VM-UNet: Vision Mamba UNet for Medical Image Segmentation

Jiacheng Ruan, Suncheng Xiang

Shanghai Jiao Tong University
jackchenruan@sjtu.edu.cn, xiangsuncheng17@sjtu.edu.cn

Abstract. In the realm of medical image segmentation, both CNN-based and Transformer-based models have been extensively explored. However, CNNs exhibit limitations in long-range modeling capabilities, whereas Transformers are hampered by their quadratic computational complexity. Recently, State Space Models (SSMs), exemplified by Mamba, have emerged as a promising approach. They not only excel in modeling long-range interactions but also maintain a linear computational complexity. In this paper, leveraging state space models, we propose a U-shape architecture model for medical image segmentation, named Vision Mamba UNet (VM-UNet). Specifically, the Visual State Space (VSS) block is introduced as the foundation block to capture extensive contextual information, and an asymmetrical encoder-decoder structure is constructed. We conduct comprehensive experiments on the ISIC17, ISIC18, and Synapse datasets, and the results indicate that VM-UNet performs competitively in medical image segmentation tasks. To our best knowledge, this is the first medical image segmentation model constructed based on the pure SSM-based model. We aim to establish a baseline and provide valuable insights for the future development of more efficient and effective SSM-based segmentation systems. Our code is available at https://github.com/JCruan519/VM-UNet.

Keywords: Medical image segmentation · State Space Models

1 Introduction

Automated medical image segmentation techniques assist physicians in faster pathological diagnosis, thereby improving the efficiency of patient care. Recently, CNN-based and Transformer-based models have demonstrated remarkable performance in various visual tasks, particularly in medical image segmentation. UNet [27], as a representative of CNN-based models, is known for its simplicity of structure and strong scalability, and many subsequent improvements are based on this U-shaped architecture [11,37,28,29,30]. TransUnet [10], a pioneer among Transformer-based models, is the first to employ Vision Transformer (ViT) [13] for feature extraction during the encoding phase and utilizes CNN in the decoding phase, demonstrating the significant capability for global information acquisition. Subsequently, TransFuse [36] incorporates a parallel architecture of ViT and CNN, capturing both local and global features simultaneously. Furthermore, Swin-UNet [9] combines Swin Transformer [21] with the U-shaped architecture, introducing a pure Transformer-based U-shaped model for the first time.

Nevertheless, both CNN-based models and Transformer-based models have inherent limitations. CNN-based models are constrained by their local receptive field, considerably hindering their ability to capture long-range information. This often leads to the extraction of inadequate features, resulting in suboptimal segmentation outcomes. Although Transformer-based models demonstrate superior performance for global modeling, the self-attention mechanism demands quadratic complexity in terms of image sizes, leading to a high computational burden [31,13], particularly for tasks requiring dense predictions like medical image segmentation. The current shortcomings in these models compel us to develop a novel architecture for medical image segmentation, capable of capturing strong long-range information and maintaining linear computational complexity.

Recently, State Space Models (SSMs) have attracted considerable interest among researchers. Building on the foundation of classical SSM [18] research, the modern SSMs (e.g., Mamba [16]) not only establish long-distance dependencies but also exhibit linear complexity with respect to input size. Additionally, SSM-based models have received substantial research across many fields, including language understanding [17,16], general vision [38,20], etc. Particularly, U-Mamba [24] has recently introduced a novel SSM-CNN hybrid model, marking its first application in medical image segmentation tasks. SegMamba [35] incorporates SSM in the encoder part, while still using CNN in the decoder part, suggesting a SSM-CNN hybrid model for 3D brain tumor segmentation tasks. Although aforementioned works have utilized SSM for medical image segmentation tasks, the performance of the pure SSM-based model has yet to be explored.

Influenced by the success of VMamba [20] in image classification tasks, this paper introduces the Vision Mamba UNet (VM-UNet) for the first time, a pure SSM-based model designed to showcase the potential in medical image segmentation tasks. Specifically, VM-UNet is composed of three main parts: the encoder, the decoder, and the skip connection. The encoder consists of VSS blocks from VMamba for feature extraction, along with patch merging operations for downsampling. Conversely, the decoder comprises VSS blocks and patch expanding operations to restore the size of the segmentation results. For the skip connection component, to highlight the segmentation performance of the most original pure SSM-based model, we adopt the simplest form of additive operation.

Comprehensive experiments are conducted on organ segmentation and skin lesion segmentation tasks to demonstrate the potential of pure SSM-based models in medical image segmentation. Specifically, we conduct extensive experiments on the Synapse [19], ISIC17 [8], and ISIC18 [12] datasets, the results of which indicate that VM-UNet can achieve competitive performance. Moreover, it is important to note that VM-UNet represents the most basic form of a pure SSM-based segmentation model, as it does not include any specially designed modules.

The main contributions of this paper can be summarized as follows: 1) We propose VM-UNet, marking the first occasion of exploring the potential applications of purely SSM-based models in medical image segmentation. 2) Comprehensive experiments are conducted on three datasets, with results indicating that VM-UNet exhibits considerable competitiveness. 3) We establish a baseline for pure SSM-based models in medical image segmentation tasks, providing valuable insights that pave the way for the development of more efficient and effective SSM-based segmentation methods.

2 Preliminaries

In modern SSM-based models, i.e., Structured State Space Sequence Models (S4) and Mamba, both rely on a classical continuous system that maps a one-dimensional input function or sequence, denoted as $x(t) \in \mathcal{R}$, through intermediate implicit states $h(t) \in \mathcal{R}^N$ to an output $y(t) \in \mathcal{R}$. The aforementioned process can be represented as a linear Ordinary Differential Equation (ODE):

$$\begin{aligned} h'(t) &= \mathbf{A}h(t) + \mathbf{B}x(t) \\ y(t) &= \mathbf{C}h(t) \end{aligned} \tag{1}$$

where $\mathbf{A} \in \mathcal{R}^{N \times N}$ represents the state matrix, while $\mathbf{B} \in \mathcal{R}^{N \times 1}$ and $\mathbf{C} \in \mathcal{R}^{N \times 1}$ denote the projection parameters.

S4 and Mamba discretize this continuous system to make it more suitable for deep learning scenarios. Specifically, they introduce a timescale parameter $\mathbf{\Delta}$ and transform $\mathbf{A}$ and $\mathbf{B}$ into discrete parameters $\overline{\mathbf{A}}$ and $\overline{\mathbf{B}}$ using a fixed discretization rule. Typically, the zero-order hold (ZOH) is employed as the discretization rule and can be defined as follows:

$$\begin{aligned} \overline{\mathbf{A}} &= \exp(\mathbf{\Delta A}) \\ \overline{\mathbf{B}} &= (\mathbf{\Delta A})^{-1}(\exp(\mathbf{\Delta A}) - \mathbf{I}) \cdot \mathbf{\Delta B} \end{aligned} \tag{2}$$

After discretization, SSM-based models can be computed in two ways: linear recurrence or global convolution, defined as equations 3 and 4, respectively.

$$\begin{aligned} h'(t) &= \overline{\mathbf{A}}h(t) + \overline{\mathbf{B}}x(t) \\ y(t) &= \mathbf{C}h(t) \end{aligned} \tag{3}$$

$$\begin{aligned} \overline{K} &= (\mathbf{C}\overline{\mathbf{B}}, \mathbf{C}\overline{\mathbf{AB}}, \ldots, \mathbf{C}\overline{\mathbf{A}}^{L-1}\overline{\mathbf{B}}) \\ y &= x * \overline{\mathbf{K}} \end{aligned} \tag{4}$$

where $\overline{\mathbf{K}} \in \mathcal{R}^L$ represents a structured convolutional kernel, and L denotes the length of the input sequence x.

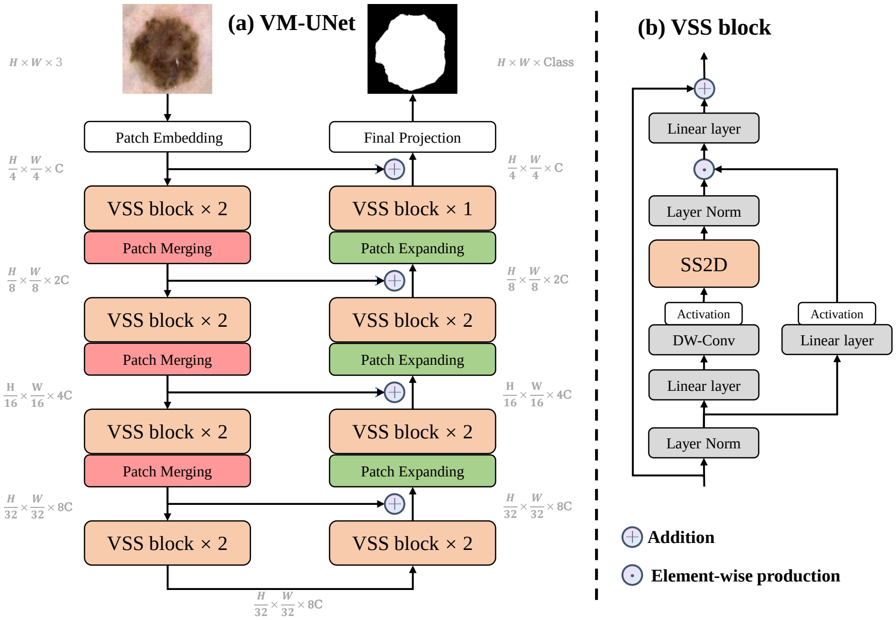

Fig. 1. (a) The overall architecture of VM-UNet. (b) VSS block is the main construction block of VM-UNet, and SS2D is the core operation in VSS block.

3 Methods

In this section, we initially introduce the overall structure of VM-UNet. Subsequently, we elaborate on the core component, the VSS block. Finally, we describe the loss function utilized during the training process.

3.1 Vision Mamba UNet (VM-UNet)

As depicted in Figure 1 (a), the overall architecture of VM-UNet is presented. Specifically, VM-UNet comprises a Patch Embedding layer, an encoder, a decoder, a Final Projection layer, and skip connections. Unlike previous methods [9], we have not adopted a symmetrical structure but instead utilized an asymmetric design.

The Patch Embedding layer divides the input image $x \in \mathcal{R}^{H \times W \times 3}$ into non-overlapping patches of size 4×4, subsequently mapping the dimensions of the image to C, with C defaulting to 96. This process results in the embedded image $x' \in \mathcal{R}^{\frac{H}{4} \times \frac{W}{4} \times C}$. Finally, we normalize x' using Layer Normalization [7] before feeding it into the encoder for feature extraction. The encoder is composed of four stages, with a patch merging operation applied at the end of first three stages to reduce the height and width of the input features while increasing the number of channels. We employ [2, 2, 2, 2] VSS blocks across four stages, with the channel counts for each stage being [C, 2C, 4C, 8C].

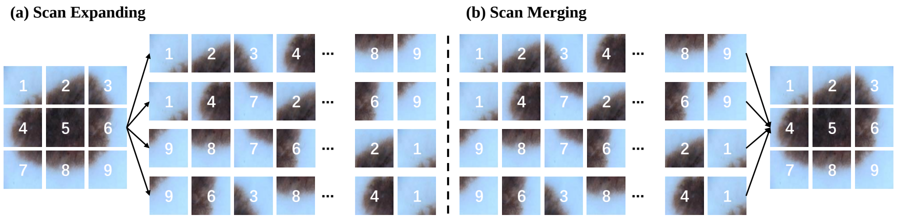

Fig. 2. (a) The scan expanding operation in SS2D. (b) The scan merging operation in SS2D.

Similarly, the decoder is organized into four stages. At the beginning of last three stages, a patch expanding operation is utilized to decrease the number of feature channels and increase the height and width. Across the four stages, we utilize [2, 2, 2, 1] VSS blocks, with the channel counts for each stage being [8C, 4C, 2C, C]. Following the decoder, a Final Projection layer is employed to restore the size of the features to match the segmentation target. Specifically, a 4-times upsampling is conducted via patch expanding to recover the height and width of the features, followed by a projection layer to restore the number of channels.

For the skip connections, a straightforward addition operation is adopted without bells and whistles, thereby not introducing any additional parameters.

3.2 VSS block

The VSS block derived from VMamaba [20] is the core module of VM-UNet, as depicted in Figure 1 (b). After undergoing Layer Normalization, the input is split into two branches. In the first branch, the input passes through a linear layer followed by an activation function. In the second branch, the input undergoes processing through a linear layer, depthwise separable convolution, and an activation function, before being fed into the 2D-Selective-Scan (SS2D) module for further feature extraction. Subsequently, the features are normalized using Layer Normalization, and then an element-wise production is performed with the output from the first branch to merge the two pathways. Finally, the features are mixed using a linear layer, and this outcome is combined with a residual connection to form the VSS block's output. In this paper, SiLU [14] is employed as the activation function by default.

The SS2D consists of three components: a scan expanding operation, an S6 block, and a scan merging operation. As shown in Figure 2(a), the scan expanding operation unfolds the input image along four different directions (top-left to bottom-right, bottom-right to top-left, top-right to bottom-left, and bottom-left to top-right) into sequences. These sequences are then processed by the S6 block for feature extraction, ensuring that information from various directions is thoroughly scanned, thus capturing diverse features. Subsequently, as illustrated in Figure 2(b), the scan merging operation sums and merges the sequences from

Algorithm 1 Pseudo-code for S6 block in SS2D

```
Input: $x$, the feature with shape [B, L, D] (batch size, token length, dimension)
Params: $\mathbf{A}$, the nn.Parameter; $\mathbf{D}$, the nn.Parameter
Operator: Linear(.), the linear projection layer
Output: $y$, the feature with shape [B, L, D]
$\mathbf{\Delta}, \mathbf{B}, \mathbf{C} = \text{Linear}(x), \text{Linear}(x), \text{Linear}(x)$
$\overline{\mathbf{A}} = \exp(\mathbf{\Delta A})$
$\overline{\mathbf{B}} = (\mathbf{\Delta A})^{-1}(\exp(\mathbf{\Delta A}) - \mathbf{I}) \cdot \mathbf{\Delta B}$
$h_t = \overline{\mathbf{A}} h_{t-1} + \overline{\mathbf{B}} x_t$
$y_t = \mathbf{C} h_t + \mathbf{D} x_t$
$y = [y_1, y_2, \cdots, y_t, \cdots, y_L]$
return $y$
```

the four directions, restoring the output image to the same size as the input. The S6 block, derived from Mamba [16], introduces a selective mechanism on top of S4 [17] by adjusting the SSM's parameters based on the input. This enables the model to distinguish and retain pertinent information while filtering out the irrelevant. The pseudo-code for the S6 block is presented in Algorithm 1.

3.3 Loss function

The introduction of VM-UNet is aimed at validating the application potential of pure SSM-based models in medical image segmentation tasks. Consequently, we exclusively utilize the most fundamental Binary Cross-Entropy and Dice loss (BceDice loss) and Cross-Entropy and Dice loss (CeDice loss) as the loss functions for binary and multi-class segmentation tasks, respectively, as denoted by Equations 5 and 6.

$$L_{\text{BceDice}} = \lambda_1 L_{\text{Bce}} + \lambda_2 L_{\text{Dice}} \tag{5}$$

$$L_{\text{CeDice}} = \lambda_1 L_{\text{Ce}} + \lambda_2 L_{\text{Dice}} \tag{6}$$

$$\begin{cases} L_{\text{Bce}} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \\ L_{\text{Ce}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log(\hat{y}_{i,c}) \\ L_{\text{Dice}} = 1 - \frac{2|X \cap Y|}{|X| + |Y|} \end{cases} \tag{7}$$

where, N denotes the total number of samples, and C represents the total number of categories. $y_i, \hat{y}_i$ respectively signify the true label and prediction. $y_{i,c}$ is an indicator that equals 1 if sample i belongs to category c, and 0 otherwise. $\hat{y}_{i,c}$ is the probability that the model predicts sample i as belonging to category c. $|X|$

and $|Y|$ represent the ground truth and prediction, respectively. λ_1, λ_2 refer to the weights of loss functions, which are both set to 1 by default.

4 Experiments

In this section, we conduct comprehensive experiments on VM-UNet for skin lesion and organ segmentation tasks. Specifically, we evaluate the performance of VM-UNet on medical image segmentation tasks on the ISIC17, ISIC18, and Synapse datasets.

4.1 Datasets

ISIC17 and ISIC18 datasets: The International Skin Imaging Collaboration 2017 and 2018 challenge datasets (ISIC17 and ISIC18) [8,1,12,2] are two publicly available skin lesion segmentation datasets, containing 2,150 and 2,694 dermoscopy images with segmentation mask labels, respectively. Following the previous work [28], we split the datasets in a 7:3 ratio for use as training and test sets. Specifically, for the ISIC17 dataset, the training set consists of 1,500 images, and the test set consists of 650 images. For the ISIC18 dataset, the training set includes 1,886 images, while the test set contains 808 images. For these two datasets, we provide detailed evaluations on several metrics, including Mean Intersection over Union (mIoU), Dice Similarity Coefficient (DSC), Accuracy (Acc), Sensitivity (Sen), and Specificity (Spe).

Synapse multi-organ segmentation dataset (Synapse): Synapse [19,3] is a publicly available multi-organ segmentation dataset comprising 30 abdominal CT cases with 3,779 axial abdominal clinical CT images, including 8 types of abdominal organs (aorta, gallbladder, left kidney, right kidney, liver, pancreas, spleen, stomach). Following the setting of previous works [10,9], 18 cases are used for training and 12 cases for testing. For this dataset, we report the Dice Similarity Coefficient (DSC) and the 95% Hausdorff Distance (HD95) as evaluation metrics.

4.2 Implementation details

Following the prior works [28,9], we resize the images in the ISIC17 and ISIC18 datasets to 256×256, and those in the Synapse dataset to 224×224. To prevent overfitting, data augmentation techniques, including random flip and random rotation, are employed. The BceDice loss function is utilized for the ISIC17 and ISIC18 datasets, while the CeDice loss function is adopted for the Synapse dataset. We set the batch size to 32 and employ AdamW [23] optimizer with an initial learning rate of 1e-3. CosineAnnealingLR [22] is utilized as the scheduler with a maximum of 50 iterations and a minimum learning rate of 1e-5. Training epochs are set to 300. For VM-UNet, we initialize the weights of both the encoder and decoder with those of VMamba-S [20], which is pre-trained on ImageNet-1k. All experiments are conducted on a single NVIDIA RTX A6000 GPU.

Table 1. Comparative experimental results on the ISIC17 and ISIC18 dataset. (**Bold** indicates the best.)

Dataset	**Model**	**mIoU(%)↑**	**DSC(%)↑**	**Acc(%)↑**	**Spe(%)↑**	**Sen(%)↑**
ISIC17	UNet [27]	76.98	86.99	95.65	97.43	86.82
	UTNetV2 [15]	77.35	87.23	95.84	98.05	84.85
	TransFuse [36]	79.21	88.40	96.17	97.98	87.14
	MALUNet [28]	78.78	88.13	96.18	**98.47**	84.78
	VM-UNet	**80.23**	**89.03**	**96.29**	97.58	**89.90**
ISIC18	UNet [27]	77.86	87.55	94.05	**96.69**	85.86
	UNet++ [37]	78.31	87.83	94.02	95.75	88.65
	Att-UNet [26]	78.43	87.91	94.13	96.23	87.60
	UTNetV2 [15]	78.97	88.25	94.32	96.48	87.60
	SANet [34]	79.52	88.59	94.39	95.97	89.46
	TransFuse [36]	80.63	89.27	94.66	95.74	**91.28**
	MALUNet [28]	80.25	89.04	94.62	96.19	89.74
	VM-UNet	**81.35**	**89.71**	**94.91**	96.13	91.12

Table 2. Comparative experimental results on the Synapse dataset. DSC of every single class is also reported. (**Bold** indicates the best.)

Model	**DSC↑**	**HD95↓**	**Aorta**	**Gallbladder**	**Kidney(L)**	**Kidney(R)**	**Liver**	**Pancreas**	**Spleen**	**Stomach**
V-Net [25]	68.81	-	75.34	51.87	77.10	80.75	87.84	40.05	80.56	56.98
DARR [4]	69.77	-	74.74	53.77	72.31	73.24	94.08	54.18	89.90	45.96
R50 U-Net [10]	74.68	36.87	87.47	66.36	80.60	78.19	93.74	56.90	85.87	74.16
UNet [27]	76.85	39.70	89.07	**69.72**	77.77	68.60	93.43	53.98	86.67	75.58
R50 Att-UNet [10]	75.57	36.97	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95
Att-UNet [26]	77.77	36.02	**89.55**	68.88	77.98	71.11	93.57	58.04	87.30	75.75
R50 ViT [10]	71.29	32.87	73.73	55.13	75.80	72.20	91.51	45.99	81.99	73.95
TransUnet [10]	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
TransNorm [5]	78.40	30.25	86.23	65.10	82.18	78.63	94.22	55.34	89.50	76.01
Swin U-Net [9]	79.13	21.55	85.47	66.53	83.28	79.61	**94.29**	56.58	**90.66**	76.60
TransDeepLab [6]	80.16	21.25	86.04	69.16	84.08	79.88	93.53	**61.19**	89.00	78.40
UCTransNet [32]	78.23	26.75	-	-	-	-	-	-	-	-
MT-UNet [33]	78.59	26.59	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81
MEW-UNet [30]	78.92	**16.44**	86.68	65.32	82.87	80.02	93.63	58.36	90.19	74.26
VM-UNet	**81.08**	19.21	86.40	69.41	**86.16**	**82.76**	94.17	58.80	89.51	**81.40**

4.3 Main results

We compare VM-UNet with some state-of-the-art models, presenting the experimental results in Table 1 and Table 2. For the ISIC17 and ISIC18 datasets, our VM-UNet outperforms other models in terms of the mIoU, DSC and Acc metrics. For the Synapse dataset, VM-UNet has also achieved competitive performance. For instance, our model surpasses Swin-UNet, which is the first pure Transformer-based model, by 1.95% and 2.34mm in DSC and HD95 metrics. The results demonstrate the superiority of the SSM-based model in medical image segmentation tasks.

Table 3. Ablation studies on Init. Weight of VM-UNet.

Init. Weight	ISIC17		ISIC18	
	mIoU(%)↑	DSC(%)↑	mIoU(%)↑	DSC(%)↑
-	77.59	87.38	78.66	88.06
VMamba-T	78.85	88.17	79.04	88.29
VMamba-S	80.23	89.03	81.35	89.71

4.4 Ablation studies

In this section, we conduct ablation experiments on the initialization of VM-UNet using the ISIC17 and ISIC18 datasets. We initialize VM-UNet with the pretrained weights from VMamba-T and VMamba-S, respectively[1] The experimental results, as shown in Table 3, reveal that more potent pretrained weights significantly enhance the downstream performance of VM-UNet, indicating that VM-UNet is substantially influenced by the pretrained weights.

5 Conclusions and Future works

Conclusions: In this paper, we introduce for the first time a pure SSM-based model for medical image segmentation, presenting VM-UNet as a baseline. To leverage the capabilities of SSM-based models, we construct VM-UNet using VSS blocks and initialize its weights with the pretrained VMamba-S. Comprehensive experiments are conducted on skin lesion and multi-organ segmentation datasets indicate that pure SSM-based models are highly competitive in medical image segmentation tasks and merit in-depth exploration in the future.

Future works: 1) Design modules that are better suited for segmentation tasks, based on the mechanisms of SSMs. 2) The parameter count of VM-UNet is approximately 30M, providing opportunities to streamline SSMs via manual design or other compression strategies, thereby fortifying their applicability in real-world medical scenarios. 3) Given the advantages of SSMs in capturing information in long sequences, it would be valuable to investigate further the segmentation performance at higher resolutions. 4) Explore the application of SSMs in other medical imaging tasks, such as detection, registration, and reconstruction, etc.

References

1. `https://challenge.isic-archive.com/data/#2017`
2. `https://challenge.isic-archive.com/data/#2018`
3. `https://www.synapse.org/#!Synapse:syn3193805/wiki/217789`

[1] VMamba-T and VMamba-S achieved Top-1 accuracy of 82.2% and 83.5% on ImageNet-1k, respectively.

4. Alom, M.Z., Yakopcic, C., Hasan, M., Taha, T.M., Asari, V.K.: Recurrent residual u-net for medical image segmentation. Journal of Medical Imaging **6**(1), 014006–014006 (2019)
5. Azad, R., Al-Antary, M.T., Heidari, M., Merhof, D.: Transnorm: Transformer provides a strong spatial normalization mechanism for a deep segmentation model. IEEE Access **10**, 108205–108215 (2022)
6. Azad, R., Heidari, M., Shariatnia, M., Aghdam, E.K., Karimijafarbigloo, S., Adeli, E., Merhof, D.: Transdeeplab: Convolution-free transformer-based deeplab v3+ for medical image segmentation. In: International Workshop on PRedictive Intelligence In MEdicine. pp. 91–102. Springer (2022)
7. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
8. Berseth, M.: Isic 2017-skin lesion analysis towards melanoma detection. arXiv preprint arXiv:1703.00523 (2017)
9. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537 (2021)
10. 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)
11. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. pp. 424–432. Springer (2016)
12. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368 (2019)
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
14. Elfwing, S., Uchibe, E., Doya, K.: Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. Neural networks **107**, 3–11 (2018)
15. Gao, Y., Zhou, M., Liu, D., Metaxas, D.: A multi-scale transformer for medical image segmentation: Architectures, model efficiency, and benchmarks. arXiv preprint arXiv:2203.00131 (2022)
16. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
17. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. arXiv preprint arXiv:2111.00396 (2021)
18. Kalman, R.E.: A new approach to linear filtering and prediction problems (1960)
19. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In: Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge. vol. 5, p. 12 (2015)
20. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Liu, Y.: Vmamba: Visual state space model. arXiv preprint arXiv:2401.10166 (2024)

21. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
22. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
23. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
24. Ma, J., Li, F., Wang, B.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722 (2024)
25. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. IEEE (2016)
26. 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)
27. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
28. Ruan, J., Xiang, S., Xie, M., Liu, T., Fu, Y.: Malunet: A multi-attention and light-weight unet for skin lesion segmentation. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 1150–1156. IEEE (2022)
29. Ruan, J., Xie, M., Gao, J., Liu, T., Fu, Y.: Ege-unet: an efficient group enhanced unet for skin lesion segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 481–490. Springer (2023)
30. Ruan, J., Xie, M., Xiang, S., Liu, T., Fu, Y.: Mew-unet: Multi-axis representation learning in frequency domain for medical image segmentation. arXiv preprint arXiv:2210.14007 (2022)
31. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
32. Wang, H., Cao, P., Wang, J., Zaiane, O.R.: Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2441–2449 (2022)
33. Wang, H., Xie, S., Lin, L., Iwamoto, Y., Han, X.H., Chen, Y.W., Tong, R.: Mixed transformer u-net for medical image segmentation. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2390–2394. IEEE (2022)
34. Wei, J., Hu, Y., Zhang, R., Li, Z., Zhou, S.K., Cui, S.: Shallow attention network for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 699–708. Springer (2021)
35. Xing, Z., Ye, T., Yang, Y., Liu, G., Zhu, L.: Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. arXiv preprint arXiv:2401.13560 (2024)
36. Zhang, Y., Liu, H., Hu, Q.: Transfuse: Fusing transformers and cnns for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 14–24. Springer (2021)
37. Zhou, Z., Rahman Siddiquee, M.M., 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, pp. 3–11. Springer (2018)

38. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. arXiv preprint arXiv:2401.09417 (2024)