A_{k^*})\right], \tag{10}$$

where $\rho_{k^*} = P_\psi(k^*|\mathbf{F})/P_{\psi_{old}}(k^*|\mathbf{F})$ denotes the importance sampling ratio, with P_ψ and $P_{\psi_{old}}$ representing the probabilities of selecting head k^* predicted by the current and old Selectors, respectively, and A_{k^*} is the clipped advantage from Eq. (6), and $\epsilon = 0.2$ is the clipping threshold that controls the update range of the policy.

To prevent mode collapse among the K heads, we maximize their structural diversity via Jensen-Shannon Divergence (JSD), which is symmetric, bounded, and distribution-aware, making it well suited for measuring differences among probabilistic segmentation hypotheses while encouraging complementary structural predictions:

$$\mathcal{L}_{div} = -\frac{1}{K(K-1)} \sum_{i=1}^{K} \sum_{j \neq i} \text{JSD}(\mathbf{M}_i || \mathbf{M}_j). \tag{11}$$

The total objective function $\mathcal{L}_{total}$ jointly optimizes all components in an end-to-end manner, balanced by hyper-parameters λ_p, λ_{kd}, and λ_{div}:

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \lambda_p \mathcal{L}_{policy} + \lambda_{kd} \mathcal{L}_{KD} + \lambda_{div} \mathcal{L}_{div}. \tag{12}$$

3 Experiments and Results

3.1 Datasets and Implementation Details

We evaluate S-GRPO on two datasets: (1) **ACDC** [2], comprising 100 cardiac MRI cases (70/10/20 split for train/val/test) for multi-structure segmentation; and (2) **OCT**, a private dataset of 4,000 retinal B-scans (640×400 resolution, 2,800/200/1,000 split for train/val/test) for choroidal vessel segmentation. Models are implemented using PyTorch on an NVIDIA RTX 4090 GPU. For preprocessing, ACDC images are resized to 224×224, while OCT images are randomly cropped to 400×400 and resized to 512×512. All baseline models are trained for 300 epochs. To ensure fairness, for S-GRPO, we initialize the Policy and Reference models with weights pre-trained for 100 epochs, followed by 200 epochs of fine-tuning with a batch size of 4. The number of parallel segmentation heads is set to $K = 4$ to balance structural diversity and computational efficiency. The reward weights are set to $\alpha = 0.7$ and $\beta = 0.3$, while the loss coefficients are $\lambda_p = 0.3$, $\lambda_{kd} = 0.1$ and $\lambda_{div} = 0.1$. Performance is evaluated using Dice, IoU, and 95% Hausdorff Distance (HD95).

3.2 Experimental Results

To verify the generalizability of S-GRPO, we integrated it into eight diverse backbones including CNNs, Transformers, Hybrid models, and the Mamba architecture. As shown in Table 1, S-GRPO consistently improves performance across all datasets and metrics. Interestingly, the performance gains on the OCT dataset are generally higher than those on ACDC. We attribute this to the fact that

Table 1. Quantitative comparison on ACDC and OCT datasets. Red values in parentheses indicate performance gains over baselines.

Type	Method	ACDC			OCT		
		Dice↑	IoU↑	HD95↓	Dice↑	IoU↑	HD95↓
CNN	U-Net[17]	91.54	84.81	1.18	80.90	68.29	6.78
	U-Net+S-GRPO	**91.74**(↑0.20)	**85.19**(↑0.38)	**1.11**(↓0.07)	**82.45**(↑1.55)	**70.50**(↑2.21)	**5.68**(↓1.10)
	U-KAN[9]	91.31	84.46	1.28	81.25	68.78	6.15
	U-KAN+S-GRPO	**91.56**(↑0.25)	**85.15**(↑0.69)	**1.18**(↓0.10)	**82.22**(↑0.97)	**70.15**(↑1.37)	**5.04**(↓1.11)
Trans.	SwinUNet[3]	88.64	80.24	1.50	80.04	67.11	5.77
	SwinUNet+S-GRPO	**90.12**(↑1.48)	**82.26**(↑2.02)	**1.21**(↓0.29)	**81.27**(↑1.23)	**68.83**(↑1.62)	**5.41**(↓0.36)
	BATFormer[10]	90.93	83.89	2.04	80.76	68.10	6.99
	BAT.+S-GRPO	**91.60**(↑0.67)	**84.96**(↑1.07)	**1.58**(↓0.46)	**81.97**(↑1.21)	**69.84**(↑1.74)	**5.92**(↓1.07)
Hybrid	MobileUViT[22]	90.22	82.82	2.63	80.66	67.13	6.78
	MUViT.+S-GRPO	**91.14**(↑0.92)	**84.25**(↑1.43)	**1.78**(↓0.85)	**81.98**(↑1.32)	**68.96**(↑1.83)	**5.58**(↓1.20)
	TransCASCADE[15]	90.73	83.58	1.20	80.11	67.19	7.44
	TCAS.+S-GRPO	**91.53**(↑0.80)	**84.85**(↑1.27)	**1.16**(↓0.04)	**82.01**(↑1.90)	**69.88**(↑2.69)	**5.70**(↓1.74)
	GCASCADE[16]	91.62	84.99	1.11	81.67	68.92	6.12
	GCAS.+S-GRPO	**92.28**(↑0.66)	**85.64**(↑0.65)	**1.02**(↓0.09)	**82.71**(↑1.04)	**70.87**(↑1.95)	**5.23**(↓0.89)
Mamba	Swin-UMamba[11]	90.71	83.54	1.94	81.10	68.51	6.57
	Swin-UM.+S-GRPO	**91.67**(↑0.96)	**85.06**(↑1.52)	**1.12**(↓0.82)	**82.58**(↑1.48)	**70.72**(↑2.21)	**5.69**(↓0.88)

OCT vessel segmentation is a harder task with lower contrast and more ambiguous boundaries, where supervised baselines easily get trapped in local optima. S-GRPO's exploration capability and direct optimization of non-differentiable metrics are highly effective in escaping sub-optimal states and unlocking further performance potential.

Visual comparisons in Fig. 3 highlight S-GRPO's superiority in handling complex structures. In the OCT dataset, baseline models (Before) often produce fragmented masks or miss fine-grained choroidal vessels due to low contrast. Conversely, S-GRPO-enhanced models (After) exhibit improved connectivity and higher sensitivity to small vessels. For ACDC, S-GRPO effectively rectifies topological errors and irregular boundaries (e.g., in Swin-Unet and GCASCADE), resulting in smoother and more anatomically consistent delineations of cardiac structures. These visual refinements substantiate that our reward-driven framework successfully captures structural dependencies beyond pixel-wise supervision.

3.3 Ablation Study

We evaluate the individual contributions of S-GRPO's components using Swin-UMamba on the OCT dataset (Table 2). Removing $\mathcal{L}_{div}$ or $\mathcal{L}_{kd}$ leads to performance degradation. Disabling the conditional EMA causes the most significant decline, proving that a stable reference anchor is indispensable for optimization. Regarding reward design, while R_{seg} and R_{bound} provide individual improvements, their integration yields the highest scores, demonstrating the efficacy of the Hybrid Reward mechanism.

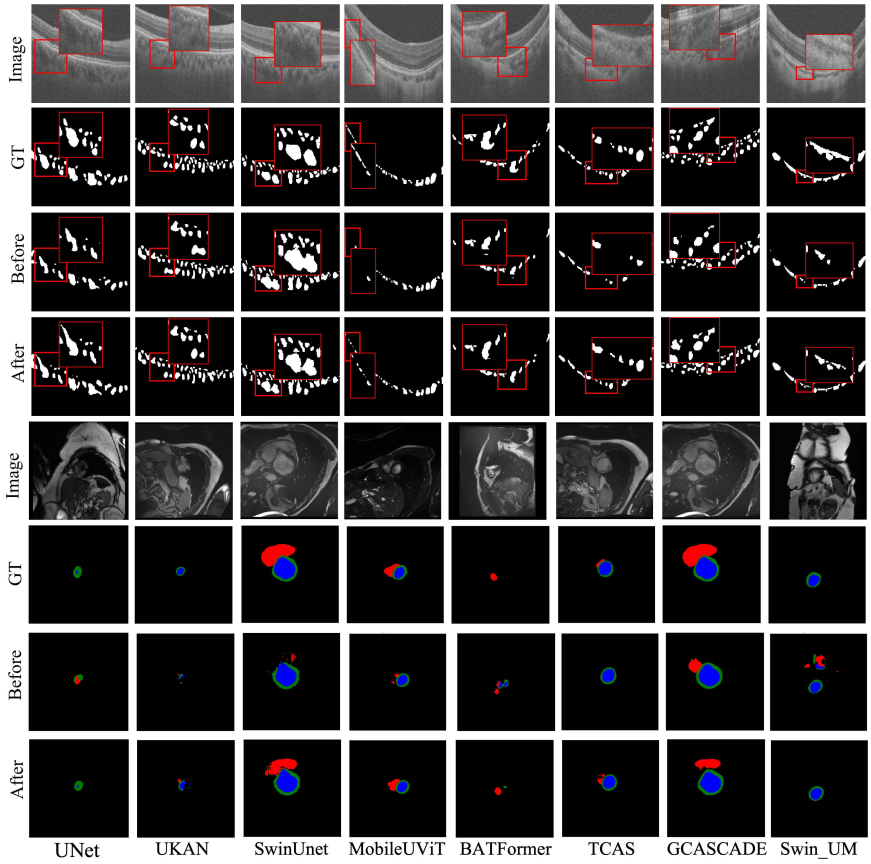

Fig. 3. Qualitative comparison between supervised baselines (Before) and S-GRPO-enhanced models (After).

Table 2. Ablation study of loss components and reward modules on OCT dataset. The baseline model is Swin-UMamba.

Method	Components			Rewards			Metrics		
	$\mathcal{L}_{div}$	$\mathcal{L}_{kd}$	EMA	R_{seg}	R_{bound}	Hybrid	Dice↑	IoU↑	HD95↓
Swin-UMamba							81.10	68.51	6.57
+ S-GRPO		✓	✓			✓	81.89	69.70	5.92
	✓		✓			✓	82.02	69.85	5.72
	✓	✓				✓	81.82	69.58	5.77
	✓	✓	✓	✓			81.97	69.73	6.07
	✓	✓	✓		✓		81.71	69.34	5.87
	✓	✓	✓			✓	**82.58**	**70.72**	**5.69**

4 Conclusion

In this paper, we propose S-GRPO, a novel reinforcement learning framework for medical image segmentation. By introducing structural parallel sampling,

we efficiently generate group-based hypotheses in a single forward pass, significantly reducing the sampling latency inherent in standard GRPO. Through the synergy of the adaptive Selector, advantage-aware distillation, and conditional EMA strategies, S-GRPO effectively optimizes both pixel-level accuracy and global anatomical consistency. Extensive experiments on ACDC and OCT datasets across eight diverse backbones demonstrate that S-GRPO serves as a powerful plug-and-play booster. Future work will explore the application of S-GRPO in 3D medical volumes and multi-modal fusion tasks.

Acknowledgments. This work was supported in part by the Major Research Plan of the National Natural Science Foundation of China (92370109), National Natural Science Foundation of China (62172223), and Natural Science Foundation of Jiangsu Province (BK20252033).

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

References

1. 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)
2. 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)
3. 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)
4. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. In: First conference on language modeling (2024)
5. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 574–584 (2022)
6. Hu, J., Chen, S., Pan, Z., Zeng, S., Yang, W.: Perspective+ unet: Enhancing segmentation with bi-path fusion and efficient non-local attention for superior receptive fields. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 499–509. Springer (2024)
7. Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.W., Wu, J.: Unet 3+: A full-scale connected unet for medical image segmentation. In: IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 1055–1059 (2020)
8. Jiangtao, W., Ruhaiyem, N.I.R., Panpan, F.: A comprehensive review of u-net and its variants: advances and applications in medical image segmentation. IET Image Processing **19**(1), e70019 (2025)

9. 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)
10. Lin, X., Yu, L., Cheng, K.T., Yan, Z.: Batformer: Towards boundary-aware lightweight transformer for efficient medical image segmentation. IEEE journal of biomedical and health informatics **27**(7), 3501–3512 (2023)
11. 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)
12. Ma, Y., Zhang, C., Cabezas, M., Song, Y., Tang, Z., Liu, D., Cai, W., Barnett, M., Wang, C.: Multiple sclerosis lesion analysis in brain magnetic resonance images: techniques and clinical applications. IEEE Journal of Biomedical and Health Informatics **26**(6), 2680–2692 (2022)
13. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al.: Human-level control through deep reinforcement learning. Nature **518**(7540), 529–533 (2015)
14. 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)
15. 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)
16. Rahman, M.M., Marculescu, R.: G-cascade: Efficient cascaded graph convolutional decoding for 2d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 7728–7737 (2024)
17. 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)
18. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
19. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
20. Shui, Z., Zhang, Y., Yao, K., Zhu, C., Zheng, S., Li, J., Li, H., Sun, Y., Guo, R., Yang, L.: Unleashing the power of prompt-driven nucleus instance segmentation. In: European conference on computer vision. pp. 288–304. Springer (2024)
21. Sutton, R.S., Barto, A.G., et al.: Reinforcement learning: An introduction, vol. 1. MIT press Cambridge (1998)
22. Tang, F., Nian, B., Ding, J., Ma, W., Quan, Q., Dong, C., Yang, J., Liu, W., Zhou, S.K.: Mobile u-vit: Revisiting large kernel and u-shaped vit for efficient medical image segmentation. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 3408–3417 (2025)
23. 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)
24. 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)