SynMamba: Synergizing Visual Mamba with Synthesized Clinical Semantics for Boundary-Aware Segmentation

Xuan Wei[1,†], Zihan Li[4,†], Mengwei Wu[5], Ziheng Zhang[1], Zeyang Wang[1], Kaiheng Li[1], and Qingqi Hong[1,2,3,*]

[1] Department of Digital Media Technology, Xiamen University, Xiamen, China
[2] National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, China
[3] Institute of Artificial Intelligence, Xiamen University, Xiamen, China
[4] University of Washington, Washington, United States
[5] Zhongshan Hospital of Xiamen University, School of Medicine, Xiamen University, Xiamen, China
hongqq@xmu.edu.cn

Abstract. Precise medical image segmentation is vital for modern clinical diagnosis. Integrating multiple modalities, such as vision and text, offers a promising solution. However, two main challenges remain: the scarcity of high-quality image-text pairs and the structural flaws in current visual backbones. Specifically, CNNs lack global context, Transformers are computationally expensive, and Mamba-based models often miss fine-grained local details. We propose SynMamba, a dual-branch architecture. It combines the global modeling of Visual Mamba with the local edge-awareness of CNNs. To solve the data shortage, we introduce a modality synthesis framework. This framework converts masks into hallucination-free clinical narratives to guide the model during training. By embedding clinical semantics directly into the visual backbone, SynMamba excels in both text-guided and fixed-text tasks. Experiments on COVID-CT and COVID-Xray datasets show that SynMamba accurately delineates large overlapping tissues and small isolated lesions. It sets a new state-of-the-art across both unimodal and multimodal benchmarks.

Keywords: Image Segmentation · CNN · State Space Model · Multimodal Segmentation · Modality Synthesis

1 Introduction

Precise segmentation of anatomical structures is vital for disease diagnosis and treatment planning. Recent studies show that multimodal integration, such as

[†] Equal contribution.
[*] Corresponding author.

combining clinical text with radiological images, significantly improves segmentation performance. However, two primary challenges remain. First, high-quality image-text pairs are scarce. High-quality medical image-mask pairs with corresponding professional reports are difficult to collect, which limits the training of robust multimodal models. Second, current visual backbones have structural flaws. Traditional CNNs [16, 22, 19, 5, 14] focus on local details but lack the global vision. In contrast, Transformers [2, 1, 10] provide global context but sacrifice local precision and carry high computational costs. Recently, visual state space models (VSSMs) [9, 7, 17] have emerged as a powerful tool for global modeling with better efficiency than Transformers. However, existing Mamba models still fail to capture the fine-grained local features vital for medical tasks. Besides, how these visual backbones efficiently interact with paired text is also a challenge.

To overcome these hurdles, we propose SynMamba. Our approach has two objectives: synergizing the global modeling of Mamba with the local precision of CNNs, and internalizing clinical semantics into the visual backbone. To address the data gap, we introduce a modality synthesis framework. This framework equips existing datasets with hallucination-free clinical narratives. By using synthesized text as guidance during training, SynMamba achieves robust performance in both multimodal and unimodal inference. It performs consistently well with or without valid text prompts.

Our main contributions are summarized as follows:

1. We propose SynMamba, a dual-branch network. It integrates the global context of VSSM with the local detail capture of CNNs. We also design a semantic interaction bottleneck for better visual-text feature integration. We utilize a boundary-aware supervision for accurate edge segementation.
2. We develop a rule-based framework to convert target masks into hallucination-free clinical narratives. During training, these narratives provide semantic guidance by injecting medical knowledge into the visual backbone. This approach ensures robust and accurate segmentation during inference. Even without valid text prompts, SynMamba maintains high-quality performance.
3. We evaluate SynMamba with and without valid text prompts on two public datasets, and SynMamba consistently outperforms all unimodal and multimodal baselines, showing its successful integration of clinical knowledge.

2 Method

2.1 Overview

As illustrated in Fig. 1, SynMamba is a framework that synergizes global anatomical context, local structural details, and synthesized clinical priors. It consists of three integral components: an edge-aware dual-branch backbone, a modality synthesis framework, and a visual-text interaction bottleneck. Formally, given an input image $I \in \mathbb{R}^{H \times W \times C}$ and a text prompt T, it predicts a segmentation mask $\hat{M} \in \mathbb{R}^{H \times W}$.

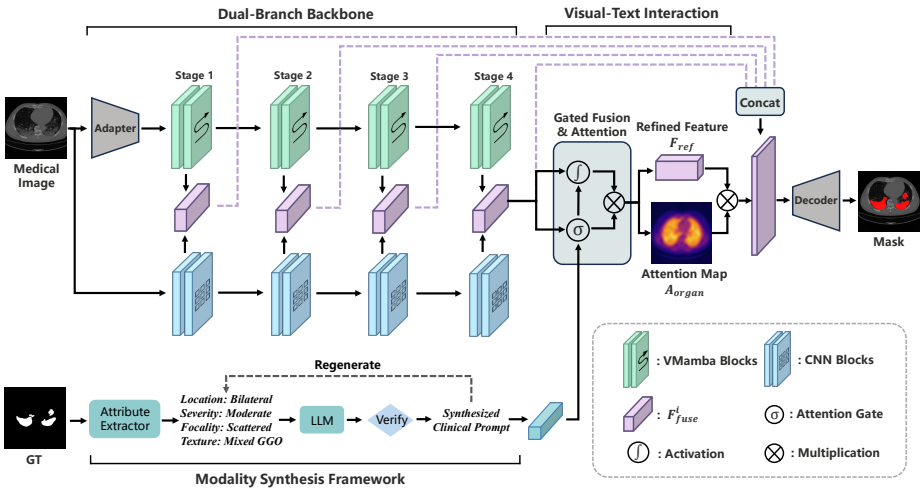

Fig. 1. Overview of the SynMamba architecture. It consists of three components: a dual-branch backbone that fuses features from both VMamba and CNN blocks, a modality synthesis framework that generates synthesized clinical prompts from GT masks for multimodal training, and a visual-text interaction bottleneck to gate and fuse multimodal features. Finally, a decoder is used to predict the segmentation mask and edge map.

Crucially, the synthesized text is utilized during training to provide rich semantic guidance. SynMamba is capable of operating with or without valid textual information by using a fixed text input like "Medical image with potential lesion".

2.2 Dual-Branch Backbone

The backbone employs a dual-branch strategy to extract complementary features. It utilizes visual state space model (VSSM) for multi-scale global context $\{f_g^i\}_{i=1}^4$ and CNN for hierarchical local details $\{f_l^i\}_{i=1}^4$. We use subscripts to denote the source branch and superscripts to indicate the spatial scale.

To integrate this visual information effectively, we concatenate and convolve the stage-wise feature pairs (f_g^i, f_l^i) from both branches:

$$F_{fuse}^i = \mathcal{F}(f_g^i \oplus f_l^i) \tag{1}$$

where $\mathcal{F}$ denotes a fusion block consisting of consecutive 1×1 and 3×3 convolutions, spatially mix the heterogeneous features. The resulting fused features F_{fuse}^i combine long-range anatomical dependencies from the VSSM branch with fine-grained textural cues from the CNN branch. This process forms a robust multi-scale representation for subsequent decoding stages.

Table 1. Examples of the Modality Synthesis pipeline. Crucially, radiological attributes (e.g., Laterality, Texture) are deterministically derived from spatial centroids and pixel-intensity thresholds prior to LLM processing, ensuring zero generative hallucination.

Image	GT	Laterality	Severity	Focality	Texture	Synthesized Clinical Prompt
		Bilateral	Moderate ($R_{inf} \approx 18.5\%$)	Diffuse	Mixed ground-glass and consolidation	Chest radiograph reveals bilateral moderate mixed ground-glass opacities and consolidation distributed in the lower lung zones.
		Bilateral	Mild ($R_{inf} \approx 4.2\%$)	Scattered	Predominant ground-glass opacities	Focal bilateral mild ground-glass opacities are observed, indicating predominantly scattered infection patterns.

2.3 Visual-Text Interaction Bottleneck

To inject semantic guidance, we extract the text embedding F_{text} from the clinical text T. We propose a visual-text interaction bottleneck to suppress noise in the deepest visual feature F^4_{fuse} and align it with textual semantics. First, we compute a refined visual feature using a residual connection:

$$F_{ref} = \mathrm{Conv}(F^4_{fuse}) + \alpha \cdot F^4_{fuse} \tag{2}$$

where $\mathrm{Conv}(\cdot)$ denotes a local refinement convolution, and α is a learnable parameter scaling the original visual connection. We then combine the spatially expanded text embedding F_{text} with F_{ref} to generate a soft semantic gating map $G \in [0,1]^{H\times W\times C}$:

$$G = \sigma(\mathcal{G}(F_{ref} \odot F_{text})) \tag{3}$$

where $\mathcal{G}$ denotes the gating network. This gate modulates the visual stream via element-wise multiplication, yielding the text-gated visual feature $F_{vis} = F_{ref} \odot G$. This mechanism effectively filters out background regions inconsistent with the clinical text. Subsequently, F_{vis} and F_{text} are fed into a cross-attention module, producing a spatial attention A_{organ} to highlight the target lesion areas.

2.4 Decode and Supervision

Based on the above features, we compute the final feature F_{final} that fed into the decoder as:

$$F_{final} = \mathrm{Conv}\Big(\mathrm{Conv}\big(\{F^i_{fuse}\}^4_{i=1} \oplus \mathcal{E}(f^4_l)\big) \oplus A_{organ}\Big) \tag{4}$$

where $\mathcal{E}$ is an edge feature extractor. Then F_{final} is processed by the decoder to output the mask $\hat{M}$ and the edge map $\hat{E}$.

Regarding supervision, standard region-based losses $\mathcal{L}_{BCE}$ and $\mathcal{L}_{Dice}$ often overlook high-frequency topological details. To address this, we enhance boundary supervision by incorporating two additional components: a gradient loss $\mathcal{L}_{grad}$ and a structural edge loss $\mathcal{L}_{edge}$. The total loss is defined as:

$$\mathcal{L}_{total} = \lambda_{BCE}\mathcal{L}_{BCE} + \lambda_{Dice}\mathcal{L}_{Dice} + \lambda_{grad}\mathcal{L}_{grad} + \lambda_{edge}\mathcal{L}_{edge} \tag{5}$$

Specifically, $\mathcal{L}_{grad}$ calculates the Mean Squared Error (MSE) between the spatial gradients of the predicted mask $\hat{M}$ and the ground truth M, both extracted via Sobel operators. Meanwhile, $\mathcal{L}_{edge}$ computes the Binary Cross-Entropy (BCE) between the predicted edge map $\hat{E}$ and the ground truth edge map E. This auxiliary task strengthens the model's ability to delineate sharp boundaries. During training, we set the loss weights to $\lambda_{BCE} = 1.0$, $\lambda_{Dice} = 1.0$, $\lambda_{grad} = 0.1$, and $\lambda_{edge} = 0.2$.

2.5 Modality Synthesis

SynMamba leverages narrative text to guide segmentation. Fluent clinical descriptions can unlock the relational knowledge within pre-trained text encoders and provide dense semantic guidance for cross-modal alignment. Since most medical datasets lack textual modalities, we synthesize text prompts from ground-truth masks and their corresponding images.

As shown in Table 1, we employ a fully rule-based attribute extractor to derive key radiological attributes, including anatomical localization, connected component counts, lesion occupancy ratio, and texture profiling. Domain-specific thresholds, such as HU ranges for CT or normalized grayscale boundaries for X-ray, are used to categorize masked regions into ground-glass opacities, consolidation, or mixed patterns. Since these attributes are deterministically computed from annotations and image statistics, no generative model is involved in producing lesion properties.

Qwen3 VL is then used only to verbalize the structured attributes into fluent clinical narratives, which better match the pre-training distribution of medical text encoders than raw attribute tokens. To suppress hallucinations, we adopt a generate-verify mechanism: each generated description is checked against the original image–attribute pair and regenerated if any mismatch is detected. Empirically, no faulty or inconsistent samples were observed after verification.

3 Experiments

3.1 Datasets and Metrics

We evaluate our model on the COVID-19 CT Scan Lesion Segmentation Dataset [12, 13] containing 2,729 radiologist-annotated slices, and the QaTa-COV19-v2 dataset [4] comprising 9,258 expert-labeled X-ray images. For clarity, we refer to these two datasets as COVID-CT and COVID-Xray, respectively, in the following sections. For model optimization and evaluation, COVID-CT is randomly partitioned into

Table 2. Performance comparison of different segmentation methods on the COVID-CT [12, 13] and the COVID-Xray [4] datasets. "Text" indicates whether valid textual descriptions are provided during inference. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	Text	COVID-CT		COVID-Xray	
		Dice (%)	IoU (%)	Dice (%)	IoU (%)
UNet [16]	×	63.07	50.72	79.13	69.08
UNet++ [22]	×	71.57	58.17	79.44	70.03
AttUNet [19]	×	65.76	51.96	79.50	70.18
nnUNet [5]	×	72.90	60.03	80.21	70.26
TransUNet [2]	×	71.24	58.44	78.63	69.13
Swin-UNet [1]	×	63.06	50.19	77.04	67.96
Rolling-Unet [10]	×	78.48	69.09	72.67	69.39
MADGNet [14]	×	79.27	**69.17**	79.22	68.35
U-Mamba [11]	×	78.86	67.60	79.23	70.18
Swin-UMamba [7]	×	<u>79.35</u>	68.34	<u>80.53</u>	<u>71.46</u>
VM-UNet [17]	×	78.12	67.56	78.12	67.56
SynMamba (ours)	×	**80.57**	<u>69.12</u>	**84.34**	**75.26**
ConVIRT [21]	✓	71.68	59.49	79.50	<u>70.65</u>
TGANet [18]	✓	70.41	57.42	80.12	70.36
CLIP [15]	✓	71.75	59.71	79.68	70.38
ViLT [6]	✓	72.61	59.76	80.00	69.99
UniLSeg [8]	✓	71.60	57.95	<u>79.99</u>	70.29
CMIRNet [20]	✓	<u>79.27</u>	<u>60.44</u>	79.22	69.98
SynMamba (ours)	✓	**83.28**	**70.96**	**88.18**	**78.86**

an 8:1:1 ratio for training, validation, and testing, while COVID-Xray follows its official 7,145/2,113 train/test split with 20% of the training data reserved for validation. Segmentation performance is quantitatively assessed using two widely adopted metrics: the Dice Similarity Coefficient (DSC) and the Intersection over Union (IoU).

3.2 Comparison with State-of-the-Art Methods

Quantitative Analysis. We evaluate SynMamba across two challenging benchmarks: COVID-CT and COVID-Xray. The comparative methods can be categorized into two primary paradigms based on their input modalities: pure vision models like MADGNet [14], TransUNet [2], VM-UNet [17]; and vision-language models like ConVIRT [21], UniLSeg [8], and CMIRNet [20]. All baselines are trained from scratch under the same data splits for fair comparison.

To ensure a fair comparison with pure-vision baselines, we provide SynMamba with a fixed, uninformative prompt—"Medical image with potential lesion"—to neutralize its semantic pathway. Even under this constrained setting, our model achieves superior results, reaching 80.57% Dice on COVID-CT and 84.34% Dice (75.26% IoU) on COVID-Xray. SynMamba outpaces both CNN-based and VSSM-based methods. This success stems from our dual-branch

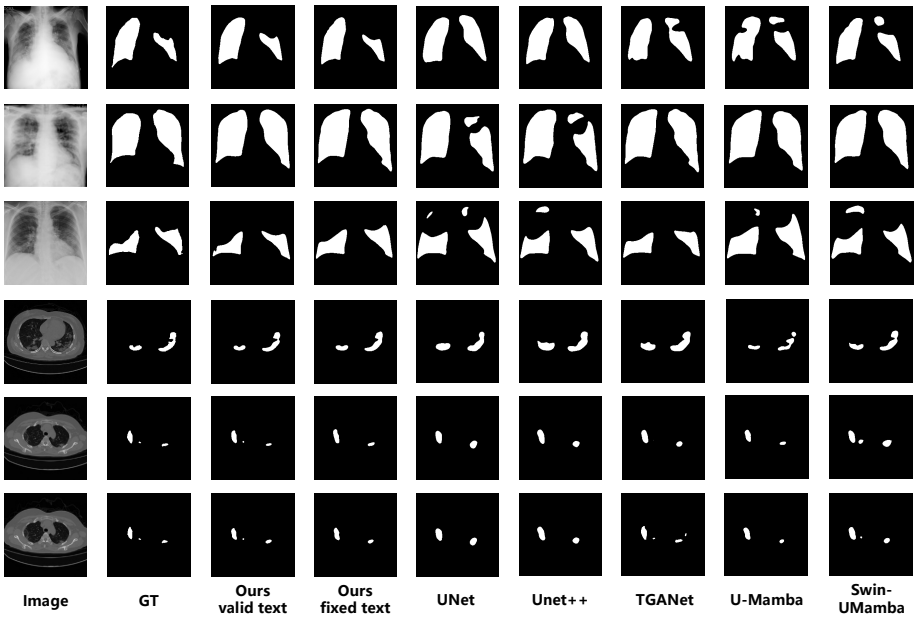

Fig. 2. Qualitative segmentation comparison on the COVID-CT [12, 13] and COVID-Xray [4] datasets.

backbone, which effectively integrates the global modeling of VSSM with the boundary precision of CNNs.

In the vision-language setting, where all models receive identical synthesized text, SynMamba establishes a new state-of-the-art. Our model reaches 83.28% Dice on COVID-CT and 88.18% on COVID-Xray, outperforming established multimodal frameworks such as CMIRNet and TGANet. This substantial performance gain underscores the model's ability to internalize text-guided clinical priors to boost segmentation accuracy.

Qualitative Analysis. As shown in Fig. 2, SynMamba demonstrates superior performance across diverse clinical scenarios. For large-scale targets with overlapping structures, CNN-based models often struggle to maintain global context due to their limited receptive fields. Conversely, for small, isolated lesions, pure SSM-based models (e.g., U-Mamba) frequently suffer from an inherent over-smoothing effect, leading to missed detections or blurred boundaries.

In contrast, SynMamba successfully captures both large tissues and tiny isolated lesions. Notably, our model produces sharper and more consistent boundaries. This confirm our dual-branch design and boundary-aware supervision successfully overcome the local feature loss in SSMs. Furthermore, our text-informed interaction prevents the misclassification of minute targets, maintaining robust segmentation even with generic, fixed text prompts.

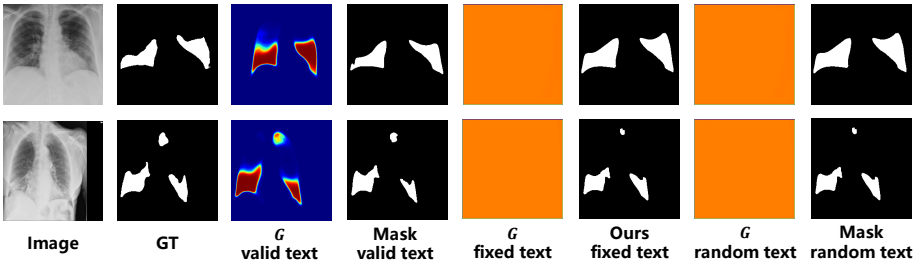

Fig. 3. Visualization of SynMamba across different textual prompts. While valid text triggers precise spatial highlighting in gate map G, fixed or random text induces a uniform pass-through state with minor degrading in segmentation accuracy.

Table 3. Ablation study of key components on the COVID-Xray [3] dataset.

Methods		Dice (%)	mIoU (%)
Baseline	Ours (valid text)	**88.18**	**78.86**
	Ours (fixed text)	84.34	75.26
Backbone	w/o CNN branch	70.92	54.95
	w/o Mamba branch	68.72	53.48
Interaction	w/o gated fusion	48.80	32.28
	w/o cross-attention	62.55	56.70
	w/o A_{organ}	82.26	69.86
Text	w/o text training	76.23	61.60

3.3 Ablation Studies

To systematically evaluate the contribution of each core component in SynMamba, we conducted ablation studies from three key perspectives, with results summarized in Table 3.

First, removing either the CNN or Mamba branch deprives the network of critical high-frequency details or global anatomical topologies, causing the Dice score to drop to 70.92% and 69.72%, respectively. Second, replacing the gated fusion mechanism with simple concatenation results in a sharp decline to 48.80%. This confirms that unmodulated textual embeddings inject severe noise into the visual features. We also tested the influence of different text prompts on SynMamba. As shown in Fig. 3, with valid text, the gate G successfully highlights lesions while suppressing background noise. However, when fed uninformative fixed text or random garbage (e.g., "123"), G degenerates into a uniform pass-through state (solid orange). This shows that our model maintains strong robustness towards text input. Furthermore, removing the cross-attention between F_{vis} and F_{text} prevents the model from bridging the vision-language gap, leading to significant degradation. Excluding the explicit attention map A_{organ} decreases performance to 82.26%, validating its role as a regional guidance map that enforces anatomical focus. Finally, omitting text guidance during training drops the Dice score to 76.23%, highlighting the necessity of distilling semantic priors into the visual backbone.

4 Conclusion

In this paper, we propose SynMamba, a text-informed dual-branch framework that integrates a Mamba branch with a lightweight CNN to effectively capture both long-range contextual dependencies and high-frequency local features. Furthermore, we introduce a modality synthesis strategy to generate fluent clinical narratives. These narratives are coupled with visual features through an interaction bottleneck, which successfully distills high-dimensional clinical priors into the visual backbone during training. Our method demonstrates superior performance in segmenting both large anatomical structures and small, isolated lesions. Moreover, it maintains remarkable robustness even in fixed-text scenarios. These advancements establish a solid foundation for adaptable and reliable clinical applications in modern medical imaging.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant 62471418 and Fujian Provincial Natural Science Foundation of China under Grant 2024J01058.

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

References

1. 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: ECCV (2022)
2. 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)
3. Chowdhury, M.E.H., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M.A., Mahbub, Z.B., Islam, K.R., Khan, M.S., Iqbal, A., Emadi, N.A., Reaz, M.B.I., Islam, M.T.: Can ai help in screening viral and covid-19 pneumonia? IEEE Access (2020)
4. Degerli, A., Ahishali, M., Yamac, M., Kiranyaz, S., Chowdhury, M.E., Hameed, K., Hamid, T., Mazhar, R., Gabbouj, M.: Covid-19 infection map generation and detection from chest x-ray images. Health information science and systems (2021)
5. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Klaus, K.H., et al.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods (2021)
6. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: ICML. pp. 5583–5594 (2021)
7. 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: MICCAI. pp. 615–625 (2024)
8. Liu, Y., Zhang, C., Wang, Y., Wang, J., Yang, Y., Tang, Y.: Universal segmentation at arbitrary granularity with language instruction. In: CVPR. pp. 3459–3469 (2024)
9. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., Liu, Y.: Vmamba: Visual state space model. Advances in neural information processing systems **37**, 103031–103063 (2024)

10. Liu, Y., Zhu, H., Liu, M., Yu, H., Chen, Z., Gao, J.: Rolling-unet: Revitalizing mlp's ability to efficiently extract long-distance dependencies for medical image segmentation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 3819–3827 (2024)
11. Ma, J., Li, F., Wang, B.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722 (2024)
12. Ma, J., Wang, Y., An, X., Ge, C., Yu, Z., Chen, J., Zhu, Q., Dong, G., He, J., He, Z., et al.: Toward data-efficient learning: A benchmark for covid-19 ct lung and infection segmentation. Medical physics (2021)
13. Morozov, S.P., Andreychenko, A.E., Blokhin, I.A., Gelezhe, P.B., Gonchar, A.P., Nikolaev, A.E., Pavlov, N.A., Chernina, V.Y., Gombolevskiy, V.A.: Mosmeddata: data set of 1110 chest ct scans performed during the covid-19 epidemic. Digital Diagnostics (2020)
14. Nam, J.H., Syazwany, N.S., Kim, S.J., Lee, S.C.: Modality-agnostic domain generalizable medical image segmentation by multi-frequency in multi-scale attention. In: CVPR. pp. 11480–11491 (2024)
15. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)
16. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
17. Ruan, J., Li, J., Xiang, S.: Vm-unet: Vision mamba unet for medical image segmentation. ACM Transactions on Multimedia Computing, Communications and Applications (2025)
18. Tomar, N.K., Jha, D., Bagci, U., Ali, S.: Tganet: Text-guided attention for improved polyp segmentation. In: MICCAI. pp. 151–160 (2022)
19. Wang, S., Li, L., Zhuang, X.: Attu-net: attention u-net for brain tumor segmentation. In: International MICCAI brainlesion workshop. pp. 302–311 (2021)
20. Xu, M., Xiao, T., Liu, Y., Tang, H., Hu, Y., Nie, L.: Cmirnet: Cross-modal interactive reasoning network for referring image segmentation. IEEE Transactions on Circuits and Systems for Video Technology **35**(4), 3234–3249 (2025)
21. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Machine learning for healthcare conference. pp. 2–25 (2022)
22. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (2018)