VM-NeXT UNet: Synergizing ConvNeXT and Visual Mamba for Robust Medical Image Segmentation

Boxuan Cao[1] and Peilin Jiang[1⋆]

Xi'an Jiaotong University, Xi'an, China
13022916069@163.com, pljiang@xjtu.edu.cn

Abstract. In recent medical image analysis, Convolutional Neural Networks (CNNs) and State Space Models (SSMs) have set benchmarks in segmentation tasks. While CNNs excel in capturing local fine-grained features, SSMs achieve remarkable global context understanding with linear complexity. However, Mamba-UNet, a pioneering pure SSM-based model, still exhibits deficiencies in feature representation, fusion efficiency, and spatial detail reconstruction. To address these limitations, we propose **VM-NeXT UNet**, an improved architecture that synergizes ConvNeXT with VSS Blocks. **VM-NeXT UNet** introduces four key optimizations: (1) a dual-encoder parallel structure for comprehensive feature extraction; (2) a channel-spatial attention gating module in skip connections for adaptive feature screening; (3) multi-scale convolution layers in the decoder to preserve spatial details; and (4) a combined FocalLoss and DiceLoss strategy to focus on hard samples. Experiments on the Synapse and ACDC datasets yielded Dice scores of 86.21% and 92.42%, respectively. The results demonstrate that **VM-NeXT UNet** significantly outperforms the original Mamba-UNet and achieves competitive performance against state-of-the-art methods, highlighting its potential for reliable clinical deployment.

Keywords: Medical Image Segmentation · State Space Models · Dual-encoder Architecture.

1 Introduction

Automated medical image segmentation is crucial for computer-aided diagnosis and treatment planning. While Convolutional Neural Networks (CNNs) [1, 4, 10, 18, 21, 25] excel at local feature extraction via their inductive bias, they are limited by the local receptive field of convolutional kernels. To overcome this, the Vision Transformer (ViT) [5] revolutionized the field by treating image patches as sequences, enabling the modeling of long-range dependencies from the very first layer. Building on this, a series of Transformer-based architectures including TransUNet [3], Swin-Unet [2], MedT [22] and MT-UNet [23] have been proposed

⋆ Corresponding author.

to bridge the gap between local and global representations in 2D medical imaging. However, ViT-based models face two significant hurdles in clinical practice: first, the patch-wise tokenization in ViT inherently discards fine-grained spatial information, which often leads to blurred boundaries for small anatomical structures; second, the quadratic computational complexity of ViT's self-attention mechanism makes it prohibitively expensive for processing high-resolution medical slices. Recently, State Space Models (SSMs), particularly Mamba [8,14], have emerged as a promising alternative, offering global modeling with linear computational complexity, making them ideal for high-resolution medical images.

Despite the success of Mamba-UNet, a pure SSM-based baseline, it faces four inherent limitations: (1) insufficient representation of local high-frequency details compared to CNNs; (2) redundant semantic noise in additive skip connections; (3) limited ability to reconstruct fine spatial details for small targets; and (4) suboptimal performance on hard-to-segment samples using standard loss functions.

To address these, we propose **VM-NeXT UNet**, a hybrid architecture synergizing the strengths of CNNs and SSMs. By integrating ConvNeXT [15] with Visual Mamba [14], we introduce four strategic enhancements:

1. **Dual-Encoder Parallel Hybridization:** A parallel structure combining ConvNeXT and VSS Blocks to simultaneously capture local textures and global dependencies, fused via residual injection.
2. **Adaptive Feature Gating (CSAG):** A Channel-Spatial Attention Gating module for skip connections that adaptively suppresses semantic noise and filters relevant features.
3. **Multi-Scale Reconstruction:** Multi-scale convolution layers in the decoder to enhance boundary delineation and small target segmentation.
4. **Optimized Loss Strategy:** A FocalLoss + DiceLoss combination to address class imbalance and focus on hard samples.

Extensive experiments on Synapse (abdominal CT) and ACDC (cardiac MRI) datasets demonstrate that **VM-NeXT UNet** significantly outperforms Mamba-UNet and achieves state-of-the-art performance.

2 Method

In this section, we elaborate on the technical details of the proposed VM-NeXT UNet, including its unified hybrid architecture, core functional modules, and the tailored loss strategy utilized for training.

2.1 Overall Architecture of VM-NeXT UNet

As illustrated in Fig. 1(a), VM-NeXT UNet follows a four-stage U-shaped hierarchical structure. An input image $\mathbf{x}$ is first projected into an embedding space ($C = 96$) via 4×4 patch embedding. Our encoder adopts a Dual-Encoder Parallel design: a ConvNeXT branch for local feature extraction and a VSS branch

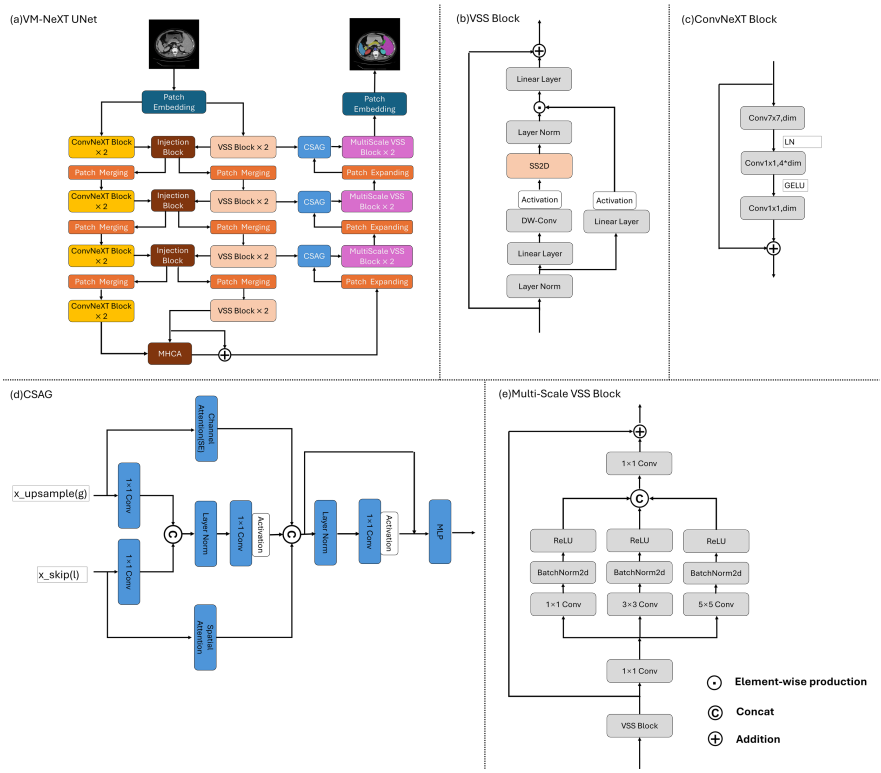

Fig. 1. (a) The overall architecture of VM-NeXT UNet. (b) The VSS Block [14] utilized for global dependency modeling. (c) The ConvNeXT Block [15] employed for local feature extraction. (d) The Channel-Spatial Attention Gating (CSAG) module for adaptive feature screening in skip connections. (e) The Multi-Scale VSS Block in the decoder for robust spatial reconstruction.

for global context modeling. Information exchange between these branches is facilitated by lightweight Injection Blocks at each stage.

At the bottleneck, a Cross-Attention mechanism deeply fuses the dual representations. The decoder then utilizes Multi-Scale VSS Block modules for progressive spatial reconstruction. To bridge the semantic gap, skip connections incorporate Channel-Spatial Attention Gating (CSAG) blocks, which adaptively filter and fuse multi-scale encoder features before passing them to the decoder.

2.2 Dual-Encoder Parallel Coding and Interaction

To address the limitations of pure SSM-based models in capturing high-frequency local details, we propose a hybrid dual-path architecture. This design synergizes the **VSS Block** (Fig. 1(b)), which utilizes the 2D Selective Scan (SS2D) mechanism for linear-complexity global modeling, with the **ConvNeXT Block**

(Fig. 1(c)), which employs an inverted bottleneck structure to provide strong local inductive bias.

Parallel Branch Design. The encoder is structured into four stages. In each stage, two parallel branches operate with aligned feature resolutions: (i) the ConvNeXT Branch, which adopts a ImageNet-pretrained ConvNeXT-Tiny [15] backbone to capture local texture patterns; and (ii) the VSS Branch, which stacks VSS Blocks initialized with ImageNet-pretrained VMamba-Tiny [14] weights to effectively model long-range dependencies. Patch Merging is applied at the end of each stage to double the channel dimensions to $[C, 2C, 4C, 8C]$.

Feature Injection Block. To prevent feature decoupling without incurring excessive computational costs, we design a lightweight Injection Block between the parallel paths. Let $\mathbf{X}_{vss} \in \mathbb{R}^{B\times H\times W\times C}$ and $\mathbf{X}_{conv} \in \mathbb{R}^{B\times C\times H\times W}$ denote the intermediate features. The injection process is formulated as:

$$\begin{aligned} \mathbf{X}'_{vss} &= \mathbf{X}_{vss} + \mathcal{P}(\phi_{c\to v}(\mathbf{X}_{conv})), \\ \mathbf{X}'_{conv} &= \mathbf{X}_{conv} + \phi_{v\to c}(\mathcal{P}^{-1}(\mathbf{X}_{vss})), \end{aligned} \tag{1}$$

where ϕ denotes a 1×1 convolution for channel alignment and $\mathcal{P}$ represents the dimension permutation.

Bottleneck Interaction via Cross-Attention. At the deepest stage (bottleneck) of the network, we introduce a Cross-Attention mechanism to achieve a high-level semantic fusion of the two representations. In this module, the global context features from the VSS branch are utilized as the Query, while the local fine-grained features from the ConvNeXT branch serve as Key and Value. This interaction allows the model to adaptively reinforce global representations with localized structural information before the decoding process.

2.3 Channel-Spatial Attention Gating (CSAG)

To bridge the semantic gap in skip connections and suppress background noise, we propose the CSAG module as shown in Fig. 1(d), which is derived from [9]. It adaptively filters local features $\mathbf{F}_l$ from the ConvNeXT branch and global features $\mathbf{F}_g$ from the VSS branch through collaborative attention pathways.

Gated Attention Pathways. CSAG employs parallel attention mechanisms to refine dual-path representations. For global features $\mathbf{F}_g$, channel attention emphasizes informative semantic categories by aggregating spatial statistics:

$$\hat{\mathbf{F}}_g = \sigma\left(\mathrm{MLP}(\mathrm{GAP}(\mathbf{F}_g)) + \mathrm{MLP}(\mathrm{GMP}(\mathbf{F}_g))\right) \odot \mathbf{F}_g. \tag{2}$$

Simultaneously, spatial attention is applied to local features $\mathbf{F}_l$ to highlight salient regions and preserve structural boundaries:

$$\hat{\mathbf{F}}_l = \sigma\left(f^{7\times 7}([\mathrm{Max}_c(\mathbf{F}_l); \mathrm{Avg}_c(\mathbf{F}_l)])\right) \odot \mathbf{F}_l. \tag{3}$$

Hybrid Feature Fusion. Following the gating process, the refined features $\hat{\mathbf{F}}_g$ and $\hat{\mathbf{F}}_l$ are concatenated with the cross-modal interaction feature $\mathbf{F}_{int}$. To achieve thorough integration, the concatenated representation $\mathbf{F}_{cat}$ undergoes a sequence of operations including Layer Normalization, a 1×1 convolution for channel alignment, and non-linear activation. Finally, an MLP layer is utilized for high-dimensional mixing:

$$\mathbf{F}_{out} = \mathrm{MLP}(\mathrm{Gated}(\hat{\mathbf{F}}_g, \hat{\mathbf{F}}_l, \mathbf{F}_{int})). \tag{4}$$

This multi-stage fusion strategy effectively suppresses redundant noise while providing the decoder with semantically enriched and spatially precise representations.

2.4 Multi-Scale Enhanced Decoder

Standard SSM-based decoders often struggle with anatomical structures of varying sizes due to single-scale processing. To address this, we propose the **Multi-Scale VSS Block**, as illustrated in Fig. 1(e), which sequentially couples global context modeling with multi-scale local spatial reconstruction. This module is designed as a holistic unit that restores high-frequency details while maintaining long-range semantic consistency.

Sequential Global-Local Coupling. The Multi-Scale VSS Block processes the input feature $\mathbf{F}_{in}$ through a nested global-to-local refinement strategy. It first models global dependencies via the Visual State Space Model (VSSM) and then adaptively captures multi-scale features through the Multi-scale Inverted Residual Block (MIRB). This sequential operation is formulated as a composite residual function:

$$\mathbf{F}_{out} = \mathcal{F}_{global}(\mathbf{F}_{in}) + \mathrm{DP}\left(\mathrm{MIRB}\left(\mathcal{F}_{global}(\mathbf{F}_{in})\right)\right), \tag{5}$$

where $\mathcal{F}_{global}(\mathbf{F}_{in}) = \mathbf{F}_{in} + \mathrm{DP}(\mathrm{VSSM}(\mathrm{LN}(\mathbf{F}_{in})))$ represents the feature map after global context injection. Here, LN and DP denote Layer Normalization and DropPath, respectively.

Multi-Scale Spatial Reconstruction. The MIRB component is specifically engineered to handle scale variations. It employs an inverted residual structure with parallel depthwise convolutions using multiple kernel sizes $\mathcal{K} \in \{1 \times 1, 3 \times 3, 5 \times 5\}$. By integrating these varying receptive fields, the decoder can effectively reconstruct both fine-grained organ boundaries and large-scale anatomical structures, significantly enhancing the robustness of the segmentation across different clinical scenarios.

2.5 Optimized Loss Function Strategy

In medical image segmentation, class imbalance between the background and small anatomical structures often leads to suboptimal convergence. To address

this, we replace the CE loss with Focal Loss to focus the training process on hard-to-segment samples and ambiguous boundaries. The Focal Loss ($\mathcal{L}_{focal}$) [13] is defined as:

$$\mathcal{L}_{focal}(p_t) = -\alpha(1-p_t)^{\gamma}\log(p_t). \tag{6}$$

In our implementation, we adopt a balanced strategy where the total objective function $\mathcal{L}_{total}$ is a weighted sum of Dice Loss ($\mathcal{L}_{Dice}$) and Focal Loss with equal importance:

$$\mathcal{L}_{total} = 0.5 \cdot \mathcal{L}_{Dice} + 0.5 \cdot \mathcal{L}_{focal}. \tag{7}$$

3 Experiments

3.1 Datasets

Synapse Multi-organ Segmentation Dataset (Synapse). The Synapse dataset comprises 30 abdominal CT cases with 3,779 axial clinical images. We segment eight abdominal organs: aorta, gallbladder, left kidney, right kidney, liver, pancreas, spleen, and stomach. Following [2, 3, 6], the dataset is split into 18 cases for training and 12 cases for validation and testing. To evaluate the segmentation performance on this dataset, we report both the Dice Similarity Coefficient (DSC) and the 95% Hausdorff Distance (HD95).

Automated Cardiac Diagnosis Challenge (ACDC). The ACDC dataset contains cine-MRI scans from 100 patients. Similar to [2, 3, 23], we partition the dataset into 70 cases for training, 10 cases for validation, and 20 cases for testing. We report the average DSC as the primary evaluation metric for this dataset.

3.2 Implementation Details

The proposed VM-NeXT UNet was implemented in PyTorch 2.8.0. All experiments were conducted on an NVIDIA GeForce RTX 5090 GPU. All input images were resized to 224×224. The training procedure was conducted with a batch size of 24, spanning 30,000 iterations for the Synapse dataset and 20,000 iterations for the ACDC dataset. We employed the AdamW [17] optimizer with an initial learning rate of 1×10^{-4} and a cosine learning rate decay strategy [16] with a linear warmup of 1500 iterations [7].

3.3 Performance on Synapse Dataset

To evaluate the effectiveness of the proposed **VM-NeXT UNet**, we conduct extensive experiments on the Synapse dataset. Table 1 presents the quantitative comparison, while Fig. 2 provides a qualitative visualization of the segmentation results.

As demonstrated in Table 1, VM-NeXT UNet achieves a state-of-the-art DSC of 86.21% and an HD95 of 16.68. Our model significantly outperforms the baseline Mamba-UNet by 3.83%. As shown in Fig. 2, our model successfully identifies the fine structures of the gallbladder and kidneys, which are often missed by previous methods.

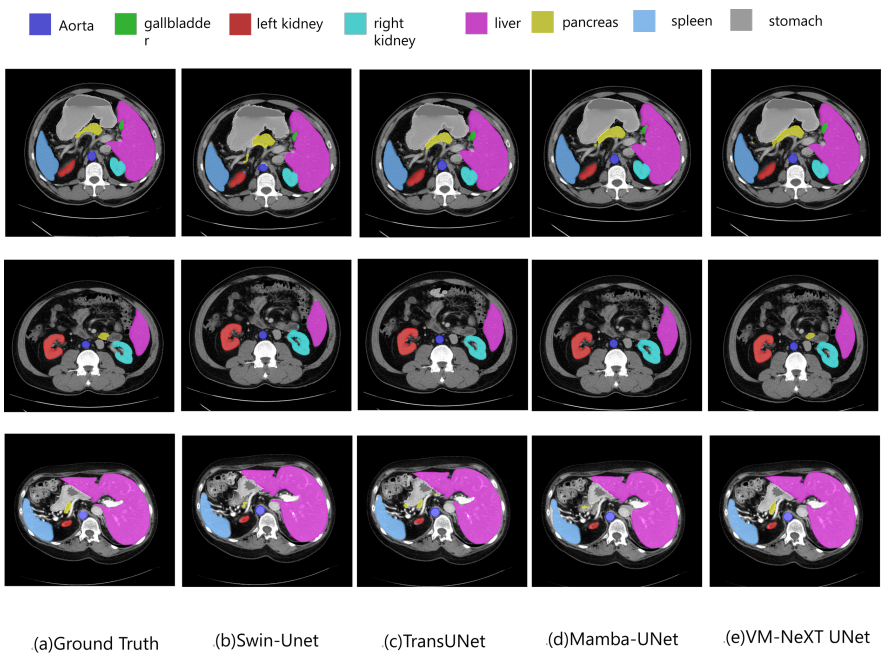

Fig. 2. Visual comparison of segmentation results on the Synapse dataset. (a) Ground Truth, (b) TransUNet, (c) Swin-Unet, (d) Mamba-UNet, and (e) our VM-NeXT UNet. Our model shows superior boundary delineation and reduces false positives in complex abdominal regions.

Table 1. Segmentation results on the Synapse dataset. The best results are in **bold**. All values except HD95 are in %. HD95 is in mm.

Methods	**DSC ↑**	**HD95 ↓**	Aor.	Gal.	Kid(L)	Kid(R)	Liv.	Pan.	Spl.	Sto.
UNet [21]	76.44	46.60	88.74	66.65	79.16	69.38	91.45	60.22	84.60	71.37
TransUNet [3]	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
Swin-Unet [2]	79.13	21.55	85.47	66.53	83.28	79.61	94.29	55.58	90.66	76.60
PVT-CASCADE [19]	81.06	20.23	83.01	70.59	82.23	80.37	94.08	64.43	90.10	83.69
Mamba-UNet [24]	82.38	18.89	87.27	69.00	88.14	82.93	95.29	64.32	88.85	83.20
TransCASCADE [19]	82.68	17.34	86.63	68.48	87.66	84.56	94.43	65.33	90.79	83.52
ReSeg-UNet [12]	83.22	15.73	89.62	71.21	88.02	81.36	**95.76**	63.36	90.61	85.82
ScaleFormer [11]	84.00	**12.30**	88.98	74.92	87.33	84.51	95.48	66.82	92.88	81.04
Parallel MERIT [20]	84.22	16.51	88.38	73.48	87.21	84.31	95.06	69.97	91.21	84.15
Cascaded MERIT [20]	84.90	13.22	87.71	74.40	87.79	84.85	95.26	**71.81**	92.01	85.38
VM-NeXT UNet	**86.21**	16.68	**90.94**	**76.72**	**90.44**	**86.11**	95.70	68.87	**93.55**	**87.36**

3.4 Performance on ACDC Dataset

The results on the ACDC dataset are summarized in Table 2. VM-NeXT UNet achieves an average DSC of 92.42% on the ACDC dataset, surpassing all compared methods.

Table 2. Segmentation results on the ACDC dataset. Values are DSC in %.

Methods	**Average DSC ↑**	RV	Myo	LV
TransUNet [3]	89.71	88.86	84.53	95.73
Swin-Unet [2]	90.00	88.55	85.62	95.83
Mamba-UNet [24]	91.18	89.40	88.49	95.64
Cascaded MERIT [20]	91.85	90.23	89.53	95.80
ReSeg-UNet [12]	91.85	90.30	89.58	95.67
ScaleFormer [11]	91.92	90.08	89.90	95.78
Parallel MERIT [20]	92.32	90.87	**90.00**	96.08
VM-NeXT UNet	**92.42**	**91.18**	89.98	**96.09**

3.5 Ablation Study

We conduct a detailed ablation study on the Synapse dataset to quantify the contribution of each proposed component. Specifically, starting from the Mamba-UNet baseline, we progressively integrate: (A) the Dual-Encoder Parallel structure (without interaction), (B) the Injection Block for feature interaction, (C) the Multi-Scale Enhanced Decoder, (D) the CSAG module in skip connections, and (E) the optimized FocalLoss and DiceLoss strategy. Note that all intermediate architectural ablations (Base to Base+A+B+C+D) adopt the standard Dice+CE Loss.

Table 3. Ablation study of VM-NeXT UNet on the Synapse dataset. Metrics except HD and ASD are in %.

Model	DSC ↑	IoU ↑	Acc ↑	Pre ↑	Sen ↑	HD (mm) ↓	ASD (mm) ↓
Base (Mamba-UNet [24])	82.38	73.22	99.91	83.30	81.79	18.89	5.70
Base + A	83.24	74.62	99.91	82.86	84.47	24.31	6.73
Base + A + B	83.81	74.92	99.91	84.64	82.68	18.52	5.31
Base + A + B + C	84.66	76.05	99.92	85.37	83.74	20.06	5.15
Base + A + B + C + D	85.61	77.39	99.93	85.78	84.95	17.60	4.64
Base+A+B+C+D+E	**86.21**	**78.25**	**99.93**	**86.48**	**85.45**	**16.68**	**4.36**

3.6 Model Efficiency

We evaluate the computational overhead of our dual-encoder on a single RTX 5090 GPU (224×224 input, batch size 1). VM-NeXT UNet requires 72.46M parameters and 16.51 GFLOPs with an inference latency of 21.76 ms (> 40 FPS), compared to Mamba-UNet's 35.86M parameters and 9.12 GFLOPs. This modest trade-off is highly justified by a 3.23% DSC gain (from 82.38% to 85.61%) derived purely from architectural optimizations, while comfortably satisfying the clinical real-time threshold (> 30 FPS).

4 Conclusion

In this paper, we introduced VM-NeXT UNet, a hybrid architecture that synergizes ConvNeXT with Visual Mamba for medical image segmentation. By leveraging a dual-encoder design, adaptive attention-gated skip connections, and a multi-scale decoder, the model successfully integrates local textural features with global context. Furthermore, to mitigate the impact of class imbalance and improve segmentation accuracy on ambiguous boundaries, we employed a composite loss function that balances Dice Loss and Focal Loss. Extensive experiments demonstrate its robustness and superiority across multiple datasets.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (No. 82400368), the Natural Science Basic Research Program of Shaanxi Province (No. 2024JC-YBQN-0872), and the IIT Clinical Research Fund of The Second Affiliated Hospital of Xi'an Jiaotong University (No. NO-Z03).

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

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(12), 2481–2495 (2017)
2. 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 (ECCV) Workshops. Lecture Notes in Computer Science, vol. 13803, pp. 205–218. Springer (2022)
3. Chen, J., Lu, Y., Yu, Q.E., 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)
4. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 801–818 (2018)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR). OpenReview.net (2021)
6. Fu, S., Lu, Y., Wang, Y.E., Zhou, Y.E., Shen, W., Fishman, E., Yuille, A.: Domain adaptive relational reasoning for 3D multi-organ segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 656–666. Springer (2020)
7. Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., He, K.: Accurate, large minibatch SGD: Training ImageNet in 1 hour. arXiv preprint arXiv:1706.02677 (2017)
8. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)

9. Huo, X., Sun, G., Tian, S., Wang, Y., Yu, L., Long, J., Zhang, W., Li, A.: HiFuse: Hierarchical multi-scale feature fusion network for medical image classification. Biomedical Signal Processing and Control **87**, 105534 (2024)
10. Jha, D., Smedsrud, P.H., Riegler, M.A., Johansen, D., de Lange, T., Halvorsen, P., Johansen, H.D.: ResUNet++: An advanced architecture for medical image segmentation. In: IEEE International Symposium on Multimedia (ISM). pp. 225–230. IEEE (2019)
11. Ji, Z., Chen, Z., Ma, X.: Grouped multi-scale vision transformer for medical image segmentation. Scientific Reports **15**(1), 11122 (2025)
12. Li, L., Tang, D., Chu, X., Yang, X., Yu, F.: ReSeg-UNet: A reconstruction-guided optimization framework for enhanced medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 544–554. Springer (2025)
13. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2980–2988 (2017)
14. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., Liu, Y.: VMamba: Visual state space model. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 37 (2024)
15. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11976–11986 (2022)
16. Loshchilov, I., Hutter, F.: SGDR: Stochastic gradient descent with warm restarts. In: International Conference on Learning Representations (ICLR). OpenReview.net (2017)
17. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR). OpenReview.net (2019)
18. 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)
19. 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 (WACV). pp. 6222–6231 (2023)
20. Rahman, M.M., Marculescu, R.: Multi-scale hierarchical vision transformer with cascaded attention decoding for medical image segmentation. In: Medical Imaging with Deep Learning. pp. 1526–1544. PMLR (2024)
21. 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 (MICCAI). pp. 234–241. Springer (2015)
22. Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M.: Medical transformer: Gated axial-attention for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 36–46. Springer (2021)
23. Wang, H., Xie, S., Lin, L., Iwamoto, Y., Han, X., Chen, Y., Tong, R.: Mixed transformer U-Net for medical image segmentation. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2390–2394. IEEE (2022)
24. 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)

25. 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)