nnMamba: 3D Biomedical Image Segmentation, Classification and Landmark Detection with State Space Model

Haifan Gong[1, 2], Luoyao Kang[1,2], Yitao Wang[1,2], Xiang Wan[1], Haofeng Li[1] *

[1]Shenzhen Research Institute of Big Data
[2]SSE, The Chinese University of Hong Kong (Shenzhen)

Abstract

In the field of biomedical image analysis, the quest for architectures capable of effectively capturing long-range dependencies is paramount, especially when dealing with 3D image segmentation, classification, and landmark detection. Traditional Convolutional Neural Networks (CNNs) struggle with locality respective field, and Transformers have a heavy computational load when applied to high-dimensional medical images. In this paper, we introduce nnMamba, a novel architecture that integrates the strengths of CNNs and the advanced long-range modeling capabilities of State Space Sequence Models (SSMs). Specifically, we propose the Mamba-In-Convolution with Channel-Spatial Siamese learning (MICCSS) block to model the long-range relationship of the voxels. For the dense prediction and classification tasks, we also design the channel-scaling and channel-sequential learning methods. Extensive experiments on 6 datasets demonstrate nnMamba's superiority over state-of-the-art methods in a suite of challenging tasks, including 3D image segmentation, classification, and landmark detection. nnMamba emerges as a robust solution, offering both the local representation ability of CNNs and the efficient global context processing of SSMs, setting a new standard for long-range dependency modeling in medical image analysis. All code and data will be made available after acceptance. Code is available at `https://github.com/lhaof/nnMamba`.

1 Introduction

Biomedical image analysis plays a pivotal role in healthcare, encompassing key tasks such as segmentation, classification, and landmark detection, which have been extensively studied in the literature [1, 2]. Image segmentation, in particular, is crucial for converting unstructured biomedical images into structured and informative representations, thereby propelling scientific discovery and clinical applications forward [3, 4]. Semantic segmentation is indispensable across various AI-enabled clinical applications, from diagnostic assistance and therapeutic planning to intra-operative support and monitoring tumor evolution [5]. Similarly, image based classification [6, 7, 8] and landmark detection tasks [9, 10] are vital for a range of downstream biomedical applications. For instance, Alzheimer's disease (AD), a prevalent neurological disorder, can be better managed through classification models that help in understanding its progression [11]. Automated landmark detection plays a critical role in assisting radiologists by providing key reference points, thereby reducing workload and the potential for diagnostic errors [9, 12].

In the field, CNNs [13] have been a dominant force, particularly with FCN-based methods [14, 15, 16, 17] that are adept at hierarchical feature extraction. Transformers, with their origins in NLP and subsequent adaptation for visual tasks via architectures like ViT [18] and SwinTransformer [19],

---

* denotes corresponding authors.
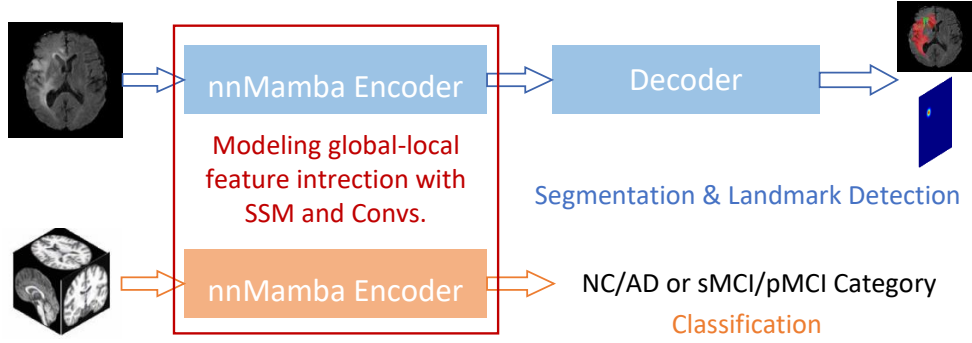
Preprint. Under review.

Figure 1: The nnMamba framework is designed for 3D biomedical tasks, focusing on dense prediction and classification. Our approach seeks to tackle the challenge of long-range modeling by leveraging the lightweight and robust long-range modeling capabilities of State Space Models.

excel in global information aggregation. Their integration into CNN frameworks has led to hybrid models such as TransUNet [20], UNETR[21], and nnFormer [22], which significantly enhance the modeling of long-range dependencies.

However, the computational intensity of Transformers, particularly due to the self-attention mechanism's scaling with input size, poses challenges in high-dimensional biomedical image analysis. State Space Sequence Models (SSMs), and specifically Structured State Space (S4) models [23], have emerged as efficient alternatives, delivering cutting-edge results in long-sequence data analysis [24, 23]. The Mamba model [25] refines the S4 approach by introducing an input-adaptive mechanism, outperforming Transformers in dense data scenarios. Recent works have begun exploring Mamba's application in medical image segmentation [26, 27, 28] and visual recognition [29]. Despite these developments, a comprehensive evaluation of Mamba's effectiveness in 3D medical segmentation, classification, and landmark detection tasks remains an open area of investigation.

Inspired by the promising results of SSMs in vision tasks and the capabilities of Mamba to handle long sequences efficiently, we explore the integration of Mamba blocks within CNNs to enhance long-range dependency modeling. Thus we proposed the nnMamba, a various of structure for both segmentation, classification, and landmark detection. Inspired by SSMs' efficacy in managing long sequences, our work introduces nnMamba—a versatile architecture designed for these 3D medical imaging applications. The main contributions of our work are summarized below:

1. We present nnMamba, a new generation backbone that takes both local and global relationship modeling into account. We established a benchmark comprising six datasets to evaluate backbones across 3 key tasks in 3D medical imaging analysis: segmentation, classification, and landmark detection.

2. We build the MICCSS (Mamba-In-Convolution with Channel-Spatial Siamese input) module as the basic module in our nnMamba, which enables the long-range relationship modeling ability at both channel and spatial levels.

3. For dense prediction tasks, we build the encoder with the MICCSS module and use skip scaling to stabilize the training. For the classification task, we add the MICCSS module to the stem layer and tailor-design a hierarchical sequential learning method based on Mamba.

4. Through extensive experimentation, we demonstrate the superior performance of the nn-Mamba framework over existing methodologies. Our results indicate that nnMamba achieves state-of-the-art effectiveness across the board, setting a new method for future research and applications in the field of 3D medical image analysis.
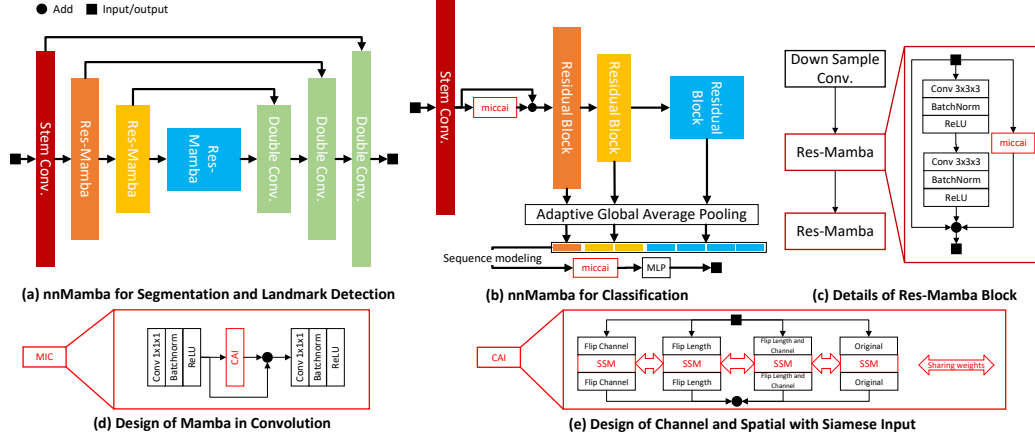
Figure 2: Illustrative diagrams of the nnMamba framework architectures. (a) Presents the network structure for segmentation and landmark detection tasks. (b) Depicts the architecture tailored for classification tasks. Detailed structures of the blocks utilized within our networks are shown in (c), (d), and (e).

2 Methodology

We first briefly review the State Space Sequence Models (SSMs) before introducing our nnMamba. SSMs are introduced as mechanisms for mapping one-dimensional input sequences $u(t)$ to outputs $y(t) \in \mathbb{R}$ through linear ordinary differential equations:

$$x'(t) = Ax(t) + Bu(t); \quad y(t) = Cx(t), \tag{1}$$

where $x(t) \in \mathbb{R}^N$, and $A \in \mathbb{R}^{N \times N}$, $B, C \in \mathbb{R}^N$ are the system parameters. SSMs suffer from memory and gradient issues, which Structured State Space models address by structuring A with a projection operator for improved long-range dependency capture. Mamba extends SSMs with adaptive parameter tuning and efficient computation, thus providing enhanced performance on long sequences.

2.1 MICCSS: Mamba-In-Convolution with Channel-Spatial Siamese Learning

Parallel CNN and SSM Design. Despite the success of sequential modeling in natural image processing [19], its application in medical imaging is challenged by the limited sample sizes (typically less than 1,000) [5] compared to those available for natural images (over 1 million in ImageNet). This makes it difficult to achieve performance comparable to Convolutional Neural Networks (CNNs). Additionally, CNNs' inductive biases are suited to medical imaging, where patterns often consist of similar intensity repeats in local regions. Thus, we introduce Mamba as a parallel module enhancing CNNs' ability to model long-range dependencies, which is shown in Fig.2 (c).

MIC: Rethinking Network-In-Network. Drawing inspiration from Network-In-Network (NIN) [30], we propose to first filter out the useful features through the conventional layers, then send the feature to mamba for better non-linear long-range modeling. In detail, we develop the Mamba-In-Convolution (MIC) module, formalized as:

$$\mathbf{F}_{\text{out}} = \text{Convs.O}(\text{SSM}(\text{Convs.I}(\mathbf{F}_{\text{in}})) + \text{Convs.I}(\mathbf{F}_{\text{in}})). \tag{2}$$

As depicted in Fig. 2(d), the MIC block integrates SSM with Convolutional layers, denoted as Convs.O and Convs.I, which include a 1×1 convolution, batch normalization, and ReLU activation. Feature maps of shape $B \times C \times H \times W \times D$ are initially processed to filter features, then reshaped to $B \times C \times L$ for global interaction modeling via the SSM.

CSS: Leveraging SSM for Channel and Spatial Feature Interaction. The core strength of the Network-In-Network (NIN) approach lies in the intermediate MLP layer's nonlinear capacity to encapsulate features. However, this MLP layer struggles with flattened features of dimensionality

3

$C \times H \times W \times D$. To address this, we investigate the potential of State Space Models (SSMs), which are inherently adept at modeling long-range relationships within sequences. Drawing inspiration from Siamese Networks[31, 32], we conceive the Channel-Spatial Siamese (CSS) input module, which fully harnesses the representational capabilities of SSMs. In this design, we augment the reshaped input feature of dimension $B \times C \times L$ across both channel and spatial dimensions. These augmented features are then processed by an SSM with shared weights, as detailed in Algorithm 1.

Algorithm 1 CSS: Channel-Spatial Siamese inputting for long-range modeling

---
1: SiamSSM   ▷ SSM with shared parameters
2: $x_{\text{flat}} \leftarrow$ input feature with shape $[B, L, C]$
3: $x_{\text{mamba}} \leftarrow \text{SiamSSM}(x_{\text{flat}})$ ▷ Iterate over combinations of flipping dimensions
4: **for** d in $\{[1], [2], [1, 2]\}$ **do**
5:    $x_{\text{flip}} \leftarrow \text{flip}(x_{\text{flat}}, \text{dims} = d)$
6:    $x_{\text{mamba}} \leftarrow x_{\text{mamba}} + \text{flip}(\text{SiamSSM}(x_{\text{flip}}), \text{dims} = d)$
7: **end for**
8: $x_{\text{mamba}} \leftarrow \frac{1}{4} x_{\text{mamba}}$

2.2 nnMamba for Segmentation and Landmark Detection

For dense prediction tasks such as segmentation or heatmap-based landmark detection, we implement a UNet architecture to capture the long-range dependencies between the encoder and decoder. Our framework utilizes a residual-based encoder paired with a convolutional decoder. We maintain the same architectural structure for both segmentation and landmark detection tasks. Notably, to stabilize the training of the UNet, we apply a learning-based scaling method [33] during the concatenation operation. Let X_h be the feature map from the h-th block of the encoder, X_h' be the corresponding feature map from the decoder, and X_{cat} be the post-concatenation feature map. Let SE denote the squeeze-extraction module as described in [33]. The process is formulated as:

$$X_{\text{cat}} = [X_h', X_h \cdot SE(\text{pooling}(X_h))]. \tag{3}$$

2.3 nnMamba for Classification

For classification tasks, we adopt a residual encoder-centric architecture. We posit that reducing the size of feature maps amplifies the ability of convolutional operations to capture long-range dependencies. With this in mind, we strategically incorporate the Mamba layer early in the network, immediately following initial feature extraction. Positioning the Mamba layer at this stage ensures that features are endowed with global context early on, potentially diminishing the necessity for complex operations later in the network for abstract reasoning.

To further refine supervision across various feature map resolutions, we propose treating hierarchical features as a sequence, which are then processed by the Mamba to extract multi-level characteristics. Specifically, let X_h (for $h = 2, 3, 4$) denote the feature map from the second to the fourth block. We first apply a max-pooling operation to these feature maps, yielding P_h with a shape of $B \times C \times 2^h$. These are then reshaped into a sequence with dimensions $B \times 2C \times (1 + 2 + 3)$. Lastly, the SSM is employed to distill the hierarchical sequential features, which are subsequently fed into the MLP.

3 Experiments

3.1 Implementation and Metric

All the models are trained with NVIDIA V100 GPU with 24GB memory. The framework is implemented in PyTorch 2.0.1 and CUDA 11.6. We train the models with the Adam optimizer at a learning rate of 0.002, batch size at 2, training epoch at 1000, and weight decay at 0.001 for segmentation tasks. For other tasks, we only train 100 epochs. For the segmentation task, we use the Dice, NSD, and HD95 for evaluation by following[34]. For the landmark detection task, we take the MRE as the metric for evaluation following [10]. For the classification task, we use the AUC, Accuracy, and F1-score by following [11]. Details of the datasets and data split used in this work are available in the supplementary material. If there is a validation set, we select the best-performed

Table 1: Evaluation of nnMamba and other state-of-the-art methods on the BraTS 2023 GIL track.

BraTS 2023 GIL	Dice				HD95			
Methods	WT	TC	ET	Average	WT	TC	ET	Average
UniMiss[36]	93.48	90.06	84.40	89.31	4.55	6.80	13.70	8.38
DIT[37]	93.49	90.22	84.38	89.36	4.21	5.27	13.64	7.71
UNETR[21]	93.33	89.89	85.19	89.47	4.76	7.27	12.78	8.27
nnUNet[5]	93.31	90.24	85.18	89.58	4.49	4.95	11.91	7.12
nnMamba	93.46	90.74	85.72	89.97	4.18	5.12	10.31	6.53

Table 2: Evaluation of nnMamba on the AMOS2022 dataset [35]. The results of the test sets are obtained by submitting them to the official online testing site.

AMOS	Effiency		CT-Test		MRI-Test	
Metrics	Parameter(M)	Flops(G)	mDice	mNSD	mDice	mNSD
nnUNet[5]	31.18	680.31	89.04	78.32	67.63	59.02
VNet[38]	45.65	849.96	82.92	67.56	65.64	57.37
CoTr [39]	41.87	668.15	80.86	66.31	60.49	51.18
nnFormer[22]	150.14	425.78	85.61	72.48	62.92	54.06
UNETR[21]	93.02	177.51	79.43	60.84	57.91	47.25
SwinUNetr[40]	62.83	668.15	86.32	73.83	57.50	47.04
nnMamba	15.55	141.14	89.63	79.73	73.98	65.13

model on the validation set and evaluate it on the test set. Otherwise, the results are obtained by averaging the results of three different seeds. The visualization results are also available in the appendix.

3.2 Segmentation Evaluation

Table 1 illustrates the competitive performance of nnMamba in comparison to other state-of-the-art methods on the BraTS 2023 GIL task [7]. nnMamba consistently achieves top-tier results across all evaluated metrics for brain tumor segmentation, demonstrating its robustness and accuracy. The method's effectiveness is further highlighted by its performance in the challenging Hausdorff distance metric, showcasing its precision in delineating tumor boundaries. It can be seen that our method can significantly boost the performance on the boundary modeling of the tumor, especially for the long-range relationship modeling of enhanced tumors.

In Table 2, nnMamba showcases superior performance on the AMOS2022 dataset[35], particularly in the MRI-Test category, where it achieves a mean Dice coefficient (mDice) of 73.98% and a mean Normalized Surface Dice (mNSD) of 65.13%, outperforming all other compared methods. It might be our lightweight design and long-range modeling to boost the performance. Notably, nnMamba achieves these results with significantly lower parameters and computational complexity, boasting only 15.55 MB parameters and 141.14 GFLOPS. These results underline nnMamba's efficiency and effectiveness in medical image segmentation tasks.

3.3 Classification Evaluation

We evaluate the classification task on the ADNI dataset[41, 42] with the AD and NC classification task. We use the transfer learning setting in [11] to evaluate the methods by first training the networks on the NC AD classification task, then we transfer the weights and make the evaluation on the sMCI pMCI classification task. Table 2 shows the performance of nnMamba and benchmark methods on the ANDI classification task, comparing non-converters (NC) to Alzheimer's Disease (AD) and stable Mild Cognitive Impairment (sMCI) to progressive MCI (pMCI). In both tasks, nnMamba outperforms the other models, achieving the highest Accuracy (ACC), F1 score, and Area Under the Curve (AUC) metrics. The results indicate nnMamba's effectiveness, particularly in the NC VS AD task with an ACC of $89.41\% \pm 0.85$, an F1 score of $88.68\% \pm 0.77$, and an AUC of $95.81\% \pm 0.59$. Moreover,

Table 3: Evaluation of nnMamba and other state-of-the-art methods on the ANDI classification task.

Task	NC VS AD			sMCI VS pMCI		
Methods	ACC	F1	AUC	ACC	F1	AUC
ResNet [43]	$88.40_{\pm3.41}$	$88.00_{\pm2.81}$	$94.93_{\pm0.72}$	$67.96_{\pm1.50}$	$52.14_{\pm1.51}$	$74.94_{\pm2.18}$
DenseNet [44]	$87.95_{\pm0.70}$	$86.93_{\pm0.87}$	$94.86_{\pm0.40}$	$73.12_{\pm3.10}$	$53.30_{\pm2.99}$	$76.31_{\pm3.09}$
ViT [18]	$88.85_{\pm1.17}$	$87.66_{\pm1.72}$	$94.12_{\pm1.29}$	$67.16_{\pm3.16}$	$51.68_{\pm5.72}$	$75.08_{\pm6.88}$
CRATE [45]	$84.69_{\pm2.53}$	$82.66_{\pm3.47}$	$91.42_{\pm1.43}$	$70.63_{\pm2.60}$	$53.41_{\pm2.53}$	$76.06_{\pm2.98}$
nnMamba	$89.41_{\pm0.85}$	$88.68_{\pm0.77}$	$95.81_{\pm0.59}$	$75.79_{\pm1.79}$	$56.55_{\pm2.37}$	$76.84_{\pm0.84}$

Table 4: Evaluation of nnMamba and other state-of-the-art methods on the landmark detection task.

Landmark Detection	TCD 1	TCD 2	HDV 1	HDV 2	ADV 1	ADV 2	Average
ResUNet[43]	2.23	2.15	2.32	2.16	2.45	2.05	2.23
VitPose[46]	3.25	3.24	2.98	3.34	3.32	2.59	3.12
nnMamba	2.21	2.18	2.10	2.14	2.06	1.99	2.11

these results also show that our method can successfully boost the performance of transferring the learned weights to the downstream tasks on classification.

3.4 Landmark Detection Evaluation

For the landmark detection task, we evaluate our method on our private fetal cerebellum landmark detection dataset. Table 4 compares the performance of nnMamba with ResUNet and VitPose on a landmark detection task across six test conditions (TCD 1, TCD 2, HDV 1, HDV 2, ADV 1, ADV 2) and reports the average error across all conditions. nnMamba consistently exhibits lower error rates than the other methods in individual test conditions and maintains the lowest average error at 2.11. This underscores nnMamba's superior accuracy in landmark detection compared to the state-of-the-art methods listed.

3.5 Ablation Study

Table 5 presents the results of an ablation study, demonstrating the incremental benefits of MIC and MICCSS on various medical image analysis tasks. The integration of MIC and MICCSS into the baseline method yields consistent enhancements across segmentation metrics on the Brats2023 dataset, as evidenced by improved Dice and HD95 scores. Additionally, in landmark detection, the incorporation of these techniques leads to a decreased MRE, signifying more accurate localization capabilities. The ADNI classification task also shows a trend of increased AUC scores with each added component, highlighting the potential of MIC and MICCSS to boost the performance of classification models.

3.6 Visualization Results

We also provides the visualization results in Figure 3 and Figure 4, from these figures, we can observe that our method can get better results by modeling long range relationship.

Table 5: The ablation study on through evaluation metrics of segmentation, landmark detection, and classification tasks. "MIC" indicates directly using the SSM module.

Methods	Brats2023		Landmark	ADNI Classification	
Metric	Dice	HD95	MRE	-	AUC
Baseline	89.58	7.12	2.23	Baseline	94.93
+MIC	89.92	6.94	2.14	+MICCSS	95.35
+MICCSS	89.97	6.53	2.11	+Sequence	95.81

Table 6: Details of the dataset used in our study.

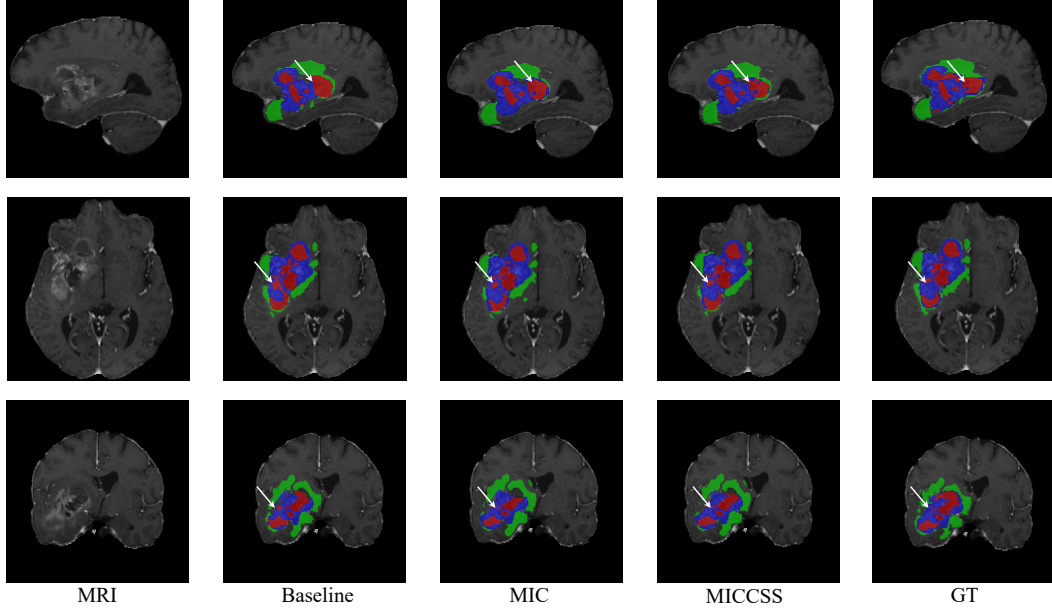| Task	Dataset	Dataset Type	Modality	Spacing	Train	Valid.	Test
	Brats	Tumor	MR	$1 \cdot 1 \cdot 1mm^3$	875	125	251
Segmentation	AMOS-CT	Organ	CT	$1.25 \cdot 5 \cdot 2mm^3$	200	100	200
	AMOS-MR	Organ	MR	$0.86 \cdot 3 \cdot 3mm^3$	40	20	40
Classification	ADNI	AD/NC	MR	$1 \cdot 1 \cdot 1mm^3$	351	-	296
	ADNI	sMCI/pMCI	MR	$1 \cdot 1 \cdot 1mm^3$	356	-	337
Landmark	Private	Landmark	MR	$0.6 \cdot 0.6 \cdot 0.6mm^3$	120	-	60

| MRI | Baseline | MIC | MICCSS | GT |

Figure 3: Visualization of segmentation predictions on the BraTS2023 dataset. From left to right, the columns display the MRI image, the prediction result of nnUNet, the prediction result of our nnMamba (MIC) model, the prediction result of our nnMamba (MICCSS) model, and the ground truth segmentation. This side-by-side comparison shows that our model can better capture the discontinue region of the tumor.

4 Conclusion

Our work presents nnMamba, a new framework combining the detailed feature extraction of CNNs with the broad dependency modeling of SSMs. Designed for 3D medical image analysis, nnMamba excels in segmentation, classification, and landmark detection. It outperforms current methodologies, offering a blend of local and global image context understanding. nnMamba's robust performance has significant implications for improving medical diagnostics and therapies.

Acknowledgments and Disclosure of Funding

References

[1] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248, 2017.
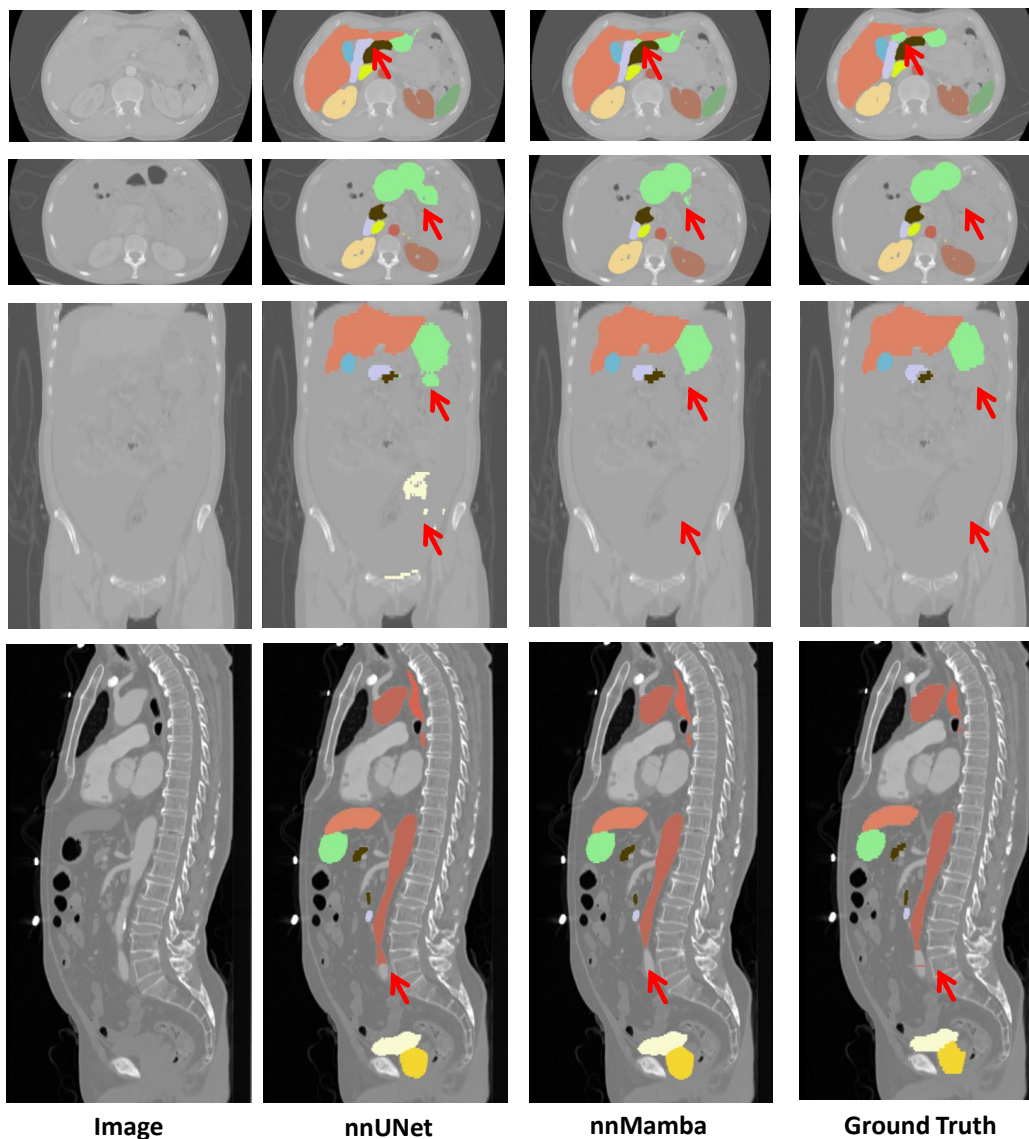
Figure 4: Visualization and comparison of segmentation predictions on the AMOS22 CT validation dataset. The four columns display, respectively, the CT image, the ground truth segmentation, the prediction result of nnUNet, and the prediction result of our nnMamba model. By capturing long-range dependencies, our nnMamba model effectively reduces over-segmentation and missed segmentation issues, particularly over long distances.

[2] Xiang He, Sibei Yang, Guanbin Li, Haofeng Li, Huiyou Chang, and Yizhou Yu. Non-local context encoder: Robust biomedical image segmentation against adversarial attacks. In *AAAI*, volume 33, pages 8417–8424, 2019.

[3] Junjia Huang, Haofeng Li, Guanbin Li, and Xiang Wan. Attentive symmetric autoencoder for brain mri segmentation. In *MICCAI*, pages 203–213. Springer, 2022.

[4] Zihang Xu, Haifan Gong, Xiang Wan, and Haofeng Li. Asc: Appearance and structure consistency for unsupervised domain adaptation in fetal brain mri segmentation. In *MICCAI*, pages 325–335. Springer, 2023.

[5] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

[6] Hee E Kim, Alejandro Cosa-Linan, Nandhini Santhanam, Mahboubeh Jannesari, Mate E Maros, and Thomas Ganslandt. Transfer learning for medical image classification: a literature review. *BMC medical imaging*, 22(1):69, 2022.

[7] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE TMI*, 34(10):1993–2024, 2014.

[8] Haifan Gong, Hui Cheng, Yifan Xie, Shuangyi Tan, Guanqi Chen, Fei Chen, and Guanbin Li. Less is more: Adaptive curriculum learning for thyroid nodule diagnosis. In *MICCAI*, pages 248–257. Springer, 2022.

[9] Junshen Xu, Molin Zhang, Esra Abaci Turk, Larry Zhang, P Ellen Grant, Kui Ying, Polina Golland, and Elfar Adalsteinsson. Fetal pose estimation in volumetric MRI using a 3d convolution neural network. In *MICCAI*, pages 403–410. Springer, 2019.

[10] Xiang Li, Songcen Lv, Minglei Li, Jiusi Zhang, Yuchen Jiang, Yong Qin, Hao Luo, and Shen Yin. Sdmt: Spatial dependence multi-task transformer network for 3d knee MRI segmentation and landmark localization. *IEEE TMI*, 2023.

[11] Luoyao Kang, Haifan Gong, Xiang Wan, and Haofeng Li. Visual-attribute prompt learning for progressive mild cognitive impairment prediction. In *MICCAI*, pages 547–557. Springer, 2023.

[12] Netanell Avisdris, Leo Joskowicz, Brian Dromey, Anna L David, Donald M Peebles, Danail Stoyanov, Dafna Ben Bashat, and Sophia Bano. Biometrynet: Landmark-based fetal biometry estimation from standard ultrasound planes. In *MICCAI*, pages 279–289. Springer, 2022.

[13] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[14] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

[15] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *CVPR*, pages 5455–5463, 2015.

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015.

[17] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pages 801–818, 2018.

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021.

[20] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[21] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *WACV*, pages 574–584, 2022.

[22] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Xiaoguang Han, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Volumetric medical image segmentation via a 3d transformer. *IEEE TIP*, 2023.

[23] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *ICLR*, 2022.

[24] Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *NeurIPS*, 34:572–585, 2021.

[25] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[26] Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024.

[27] Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. *arXiv preprint arXiv:2401.13560*, 2024.

[28] Yijun Yang, Zhaohu Xing, and Lei Zhu. Vivim: a video vision mamba for medical video object segmentation. *arXiv preprint arXiv:2401.14168*, 2024.

[29] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.

[30] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. 2013.

[31] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *ECCV*, pages 850–865. Springer, 2016.

[32] Haifan Gong, Yumeng Zhang, Chenhe Dong, Yue Wang, Guanqi Chen, Bilin Liang, Haofeng Li, Lanxuan Liu, Jie Xu, and Guanbin Li. Unbiased curriculum learning enhanced global-local graph neural network for protein thermodynamic stability prediction. *Bioinformatics*, page btad589, 09 2023.

[33] Zhongzhan Huang, Zhou Pan, Shuicheng Yan, and Liang Lin. Scalelong: Towards more stable training of diffusion model via scaling network long skip connection. In *NeurIPS*, 2023.

[34] Annika Reinke, Minu D Tizabi, Michael Baumgartner, Matthias Eisenmann, Doreen Heckmann-Nötzel, A Emre Kavur, Tim Rädsch, Carole H Sudre, Laura Acion, Michela Antonelli, et al. Understanding metric-related pitfalls in image analysis validation. *Nature methods*, pages 1–13, 2024.

[35] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *NeurIPS*, 35:36722–36732, 2022.

[36] Yutong Xie, Jianpeng Zhang, Yong Xia, and Qi Wu. Unimiss: Universal medical self-supervised learning via breaking dimensionality barrier. In *ECCV*, pages 558–575. Springer, 2022.

[37] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023.

[38] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, Oct 2016.

[39] Yutong Xie et al. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *MICCAI*, volume 24. Springer International Publishing, 2021.

[40] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in MRI images. In *International MICCAI Brainlesion Workshop*, pages 272–284. Springer, 2021.

[41] Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, 27(4):685–691, 2008.

[42] Chunfeng Lian, Mingxia Liu, Jun Zhang, and Dinggang Shen. Hierarchical fully convolutional network for joint atrophy localization and alzheimer's disease diagnosis using structural MRI. *TPAMI*, 42(4):880–893, 2020.

[43] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[44] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017.

[45] Yaodong Yu, Sam Buchanan, Druv Pai, Tianzhe Chu, Ziyang Wu, Shengbang Tong, Benjamin D Haeffele, and Yi Ma. White-box transformers via sparse rate reduction. In *NeurIPS*, 2023.

[46] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *NeurIPS*, 35:38571–38584, 2022.