Displays 85 (2024) 102890

Contents lists available at ScienceDirect

Displays

journal homepage: www.elsevier.com/locate/displa

Mambav3d: A mamba-based virtual 3D module stringing semantic information between layers of medical image slices☆

Xiaoxiao Liu, Yan Zhao *, Shigang Wang, Jian Wei

School of Communication Engineering, Jilin University, Changchun 130012, China

ARTICLE INFO

Keywords:
Medical image segmentation
Mamba
Vision state space models
Unet
Deep learning

ABSTRACT

High-precision medical image segmentation provides a reliable basis for clinical analysis and diagnosis. Researchers have developed various models to enhance the segmentation performance of medical images. Among these methods, two-dimensional models such as Unet exhibit a simple structure, low computational resource requirements, and strong local feature capture capabilities. However, their spatial information utilization is insufficient, limiting their segmentation accuracy. Three-dimensional models, such as 3D Unet, utilize spatial information more fully and are suitable for complex tasks, but they require high computational resources and have limited real-time performance. In this paper, we propose a virtual 3D module (Mambav3d) based on mamba, which introduces spatial information into 2D segmentation tasks to more fully integrate the 3D information of the image and further improve segmentation accuracy under conditions of low computational resource requirements. Mambav3d leverages the properties of hidden states in the state space model, combined with the shift of visual perspective, to incorporate semantic information between different anatomical planes in different slices of the same 3D sample. The voxel segmentation is converted to pixel segmentation to reduce model training data requirements and model complexity while ensuring that the model integrates 3D information and enhances segmentation accuracy. The model references the information from previous layers when labeling the current layer, thereby facilitating the transfer of semantic information between slice layers and avoiding the high computational cost associated with using structures such as Transformers between layers. We have implemented Mambav3d on Unet and evaluated its performance on the BraTs, Amos, and KiTs datasets, demonstrating superiority over other state-of-the-art methods.

1. Introduction

Medical image segmentation is a pivotal technology in medical image analysis, crucial for the accurate diagnosis and treatment of diseases [1]. More accurate and high-precision medical image segmentation methods are in demand in clinical practice, directly impacting the life and health of patients [2]. As discussed in various surveys and studies [3–7], high-quality images are essential for the successful implementation of these intelligent systems, whereas low-quality images may degrade their performance. However, medical images exhibit characteristics such as ambiguity, complexity, and variability, which pose challenges for accurate segmentation [8,9]. Owing to the significant success of deep learning and its extensive research and application in medical image segmentation tasks, it has become a popular research focus in recent years to further enhance the accuracy of target segmentation for medical image tasks [10].

The convolutional neural network (CNN) is a classical model for deep learning to process images [11]. Ronneberger et al. [12] proposed Unet, which combines low-resolution features with high-resolution features through skip connections between the encoder and decoder, thereby enhancing segmentation capability and achieving a significant breakthrough in medical image segmentation tasks. Unet has since become one of the most successful applications of CNN in medical image tasks. Consequently, numerous Unet-based methods [13–18] have been developed to address the growing demands in clinical practice. With the deepening of research, CNN-based medical image segmentation methods are also becoming more mature. However, constrained by the inherent locality of convolutional operations, there are limitations in capturing global contextual relationships. The attention mechanism significantly enhances the model's ability to capture global contextual

☆ This paper was recommended for publication by Prof Guangtao Zhai.

* Corresponding author.
E-mail address: zhao_y@jlu.edu.cn (Y. Zhao).

https://doi.org/10.1016/j.displa.2024.102890
Received 31 July 2024; Received in revised form 8 November 2024; Accepted 11 November 2024
Available online 15 November 2024

relationships by dynamically focusing on different parts of the input data, while also improving the model's expressive power and performance [19–21]. The advantages of the attention mechanism-based Transformer in long-range dependency tasks continue to be explored; its multi-head self-attention and multilayer perceptron are used to capture long-range semantic information [22]. Researchers have combined Transformer with CNN, attempting to exploit their complementary nature to enable the model to capture both global contextual semantics and local features, thereby improving segmentation accuracy. Chen et al. [23] proposed TransUNet, which encodes labeled patches in CNN feature maps as input sequences, thus extracting global contextual relations and recovering local spatial information to enhance details. A common feature of TransUNet and most of its variants [24–32] is to consider the convolutional network as the primary subject and further apply transformers on its basis to capture long-range contextual relationships. From the practical results of these studies, Transformer and CNN, as foundational medical image methods, have performed very well on 2D medical image segmentation tasks. However, for certain scenarios, they are inferior to 3D network models. For instance, in the ground truth of some single slices, the target is very small and cannot be recognized or has incorrect annotations. If three-dimensional information is introduced at this point and rich contextual connections between spatial slices are utilized, these problems can be effectively addressed, and segmentation performance can be improved.

Therefore, researchers began to enrich spatial information for the models. 3D U-net [33] and V-net [34] were initially investigated for medical image volume segmentation tasks. Subsequently, an increasing number of models were proposed to address 3D medical image segmentation tasks [35–38]. Through these studies, it has been observed that the accuracy performance of 3D networks is actually decreased compared to most 2D networks. On one hand, this is due to the introduction of more parameters into the 3D network itself. On the other hand, hardware constraints necessitate that the weights assigned to each voxel in the 3D model within the neural network are inevitably much smaller than those assigned to each pixel in the 2D network. Additionally, the sample size of the dataset is limited. When the weight of each voxel in the 3D network is significantly smaller than that of each pixel in the 2D network, and the total number of parameters in the model is much larger than that of the 2D network, this may result in each parameter in the 3D network not being adequately trained.

The existing commonly used medical image datasets are predominantly composed of a series of 2D slices, such as the BraTs dataset [39,40], the KiTs dataset [41], and the AMOS dataset [42], which contain a substantial amount of 3D information. Compared with 2D methods, 3D methods focus more on the correlation between slices and provide additional spatial information for image segmentation. Consequently, several studies have proposed the use of 2.5D [43,44] and 2D multi-view [45,46] methods to address this issue. The 2.5D methods treat adjacent slices as different channels of the input image, enabling the capture of inter-slice information within a small area. The 2D multi-view methods split the voxel data along three axes and then process the slices individually. Although these methods supplement some small-scale inter-slice information, there is still significant room for improvement.

To integrate more 3D information into the model segmentation process with lower computational resource requirements, thereby supplementing spatial slice information and further enhancing segmentation performance, this paper investigates a semantic extraction module called Mambav3d for extracting inter-slice information. It can be incorporated into Unet (Mambav3d-Unet) and its variants to enrich the model's understanding of inter-slice semantic information, thus improving segmentation results. By shifting the visual perspective, we introduce a virtual-3D concept, fully leveraging the property that the hidden state in Mamba pays attention to the previous state when utilized, and incorporating some memories from other layers on the same sample into the current layer. The primary segmentation tasks of the model remain centered on the 2D Unet, but in the connections between spatial slices, we introduce some bias on the interrelationships between anatomical surfaces at different locations. The closer the slices are, the deeper the memories they carry, guiding the model to refer to the overview of information encountered in the previous layers simultaneously to complete the annotation of the current layer. The essence is to achieve a virtual-3D effect by using bidirectional time series instead of employing a 3D network with a large number of parameters. Mambav3d is designed to be easy to train, avoiding the use of RNN or Transformer structures between layers of different heights. The Mambav3d modules are incorporated into Unet, and the model performance is validated on three datasets, yielding results that are more competitive than the state-of-the-art.

In summary, our contributions can be summarized as follows:

- For the first time, the concept of Mamba hidden state is introduced into medical image segmentation models to enrich the three-dimensional information between slices. By guiding the network to pay attention to the previous state and summarizing all the information encountered before, the model can achieve inter-layer information association to enrich semantics and improve segmentation performance.
- The Mambav3d module is proposed, which can be incorporated into Unet and Swin Unet. Compared with 3D network models, it introduces more 3D information to the model through visual perspective shift without using a large number of parameters, thereby improving the target segmentation effect within the slice.
- The proposed Mambav3d module is integrated into Unet (Mambav3d-Unet), and competitive results are obtained on the BraTs, KiTs, and AMOS datasets.
- The hyperparameter settings of the Mambav3d module are discussed on the BraTs dataset to achieve the best segmentation results.

2. Related work

2.1. Unet-based methods

In recent years, with the proposal and application of U-net [12], Unet-based methods have been widely used in medical image segmentation, and numerous variants have been developed. Zhou et al. [13] proposed U-net++, which redesigns the skip connections and aggregates semantic features at different scales. Oktay et al. [15] introduced Attention-Unet, which can focus on target structures of varying shapes and sizes to achieve automatic learning. Isensee et al. [17] simplified the model and focused on its performance and generalizability. Chen et al. [47] proposed MU-Net, a novel multi-path up-sampling convolutional network designed to retain more high-level information. Johansen et al. [18] proposed DoubleU-Net, incorporating an additional Unet to capture more semantic information.

Although these methods have achieved some success in the field of medical image segmentation, they still have shortcomings in capturing global semantic information due to the limitations of CNNs. Transformer has emerged as a widely used coping strategy for this problem. Wang et al. [48] proposed TransBTS, which uses CNN to extract feature maps, which are then labeled as tokens and input into the Transformer for global feature modeling. Cao et al. [32] introduced Swin-Unet, which employs a hierarchical Swin Transformer and Unet architecture to extract contextual features. The nnFormer [28] features an interleaved architecture based on self-attention and convolutional operations. Li et al. [29] proposed ScribFormer, aiming to combine the local features learned by CNN with the global semantics obtained by Transformer. Considering that cross-modal fusion strategies can construct models with richer features [49–51], some studies have also attempted to use Transformers to build frameworks for cross-modal segmentation methods [30], enabling the learning of long-range contextual relationships. The UCFilTransNet [31] was proposed to fuse multi-scale feature

information extracted from encoders, enhancing features while establishing long-distance contextual relationships.

Medical image segmentation methods for 2D are gradually becoming mature. To further enhance the segmentation effect, some researchers have been deeply investigating the assessment and optimization of medical image quality [52–54], while others are attempting to incorporate 3D information from the dataset into the 2D segmentation model to better enrich the model's semantic information. Xiong et al. [43] proposed the weight box fusion algorithm based on adaptive 3D CNN and applied it to a 2.5D lung nodule detection network. Li et al. [55] introduced H-DenseUNet, which combines 2D DenseUNet for extracting intra-slice features with 3D counterparts for aggregating volume context, aiming to learn both intra-slice information and inter-slice features simultaneously. Yang et al. [45] proposed axial-coronal stochastic convolutions to perform 3D representation learning. Ding et al. [46] proposed a multi-view dynamic fusion model to improve the performance of brain tumor segmentation. Although these methods to some extent supplement the information between adjacent slices, there is still significant room for improvement in fully integrating 3D information into 2D segmentation tasks.

2.2. Mamba-based methods

The state space model (SSM) is a mathematical model used to describe and analyze the behavior of dynamic systems. In the field of deep learning, mapping sequence data to the state space allows for better capture of long-distance contextual relationships in data. Mamba [56] is a novel SSM that incorporates time-varying parameters into the SSM and relies on hardware-aware algorithms for efficient training and inference. Vision Mamba [57] first applied SSM to the general backbone in vision, addressing the issues of unidirectionality and lack of position identification in Mamba modeling. Due to its excellent computational efficiency and long-range context modeling capability, many Mamba-based methods have been rapidly proposed for medical image analysis tasks.

U-Mamba [58] was proposed to combine U-net with Mamba, leveraging the advantages of CNN in extracting local features and SSM in capturing global information. Its performance surpasses that of state-of-the-art segmentation networks based on CNN and Transformer in medical image tasks. VM Unet [59], Mamba Unet [60], LightM Unet [61], Mamba Unet [62], and LKM Unet [63] all share similar core ideas. They achieve good performance in medical image segmentation tasks by constructing an encoder-decoder structure to extract local features, while introducing a visual state space (VSS) module to capture long-distance contextual information. Liu et al. [64] proposed Swin-Umamba, demonstrating that the pre-trained network of Mamba helps to improve medical image segmentation. Xie et al. [65] introduced ProMamba, which combines Vision Mamba and prompt techniques, exhibiting good generalization ability. Ye et al. [66] proposed P-Mamba, which combines the noise suppression and local feature extraction capabilities of Perona Malik Diffusion with Mamba's ability to capture long-range contextual relationships, thereby reducing the impact of noise on segmentation performance in echocardiography.

Almost all existing medical image research based on Mamba adopts an encoder structure similar to Transformer, attempting to replace Transformer with its current scaling characteristics in 2D tasks to improve segmentation performance. However, based on the hidden states and bidirectional time series characteristics of Mamba, we propose a method that can be applied to spatial information association by shifting the visual perspective. The reasonable design of the model can solve the problem of spatial information association. This is because the Mamba network pays attention to the previous state when in use, it can summarize all the information encountered before, and the closer the distance, the more profound the memory carried. Mamba is also easy to train, and it can avoid the use of RNN or Transformer between different layers in space.

3. Method

3.1. Overall architecture

U-Net is a highly successful image segmentation model, and its design philosophy and performance have made it the preferred choice for many image segmentation tasks. The structure of U-Net is similar to the encoder-decoder architecture, but it is unique in that it contains 'U' shaped skip connections that pass the feature maps from the encoder stage directly to the decoder stage. This helps to preserve the detailed information of the image and is essential for accurate image segmentation. The encoder is a typical convolutional neural network containing multiple convolutional and pooling layers to capture the contextual information of the image. Subsequently, the decoder passes high-resolution features through upsampling operations and skip connections, achieving accurate localization. The traditional U-net performs well in 2D medical image segmentation tasks and has a very large number of improved variants. We choose two typical Unet-based networks, U-Net and Swin Unet, and add Mambav3d to achieve the incorporation of interlayer information between different slices of the same sample in the 2D segmentation task, thereby improving the segmentation effectiveness. The structure is shown in Fig. 1.

3.2. Mamba

Mamba is a state space model (SSM) designed with a simple selection mechanism that allows the model to selectively process information by 'parameterizing the inputs to the SSM'. The model is capable of filtering out irrelevant information and retaining relevant information over the long term. In general terms, Mamba refers to a generalization of all previous content each time, with the further it progresses, the more concise the generalization of the previous content becomes, dropping details and retaining the general idea. In terms of long-range contextual association capability, it is more efficient than Transformer and more effective than RNN. The following is a brief description of its key concepts.

The state space model is defined by Eq. (1). It maps a one-dimensional input signal $x(t)$ to a hidden state $h(t)$, which is then projected to a one-dimensional output signal $y(t)$.

$$\begin{aligned} h'(t) &= Ah(t) + Bx(t) \\ y(t) &= Ch(t) + Dx(t) \end{aligned} \tag{1}$$

SSM is widely used in many subject areas, and its role in deep sequence models is simply represented in Fig. 2, where A, B, C, D are the parameters learnt by gradient descent method, $A \in R^{N\times N}$, $B \in R^{N\times 1}$, $C \in R^{1\times N}$, D can be ignored.

To address the problem that SSM may suffer from exponential scaling of the gradient over the length of the sequence, Hippo theory [67] is applied to Eq. (1), allowing the state $x(t)$ to remember the history of the input $h(t)$. Hippo specifies a particular class of matrix $A \in R^{N\times N}$.

$$A_{nk} = -\begin{cases} (2n+1)^{1/2}(2k+1)^{1/2} & \text{if } n > k \\ n+1 & \text{if } n = k \\ 0 & \text{if } n < k \end{cases} \tag{2}$$

In order to process the discrete input sequence $x = (x_0, x_1, \cdots) \in R^L$, the structured state spaces [68] discretize these parameters in Eq. (1) using a step Δ, this can be considered as the resolution of the continuous input $x(t)$. The continuous parameters A and B are converted to discrete parameters $\overline{A}$ and $\overline{B}$ by means of a zero-order holder (ZOH), which is defined as:

$$\begin{aligned} \overline{A} &= \exp(\Delta A) \\ \overline{B} &= (\Delta A)^{-1}(\exp(\Delta A)\text{-I}) \bullet \Delta B \end{aligned} \tag{3}$$

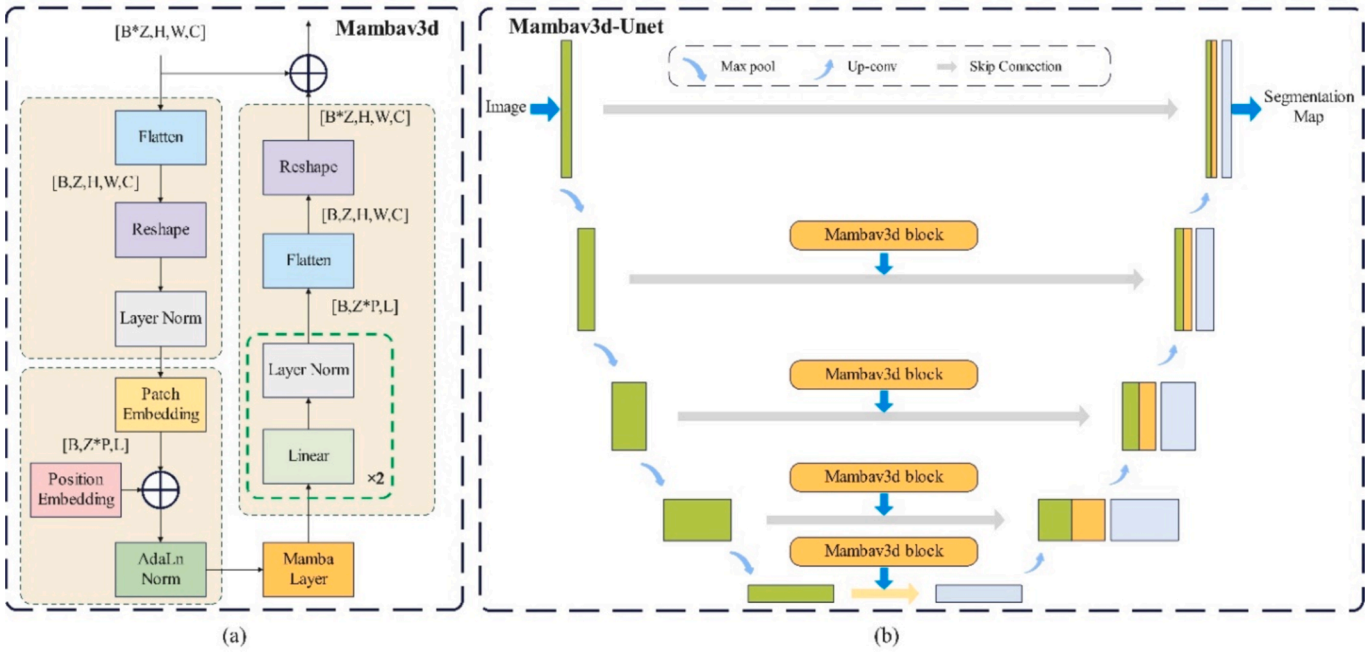

Fig. 1. Structural schematic. (a) The Mambav3d structure with two visual perspective shifts and an SSM-based Mamba layer for capturing semantic information between sliced layers. (b) Overview of the Mambav3d-Unet architecture.

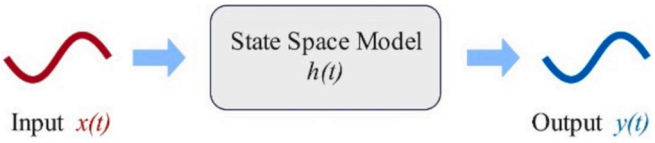

Fig. 2. Illustration of the state space model.

After discretizing A, B into $\overline{A}$, $\overline{B}$, Eq. (1) can be reformulated as:

$$h'(t) = \overline{A}h(t) + \overline{B}x(t)$$

$$y(t) = Ch(t) \tag{4}$$

The SSM can then be efficiently computed using RNN. This recursive process can also be reformulated and computed as a convolution:

$$\overline{K} = \left(C\overline{B}, C\overline{AB}, \cdots, C\overline{A}^{L-1}\overline{B}\right)$$

$$y = x * \overline{K} \tag{5}$$

where L is the length of the input sequence x and $\overline{K} \in R^L$ is the convolutional kernel of SSM.

Mamba is a simplified SSM architecture that inherits the linear scalability of the sequence length in state-space models, combines the advantages of CNN and Transformer, making it a promising base model for computer vision. The parameters in the SSM, denoted by Eq. (1), Eq. (4) and Eq. (5) remain invariant with respect to the input or time dynamics. This linear time-invariant property is a fundamental limitation of SSM when it comes to context-based inference [56]. Mamba sets the parameters of the input SSM function through a selection mechanism that achieves input dependency interaction along the sequence. Among them, parameters B, C and Δ depend on the input sequence x:

$$B,\ C,\ \Delta = \text{Linear}(x) \tag{6}$$

Mamba uses hardware-aware technology for computation and is capable of parallel scanning to improve efficiency and reduce cost. In terms of the architecture, Mamba differs from the commonly used SSM architecture in that it integrates a linear class of attention blocks and a multilayer perceptron to construct the Mamba blocks, as shown in Fig. 3.

First, an image of shape ($B \times Z, P, L$) is fed into the Mamba layer, where $L = d^2 \times C$, d is the patch size and C is the number of channels. After passing the Layer Normalization, the features enter the Mamba block containing two parallel branches. In the first branch, the features are extended to ($B \times Z, P, 2L$) through linear layers, and then sequentially through one-dimensional convolutional layers and SiLU activation functions, and together with the SSM layer. In the second branch, the features are also extended to ($B \times Z, P, 2L$), and then through SiLU activation functions. After that, the features from the two branches are merged with the Hadamard product. Finally, the features are projected back to the original shape ($B \times Z, P, L$).

3.3. Mambav3d module

The purpose of the Mambav3d module is to add contextual semantic information between layers of medical image slices for the Unet model. To achieve this goal, we design the Mambav3d module as shown in Fig. 1 and load it into the skip connections and bottleneck part of the U-shaped structure. The images are first divided into patches, embedded, and positionally encoded after shifting the visual perspective. The encoded features are then normalized by a linear layer and fed into the Mamba layer for parallel processing, which extracts and remembers the previous inter-layer correlation information, guiding the target segmentation in the current layer. The output layer of the Mambav3d module consists of Layer Normalization.

The specific process is described as follows. The model introduces a random monotonically increasing or decreasing sequence that satisfies the distribution of $(Z_1, Z_2, \cdots, Z_n)$, where Z_n is in $[0, Z_{max}]$ and does not repeat. The model introduces a relative position variable $\sin(\pi/2 * Z/Z_{max})$, where Z_{max} is the total number of slices in the coronal or sagittal image of a certain raw sample in the dataset, and Z is the ordering of slices in the sample. Thus, a dataset with dimensions [*B, Z, C, W, H*] is obtained, where B is the batch size, Z is the monotonic sequence, denotes the number of layers in different slices of the same sample, C is the number of channels, and W and H are the width and height of the image. As shown in Fig. 4, the image set with size [*B*Z, W, H, C*] is divided into patches and shifted the visual perspective to [*B,*

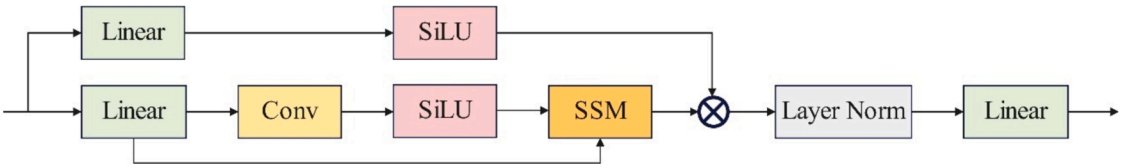

Fig. 3. The structure of Mamba.

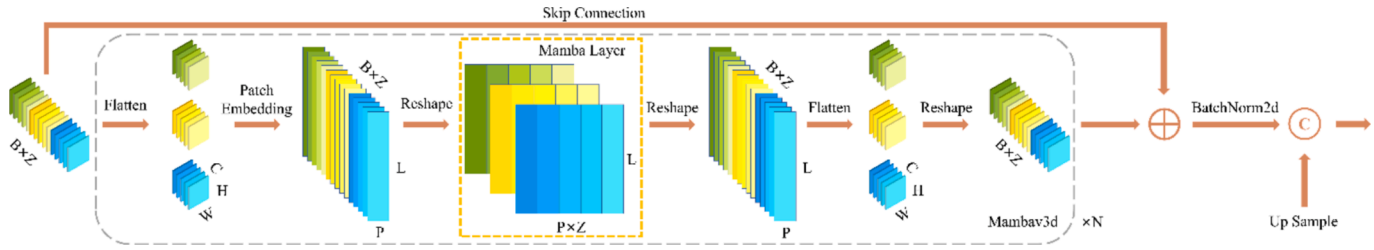

Fig. 4. The schematic diagram of the visual perspective shift process.

*Z*P, L*], where *P* represents the number of slices on the 2D image and *L* represents the number of parameters in the hidden layer. The features in the encoder are normalized by adaptive layer normalization after transforming the viewpoint and being added with positional information. The mamba layer captures interlayer information for slices. There are two linear normalizations in the output layer to correct the size, while shifting the visual perspective to [*B*Z, P, L*], which is [*B*Z, P,* d^2∗C], where *d* is the patch size. Finally, the patch is unpatchified and visual perspective transformed to return [*B*Z, W, H, C*]. The obtained interlayer features are concatenated with the input information of mambav3d to obtain richer image features containing interlayer information. The code for this process is detailed in Appendix.

3.4. Skip connection and bottle neck

In order to fully integrate the interlayer semantic information into the 2D segmentation tasks and introduce the memory of the previous layers into the current image to be segmented, we load the Mambav3d module in the skip connections and bottle neck part of the Unet model, as shown in Fig. 5. The model introduces N Mambav3d blocks in parallel to fully extract inter layer information of the image and concatenate them. Finally, it is concatenated with the features passed by the skip connections after batchnorm2d. Thus, a Z-space axial correlation bias is introduced to the image, which can be interpreted as the contextual semantic information between different layers in the same sample.

4. Experiments

4.1. Datasets and preprocessing

In order to objectively evaluate the effectiveness of loading the Mambav3d module into the Unet model, we conducted experiments on three medical image datasets: BraTs [39,40], AMOS [41], and KiTs dataset [42]. A brief description of these datasets is as follows. The BraTs2019 dataset provides 368 cases, including brain MRI images of four modalities: T1, T1ce, T2, and FLAIR. The BraTs2023 dataset provides 1,000 cases, including four modalities: T1c, T1n, T2f, and T2w. AMOS is an MRI dataset of abdominal organ segmentation including 565 slices from 60 cases. KiTs is a CT dataset of renal tumors including 599 cases. The dataset information is shown in Table 1. We consider that the relationship between layers is slightly different from the usual 3D or 2D dataset sampling methods. So according to the introduction in section 3.3, we correspond to take three datasets with dimensions of [*B, Z, C, W, H*]. In this case, *B* is the batch size, *Z* represents the number of slicing layers, *C* is the number of channels, and *W* and *H* are the width and height of the image. In order to validate the superiority of the proposed method, the experiments compare it with eight advanced and representative medical image segmentation methods, including Unet [12], Attention Unet [15], Swin Unet [32], ViT [22], SegDiff [69], MedsegDiff [70], Umamba [58], Swin Umamba [64]. To verify that the Mambav3d module can be loaded into the Unet-based model, the experiments are conducted on Unet [12] and Swin Unet [32], respectively.

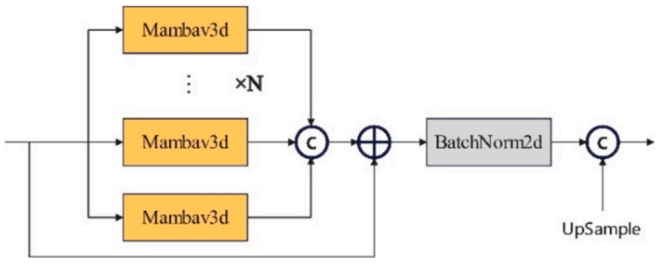

Fig. 5. The structure of the Mambav3d blocks on the skip connections and bottle neck.

4.2. Evaluation metrics

For quantitatively evaluating the performance of the proposed method, three widely used medical image segmentation task metrics are selected as evaluation criteria for the experiments, specifically: the Dice Similarity Coefficient (Dice), the Intersection over Union (IoU), and the Hausdorff distance (HD95). The specific definitions are as follows.

$$\text{Dice} = \frac{2\text{TP}}{2\text{TP} + \text{FN} + \text{FP}} \tag{7}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}} \tag{8}$$

Table 1
Dataset information.

Dataset	Modality	Training set	Testing set
BraTs2019[39]	MRI	320	48
BraTs2023[40]	MRI	900	100
AMOS2022[41]	CT/MRI	50	10
KiTs19[42]	CT	300	70

$$HD95 = \max_{95\%}(h(G,P), h(P,G)) \quad (9)$$

where *TP*, *TN*, *FP*, and *FN* denote the number of true positives, true negatives, false positives, and false negatives, respectively, G is the contour point corresponding to the ground truth, and P is the contour point corresponding to the prediction. Dice and IoU reflect the similarity between the prediction and the ground truth, and HD95 is the segmentation accuracy assessment of the quantitative metrics.

4.3. Implementation details and training process

Mambav3d-Unet is implemented based on the PyTorch framework and trained using two RTX4090 GPUs with 24 g of memory. The system used is win11 and the chip used is Intel Core I7-12700. The dataset is

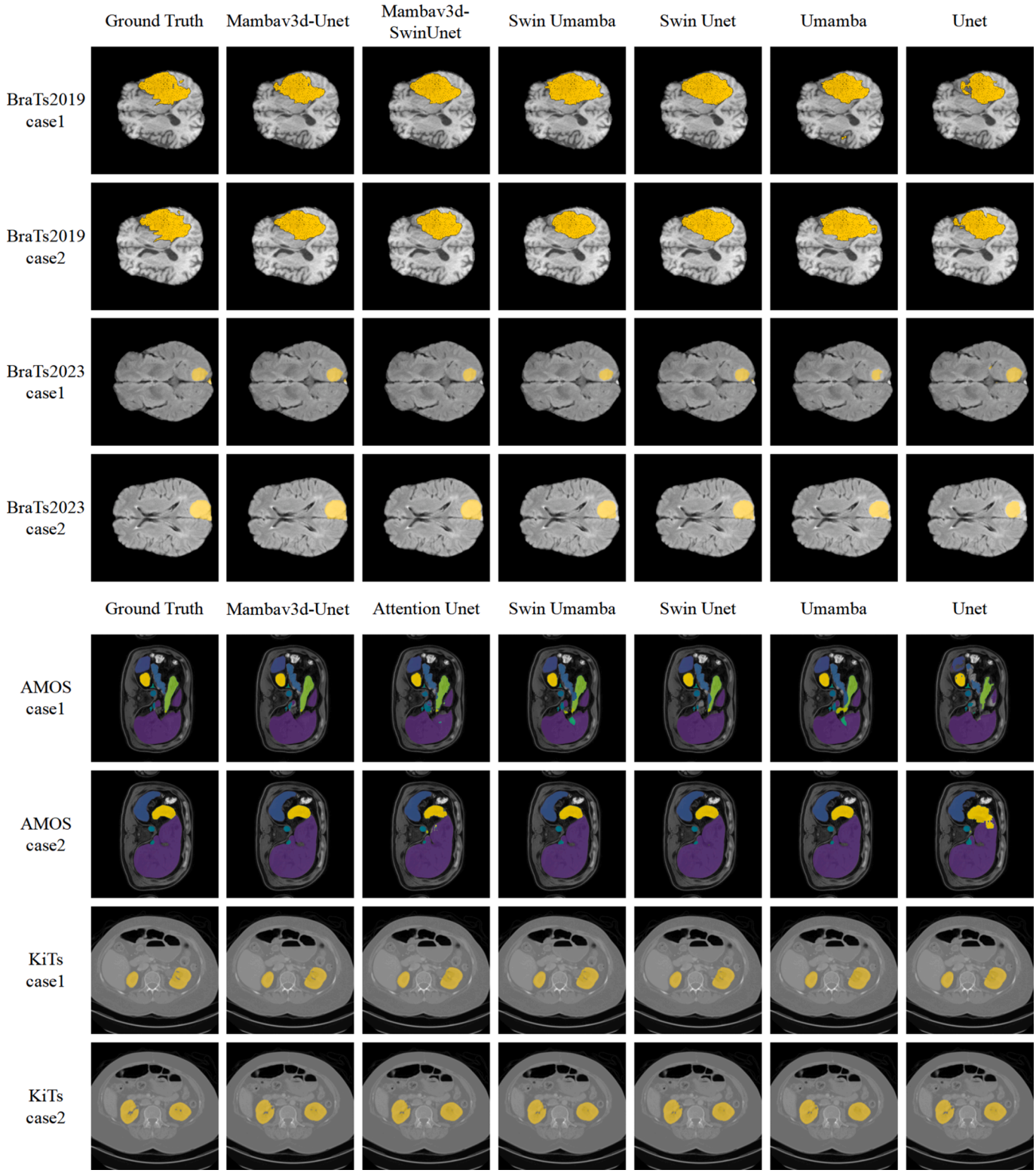

Fig. 6. Segmentation results of different methods on BraTs, AMOS and KiTs datasets.

first sampled and organized and then trained for 1000 epoch on the environment cuda 11.8. The optimizer used is adamW [71], with a momentum of 0.9 and a learning rate of 1e-4.

4.4. Experimental results

The experiments are conducted on three different medical image segmentation tasks: segmenting brain tumors from MR images, achieving multi-category segmentation from abdominal images, and segmenting kidney tumors from CT images. To compare the proposed method with the most popular and latest medical image segmentation methods, the experiments have included the test results of some popular methods on the same datasets and in the same experimental environment. The results are shown in Fig. 6.

On the BraTs dataset, both Unet and Swin Unet loaded with Mambav3d are displayed simultaneously. Comparing the methods in the figure, Mambav3d-Unet and Mambav3d-SwinUnet are the closest to the ground truth. The Swin Unet model is complex and overly sensitive to details, resulting in a certain degree of over-segmentation. The Unet has the weakest segmentation as a baseline for these methods. However, the segmentation results of Umamba and Swin Umamba are unstable, sometimes over-segmented and sometimes under-segmented. This may be because, although Umamba has strong context-capturing ability, the farther away the distance is, the more condensed the relevant information is compressed, which may negatively affect the segmentation effect.

In particular, a comparison of Mambav3d-Unet and Unet shows that the segmentation of Unet with Mambav3d loaded is significantly improved. The same situation occurs on Swin Unet. This demonstrates that Mambav3d effectively introduces inter-layer information into the segmentation process of the current layer, fully complementing the semantic information. This is crucial for improving the segmentation performance of the current layer and also confirms the effectiveness of the proposed strategy.

Similar results are also obtained on the AMOS and KiTs datasets. In AMOS case 1, the relatively irregular shapes of the stomach (green) and pancreas (blue) have a more pronounced visual effect. Compared to Attention Unet and Swin Unet, Mambav3d-Unet has a more accurate segmentation effect, while Swin Umamba and Umamba have unstable performance. After enlarging KiTs case 1, it is found that Mambav3d-Unet has the best segmentation of details in the kidneys.

To further illustrate the positive role of Mambav3d in the model segmentation process, we have compared the experimental results of Unet, Mambav3d-Unet, and 3D Unet on the KiTs dataset. The segmentation results of three consecutive slices are shown in Fig. 7. Through the analysis of the segmentation effect, it is found that the segmentation effect of Mambav3d-Unet is closer to the ground truth. This is because the Unet model can complete the segmentation task well, but it is easy to lose some detailed segmentation, resulting in low segmentation accuracy. The 3D Unet has a huge number of parameters, it needs a large amount of 3D training data to achieve perfect training, which is very demanding in terms of computational cost and hardware equipment requirements. Under general experimental and data volume conditions, the expected segmentation effect cannot be achieved. The advantage of Mambav3d-Unet is that it does not introduce spatial convolution or Transformer, but adds spatial information between slices to the model analysis process through Mambav3d in the skip connection part. This helps the model to control computational cost while improving segmentation performance.

To quantitatively evaluate the performance of these methods, the

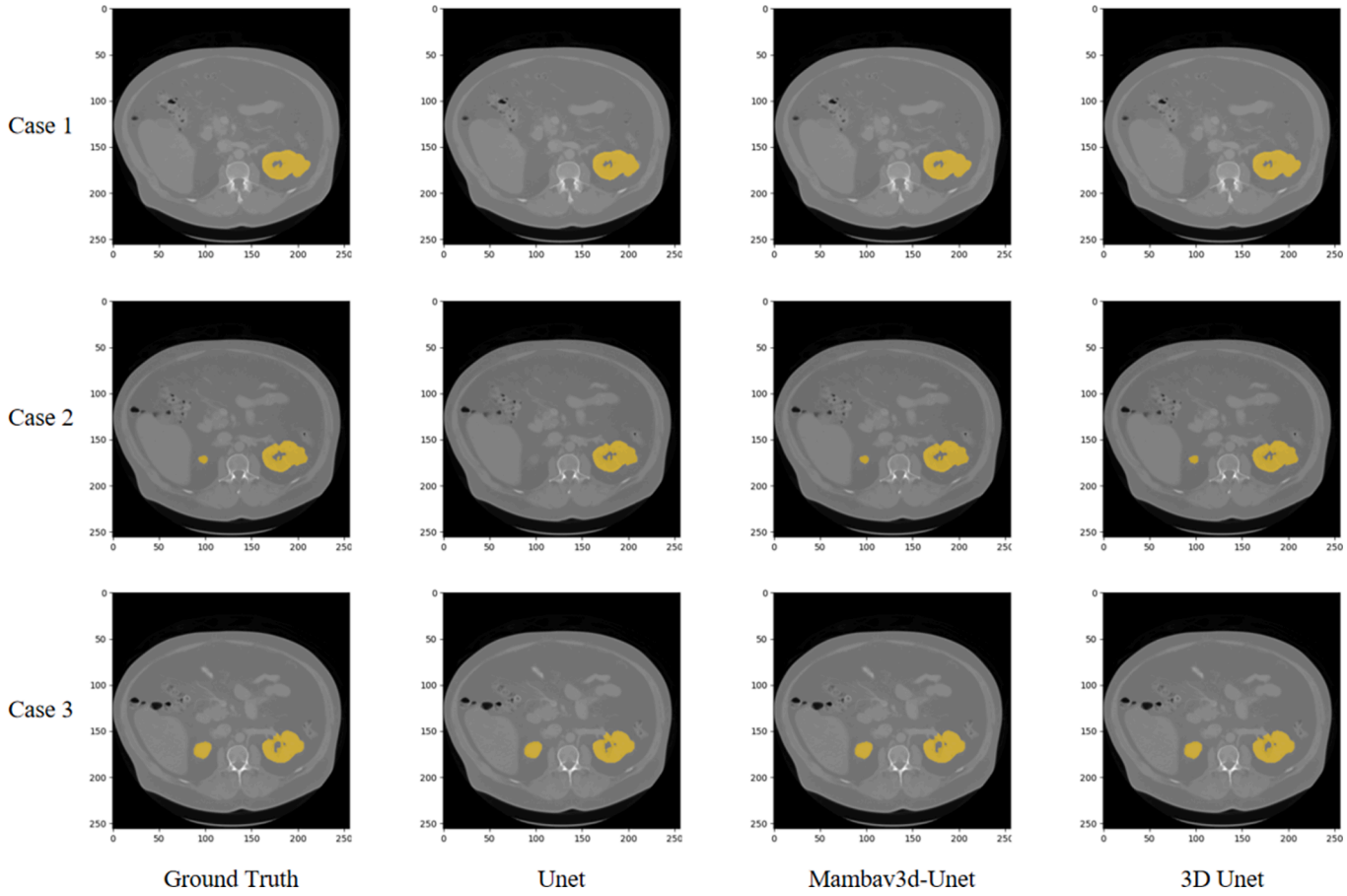

Fig. 7. Segmentation results of Unet, Mambav3d-Unet and 3D Unet on KiTs dataset.

evaluation metrics introduced in section 4.2 are used in the experiment to assess their performance on three datasets.

The BraTs dataset consists of image data from multiple cases of brain MRI. It includes four common types of brain tumors: glioblastoma, astrocytoma, oligodendroglioma, and meningioma. These types of tumors differ in morphology, location, and size, while the segmentation task is to distinguish the region of interest from the background, so the accuracy of the segmentation method is more demanding.

The experimental results on the BraTs dataset are shown in Table 2. In the experiment, the most popular and latest medical image segmentation methods are compared, including Unet with an attention mechanism, transformer-based methods, diffusion probabilistic model-based methods, and the method of directly applying Mamba to Unet. Comparing Mambav3d-Unet and Unet, it is found that there is a significant improvement in various evaluation metrics, with Dice increasing by nearly 3 percentage points, IoU increasing by 4 percentage points, and HD95 decreasing by more than 30 percent. The same is true between Mambav3d-SwinUnet and Swin Unet, and while there is a clear optimization effect, the improvement is not as pronounced as with Unet. Comparing Mambav3d-Unet with diffusion probabilistic-based models and other Mamba-based methods, it is found that Mambav3d-Unet has a slight advantage in the evaluation metrics, typically with a performance improvement of 0.2 to 2 percentage points. In the experiment, the complexity of each model is discussed by calculating Params and GFLOPs. Within the appropriate range, relatively high Params reflect that the model has strong learnability and can better learn the feature information of three-dimensional space than other two-dimensional models. Although the GFLOPs of Mambav3d-Unet is several times that of Unet and other models, it realizes the connection of semantic information between three-dimensional space slices. Under the experimental conditions, the Params of 3D Unet is 244.13 M, the GFLOPs is 10132, and the Params of 3D U-Mamba is 602 M, and the GFLOPs is 22513. Their complexity and computational cost are higher than Mambav3d-Unet, but they do not achieve better segmentation results. Additionally, by comparing Mambav3d-Unet and Mambav3d-SwinUnet, it is found that the effects of the two are similar, but the model structure of Mambav3d-SwinUnet is more complex, and the computational and time costs are higher than Mambav3d-Unet.

In clinical practice, not only is a single segmentation task required, but also accurate segmentation with multiple targets. Therefore, the experiments choose to conduct further research on the AMOS dataset, an abdominal multimodality and multiorgan dataset, which aims at accurate segmentation for multi-target tasks.

The results of the experiments on the AMOS dataset are shown in Table 3. From the overall task of multi-target segmentation, Mambav3d-Unet has the highest Dice and the lowest HD95, far ahead of the six other methods compared in the experiment. The Dice has increased by 1 to 6 percentage points, and HD95 has decreased by 10 to 30 percent. However, this does not mean that Mambav3d-Unet achieves the best results on every target. For example, in the segmentation tasks of Aorta, Kidney, and Spleen, Attention Unet, Swin Umamba, and Swin Unet achieve the best results, respectively, while Mambav3d-Unet ranks second with a small gap from the first place. In the segmentation tasks of Gallbladder and Pancreas, Mambav3d-Unet achieves the best results.

The KiTs dataset is a commonly used CT image, different from the image modality of the previous two datasets. In the experiment, selecting the KiTs dataset helps to verify the wider applicability of the proposed model. The experimental results on the KiTs dataset are shown in Table 4. It can be seen intuitively that the evaluation metrics of Mambav3d-Unet are better than the other six comparison methods. This indicates that Mambav3d-Unet obtains good results in terms of segmentation accuracy and edge details, and the segmentation performance is significantly improved.

4.5. Ablation experiments

In this section, on the BraTs dataset, under the same experimental setting, the experiments investigate the effects of loading the number and position of Mambav3d blocks, the number of patches and the size of the hidden layer in the Mambav3d blocks on the segmentation effect, respectively.

As shown in Fig. 5 and Table 5, the number of Mambav3d blocks can be loaded on the Unet to process interlayer information in parallel. It is found that the evaluation metrics such as Dice are first improved with the increase in the number of parallel Mambav3d blocks. The best results are obtained when three Mambav3d blocks are parallelized, and then the evaluation metrics of segmentation decrease with the increase of the number. The more the number of parallel Mambav3d blocks, the more complex the model becomes, and the higher the time and computational cost. Therefore, this experiment chooses to parallelize three Mambav3d blocks for subsequent research.

According to the experimental design, Mambav3d blocks can be loaded to the four skip connections and bottle neck parts of the Unet. As shown in Table 6, the best segmentation results are achieved when Mambav3d blocks are loaded to three deep-layer skip connections and bottle neck at the same time.

The patch size is an important hyperparameter in the network, which is directly related to the input sequence length, computational complexity, feature details and receptive field, and has an important impact on the model performance. As shown in Table 7, comparative experiments are conducted using three different patch sizes. In the experiment, the number of patches determines the patch size. The experiments show that as the number of patches increases, which is equal to the patch getting smaller, the segmentation effectiveness first gets better and then weakens. Considering factors such as overfitting risk and computational resources, the number of patches is chosen as four in this experiment for subsequent studies.

The hidden layer size has a certain impact on the complexity, learning ability, computational cost and performance of the model. As shown in Table 8, comparative experiments are conducted with four different sizes of hidden layers. It is found that with the increase in the size of hidden layers, the evaluation metrics are slightly improved and

Table 2
Segmentation evaluation of different methods on BraTs dataset.

	2019			2023			Params(M)	GFLOPs
	DICE	IOU	HD95	DICE	IOU	HD95		
Unet [12]	0.8659	0.7635	0.1498	0.8987	0.8160	0.1309	15.46	1011
Attention Unet [15]	0.8682	0.7687	0.1226	0.9003	0.8186	0.1187	34.87	1064
Swin Unet [32]	0.8877	0.7945	0.1137	0.9104	0.8355	0.1074	61.99	1530
ViT [22]	0.8456	0.7323	0.1901	0.8826	0.7898	0.1713	35.16	1449
SegDiff [69]	0.8795	0.7849	0.1172	0.9035	0.8240	0.1160	22.50	1232
MedSegDiff [70]	0.8893	0.8007	0.1075	0.9110	0.8366	0.1061	68.71	1569
Umamba [58]	0.8754	0.7784	0.1337	0.9022	0.8218	0.1289	173.53	10,685
Swin Umamba [64]	0.8852	0.794	0.1128	0.9049	0.8263	0.1092	274.87	18,130
Mambav3d-Unet *	0.8901	0.8019	0.1053	0.9137	0.8411	0.1037	159.13	9530
Mambav3d-SwinUnet	0.8913	0.8039	0.1031	0.9115	0.8373	0.1035	205.57	11,077

Table 3
Segmentation evaluation of different methods on AMOS dataset.

Datas	AMOS								
Evaluation Metric	Dice	HD95	Aorta	Gallbladder	Kidney	Liver	Pancreas	Spleen	Stomach
Unet [12]	0.7347	0.4233	0.8703	0.6422	0.7551	0.9457	0.5286	0.8667	0.7655
Attention Unet [15]	0.7557	0.3802	0.9357	0.6301	0.7895	0.9263	0.5804	0.8522	0.7466
Swin Unet [32]	0.7685	0.3079	0.8998	0.6635	0.8137	0.9415	0.5921	0.8942	0.7831
ViT [22]	0.7176	0.3612	0.7385	0.5140	0.7927	0.9151	0.4377	0.8371	0.7014
Umamba [58]	0.7419	0.3755	0.8512	0.6431	0.7990	0.9222	0.5327	0.8690	0.7789
Swin Umamba [64]	0.7603	0.3216	0.8724	0.6814	0.8471	0.9359	0.5766	0.8775	0.7370
Mambav3d-Unet*	0.7773	0.2864	0.9216	0.6925	0.8382	0.9390	0.6143	0.8906	0.7753

Table 4
Segmentation evaluation of different methods on KiTs dataset.

Datas	KiTs		
Evaluation Metric	DICE	IoU	HD95
Unet [12]	0.8945	0.8089	0.1057
Attention Unet [15]	0.9177	0.8467	0.0976
Swin Unet [32]	0.9194	0.8508	0.0925
ViT [22]	0.8851	0.7940	0.1195
Umamba [58]	0.9034	0.8238	0.0996
Swin Umamba [64]	0.9166	0.8461	0.0957
Mambav3d-Unet*	0.9239	0.8586	0.0913

Table 5
The effect of the number of Mambav3d blocks loaded on the segmentation performance.

Number of parallel Mambav3d blocks	DICE	IoU	HD95
N = 4	0.8871	0.7971	0.1159
N = 3*	0.8901	0.8019	0.1053
N = 2	0.8866	0.7963	0.1102
N = 1	0.8772	0.7810	0.1211

Table 6
The effect of the position of Mambav3d blocks loaded on the segmentation performance.

Position for loading Mambav3d	DICE	IoU	HD95	Params	GFLOPs
Bottleneck-only	0.8747	0.7773	0.1201	38.72	1337
Skip Connection-only	0.8803	0.7862	0.1159	159.13	9659
Skip Connection and Bottleneck	0.8892	0.8005	0.1093	185.94	10,731
Skip Connection and Bottleneck (without-first skip connection) *	0.8901	0.8019	0.1053	159.13	9530

Table 7
The effect of the patch number on the segmentation performance.

Patch number	DICE	IOU	HD95	Params	GFLOPs
patches_per_picture = 1	0.8685	0.7676	0.1535	66.35	4520
patches_per_picture = 4*	0.8901	0.8019	0.1053	159.13	9530
patches_per_picture = 16	0.8882	0.7989	0.1152	453.16	19,033

Table 8
The effect of the Mambav3d blocks hidden layer size on the segmentation performance.

Hidden size in Mambav3d	DICE	IoU	HD95	Params	GFLOPs
hidden_size = 384	0.8869	0.7968	0.1180	115.13	7920
hidden_size = 512	0.8890	0.8001	0.1096	133.76	8552
hidden_size = 768*	0.8901	0.8019	0.1053	159.13	9530
hidden_size = 1024	0.8885	0.7994	0.1385	212.46	13,220

then show a decreasing trend. By considering factors such as overfitting risk and computational resources, this experiment chooses the size of hidden layers to be 768 for subsequent research.

4.6. Discussion

In this paper, we propose a Mamba-based module for enriching semantic information between slices to address the problem that 2D models cannot fully utilize spatial information. The experimental results show that the segmentation performance of Unet loaded with Mambav3d is superior to popular and latest commonly used segmentation networks in different modalities and segmentation tasks. It indicates that Mambav3d loaded on Unet effectively captures the interlayer semantic information and provides many interrelationships between different anatomical surfaces in space for the segmentation process, which in turn enhances the segmentation performance.

Unet is a very successful architecture, which has been widely used in medical image segmentation tasks due to its clever architecture design, outstanding performance, and wide applicability. However, its direct processing of 3D has the disadvantages such as high computational cost, less effective segmentation than 2D, and fewer training samples. By loading Mambav3d on Unet, on the one hand, it can introduce 3D information into the model to obtain better segmentation results, and on the other hand, it also avoids high computational cost. This is attributed to one of the inherent properties of Mamba, having hidden states. When we connect different layers of the same sample with Mamba, the state of the current layer can be predicted with reference to the state of the previous layers, which introduces interrelationships between the two layers. Meanwhile, based on the Hippo matrix, Mamba can generate a hidden state to remember its history, and the closer the distance, the more sufficient relevant information is retained. This is exactly in accordance with the objective phenomenon that the slices with closer distances have greater mutual influence. Comparison of Unet and Mambav3d-Unet shows that the improved segmentation results are significantly improved and slightly higher than other intra-2D segmentation methods. This is due to the fact that Mambav3d-Unet successfully introduces spatial interlayer semantic information, and it has the advantage of finer feature extraction due to the smaller patches within the Mambav3d module. When Mambav3d is loaded onto Swin Unet, similar results are obtained.

More accurate multi-objective segmentation is also one of the challenges facing medical clinical practice. The experiments from the AMOS dataset show that Mambav3d-Unet is suitable for the CT image task and achieves the best results in the overall segmentation task. However, it is not the case that Mambav3d-Unet achieves the best results on every objective. The reason for this is that Mambav3d-Unet only enhances the inter-slice semantic information and is still missing for contextual connections in the 2D plane. Other methods lack three-dimensional spatial information, but with the advantages of structures such as Transformers, the model has a certain contextual connection in the two-dimensional plane. From this analysis, it can be concluded that contextual connections between spatial slices are equally important as contextual connections within a plane. Due to the differences in the three-dimensional

shape of the target in various spatial directions, the contextual connections of cross-sections at different angles have varying effects on the segmentation results.

The experiments demonstrate that Mambav3d can effectively introduce inter-slice semantic information on Unet-based models. In order to achieve better segmentation performance, the experiments discuss the effects of hyperparameters such as the number of Mambav3d parallels and loading positions, the size of hidden layers in the Mambav3d block, and the patch size on the segmentation effectiveness. The more the number of parallel Mambav3d is, the more parameters can be learned and trained. Too few can lead to poor learning results, and too many can lead to a complex and overfitted network. There are four skip connections and a bottleneck available for loading Mambav3d. The shallow-layer skip connections usually have the lowest semantic level, high image resolution, and the main function is to preserve the details of the image. The deep-layer skip connections usually have the highest semantic level, mainly helping decoders understand the global structure of the image. Mambav3d plays the role of helping the model to understand the correlation information between the slice layers, so it is more effective to load in the deep-layer skip connections. The patch size is also a key hyperparameter, which affects the model complexity, receptive field, and segmentation performance. Mambav3d is loaded into skip connections of different depths, and its input patch size varies. Therefore, it is necessary to find an appropriate size to ensure that the model complexity is within an acceptable range and has sufficient inter-slice semantic learning ability, in order to achieve better segmentation results. The size of hidden layers affects the model complexity, learning ability, computational cost, and risk of overfitting, which directly affects the segmentation effectiveness. If the hidden layer is too large, the model has a strong learning ability but the computational cost and overfitting risk increase, and vice versa.

Mambav3d is designed as a module that achieves 3D information fusion at low computational cost, aiming to improve segmentation performance by enriching spatial features. However, the adaptability has only been tried in Unet and Swin Unet with competitive segmentation results. In future work, on one hand, we will continue to verify the feasibility and effectiveness of loading mambav3d on other Unet-based models through skip connections. On the other hand, we will continue to explore the parallel connection of a mambav3d spatial feature extraction module on the feature extraction channels of other models, making it applicable to more advanced segmentation models to achieve better medical image segmentation results.

5. Conclusion

In this paper, we propose a Mamba-based module for enriching semantic information between slice layers, termed Mambav3d, which can be loaded onto Unet (Mambav3d-Unet) and Swin Unet (Mambav3d-SwinUnet) to improve the segmentation performance of medical images. This module effectively introduces memory information from nearby slices of the same 3D sample into the current 2D segmentation layer, thereby introducing interrelationships between anatomical surfaces at various locations. With Mambav3d, the model can fully refer to the information of the previous layers to complete the labeling of the current layer, effectively enhancing the segmentation performance. This approach effectively addresses the limitations of 2D convolution in not considering volume context and the high computational cost associated with 3D convolution. The experiments were conducted on datasets of BraTs, AMOS, and KiTs with different modalities and task objectives, and competitive results were obtained compared to popular and advanced methods such as Unet-based, transformer-based, and diffusion probabilistic model-based approaches.

CRediT authorship contribution statement

Xiaoxiao Liu: Conceptualization, Investigation, Methodology, Validation, Visualization, Writing – original draft. **Yan Zhao:** Funding acquisition, Project administration, Writing – review & editing. **Shigang Wang:** Resources, Supervision. **Jian Wei:** Funding acquisition, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A

```
class Mambap3d(nn.Module):
    def __init__(self, image_size, hidden_size, patch_size, channels, znum):
        super().__init__()
        self.channels = channels
        self.znum = znum
        self.tokens_num = (image_size // patch_size) ** 2
        self.embed1 = PatchEmbed(image_size, patch_size, channels, hidden_size, bias = True)
        self.mamba1 = Mamba2(hidden_size)
        self.zout1 = ReshapeLayer(hidden_size, patch_size, channels)
        self.zoutnorm1 = nn.BatchNorm1d(self.tokens_num)
        self.embed2 = PatchEmbed(image_size, patch_size, channels, hidden_size, bias = True)
        self.mamba2 = Mamba2(hidden_size)
        self.zout2 = ReshapeLayer(hidden_size, patch_size, channels)
        self.zoutnorm2 = nn.BatchNorm1d(self.tokens_num)
        self.embed3 = PatchEmbed(image_size, patch_size, channels, hidden_size, bias = True)
        self.mamba3 = Mamba2(hidden_size)
        self.zout3 = ReshapeLayer(hidden_size, patch_size, channels)
        self.zoutnorm3 = nn.BatchNorm1d(self.tokens_num)
        self.zoutnormend = nn.BatchNorm2d(channels)
    def unpatchify(self, x):
        """
        x: (N, T, patch_size**2 * C)
        imgs: (N, H, W, C)
        """
        c = self.channels
        p = self.embed1.patch_size[0]
```

(continued on next page)

(*continued*)

```
class Mambap3d(nn.Module):
    h = w = int(x.shape[1] ** 0.5)
    assert h * w == x.shape[1]
    x = x.reshape(shape=(x.shape[0], h, w, p, p, c))
    x = torch.einsum('nhwpqc->nchpwq', x)
    imgs = x.reshape(shape=(x.shape[0], c, h * p, h * p))
    return imgs
  def forward(self, x):
    xres = x
    # D is the number of images on different Z axes in each batch
    D = self.znum
    # BD, CN, HN and WN are batch*depth, number of channels, image height and width respectively
    BD, CN, HN, WN = x.shape
    # Patch the image, each size is patch_size, and can be divided into (image_size//patch_size)**2 pieces
    x1 = self.embed1(x)
    # Regroup the patches and group D and the number of patches together
    x1 = x1.reshape(shape=(BD//D, self.tokens_num * D, x1.shape[2]))
    # Such a group can be trained by the Mamba structure
    x1 = self.mamba1(x1)
    # The FINAL LAYER part of the token-level regularization output
    x1 = self.zout1(x1)
# Regrouping and Standardization
    x1 = x1.reshape(shape=(BD, self.tokens_num, x1.shape[2]))
    x1 = self.zoutnorm1(x1)
    # Reorganize the output to BD, CN, HN, WN
    x1 = self.unpatchify(x1)
    x1 = x1.view(BD, CN, HN, WN)
    # x2, x3 have a multihead relationship with x1
    x2 = self.embed2(x)
    x2 = x2.reshape(shape=(BD//D, self.tokens_num * D, x2.shape[2]))
    x2 = self.mamba2(x2)
    x2 = self.zout2(x2)
    x2 = x2.reshape(shape=(BD, self.tokens_num, x2.shape[2]))
    x2 = self.zoutnorm2(x2)
    x2 = self.unpatchify(x2)
    x2 = x2.view(BD, CN, HN, WN)
    x3 = self.embed3(x)
    x3 = x3.reshape(shape=(BD//D, self.tokens_num * D, x3.shape[2]))
    x3 = self.mamba3(x3)
    x3 = self.zout3(x3)
    x3 = x3.reshape(shape=(BD, self.tokens_num, x3.shape[2]))
    x3 = self.zoutnorm3(x3)
    x3 = self.unpatchify(x3)
    x3 = x3.view(BD, CN, HN, WN)
    x = xres + self.zoutnormend(x1 + x2 + x3)
    return x
```

Data availability

Data will be made available on request.

References

[1] S. Wang, C. Li, R. Wang, Z. Liu, M. Wang, H. Tan, Y. Wu, X. Liu, H. Sun, R. Yang, et al., Annotation-efficient deep learning for automatic medical image segmentation, Nat. Commun. 12 (1) (2021) 5915, https://doi.org/10.1038/s41467-021-26216-9.

[2] L.K. Lee, S.C. Liew, W.J. Thong, A review of image segmentation methodologies in medical image, in: Advanced Computer and Communication Engineering Technology, in: Proceedings of the 1st International Conference on Communication and Computer Engineering, 2015, pp. 1069–1080, https://doi.org/10.1007/978-3-319-07674-4_99.

[3] X. Min, H. Duan, W. Sun, Y. Zhu, G. Zhai, Perceptual video quality assessment: A survey, arXiv preprint arXiv:2402.03413 (2024). doi: 10.48550/arXiv.2402.03413.

[4] Y. Zhang, J. Wang, Y. Zhu, R. Xie, Subjective and objective quality evaluation of ugc video after encoding and decoding, Displays 83 (2024) 102719, https://doi.org/10.1016/j.displa.2024.102719.

[5] X. Min, K. Gu, G. Zhai, X. Yang, W. Zhang, P. Le Callet, C.W. Chen, Screen content quality assessment: Overview, benchmark, and beyond, ACM Computing Surveys (CSUR) 54 (9) (2021) 1–36, https://doi.org/10.1145/3470970.

[6] Y. Zhu, Y. Li, W. Sun, X. Min, G. Zhai, X. Yang, Blind image quality assessment via cross-view consistency, IEEE Trans. Multimedia 25 (2022) 7607–7620, https://doi.org/10.1109/TMM.2022.3224319.

[7] X. Min, K. Gu, G. Zhai, J. Liu, X. Yang, C.W. Chen, Blind quality assessment based on pseudo-reference image, IEEE Trans. Multimedia 20 (8) (2017) 2049–2062, https://doi.org/10.1109/TMM.2017.2788206.

[8] M. Eslami, S. Tabarestani, S. Albarqouni, E. Adeli, N. Navab, M. Adjouadi, Image-to-images translation for multi-task organ segmentation and bone suppression in chest x-ray radiography, IEEE Trans. Med. Imaging 39 (7) (2020) 2553–2565, https://doi.org/10.1109/TMI.2020.2974159.

[9] X. Min, Y. Gao, Y. Cao, G. Zhai, W. Zhang, H. Sun, C. W. Chen, Exploring rich subjective quality information for image quality assessment in the wild, arXiv preprint arXiv:2409.05540 (2024). doi: 10.48550/arXiv.2409.05540.

[10] M. Nagendran, Y. Chen, C. A. Lovejoy, A. C. Gordon, M. Komorowski, H. Harvey, E. J. Topol, J. P. Ioannidis, G. S. Collins, M. Maruthappu, Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies, bmj 368 (2020). doi: 10.1136/bmj.m689.

[11] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Commun. ACM 60 (6) (2017) 84–90, https://doi.org/10.1145/3065386.

[12] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, Springer, 2015, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.

[13] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: Redesigning skip connections to exploit multiscale features in image segmentation, IEEE Trans. Med. Imaging 39 (6) (2019) 1856–1867, https://doi.org/10.1109/TMI.2019.2959609.

[14] M. Marhamati, A.A.L. Zadeh, M.M. Fard, M.A. Hussain, K. Jafarnezhad, A. Jafarnezhad, M. Bakhtoor, M. Momeny, Laiu-net: a learning-to-augment incorporated robust u-net for depressed humans' tongue segmentation, Displays 76 (2023) 102371, https://doi.org/10.1016/j.displa.2023.102371.

[15] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, et al., Attention u-net: Learning where to look for the pancreas, arXiv preprint arXiv:1804.03999 (2018). doi: 10.48550/arXiv.1804.03999.

[16] F. Bougourzi, C. Distante, F. Dornaika, A. Taleb-Ahmed, Pdatt-unet: Pyramid dual-decoder attention unet for covid-19 infection segmentation from ct-scans, Med. Image Anal. 86 (2023) 102797, https://doi.org/10.1016/j.media.2023.102797.

[17] F. Isensee, P.F. Jaeger, S.A. Kohl, J. Petersen, K.H. Maier-Hein, nnu-net: a self-configuring method for deep learning-based biomedical image segmentation, Nat. Methods 18 (2) (2021) 203–211, https://doi.org/10.1038/s41592-020-01008-z.

[18] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, H. D. Johansen, Doubleu-net: A deep convolutional neural network for medical image segmentation, in: 2020 IEEE 33rd International symposium on computer-based medical systems (CBMS), IEEE, 2020, pp. 558–564. doi: 10.1109/CBMS49503.2020.00111.

[19] Y. Cao, X. Min, W. Sun, G. Zhai, Attention-guided neural networks for full-reference and no-reference audio-visual quality assessment, IEEE Trans. Image Process. 32 (2023) 1882–1896, https://doi.org/10.1109/TIP.2023.3251695.

[20] X. Min, G. Zhai, K. Gu, Y. Zhu, J. Zhou, G. Guo, X. Yang, X. Guan, W. Zhang, Quality evaluation of image dehazing methods using synthetic hazy images, IEEE Trans. Multimedia 21 (9) (2019) 2319–2333, https://doi.org/10.1109/TMM.2019.2902097.

[21] X. Min, G. Zhai, K. Gu, X. Yang, X. Guan, Objective quality evaluation of dehazed images, IEEE Trans. Intell. Transp. Syst. 20 (8) (2018) 2879–2892, https://doi.org/10.1109/TITS.2018.2868771.

[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020). doi: 10.48550/arXiv.2010.11929.

[23] J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang, et al., Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers, Med. Image Anal. (2024) 103280, https://doi.org/10.1016/j.media.2024.103280.

[24] W. Wu, J. Yan, Y. Zhao, Q. Sun, H. Zhang, J. Cheng, D. Liang, Y. Chen, Z. Zhang, Z.-C. Li, Multi-task learning for concurrent survival prediction and semi-supervised segmentation of gliomas in brain mri, Displays 78 (2023) 102402, https://doi.org/10.1016/j.displa.2023.102402.

[25] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, V. M. Patel, Medical transformer: Gated axial-attention for medical image segmentation, in: Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, part I 24, Springer, 2021, pp. 36–46. doi: 10.1007/978-3-030-87193-2_4.

[26] P. Wu, L. Jiang, Z. Hua, J. Li, Multi-focus image fusion: Transformer and shallow feature attention matters, Displays 76 (2023) 102353, https://doi.org/10.1016/j.displa.2022.102353.

[27] J. Yin, J. Jiang, W. Li, E. Chen, L. Chen, L. Tong, B. Huang, Muptnet: Multi-scale u-shape pyramid transformer network for infrared small target detection, Displays 83 (2024) 102681, https://doi.org/10.1016/j.displa.2024.102681.

[28] H.-Y. Zhou, J. Guo, Y. Zhang, X. Han, L. Yu, L. Wang, Y. Yu, nnformer: Volumetric medical image segmentation via a 3d transformer, IEEE Trans. Image Process. (2023), https://doi.org/10.1109/TIP.2023.3293771.

[29] Z. Li, Y. Zheng, D. Shan, S. Yang, Q. Li, B. Wang, Y. Zhang, Q. Hong, D. Shen, Scribformer: Transformer makes cnn work better for scribble-based medical image segmentation, IEEE Trans. Med. Imaging (2024), https://doi.org/10.1109/TMI.2024.3363190.

[30] W. Ji, A.C. Chung, Unsupervised domain adaptation for medical image segmentation using transformer with meta attention, IEEE Trans. Med. Imaging (2023), https://doi.org/10.1109/TMI.2023.3322581.

[31] L. Li, Q. Liu, X. Shi, Y. Wei, H. Li, H. Xiao, Ucfiltransnet: Crossfiltering transformer-based network for ct image segmentation, Expert Syst. Appl. 238 (2024) 121717, https://doi.org/10.1016/j.eswa.2023.121717.

[32] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang, Swin-unet: Unet-like pure transformer for medical image segmentation, in: European conference on computer vision, Springer, 2022, pp. 205–218. doi: 10.1007/978-3-031-25066-8_9.

[33] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, O. Ronneberger, 3d u-net: learning dense volumetric segmentation from sparse annotation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19, Springer, 2016, pp. 424–432. doi: 10.1007/978-3-319-46723-8_49.

[34] F. Milletari, N. Navab, S.-A. Ahmadi, V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: 2016 fourth international conference on 3D vision (3DV), Ieee, 2016, pp. 565–571. doi: 10.1109/3DV.2016.79.

[35] L. Heinrich, D. Bennett, D. Ackerman, W. Park, J. Bogovic, N. Eckstein, A. Petruncio, J. Clements, S. Pang, C.S. Xu, et al., Wholecell organelle segmentation in volume electron microscopy, Nature 599 (7883) (2021) 141–146, https://doi.org/10.1038/s41586-021-03977-3.

[36] J. Yang, L. Jiao, R. Shang, X. Liu, R. Li, L. Xu, Ept-net: Edge perception transformer for 3d medical image segmentation, IEEE Trans. Med. Imaging 42 (11) (2023) 3229–3243, https://doi.org/10.1109/TMI.2023.3278461.

[37] S. Nikan, K. Van Osch, M. Bartling, D.G. Allen, S.A. Rohani, B. Connors, S. K. Agrawal, H.M. Ladak, Pwd-3dnet: a deep learning-based fully-automated segmentation of multiple structures on temporal bone ct scans, IEEE Trans. Image Process. 30 (2020) 739–753, https://doi.org/10.1109/TIP.2020.3038363.

[38] J. Li, N. Chen, H. Zhou, T. Lai, H. Dong, C. Feng, R. Chen, C. Yang, F. Cai, L. Wei, Mcrformer: Morphological constraint reticular transformer for 3d medical image segmentation, Expert Syst. Appl. 232 (2023) 120877, https://doi.org/10.1016/j.eswa.2023.120877.

[39] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, et al., "Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge", arXiv preprint arXiv:1811.02629 (2018).

[40] D. LaBella, M. Adewole, M. Alonso-Basanta, T. Altes, S. M. Anwar, U. Baid, T. Bergquist, R. Bhalerao, S. Chen, V. Chung, et al., The asnr-miccai brain tumor segmentation (brats) challenge 2023: Intracranial meningioma, arXiv preprint arXiv:2305.07642 (2023).

[41] N. Heller, N. Sathianathen, A. Kalapara, E. Walczak, K. Moore, H. Kaluzniak, J. Rosenberg, P. Blake, Z. Rengel, M. Oestreich, et al., The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes, arXiv preprint arXiv:1904.00445 (2019). doi: 10.48550/arXiv.1904.00445.

[42] Y. Ji, H. Bai, C. Ge, J. Yang, Y. Zhu, R. Zhang, Z. Li, L. Zhanng, W. Ma, X. Wan, et al., Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation, Adv. Neural Inf. Proces. Syst. 35 (2022) 36722–36732, https://doi.org/10.48550/arXiv.2206.08023.

[43] Y. Xiong, L. Deng, Y. Wang, Pulmonary nodule detection based on model fusion and adaptive false positive reduction, Expert Syst. Appl. (2023) 121890, https://doi.org/10.1016/j.eswa.2023.121890.

[44] X. Lu, A. K. Jain, D. Colbry, Matching 2.5 d face scans to 3d models, IEEE transactions on pattern analysis and machine intelligence 28 (1) (2005) 31–43. doi: 10.1109/TPAMI.2006.15.

[45] J. Yang, X. Huang, Y. He, J. Xu, C. Yang, G. Xu, B. Ni, Reinventing 2d convolutions for 3d images, IEEE J. Biomed. Health Inform. 25 (8) (2021) 3009–3018, https://doi.org/10.1109/JBHI.2021.3049452.

[46] Y. Ding, W. Zheng, J. Geng, Z. Qin, K.-K.-R. Choo, Z. Qin, X. Hou, Mvfusfra: A multi-view dynamic fusion framework for multimodal brain tumor segmentation, IEEE J. Biomed. Health Inform. 26 (4) (2021) 1570–1581, https://doi.org/10.1109/JBHI.2021.3122328.

[47] J. Chen, D. Zhu, B. Hui, R. Y. M. Li, X.-G. Yue, et al., Mu-net: Multi-path upsampling convolution network for medical image segmentation., CMES-Computer Modeling in Engineering & Sciences 131 (1) (2022). doi: 10.32604/cmes.2022.018565.

[48] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, J. Li, Transbts: Multimodal brain tumor segmentation using transformer, in: M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, C. Essert (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Springer International Publishing, Cham, 2021, pp. 109–119. doi: 10.1007/978-3-030-87193-2_11.

[49] X. Min, G. Zhai, J. Zhou, M.C. Farias, A.C. Bovik, Study of subjective and objective quality assessment of audio-visual signals, IEEE Trans. Image Process. 29 (2020) 6054–6068, https://doi.org/10.1109/TIP.2020.2988148.

[50] X. Min, G. Zhai, K. Gu, X. Yang, Fixation prediction through multimodal analysis, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 13 (1) (2016) 1–23. doi: 10.1145/2996463.

[51] X. Min, G. Zhai, J. Zhou, X.-P. Zhang, X. Yang, X. Guan, A multimodal saliency model for videos with high audio-visual correspondence, IEEE Trans. Image Process. 29 (2020) 3805–3819, https://doi.org/10.1109/TIP.2020.2966082.

[52] X. Min, G. Zhai, K. Gu, Y. Liu, X. Yang, Blind image quality estimation via distortion aggravation, IEEE Trans. Broadcast. 64 (2) (2018) 508–517, https://doi.org/10.1109/TBC.2018.2816783.

[53] Q. Chen, X. Min, H. Duan, Y. Zhu, G. Zhai, Muiqa: Image quality assessment database and algorithm for medical ultrasound images, in: 2021 IEEE International Conference on Image Processing (ICIP), IEEE, 2021, pp. 2958–2962. doi: 10.1109/ICIP42928.2021.9506431.

[54] X. Min, K. Ma, K. Gu, G. Zhai, Z. Wang, W. Lin, Unified blind quality assessment of compressed natural, graphic, and screen content images, IEEE Trans. Image Process. 26 (11) (2017) 5462–5474, https://doi.org/10.1109/TIP.2017.2735192.

[55] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, P.-A. Heng, H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes, IEEE Trans. Med. Imaging 37 (12) (2018) 2663–2674, https://doi.org/10.1109/TMI.2018.2845918.

[56] A. Gu, T. Dao, Mamba: Linear-time sequence modeling with selective state spaces, arXiv preprint arXiv:2312.00752 (2023). doi: 10.48550/arXiv.2312.00752.

[57] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, X. Wang, Vision mamba: Efficient visual representation learning with bidirectional state space model, arXiv preprint arXiv:2401.09417 (2024). doi: 10.48550/arXiv.2401.09417.

[58] J. Ma, F. Li, B. Wang, U-mamba: Enhancing long-range dependency for biomedical image segmentation, arXiv preprint arXiv:2401.04722 (2024). doi: 10.48550/arXiv.2401.04722.

[59] J. Ruan, S. Xiang, Vm-unet: Vision mamba unet for medical image segmentation, arXiv preprint arXiv:2402.02491 (2024). doi: 10.48550/arXiv.2402.02491.

[60] Z. Wang, J.-Q. Zheng, Y. Zhang, G. Cui, L. Li, Mamba-unet: Unet-like pure visual mamba for medical image segmentation, arXiv preprint arXiv:2402.05079 (2024). doi: 10.48550/arXiv.2402.05079.

[61] W. Liao, Y. Zhu, X. Wang, C. Pan, Y. Wang, L. Ma, Lightm-unet: Mamba assists in lightweight unet for medical image segmentation, arXiv preprint arXiv:2403.05246 (2024). doi: 10.48550/arXiv.2403.05246.

[62] K. S. Sanjid, M. T. Hossain, M. S. S. Junayed, D. M. M. Uddin, Integrating mamba sequence model and hierarchical upsampling network for accurate semantic segmentation of multiple sclerosis legion, arXiv preprint arXiv:2403.17432 (2024). doi: 10.48550/arXiv.2403.17432.

[63] H. Tang, L. Cheng, G. Huang, Z. Tan, J. Lu, K. Wu, Rotate to scan: Unet-like mamba with triplet ssm module for medical image segmentation, arXiv preprint arXiv: 2403.17701 (2024). doi: 10.48550/arXiv.2403.17701.
[64] J. Liu, H. Yang, H.-Y. Zhou, Y. Xi, L. Yu, Y. Yu, Y. Liang, G. Shi, S. Zhang, H. Zheng, et al., Swin-umamba: Mamba-based unet with imagenet-based pretraining, arXiv preprint arXiv:2402.03302 (2024). doi: 10.48550/arXiv.2402.03302.
[65] J. Xie, R. Liao, Z. Zhang, S. Yi, Y. Zhu, G. Luo, Promamba: Prompt-mamba for polyp segmentation, arXiv preprint arXiv:2403.13660 (2024). doi: 10.48550/ arXiv.2403.13660.
[66] Z. Ye, T. Chen, P-mamba: Marrying perona malik diffusion with mamba for efficient pediatric echocardiographic left ventricular segmentation, arXiv preprint arXiv:2402.08506 (2024). doi: 10.48550/arXiv.2402.08506.
[67] A. Gu, T. Dao, S. Ermon, A. Rudra, C. Ré, Hippo: Recurrent memory with optimal polynomial projections, Adv. Neural Inf. Proces. Syst. 33 (2020) 1474–1487.
[68] A. Gu, K. Goel, C. Ré, Efficiently modeling long sequences with structured state spaces, arXiv preprint arXiv:2111.00396 (2021). doi: 10.48550/arXiv.2111.00396.
[69] T. Amit, T. Shaharbany, E. Nachmani, L. Wolf, Segdiff: Image segmentation with diffusion probabilistic models, arXiv preprint arXiv:2112.00390 (2021). doi: 10.48550/arXiv.2112.00390.
[70] J. Wu, R. Fu, H. Fang, Y. Zhang, Y. Yang, H. Xiong, H. Liu, Y. Xu, Medsegdiff: Medical image segmentation with diffusion probabilistic model, in: Medical Imaging with Deep Learning, PMLR, 2024, pp. 1623–1639.
[71] I. Loshchilov, F. Hutter, et al., Fixing weight decay regularization in adam, arXiv preprint arXiv:1711.05101 5, 2017.