Autoregressive Medical Image Segmentation via Next-Scale Mask Prediction

Tao Chen[1], Chenhui Wang[1], Zhihao Chen[1], and Hongming Shan[1,2,3](✉)[0000−0002−0604−3197]

[1] Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai, China
[2] MOE Frontiers Center for Brain Science, Fudan University, Shanghai, China
[3] Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence (Fudan University), Ministry of Education, China
hmshan@fudan.edu.cn

Abstract. While deep learning has significantly advanced medical image segmentation, most existing methods still struggle with handling complex anatomical regions. Cascaded or deep supervision-based approaches attempt to address this challenge through multi-scale feature learning but fail to establish sufficient inter-scale dependencies, as each scale relies solely on the features of the immediate predecessor. Towards this end, we propose the AutoRegressive Segmentation framework via next-scale mask prediction, termed `AR-Seg`, which progressively predicts the next-scale mask by explicitly modeling dependencies across all previous scales within a unified architecture. `AR-Seg` introduces three innovations: (1) a multi-scale mask autoencoder that quantizes the mask into multi-scale token maps to capture hierarchical anatomical structures, (2) a next-scale autoregressive mechanism that progressively predicts next-scale masks to enable sufficient inter-scale dependencies, and (3) a consensus-aggregation strategy that combines multiple sampled results to generate a more accurate mask, further improving segmentation robustness. Extensive experimental results on two benchmark datasets with different modalities demonstrate that `AR-Seg` outperforms state-of-the-art methods while explicitly visualizing the intermediate coarse-to-fine segmentation process. Source code is made available at https://github.com/takimailto/AR-Seg.

Keywords: Next-scale prediction· Autoregressive model · Medical image segmentation.

1 Introduction

Medical image segmentation is crucial for clinical diagnosis and treatment planning [10]. Despite advancements in deep learning, most existing methods [5,7,11] struggle with the variability of complex anatomical structures—differing in size, shape, and appearance—making accurate segmentation particularly challenging in intricate and ambiguous regions. To address these challenges, cascaded and

deep supervision-based approaches [1,32] have focused on multi-scale feature integration or loss constraints to extract more robust representations. However, these methods still struggle to establish sufficient inter-scale dependencies, as each scale relies solely on the features of the immediate predecessor.

Drawing from the success of autoregressive models via next-token prediction in natural language processing [25], we recognize their potential to address the inherent variability in discrete segmentation tasks. To this end, we propose a novel AutoRegressive Segmentation framework via next-scale mask prediction, termed `AR-Seg`, which progressively predicts the next-scale mask by explicitly modeling dependencies across all previous scales within a unified architecture. This holistic modeling of multi-scale dependencies leads to a more coherent and accurate segmentation, especially in regions with varying anatomical features, thus improving both segmentation robustness and clinical usability. To the best of our knowledge, `AR-Seg` is the first time to utilize next-scale prediction [27] for the discrete segmentation task.

Our contributions are summarized as follows. **(i)** We propose `AR-Seg`, an autoregressive framework via next-scale mask prediction for medical image segmentation, explicitly modeling dependencies across all previous scales within a unified architecture. **(ii)** We introduce a multi-scale mask autoencoder that quantizes the mask into multi-scale token maps, capturing hierarchical anatomical structures. **(iii)** We present a next-scale autoregressive mechanism that progressively predicts next-scale mask, enabling sufficient dependencies across scales. **(iv)** We introduce a consensus-aggregation strategy that combines multiple sampled results into a final mask, further improving segmentation robustness. **(v)** Extensive experiments on two benchmark datasets with different modalities demonstrate that `AR-Seg` outperforms state-of-the-art (SOTA) methods while explicitly visualizing the intermediate coarse-to-fine segmentation process.

2 Methodology

Let $\boldsymbol{x}$ and $\boldsymbol{y}$ denote the medical image and its corresponding mask with C target classes, respectively. `AR-Seg` predicts the mask $\hat{\boldsymbol{y}}$ in a next-scale autoregressive manner, as shown in Fig. 1. Specifically, we first quantize $\boldsymbol{y}$ into K multi-scale token maps via a multi-scale mask autoencoder, then apply a next-scale autoregressive mechanism to predict the next-scale mask progressively. Finally, a consensus-aggregation strategy produces the final mask by combining multiple sampled results. We then detail each component along with the training and inference processes.

2.1 Multi-Scale Mask Autoencoder

To effectively capture hierarchical anatomical structures while preserving critical discriminative information across scales, we propose a multi-scale mask autoencoder to progressively quantize the mask to token maps instead of aggressively

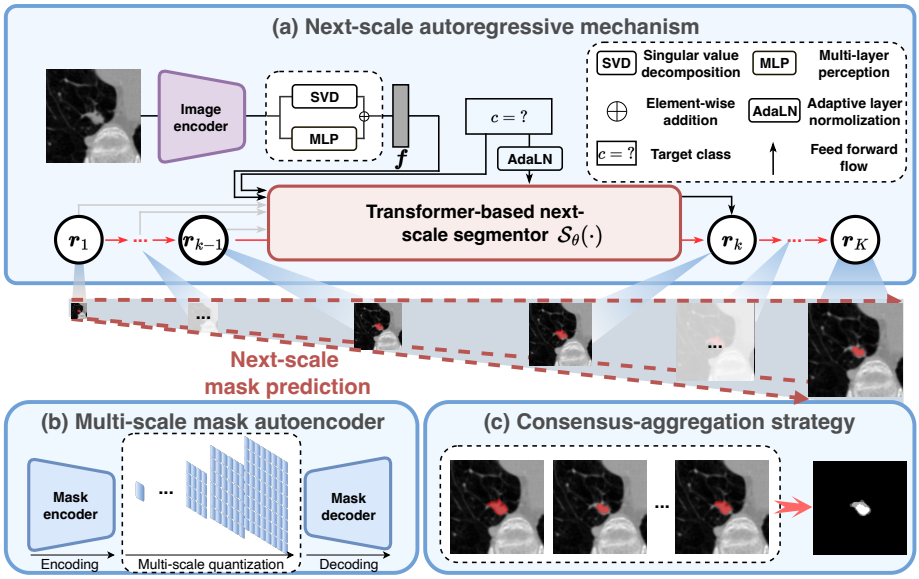

Fig. 1. Illustration of the proposed `AR-Seg`.

downsampling it. The autoencoder consists of a mask encoder $\mathcal{E}_{\text{mask}}(\cdot)$, a mask decoder $\mathcal{D}_{\text{mask}}(\cdot)$, and a token quantizer $\mathcal{Q}(\cdot)$.

Mask quantized process. We first encode the mask $\boldsymbol{y}$ into a feature map $\boldsymbol{m} = \mathcal{E}_{\text{mask}}(\boldsymbol{y})$. Then, we dynamically update the feature map $\boldsymbol{m}$ in a residual manner [18,27] to enhance the token differences across K scales from coarse to fine, and progressively quantize $\boldsymbol{m}$ into K multi-scale token maps, $\mathcal{R} = (\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_K)$. Specifically, the k-scale quantized process maps each element of the mask residual feature map $\boldsymbol{m}$ to the nearest code index $\boldsymbol{r}_k^{(i,j)}$ from a learnable codebook $\boldsymbol{Z}$ of size V:

$$\boldsymbol{r}_k{}^{(i,j)} = \mathcal{Q}(\boldsymbol{m}^{(i,j)}) = \arg\min_{v\in[V]} \| \text{lookup}(\boldsymbol{Z}, v) - \boldsymbol{m}^{(i,j)} \|_2, \tag{1}$$

where $\text{lookup}(\boldsymbol{Z}, v)$ means taking the v-th vector in codebook $\boldsymbol{Z}$. The detailed quantized process is shown in Alg. 1.

Mask dequantized process. Based on the multi-scale token maps $\mathcal{R}$, the dequantized process first progressively dequantizes the token map to the corresponding mask residual feature map $\hat{\boldsymbol{m}}$ and then uses the mask decoder $\mathcal{D}_{\text{mask}}(\cdot)$ to reconstruct the mask $\hat{\boldsymbol{y}}$, as shown in Alg. 2.

2.2 Next-Scale Autoregressive Mechanism

To explicitly model inter-scale dependencies in a unified architecture, we propose a next-scale autoregressive mechanism based on multi-scale token maps. This

Algorithm 1 Quantized process

Inputs: mask $\boldsymbol{y}$, scales K, resolutions $(h_k, w_k)_{k=1}^{K}$;
$\boldsymbol{m} = \mathcal{E}_{\text{mask}}(\boldsymbol{y})$, $\mathcal{R} = []$;
for $k = 1$ to K **do**
  $\boldsymbol{r}_k = \mathcal{Q}(\text{interpolate}(\boldsymbol{m}, h_k, w_k))$;
  $\mathcal{R} = \text{queue\_push}(\mathcal{R}, \boldsymbol{r}_k)$;
  $\boldsymbol{z}_k = \text{lookup}(\boldsymbol{Z}, \boldsymbol{r}_k)$;
  $\boldsymbol{z}_k = \text{interpolate}(\boldsymbol{z}_k, h_K, w_K)$;
  $\boldsymbol{m} = \boldsymbol{m} - \phi_k(\boldsymbol{z}_k)$;
end for
return multi-scale token maps $\mathcal{R}$

Algorithm 2 Dequantized process

Inputs: multi-scale token maps $\mathcal{R}$, scales K, resolutions $(h_k, w_k)_{k=1}^{K}$;
$\hat{\boldsymbol{m}} = \boldsymbol{0}$;
for $k = 1$ to K **do**
  $\boldsymbol{r}_k = \text{queue\_pop}(\mathcal{R})$;
  $\boldsymbol{z}_k = \text{lookup}(\boldsymbol{Z}, \boldsymbol{r}_k)$;
  $\boldsymbol{z}_k = \text{interpolate}(\boldsymbol{z}_k, h_K, w_K)$;
  $\hat{\boldsymbol{m}} = \hat{\boldsymbol{m}} + \phi_k(\boldsymbol{z}_k)$;
end for
$\hat{\boldsymbol{y}} = \mathcal{D}_{\text{mask}}(\hat{\boldsymbol{m}})$;
return reconstructed mask $\hat{\boldsymbol{y}}$

mechanism employs a transformer-based [28] next-scale segmentor to reformulate the medical image segmentation task as a conditional autoregressive process.

Specifically, we first use the pretrained MedSAM [21] as the image encoder backbone, $\mathcal{E}_{\text{image}}(\cdot)$, followed by a parameter-efficient adapter to extract the corresponding image features. Instead of using a non-linear multi-layer perceptron (MLP) directly, our adapter enhances the feature extraction through linear singular value decomposition (SVD) [15] to preserve more global image information, producing the final image embedding $\boldsymbol{f}$ as:

$$\boldsymbol{f} = \text{MLP}\left(\mathcal{E}_{\text{image}}(\boldsymbol{x})\right) + \text{SVD}\left(\mathcal{E}_{\text{image}}(\boldsymbol{x})\right). \tag{2}$$

Then, for each scale k, the next-scale segmentor $\mathcal{S}_\theta(\cdot)$ with parameters θ takes the target class c, the image embedding $\boldsymbol{f}$, and all previously predicted mask token maps as input to predict the k-scale token map. This process defines the autoregressive likelihood of the segmentation as follows:

$$p(\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_K) = \textstyle\prod_{k=1}^{K} p_\theta(\boldsymbol{r}_k | \boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_{k-1}, c, \boldsymbol{f}). \tag{3}$$

Note that the target class is also passed through adaptive layer normalization (AdaLN) [13] to condition the segmentor. We highlight that this next-scale autoregressive mechanism can explicitly visualize the intermediate coarse-to-fine segmentation process, helping clinicians improve AI confidence.

2.3 Consensus-Aggregation Strategy

To further improve segmentation performance and resolve ambiguity in challenging regions, we introduce a consensus-aggregation strategy. Specifically, we first sample N multi-scale token maps, $\mathcal{R}^n|_{n=1}^{N}$. For each element $\boldsymbol{r}_k^n$ in $\mathcal{R}^n$, we sample it from a multinomial distribution $\mathcal{M}$, with probabilities generated by the next-scale segmentor:

$$\boldsymbol{r}_k^n \sim \mathcal{M}\Big(\boldsymbol{r}_k^n; \mathcal{S}_\theta(\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_{k-1}, c, \boldsymbol{f})\Big). \tag{4}$$

We then dequantize these N multi-scale token maps $\mathcal{R}^n|_{n=1}^N$ into the corresponding masks $\hat{\boldsymbol{y}}^n|_{n=1}^N$ through Alg. 2. Finally, these multiple sampled results are aggregated using a mean operation: $\hat{\boldsymbol{y}} = \frac{1}{N}\sum_{n=1}^{N}\hat{\boldsymbol{y}}^n$, which enhances the model's ability to handle ambiguous regions.

2.4 Detailed Procedure

Here, we provide the detailed training and inference procedures of our `AR-Seg`.

The training of `AR-Seg` consists of two stages: multi-scale mask learning and next-scale autoregressive learning. In the first stage, the multi-scale mask autoencoder is trained by minimizing a compound loss $\mathcal{L}$, which combines a quantization constraint and a segmentation objective:

$$\mathcal{L} = \underbrace{\|\boldsymbol{m}-\mathrm{sg}(\hat{\boldsymbol{m}})\|_2+\beta\|\mathrm{sg}(\boldsymbol{m})-\hat{\boldsymbol{m}}\|_2}_{\text{quantization constraint}} + \underbrace{\lambda_{\text{Dice}}\mathcal{L}_{\text{Dice}}(\boldsymbol{y},\hat{\boldsymbol{y}})+\lambda_{\text{BCE}}\mathcal{L}_{\text{BCE}}(\boldsymbol{y},\hat{\boldsymbol{y}})}_{\text{segmentation objective}} \quad (5)$$

where β is the weight for the commitment loss [9], $\mathrm{sg}(\cdot)$ denotes the stop-gradient operation, while $\mathcal{L}_{\text{Dice}}$ and $\mathcal{L}_{\text{BCE}}$ are the Dice [22] and binary cross-entropy losses, with λ_{Dice} and λ_{BCE} as their respective weights. In the second stage, we train the next-scale segmentor optimized with cross-entropy loss [25] with the frozen multi-scale mask autoencoder. A block-wise causal attention mask is applied during training to ensure causality.

During inference, the next-scale segmentor iteratively generates token maps based on the class and image embedding until the multi-scale token maps are formed, which are then dequantized using Alg. 2 and passed through the consensus-aggregation strategy to produce the final segmentation mask.

3 Experiment

3.1 Experimental Setup

Dataset and preprocessing. The experiments are conducted on two datasets: LIDC-IDRI [8] and BRATS 2021 [3]. The LIDC-IDRI dataset comprises 1,018 lung CT scans, each annotated by four radiologists. These CT scans are preprocessed using a standard pipeline and follow a train-validation-test split consistent with prior studies [7,16]. Note that we randomly select one of four annotated masks for each image during training. The BRATS 2021 dataset provides four MRI modalities (T1, T2, FLAIR, and T1CE) per patient, which are concatenated along the channel axis to form a combined input for our experiments. The preprocessing and train-test split follow previous studies [6]. Segmentation performance is evaluated on three tumor types: necrotic tumor core (NT), peritumoral edema (ED), and enhancing tumor (ET).

Implementation details. All experiments are implemented by PyTorch [24] and trained on NVIDIA V100 GPUs. The multi-scale mask autoencoder is trained for 100 epochs with a batch size of 128. While the next-scale segmentor is

Table 1. Results on LIDC-IDRI.

Methods	GED↓ 16	HM-IoU↑ 16	Soft-Dice↑ 16
Prob.U-net [16]	0.320	0.500	-
Hprob.U-net [17]	0.270	0.530	0.624
D-Personal [30]	0.346	0.555	0.604
CAR [14]	0.264	0.592	0.633
MR-Net [12]	0.658	0.447	0.616
PADL [19]	0.544	0.462	0.595
MedSegDiff [29]	0.420	0.413	0.453
BerDiff [7]	0.238	0.596	0.644
PixelSeg [31]	0.260	0.587	-
AR-Seg (ours)	**0.232**	**0.616**	**0.658**

Table 2. Results on BRATS 2021.

Methods	Dice↑
U-net [26]	75.48
nnU-net [11]	84.57
TransU-net [5]	83.50
SwinU-net [4]	83.94
MERIT [23]	83.02
MedSegDiff [2]	82.81
BerDiff [7]	85.42
HiDiff [6]	85.80
AR-Seg (ours)	**86.97**

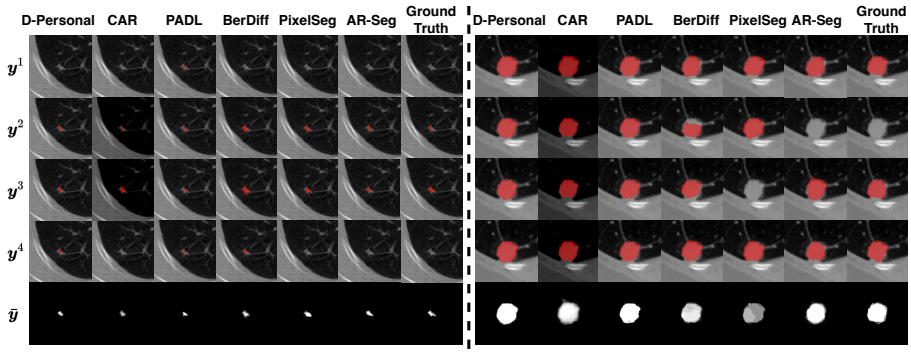

Fig. 2. Qualitative results of two lung nodules from LIDC-IDRI. y^i and $\bar{y}$ refer to the i-th segmentation masks and the final consensus-aggregated masks, respectively.

trained for 300 epochs with a batch size of 32. Both networks employ the AdamW optimizer [20] with an initial learning rate of 1×10^{-4} and a cosine annealing schedule. The loss weights β, λ_{Dice}, and λ_{BCE}, are empirically set to 0.25, 1, and 1, respectively. The number of scale K is set to 8. For performance evaluation, we employ four metrics: Generalized Energy Distance (GED) [16], Hungarian-Matched Intersection over Union (HM-IoU) [17], Soft-Dice [12], and Dice coefficient. GED is computed using varying numbers of segmentation samples (1, 4, 8, and 16), while HM-IoU and Soft-Dice are calculated using 16 samples.

3.2 Experimental Results

Comparison to SOTA methods. We evaluate our AR-Seg against other SOTA methods on two datasets: LIDC-IDRI, to assess performance under inter-radiologist variability, and BRATS 2021, to evaluate multi-class segmentation of small lesions and ambiguous boundaries, using different performance metrics.
(i) Results on LIDC-IDRI. We compare AR-Seg with conditional variational autoencoder-based methods like Prob. U-net [16], Hprob. U-net [16], and D-

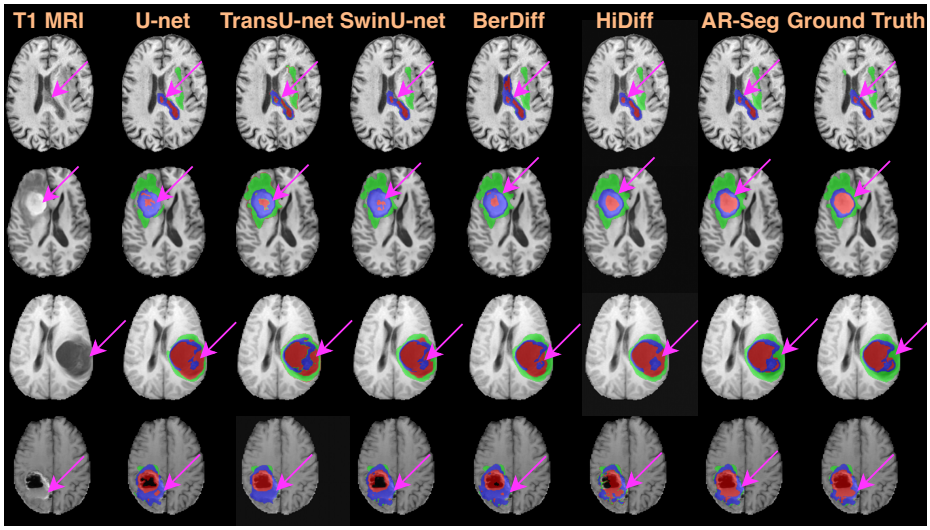

Fig. 3. Qualitative results of four MRI images from BRATS 2021. Only T1-weighted images are shown for convenience.

Table 3. Ablation results of multi-scale mask autoencoder.

	Single-scale				Multi-scale			
Class	Nodules	NT	ED	ET	Nodules	NT	ED	ET
Dice↑	0.997	0.991	0.976	0.987	**0.999**	**0.996**	**0.988**	**0.993**

Personal [30], generative adversarial network-based CAR [14], multi-annotator variability modeling techniques like MR-Net [12] and PADL [19], diffusion-based models like MedSegDiff [29] and BerDiff [7], and pixel-wise autoregressive PixelSeg [31]. From the quantitative results in Table 1, we draw three key observations as follows. (1) While many SOTA methods leverage U-shape cascaded architectures to capture multi-scale features, their lack of inter-scale dependencies hampers the segmentation. (2) Diffusion based methods, despite their explicit mask distribution modeling, still overlook multi-scale feature integration, resulting in lower accuracy compared to `AR-Seg`. (3) By incorporating next-scale prediction and explicit inter-scale dependency modeling, `AR-Seg` outperforms pixel-wise PixelSeg. Qualitative results are shown in Fig. 2. Compared to other models, `AR-Seg` accurately captures segmentation patterns. For the left example, `AR-Seg` generates masks that align more closely with the ground truth, particularly in capturing subtle prominences at the lower right of the nodules.

(ii) Results on BRATS 2021. We present quantitative and qualitative results in Table 2 and Fig. 3, respectively, comparing `AR-Seg` with convolution-based methods (U-net [26] and nnU-net [11]), transformer-based approaches (TransU-net [5], SwinU-net [4], and MERIT [23]), diffusion-based models (MedSegDiff [29] and BerDiff [7]) and hybrid diffusion method (HiDiff [6]). First, while transformer-

Table 4. Ablation results of next-scale autoregressive mechanism on LIDC-IDRI.

Framework	GED↓				HM-IoU↑
	16	8	4	1	16
Baseline	0.275	0.310	0.383	0.814	0.557
+ Next-scale autoregressive	0.239	0.277	0.348	0.745	0.595
+ SVD-based adapter	**0.232**	**0.267**	**0.332**	**0.736**	**0.616**

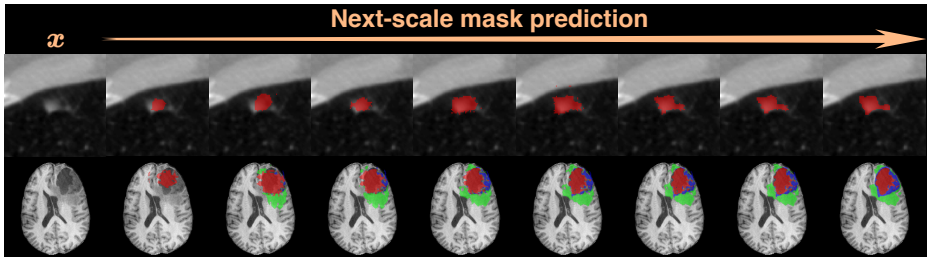

Fig. 4. Explicit segmentation process of `AR-Seg`.

based methods enhance image feature extraction, they fail to address the inherent limitation of U-shaped architectures in modeling inter-scale dependencies, resulting in inferior performance. Besides, although hybrid diffusion methods refine masks using diffusion processes, their single-scale operation limits their effectiveness, making them less accurate than `AR-Seg`. From Fig. 3, we observe that `AR-Seg` achieves improved accuracy in ambiguous regions. For instance, in the 3rd row, `AR-Seg` accurately segments a challenging tumor region.

Ablation study. We first conduct ablation experiments to evaluate the effectiveness of the multi-scale mask autoencoder in Table 3. We introduce a single-scale baseline, where the mask features are quantized into a single scale. The results demonstrate that the multi-scale mask autoencoder achieves higher Dice coefficients across all classes, owing to its multi-scale design, which enables learning a more robust representation through inter-scale mutual reinforcement.

We then conduct ablation studies on the next-scale autoregressive mechanism using LIDC-IDRI in Table 4. We compare against a baseline that uses a single-scale mask autoencoder and generates the mask via next-token prediction. When shifting the next-token to next-scale objective, our `AR-Seg` largely outperforms this baseline, demonstrating the benefits of the next-scale progression. Further gains are observed by replacing the MLP adapter with our SVD-based adapter, highlighting its contribution to more robust feature extraction.

Explicit segmentation process. We highlight the explicit segmentation advantage of `AR-Seg`, demonstrated in the qualitative results in Fig. 4, where the process starts with an initial coarse mask and iteratively refines the boundaries. For example, `AR-Seg` enhances the contextual relationships between tumor subtypes in brain tumors as the scale increases.

4 Conclusion

This paper introduces the autoregressive segmentation framework via next-scale mask prediction, a novel approach for medical image segmentation that explicitly models inter-scale dependencies. By incorporating the multi-scale mask autoencoder, next-scale autoregressive mechanism, and the consensus-aggregation strategy, `AR-Seg` not only achieves accurate segmentation but also provides visualization of the intermediate coarse-to-fine segmentation process. Experiments on the LIDC-IDRI and BRATS 2021 show that `AR-Seg` outperforms SOTA methods. Future work will focus on integrating `AR-Seg` into clinical practice by exploring clinician-in-the-loop strategies to leverage its progressive capabilities.

Acknowledgments. This work was supported in part by STI2030-Major Projects (No. 2021ZD0200204), National Natural Science Foundation of China (No. 62471148), and Shanghai Center for Brain Science and Brain-inspired Technology.

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

References

1. Aghapanah, H., Rasti, R., Tabesh, F., Pouraliakbar, H., Sanei, H., Kermani, S.: MECardNet: A novel multi-scale convolutional ensemble model with adaptive deep supervision for precise cardiac MRI segmentation. Biomed. Signal Process. Control. **100**, 106919 (2025)
2. Amit, T., Shaharbany, T., Nachmani, E., Wolf, L.: SegDiff: Image segmentation with diffusion probabilistic models. arXiv:2112.00390 (2021)
3. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., et al.: The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv:2107.02314 (2021)
4. Cao, H., et al.: Swin-Unet: Unet-like pure transformer for medical image segmentation. In: ECCV. pp. 205–218 (2022)
5. 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:2102.04306 (2021)
6. Chen, T., Wang, C., Chen, Z., Lei, Y., Shan, H.: HiDiff: Hybrid diffusion framework for medical image segmentation. IEEE Trans. Med. Imaging **43**(10), 3570–3583 (2024)
7. Chen, T., Wang, C., Shan, H.: BerDiff: Conditional Bernoulli diffusion model for medical image segmentation. In: MICCAI (2023)
8. Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al.: The cancer imaging archive (TCIA): maintaining and operating a public information repository. J. Digit. Imaging **26**(6), 1045–1057 (2013)
9. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: CVPR. pp. 12873–12883 (2021)

10. Haque, I.R.I., Neubert, J.: Deep learning approaches to biomedical image segmentation. Informatics Med. Unlock. **18**, 100297 (2020)
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. Nat. Methods **18**(2), 203–211 (2021)
12. Ji, W., Yu, S., Wu, J., Ma, K., Bian, C., Bi, Q., Li, J., Liu, H., Cheng, L., Zheng, Y.: Learning calibrated medical image segmentation via multi-rater agreement modeling. In: CVPR. pp. 12336–12346 (2021)
13. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR. pp. 4401–4410 (2019)
14. Kassapis, E., Dikov, G., Gupta, D.K., Nugteren, C.: Calibrated adversarial refinement for stochastic semantic segmentation. In: ICCV. pp. 7037–7047 (2020)
15. Klema, V., Laub, A.: The singular value decomposition: Its computation and some applications. IEEE Trans. Autom. Control **25**(2), 164–176 (1980)
16. Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J.R., Maier-Hein, K., Eslami, S., Jimenez Rezende, D., Ronneberger, O.: A probabilistic U-Net for segmentation of ambiguous images. In: NIPS (2018)
17. Kohl, S.A., Romera-Paredes, B., Maier-Hein, K.H., Rezende, D.J., Eslami, S., Kohli, P., Zisserman, A., Ronneberger, O.: A hierarchical probabilistic U-Net for modeling multi-scale ambiguities. In: medical workshop of NIPS (2019)
18. Lee, D., Kim, C., Kim, S., Cho, M., Han, W.S.: Autoregressive image generation using residual quantization. In: CVPR. pp. 11523–11532 (2022)
19. Liao, Z., Hu, S., Xie, Y., Xia, Y.: Modeling annotator preference and stochastic annotation error for medical image segmentation. Med. Image Anal. **92**, 103028 (2024)
20. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2017)
21. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nat. Commun. **15**, 654 (2024)
22. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 3DV. pp. 565–571 (2016)
23. Mostafijur Rahman, M., Marculescu, R.: Multi-scale hierarchical vision transformer with cascaded attention decoding for medical image segmentation. In: MIDL (2023)
24. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., et al.: PyTorch: An imperative style, high-performance deep learning library. In: NIPS (2019)
25. Radford, A., Narasimhan, K.: Improving language understanding by generative pre-training (2018), https://api.semanticscholar.org/CorpusID:49313245
26. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241 (2015)
27. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. In: NIPS (2024)
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NIPS (2017)
29. 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: MIDL. pp. 1623–1639 (2024)
30. Wu, Y., Luo, X., Xu, Z., Guo, X., Ju, L., Ge, Z., Liao, W., Cai, J.: Diversified and personalized multi-rater medical image segmentation. In: CVPR. pp. 11470–11479 (2024)
31. Zhang, W., Zhang, X., Huang, S., Lu, Y., Wang, K.: PixelSeg: Pixel-by-pixel stochastic semantic segmentation for ambiguous medical images. In: ACM MM. pp. 4742–4750 (2022)

32. Zhang, Y., Chung, A.C.: Deep supervision with additional labels for retinal vessel segmentation task. In: MICCAI. pp. 83–91 (2018)