Deep-learning based automated lesion segmentation in whole-body PET/CT/MRI using Att-UMamba

Yijie Gong[1*], Second Author[2†] and Third Author[2†]

[1*]Department of Computing, Imperial College London, Exhibition Rd, London, SW7 2AZ, London, United Kingdom.
[2]Department, Organization, Street, City, 10587, State, Country.
[3]Department, Organization, Street, City, 610101, State, Country.

*Corresponding author(s). E-mail(s): yijie.gong23@imperial.ac.uk;
Contributing authors: iiauthor@gmail.com; iiiauthor@gmail.com;
†These authors contributed equally to this work.

Abstract

Lesion Segmentation is a typical task in medical image segmentation. There are well-developed models for regional lesion segmentation that are reliable and accurate. However, traditional models remain struggling to predict segmentation masks with high accuracy for whole-body medical images, such as whole-body PET/CT/MRI. In this paper, we proposed Att-UMamba, a novel, general-purpose model for whole-body biomedical image segmentation. Inspired by U-Mamba, we applied SSM blocks to find the long-range dependency. Att-UMamba automatically focuses on the target lesion spots through the attention mechanism given by attention gates. The trained model will suppress irrelevant regions and highlight the high-interest regions. The proposed Att-UMamba model is evaluated on a new dataset called PSMA consisting of 600 whole-body images in both CT/PET forms. Experimental results show that Att-UMamba surpasses traditional models such as SAM, nnU-Net, and UMamba when predicting based on whole-body medical images. This proves Att-UMamba can be used as a reliable tool when conducting segmentation tasks based on whole-body image segmentation. The source code for the proposed Att-UMamba model is publicly available at https://github.com/ygong0712/Att-UMamba.

Keywords: CNN, Deep Learning, Medical Image Segmentation, Mamba, U-Net, Attention, Lesion Segmentation, Whole-Body CT/PET

1 Introduction

1.1 Background

Manual medical image diagnosis is usually time-consuming and costly. It also requires expert knowledge even though it may still be exposed to human error. The deep-learning model has gained popularity over the past decades due to its growing accuracy in various tasks. Such a powerful model makes dealing with complex medical images and extracting useful measurements from the data possible. Classification is a typical computer vision task that assigns an input image to one of a predefined set of categories or classes. The semantic segmentation task is a pixel-level classification task where the model predicts each pixel and assigns a label to each pixel. Semantic segmentation is widely adopted in medical images, supporting radiologists in tasks such as lesion detection, disease progression monitoring, etc. Medical image segmentation transforms original biological image data into processed, spatially divided image output. The goal is to partition an image into meaningful regions corresponding to different anatomical structures or regions of interest (ROIs). Medical image segmentation models can help radiologists analyze biological image data, which involve various tasks, including marking organs, lesions, and tissue by delineating contours or giving masks of these regions. Such autonomous regions-notation tools will help radiologists diagnose the potential occurrence of diseases, monitor disease phases, and aid the surgical process.

1.2 Motivation

With the development of Convolutional Neural Networks, U-Net variants are commonly adopted in semantic segmentation. Its U-shape encoder and decoder architecture achieve state-of-the-art results in medical image segmentation. The encoder extracts global feature representation and down-sampling to low-resolution embedding, while the decoder up-sampling the output of the encoder to the original resolution. Despite its representation extraction power, it can be trained on a small-scale dataset but still maintains high prediction accuracy. Cascaded frameworks extract embedding from images [1] and infer the region of interest(ROI). For whole-body medical images, the target ROI regions are smaller and more sparsely distributed than regional medical image data. However, the traditional U-Net models have limited ability to find long-range dependency in the image data, and cascaded frameworks cause redundant use of computational resources. Such a deficiency causes sub-optimal segmentation outcomes.

To address these problems, we propose Att-UMamba, specifically designed for whole-body medical image segmentation based on SSM, Attention mechanism, and nnU-Net architecture. It offers linear scaling for different input feature sizes and highlights salient features of input images, which outperform most Transformer-based or U-Net-based architectures. Moreover, it also inherits the self-configuring mechanism of nnU-Net, which requires less effort in hyperparameter tunning and enhances its

Table 1 Model Performance Comparison

Model Category	Training	Inference	Additional Issue
RNN(1986)	Slow	Fast	Gradient descent
LSTM(1997)	Slow	Fast	Forgetting
Transformer(2017)	Fast(Parallel training)	Slow	RAM:$O(N^2)$
SSMs	Fast(Parallel training)	Fast	RAM:$O(N)$

scalability and flexibility across different datasets. It mainly uses SSMs block, which is an edge-cutting model architecture that gains multiple advantages over traditional models, as shown in Table 1. For the dataset, we chose the Prostate-Specific Membrane Antigen(PSMA) [2] from LMU Hospital in Germany to provide experimental evidence for our proposed network. The results show that Att-UMamba outperforms most of the prevailing models on the PSMA dataset.

1.3 Contribution

Developing deep learning models based on whole-body medical image data presents several technical and theoretical difficulties, and the research will seek to overcome these barriers.

- Long-range dependency is a major challenge when we develop deep learning models for whole-body segmentation. 3D images are usually flattened into longer sequences compared to 2D images, and thus, long-range dependency is required to find relationships for all pixels in the image. The traditional medical image segmentation models such as U-Net [3] are mostly CNN-based [4]. However, most traditional CNN models cannot precisely find the long-range dependency. Thus, developing a novel model with new mechanisms to find long-dependency in image data is important.
- Data privacy and security is another challenge for medical image segmentation. There is limited data for training the model. The accuracy of the trained model with limited data will be affected, and the desired result will not be reached. We try to build a model that can reach state-of-art accuracy with limited training data. In this project, we mainly use the PSMA dataset. It contains 600 whole-body CT/PET images. We used a 4:1 training and testing data split, and we obtained 480 training images and 120 testing images.
- Medical image segmentation often requires specific tuning methods based on model architecture, datasets, and preprocessing methods. Manually tuning the model's hyperparameters is time-consuming and resource-intensive. We try to find an autonomous solution for hyperparameter tuning. The main idea of autonomous hyperparameter tuning is based on the nnU-Net framework [5]. The result proves that the autonomous configuration mechanism surpasses most of the manual hyperparameter tuning operations.

2 Methods

2.1 3D Attention Gate

Inspired by the paper ***Attention U-Net: Learning Where to Look for the Pancreas*** [6], the attention mechanism is added to U-Mamba architecture. The self-attention mechanism can be used to find long-range dependencies in image data. It is especially useful in semantic segmentation tasks, which can be treated as a pixel-level classification problem. In encoder-decoder model architecture such as U-Net, input data will pass through several convolution layers for feature extraction and spatial reduction. Features in different stages will pass to the corresponding stages in the decoder using the skip connection. However, these feature embeddings extracted by the encoder are usually sub-optimal and abstract due to limitations in computation resources. Consequently, Attention Gates(AGs) [6] are a novel and effective solution for enhancing the feature embeddings extracted by the encoder. It automatically focuses on target structures without additional supervision. It also highlights the salient features for specific segmentation tasks. It requires low-level computation resources. It focuses more on the high-interest regions and reduces the focus on low-interest regions in semantic segmentation tasks. It is particularly effective in whole-body lesion segmentation tasks, as most regions are low-interest regions. Therefore, we need to focus more on high-interest regions(lesion location). The architecture of Attention Gates is shown in Figure 1.

The Attention Gate will receive two inputs, x, and g, of shape (B, C, H, W, D). Two 1x1x1 convolutions followed by batch normalization apply to both x and g. The transformed gating signal and the input feature map will be added element-wise. ReLU activation will be applied for the output. Combined features will be projected to a single-channel attention map, followed by a sigmoid activation, generating the attention coefficients between 0 and 1. Finally, the input feature map x is multiplied element-wise by the attention coefficients, focusing on the most relevant features.

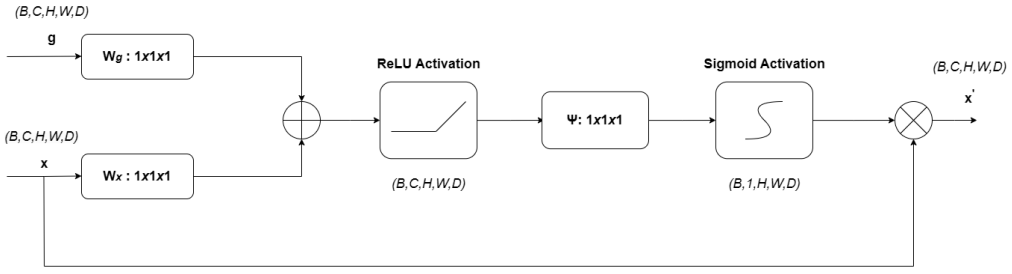

Fig. 1 Spatial Attention Gate

2.2 Automated segmentation pipeline

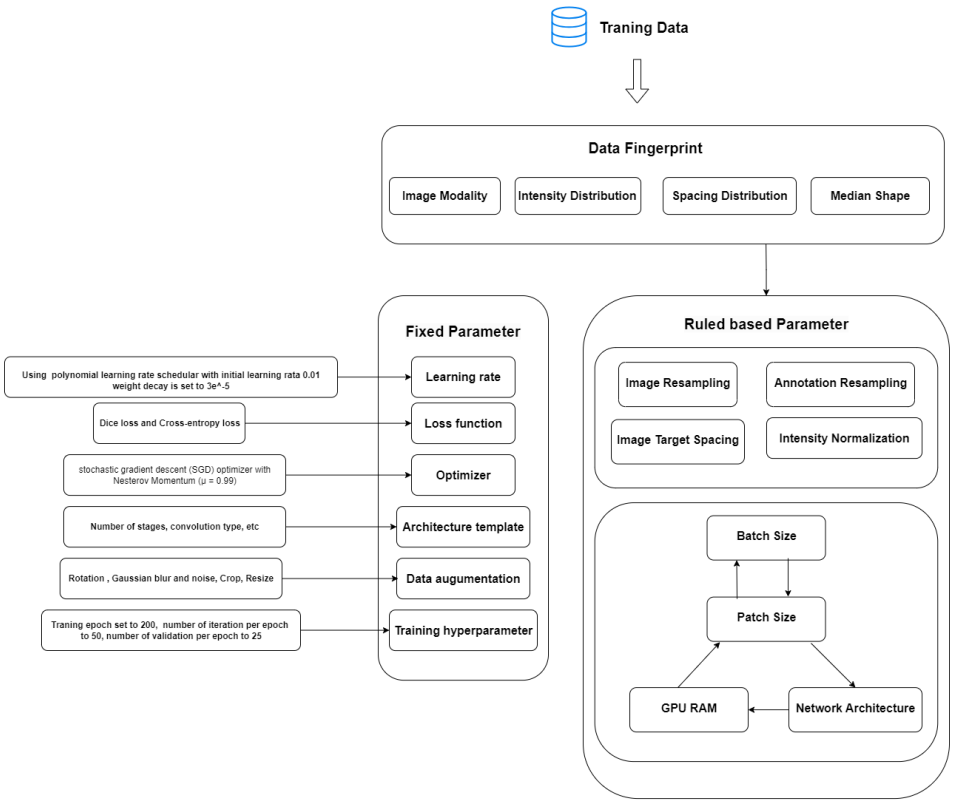

Fig. 2 Self-configuration mechanism pipeline

Inspired by nnU-Net's self-configuration pipeline, Att-UMamba also applies such a mechanism when configuring hyperparameters. The overview of the self-configuration mechanism is shown in Figure 2 [7]. Before training, the model extracts data fingerprints, including image modality, intensity distribution, spacing distribution, and median shape from the training data. The data fingerprints will be used to choose rule-based parameters. The preprocessing strategy will be adjusted based on the data fingerprint. Image resampling or annotation resampling is a process used to ensure that input images are normalized in spatial resolution and size before being fed into the neural network. The input images might have varying sizes. Resampling adjusts the size of these images, making them uniform, which allows the network to process them efficiently. Image Target spacing is also a type of image resampling strategy that ensures images are brought to a common resolution or common voxel spacing, especially for 3D medical images. This is crucial for consistent feature extraction and

model training. Intensity normalization is a preprocessing step that adjusts the pixel or voxel intensity values of medical images to a standardized scale. For the PSMA dataset as an example, we will use the Z-score normalization, which involves scaling the intensity values to have a mean of zero and a standard deviation of one given by the formula $x^{'} = \frac{x-\mu}{\sigma}$ where x is the original intensity value, μ is the mean intensity of the image and σ is the standard deviation of the image intensities. Besides, it will also automatically configure batch size, patch size, GPU RAM, and network architecture, which will be discussed in the Experiment section.

There are some fixed parameters derived empirically. We use the polynomial learning rate scheduler with an initial learning rate of 0.01 and weight decay of $3e^{-5}$. The learning rate η_t at training step t is computed using the formula: $\eta_t = \eta_0(1 - \frac{t}{T})^p$ where η_0 is the initial learning rate, t is the current epoch, T is the total number of epochs, and p is the weight decay. It provides a smooth decay of the learning rate and achieves better convergence as training progresses. The Loss function of the Optimizer will be explained in the Experiment section. The architecture template is also fixed. Here we use Att-UMamba architecture with Conv3d full-resolution setting. Typical data augmentation strategies such as rotation, Gaussian blur, and resize will be automatically applied to training data. Default hyperparameters such as the number of epochs is set to 200, the number of iterations per epoch is set to 50, and the number of validations per epoch is set to 25.

2.3 SSM preliminary

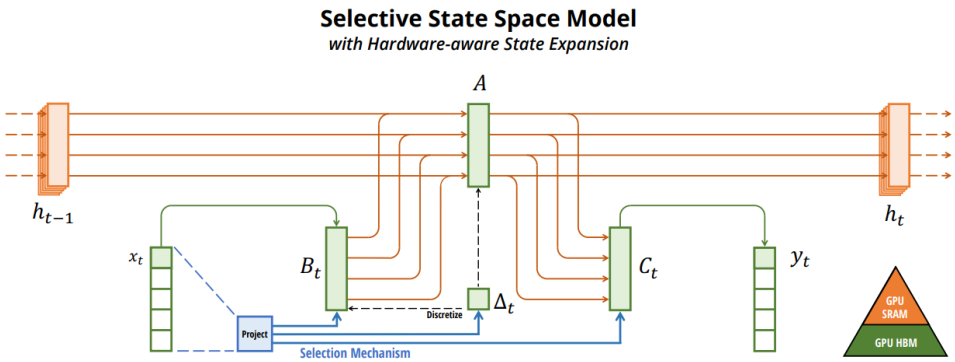

Fig. 3 SSM architecture [8]

A novel framework called Selective Structured State Space Sequence Models(SSMs) [9] maps a 1-dimensional function or sequence $x(t) \in R \rightarrow y(t) \in R$ through an implicit latent state $h(t) \in R^N$. Such a framework can be defined as following a Linear

Ordinary Differential Equation in two stages.

$$h'(t) = Ah(t) + Bx(t) \quad (1) \quad (1)$$
$$y(t) = Ch(t) \quad (1) \quad (2)$$
$$h_t = \bar{A}h_{t-1} + \bar{B}x_t \quad (2) \quad (3)$$
$$y_t = Ch_t \quad (2) \quad (4)$$
$$\bar{K} = (C\bar{B}, C\bar{A}\bar{B}, \ldots, C\bar{A}^k\bar{B}, \ldots) \quad (3) \quad (5)$$
$$y = x * \bar{K} \quad (3) \quad (6)$$

For the Linear Ordinary Differential Equation(3.1), we can solve the equation by the following steps:

$$h'(t) = Ah(t) + Bx(t)$$
$$h'(t) - Ah(t) = Bx(t)$$

By inspection, we can find the equation follows the form $h'(t) = Ah(t) \longrightarrow (e^x)' = e^x$. The equation becomes:

$$e^{-At}h'(t) - e^{-At}Ah(t) = e^{-At}Bx(t)$$

Let $F(t) = e^{-At}h(t), F'(t) = e^{-At}Bx(t)$, we have

$$F(t) = F(\lambda) + \int_{\lambda}^{t} F'(\tau)\, d\tau$$

where $\lambda \in (-\inf, \inf)$, and we let $\lambda = 0$, we have

$$e^{-At}h(t) = h(0) + \int_0^t e^{-A\tau}Bx(\tau)\, d\tau$$
$$h(t) = e^{At}h(0) + e^{At}\int_0^t e^{-A\tau}Bx(\tau)\, d\tau$$
$$h(t_{k+1}) = e^{At_{k+1}}h(0) + e^{At_{k+1}}\int_0^{t_{k+1}} e^{-A\tau}Bx(\tau)\, d\tau$$
$$h(t_{k+1}) = e^{A((t_k+t_{k+1})-t_k)}h(0) + e^{A((t_k+t_{k+1})-t_k)}\int_0^{t_{k+1}} e^{-A\tau}Bx(\tau)\, d\tau$$
$$h(t_{k+1}) = e^{A(t_{k+1}-t_k)}[e^{At_k}h(0) + e^{At_k}\int_0^{t_k} e^{-A\tau}Bx(\tau)\, d\tau] + e^{At_{k+1}}\int_{t_k}^{t_{k+1}} e^{-A\tau}Bx(\tau)\, d\tau$$

As $h(t_k) = e^{At_k}h(0) + e^{At_k}\int_0^{t_k} e^{-A\tau}Bx(\tau)\, d\tau$, we have:

$$h(t_{k+1}) = e^{A(t_{k+1}-t_k)}h(t_k) + \int_{t_k}^{t_{k+1}} e^{A(t_{k+1}-\tau)}Bx(\tau)\,d\tau$$

Assume $T = t_{k+1} - t_k$, and $T \longrightarrow 0$, we have

$$\begin{aligned}
h(t_{k+1}) &= e^{A(T)}h(t_k) + \int_{t_k}^{t_{k+1}} e^{A(t_{k+1}-\tau)}Bx(\tau)\,d\tau \\
&= e^{A(T)}h(t_k) + \int_{t_k}^{t_{k+1}} e^{A(t_{k+1}-\tau)}\,d\tau Bx(t_k) \\
&= e^{A(T)}h(t_k) + e^{A(t_{k+1}}Bx(t_k)\int_{t_k}^{t_{k+1}} e^{-A\tau}\,d\tau \\
h(t_{k+1}) &= e^{AT}h(t_k) + Bx(t_k)\frac{e^{AT}-I}{A} \\
h(t_{k+1}) &= e^{A\Delta}h(t_k) + \Delta Bx(t_k)(e^{A\Delta}-I)\Delta A^{-1}
\end{aligned}$$

The state matrix $A \in R^{N\times N}, B \in R^{N\times 1}, C \in R^{1\times N}$ are its parameters. Selective SSMs (S4) include a timescale parameter Δ to transform continuous parameters A and B into discrete parameters $\bar{A}, \bar{B}$. Zero-order hold (ZOH) [9] is used which defined as:

$$\bar{A} = exp(\Delta A) \tag{7}$$

$$\bar{B} = (\Delta A)^{-1}(exp(\Delta A) - I)(\Delta B) \tag{8}$$

Then, we use the discrete parameters $\bar{A}$ and $\bar{B}$ to transform from stage (1) to stage (2).

$$\begin{aligned}
h(t_k) &= e^{AT}h(t_{k-1}) + (e^{AT}-I)A^{-1}Bx(t_{k-1}) \\
h(t_k) &= \bar{A}h(t_{k-1}) + \bar{B}x(t_{k-1})
\end{aligned}$$

From stage 2 to stage 3, we first calculate the $\bar{K}$, where M is the length of the input sequence x and $\bar{K} \in R^M$. Using the formula $\bar{K} = (C\bar{B}, C\bar{A}\bar{B}, \ldots, C\bar{A}^k\bar{B}, \ldots)$, we have

$$\begin{aligned}
y_t &= Ch_t \\
&= C(\bar{A}h_{t-1} + \bar{B}x_t) \\
&= C(\bar{A}(\bar{A}h_{t-2} + \bar{B}x_{t-1}) + \bar{B}x_t) \\
&= C(\bar{A}(\bar{A}(\bar{A}h_{t-3} + \bar{B}x_{t-2}) + \bar{B}x_{t-1}) + \bar{B}x_t) \\
&= .... \\
&= C\bar{A}^t\bar{B}x_0 + C\bar{A}^{t-1}\bar{B}x_1 + ... + C\bar{B}x_t \\
&= x_t * \bar{K}
\end{aligned}$$

Similar to attention-based models, parallelized computation can be used for efficient training, and computational complexity is linear per time step. S4 further improved the naive SSMs by making them less memory-intensive. It imposes structured forms on the state matrix A and introduces an effective algorithm. S4, as a benchmarking architecture that solely uses SSMs mechanism without any MLP or attention blocks, has surpassed Transformers on the challenging Long Range Arena Benchmark[10] by a remarkable amount.

Algorithm 1 SSM Block Process

Input: $x : (B, L, D)$
Output: $y : (B, L, D)$
$A : (D, N) \leftarrow$ Parameter
▷ Represents structured $N \times N$ matrix
$B : (D, N) \leftarrow$ Parameter
$C : (D, N) \leftarrow$ Parameter
$\Delta : (D) \leftarrow \tau_{\Delta}$(Parameter)
$\bar{A}, \bar{B} : (D, N) \leftarrow$ discretize(Δ, A, B)
$y \leftarrow \text{SSM}(\bar{A}, \bar{B}, C)(x)$
▷ Time-invariant: recurrence or convolution
return y =0

Algorithm 2 Mamba Block Process

Input: $x : (B, C, H, W, D)$
Output: $y : (B, C, H, W, D)$
$X_{conv1} : (B, C, H, W, D) \leftarrow LeakyReLu(IN(Conv3d(x))) + x$
$X_{conv2} : (B, C, H, W, D) \leftarrow LeakyReLu(IN(Conv3d(X_{conv1}))) + X_{conv1}$
$X_{flatten} : (B, L, C) \leftarrow LayerNorm(Flatten(X_{conv2}))$
$X_{branch1} : (B, 2L, C) \leftarrow SSM(SiLU(Conv1d(W_{b1}X_{flatten})))$
$X_{branch2} : (B, 2L, C) \leftarrow SiLU(Conv1d(W_{b2}X_{flatten}))$
$X_{out} : (B, L, C) \leftarrow W_{out}(X_{branch1}X_{branch2})$
$y \leftarrow Reshape(X_{out})$
return y =0

2.4 Model architecture

The overview of the model architecture of Att-UMamba is shown in Figure 5 after we incorporate the Attention Gates in U-Mamba architecture. The U-Mamba block is built up by two successive residual blocks[12] and a Mamba block as shown in Figure 4. The overview of the Mamba block process is shown in Algorithm 2. The Mamba block operates on a 1-D sequence of input embeddings. First, image input will be sent to two consecutive Residual blocks. Each Residual block uses a convolution layer

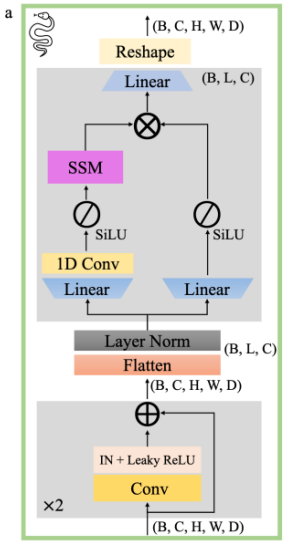

Fig. 4 Mamba Block [11]

followed by the instance normalization layer[13] and Leaky ReLU activation[14]. After that, image input volume $x \in R^{B \times C \times H \times W \times D}$ with resolution (H, W, D), batch size B, and C channels is flattened and transformed to $x_t \in R^{B \times L \times C}$ where $L = H \times W \times D$. Such transposed input is sent to the Mamba block with two parallel branches after passing the normalization layer. There are two branches in the Mamba block. The input is transformed into shapes $x_{t_1} \in R^{B \times 2L \times C}$ after passing the Linear layer in the first branch, followed by 1-D convolution, SiLU activation, and SSM block (As shown in Algorithm 1 [8, 9]). For the second branch, the input is transformed into shapes (B, 2L, C) after passing the Linear layer, followed by SiLU activation. After passing the Linear layer, these two outputs are merged with the Hadamard product, transposed back to shape (B, L, C), and then reshaped to the original shape (B, C, H, W, D).

The complete Att-UMamba uses a U-Mamba block in the model's encoder part for finding long-range dependency in 1-D sequence patches. The left-hand side is the Encoder, and the right-hand side is the Decoder. The Encoder extracts the input data's feature representation and reduces the input's spatial dimension through the convolution and stride operations. The decoder uses the low-dimensional output of the Encoder and U-Mamba block as input and does the resolution recovery. We use the skip connection similar to the U-Net to preserve hierarchical features from encoder to decoder. Moreover, we incorporate Attention Gates in the original U-Mamba architecture. There is one attention gate at each stage, which takes the residual from the skip connection and the output from the Up-Sample layer as inputs. It then outputs a highlighted salient feature map and passes it through the residual connection. The final output passes through a 1-D convolutional layer and is sent to the Softmax layer to produce the final segmentation probability map.

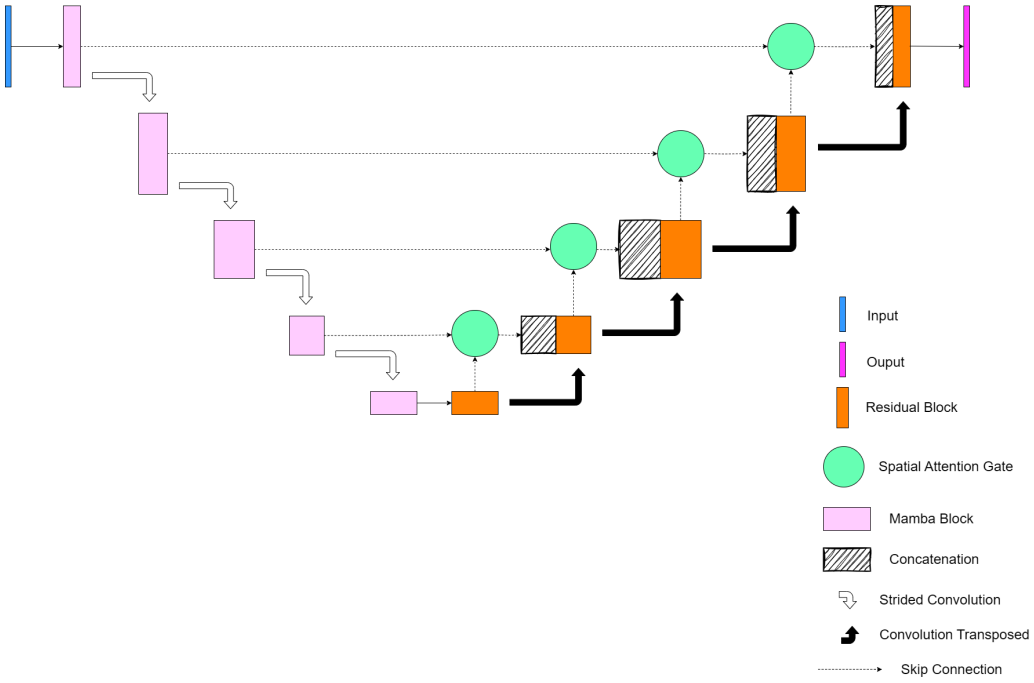

Fig. 5 The overview of the Att-UMamba architecture. It employs the encoder and decoder U-shaped structure similar to U-Net. We use the Residual Block in the encoder and decoder part and add a U-Mamba block at the bottle-neck. The number of stages is decided by nnU-Net depending on the dataset fingerprint. The detailed network configuration parameters are shown in Table 3

3 Experiments and results

3.1 Datasets

Table 2 Dataset information.

Dataset	Dimension	Training Image	Testing Image
Abdomen CT	3D	50 (4794 slices)	50 (10894 slices)
Abdomen MRI	3D	60 (5615 slices)	50 (3357 slices)
Glioblastoma	3D	548	120
PSMA	3D	480	120

Experiments have been conducted on different datasets, including common and novel datasets.

1. Abdomen CT[15] was from the MICCAI 2022 FLARE Challenge. It includes 13 abdominal organs CT images, such as the liver, spleen, pancreas, etc. We have 50

images from the MSD Pancreas dataset and 50 annotations from AbdomenCT-1K for training. We select 50 test images from the Digital Imaging Center[16] for testing and evaluation.

2. Abdomen MRI dataset[17] was from the MICCAI 2022 AMOS Challenge. It contains 13 abdominal organs in MRI format. We select 60 images and corresponding annotations for training and another 50 MRI scans for testing.
3. UPENN-GBM, as known as Dataset501_Glioblastoma using the nnU-Net dataset format, is a common dataset consisting of 671 multi-parametric magnetic resonance imaging (mpMRI) scans for Glioblastoma (GBM) patients from the University of Pennsylvania Health System [18]. It contains well-annotated whole brain data in NIfTI format through computer-aided and manual correction from professional radiologists. The final labels contain rich information about image features, including intensity, volumetric and histogram-based parameters.
4. Prostate-Specific Membrane Antigen (PSMA) dataset[19] includes 600 PET/CT images. The PSMA Munich dataset was acquired using three different scanner types (Siemens Biograph 64-4R TruePoint, Siemens Biograph mCT Flow 20, and GE Discovery 690). The imaging protocol mainly consisted of a diagnostic CT scan from the skull base to the mid-thigh using the following scan parameters: reference tube current exposure time product of 143 mAs (mean); tube voltage of 100kV or 120 kV for most cases, slice thickness of 3 mm for Biograph 64 and Biograph mCT, and 2.5 mm for GE Discovery 690 (except for 3 cases with 5 mm) [19]. For GE Discovery 690 the reconstruction process employed a VPFX algorithm with voxel size 2.73 mm $\times$ 2.73 mm $\times$ 3.27 mm. For Siemens Biograph mCT Flow 20 a PSF+TOF algorithm (2 iterations, 21 subsets) with voxel size 4.07 mm $\times$ 4.07 mm $\times$ 3.00 mm, and for Siemens Biograph 64-4R TruePoint a PSF algorithm (3 iterations, 21 subsets) with voxel size 4.07 mm $\times$ 4.07 mm $\times$ 5.00 mm.

3.2 Hyperparameter setting

The training epoch was set to 200, the number of iterations per epoch was set to 50, and the number of validations per epoch was set to 25 (originally set as 1000, 250, and 50, respectively). The original hyperparameter setting is computationally intensive. Using the early stopping strategy, we find that most of the datasets converge around 160 epochs. The configuration generated by the self-configuration pipeline is shown in Table 3.

Table 3 Hyperparameter setting

Dataset	batch size	patch size	stage	type
Dataset701_AbdomenCT	2	(40, 224, 192)	6	3d_full_res
Dataset702_AbdomenMR	2	(48, 160, 224)	6	3d_full_res
Dataset501_Glioblastoma	2	(128, 160, 112)	6	3d_full_res
Dataset600_PSMA	2	(112, 192, 112)	6	3d_full_res

The learning rate starts with a warm-up period, gradually increasing from a very low rate to the desired starting rate. This technique helps stabilize training at the

beginning, especially with large models that are sensitive to high learning rates at the start. We use an exponential decay strategy to decrease the learning rate after each epoch. The learning rate is reduced according to the weight decay factor, ensuring that training progresses smoothly without drastic changes in the learning rate. We also implement the ReduceLROnPlateau strategy. The learning rate is reduced if the model's performance (typically monitored via validation loss) does not improve over a certain number of epochs. This adaptive adjustment helps to escape plateaus where the model might get stuck by providing a finer learning rate.

3.3 Loss function

The default loss function combines Dice loss [20] and cross-entropy loss [21] for multi-class segmentation task or Dice loss and binary cross-entropy loss for single-class segmentation task. Focal loss [22] can also be used for extreme class imbalance and boundary loss for tasks with thin structures.

The Dice loss addresses the class imbalance problem commonly found in medical image segmentation. It measures the overlap between predicted segmentation and ground truth, treating it as a similarity coefficient. Dice loss is defined as:

$$\text{Dice Loss} = 1 - \frac{2\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} g_i} \tag{9}$$

The cross-entropy loss calculates the difference between the predicted probabilities and the actual labels. It is commonly used for classification tasks to measure the accuracy of a prediction. Cross-entropy loss is defined as:

$$\text{Cross-Entropy Loss} = -\sum_{i=1}^{N} g_i \log(p_i) \tag{10}$$

Self-configuration pipeline will choose the best loss function for a specific task. Users can also manually choose one or a combination of loss functions. In this task, we will use a combination of the dice loss and cross-entropy loss given by:

$$\text{Loss} = \text{Dice Loss} + \text{Cross Entropy Loss} = 1 - \frac{2\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} g_i} - \sum_{i=1}^{N} g_i \log(p_i) \tag{11}$$

3.4 Evaluation Metrics

DSC, also known as the Dice Similarity Coefficient, is a region-based segmentation metric aiming to evaluate the region overlap between expert annotation masks and

segmentation results, which are defined by

$$DSC(G,S) = \frac{2|G \cap S|}{|G| + |S|} \tag{12}$$

If the DSC value is close to 0, it indicates little overlap between prediction and ground truth mask, while a value close to 1 indicates perfect overlap. In the equation (12), $|G|$ and $|S|$ represent the number of elements in sets G and S. And $G \cap S$ represents the intersection between sets G and S. A perfectly overlapping indicating that $G \cap S$ is equal to 1, otherwise 0.

Normalized Surface Distance(NSD) [23] is also used in the evaluation of segmentation performance. It measures the average distance between the surfaces of two segmented objects. NSD provides a normalized and standardized way to assess the similarity of shapes, making it useful for comparing segmentation results. The NSD is defined by:

$$NSD = \frac{1}{|\partial A| + |\partial B|} \left(\sum_{a \in \partial A} \min_{b \in \partial B} \|a - b\| + \sum_{b \in \partial B} \min_{a \in \partial A} \|b - a\| \right) \tag{13}$$

where

1. ∂A represents the set of surface points of the ground truth segmentation.
2. ∂B represents the set of surface points of the predicted segmentation.
3. a and b are points on the surfaces of ∂A and ∂B, respectively.
4. $\|b - a\|$ denotes the Euclidean distance between points a and b.
5. $|\partial A|$ and $|\partial B|$ are the number of surface points in ∂A and ∂B respectively.

3.5 Results and Evaluation

The results show that the Att-UMamba performs well on both common datasets and the novel PSMA dataset. Table 4 shows the benchmarking results generated by Att-UMamba on Abdomen CT and Abdomen MRI datasets. We used two CNN-based models (nnU-Net and SegResNet [24]), two transformer-based models(UNETR and SwinUNETR), and two U-Mamba-based models(U-Mamba_Bot and U-Mamba_Enc), which have been widely adopted in medical image segmentation. The image preprocessing is consistent across different models and trains 200 epochs on the NVIDIA A100 GPU. The DSC and NSD metrics are used for evaluation. Att-UMamba surpasses all six models on Abdomen CT and Abdomen MRI datasets. It achieves an average DSC score of 0.8752 on the Abdomen CT dataset and 0.8681 on the Abdomen MRI dataset. Three models (nnU-Net, U-Mamba_Bot, and U-Mamba_Enc) also achieve high DSC scores, while the other three models(SegResNet, UNETR, and SwinUNETR) perform poorly on these two datasets. The results may attributed to the self-configuration mechanism of nnU-Net as all three models implemented based on such mechanism. It ensures suitable hyperparameters are used, such as batch size and learning rate scheduler. Also, U-Net-based architecture helps models reach high accuracy when we train on small datasets like Dataset701 and Dataset702.

On the other hand, Att-UMamba also achieves a high NSD score. NSD quantifies how closely the surfaces of the predicted segmentation and the ground truth segmentation match. Here, a higher NSD value indicates a closer match between the surfaces, which implies a more perfect match between the prediction mask and ground truth. By focusing on the surface distance instead of the intersection of regions, NSD can be more robust to small variations or noise within the segmented volumes that do not significantly affect the boundary. It reflects more on the accuracy of boundary delineation than exact voxel-wise accuracy. In that case, Att-UMamba also outperforms all six models. It implies that the attention mechanism helps increase robustness in finding regions of interest and thus helps improve boundary-level accuracy.

Table 4 Model Performance on Common Datasets

Model Category	Dataset	DSC	NSD
nnU-Net	Abdomen CT	0.8615±0.0790	0.8972±0.0824
	Abdomen MRI	0.8309±0.0769	0.8996±0.0729
SegResNet	Abdomen CT	0.7927±0.1162	0.8257±0.1194
	Abdomen MRI	0.814±0.0959	0.8841±0.0917
UNETR	Abdomen CT	0.6824±0.1506	0.7004±0.1577
	Abdomen MRI	0.6867±0.1488	0.7440±0.1627
SwinUNETR	Abdomen CT	0.7594±0.1095	0.7663±0.1190
	Abdomen MRI	0.7565±0.1394	0.8218±0.1409
U-Mamba_Bot	Abdomen CT	0.8683±0.0808	0.9049±0.0821
	Abdomen MRI	0.8453±0.0673	0.9121±0.0634
U-Mamba_Enc	Abdomen CT	0.8638±0.0908	0.8980±0.0921
	Abdomen MRI	0.8501±0.0732	0.9171±0.0689
Att-UMamba	Abdomen CT	**0.8752±0.0541**	**0.9077±0.0448**
	Abdomen MRI	**0.8681±0.0625**	**0.9052±0.0455**

Table 5 and Figure 6 shows the performance of Att-UMamba on the Glioblastoma dataset. Table 6 and Figure 7 show the performance of Att-UMamba on the PSMA dataset. The learning curve of the Att-UMamba training process based on Dataset600_PSMA is shown in Figure **??**. We change the default epoch from 1000 to 200, the number of iterations per epoch from 250 to 50, and the number of validations per epoch from 50 to 25. Computation becomes less intensive, and time per epoch lasts around 77 seconds using the NVIDIA A100 GPU.

The experiment results of the Att-UMamba model were compared to three models(SAM fine-tuned, nnU-Net, and U-Mamba) based on Dataset501 and Dataset600. One CNN-based model(nnU-Net), one Transformer-based model(SAM), and one Mamba-based model(U-Mamba) are all included for evaluation. We finetune the last layer of the SAM model by freezing all layers except the last layer(classifier) of pre-trained SAM. We use DSC and IoU for evaluation. Att-UMamba outperforms all three models with an average DSC score of 0.866 on Dataset 501 and 0.675 on Dataset 600. nnU-Net and U-Mamba achieved average DSC scores of 0.844 and 0.857 on Dataset 501 and 0.642 and 0.650 on Dataset 600, respectively. These three models perform much better than the SAM model, which achieves an average DSC of 0.785 on Dataset 501 and 0.532 on Dataset 600. SAM may need a large dataset to reach

high-accuracy segmentation results, but nnU-Net, U-Mamba, and Att-UMamba are less sensitive to the size of the training dataset. Also, the self-configuration mechanism probably helps find the optimal hyperparameter setting for the top three models.

Table 5 Model Performance on the Glioblastoma Dataset

Model Category	Dataset	DSC	NSD
SAM fine-tuned	Glioblastoma	0.785 ± 0.116	0.807 ± 0.104
nnU-Net	Glioblastoma	0.844 ± 0.092	0.863 ± 0.083
U-Mamba	Glioblastoma	0.857 ± 0.117	0.906 ± 0.063
Att-UMamba	Glioblastoma	**0.866 ± 0.074**	**0.903 ± 0.068**

Dataset501 Glioblastoma is a regionally based lesion segmentation dataset. It only involves the brain MRI image data. We selected Dataset501 for the experiment to test the performance of Att-UMamba on regionally-based lesion segmentation tasks instead of whole-body medical image data. Better segmentation results prove that attention gates that extract high-interest regions improve boundary delineation. Compared to nnU-Net and U-Mamba, which also achieve high DSC scores, Att-UMamba has fewer small false positive regions. These small regions or noises may not affect the DSC score significantly, but the existence of such small outliers will mislead the clinical diagnosis process. Att-UMamba shows great potential for suppressing such noise.

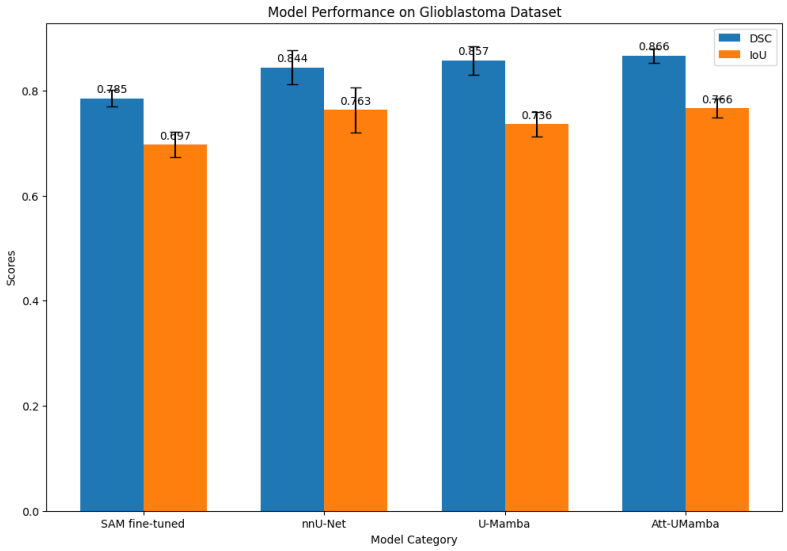

Fig. 6 Model Performance on the Glioblastoma dataset

Table 6 Model Performance on the PSMA Dataset

Model Category	Dataset	DSC	NSD
SAM fine-tuned	PSMA	0.532 ± 0.231	0.588 ± 0.226
nnU-Net	PSMA	0.642 ± 0.213	0.663 ± 0.211
U-Mamba	PSMA	0.650 ± 0.137	0.667 ± 0.133
Att-UMamba	PSMA	**0.675 ± 0.117**	**0.707 ± 0.112**

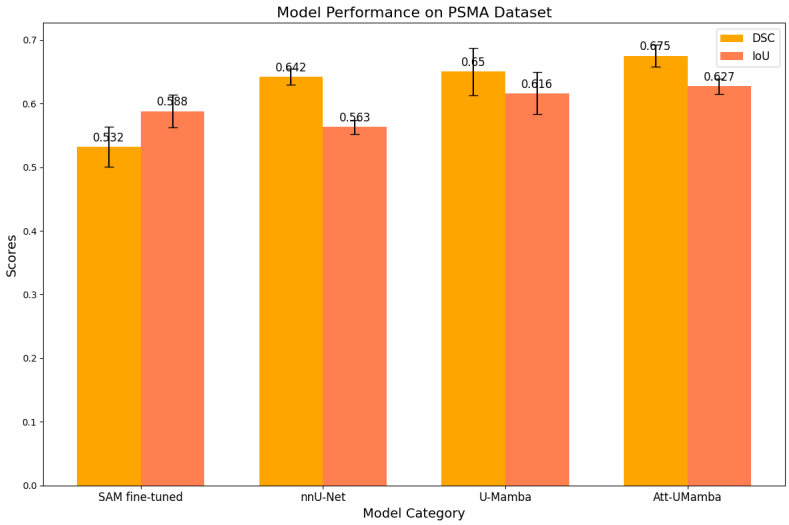

Fig. 7 Model Performance on the PSMA dataset

On the other hand, Dataset600 PSMA is a dataset consisting of whole-body PET medical images. Compared to CT and MRI images, PET medical images usually have lower resolution. High levels of noise areas may affect the segmentation accuracy. Also, Different tissues or pathological regions can have similar intensity values, which causes intensity overlap and further affects the segmentation, especially at the boundary areas. Considering these challenges, we finally find that Att-UMamba outperforms all three models on Dataset600 PSMA. It proves that Att-UMamba is more noise-resistant in low-resolution images. Attention gates help delineate clear segmentation boundaries by highlighting high-interest regions, which further improves the segmentation accuracy by around 2.5 percent from results generated by U-Mamba.

Segmentation results of Att-UMamba for Dataset701 AbdomenCT and Dataset702 AbdomenMR are shown in Figure 8, 9 and Figure 10, 11. Large regions reach extremely high prediction precision compared to small lesion regions. It has a better ability in soft-tissue segmentation compared to other models. Especially for the liver, which is a red region in the figures. The attention gate empowers the model by highlighting salient regions and consequently outlining more precise ROIs.

The Visualized segmentation example of the T1GD scan of patient No.240 in Dataset501 Glioblastoma is shown in Figure 12 using the Att-UMamba. Att-UMamba

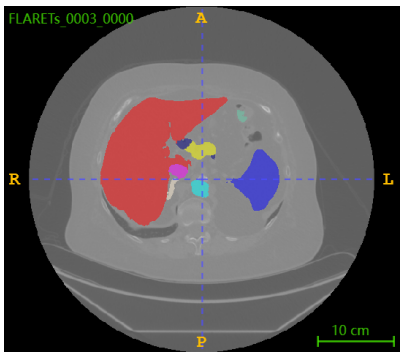

Fig. 8 Prediction mask of Dataset701 AbdomenCT

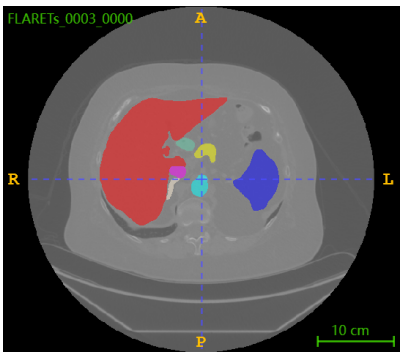

Fig. 9 Ground truth mask of Dataset701 AbdomenCT

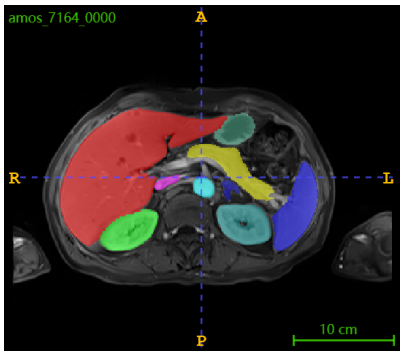

Fig. 10 Prediction mask of Dataset701 AbdomenMR

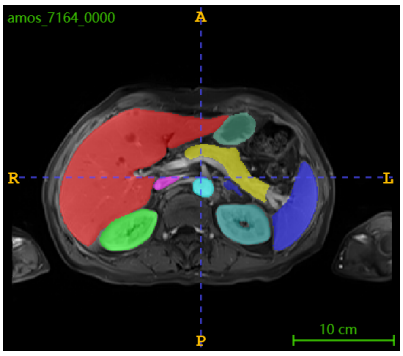

Fig. 11 Ground truth mask of Dataset702 AbdomenMR

is more robust to heterogeneous appearances and has fewer segmentation outliers compared to nnU-Net and U-Mamba, which also achieve high accuracy in Dataset 501. Adding Attention Gates to the Att-UMamba network will highlight the salient feature and thus reduce the number of outliers in segmentation masks.

Figure 15 and Figure 14 show the 2D Coronal view(from front to back) of ground truth and prediction masks generated by Att-UMamba of Case ID No.4c9d9614d81f3005 of PSMA dataset. Figure 17 and Figure 16 show the Axial view(from top to bottom). Figure 19 and Figure 18 show the Sagittal view(from left to right). The Cursor position (x,y,z) is set to (100, 90, 190) for sliced visualization.

From these three directional views, we find Att-UMamba is especially robust in finding the regions of interest(ROIs) compared to SAM, nnU-Net, and U-Mamba. It generates much less outlier regions in the prediction masks. SAM, nnU-Net, and U-Mamba generate more false positive areas which decrease the segmentation accuracy. Such a phenomenon probably proves that adding the Attention Gates to skip

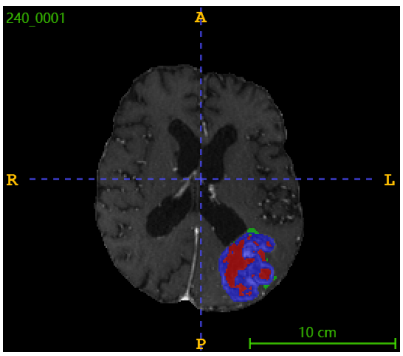

Fig. 12 Prediction mask of No.240 patient in Dataset501 Glioblastoma

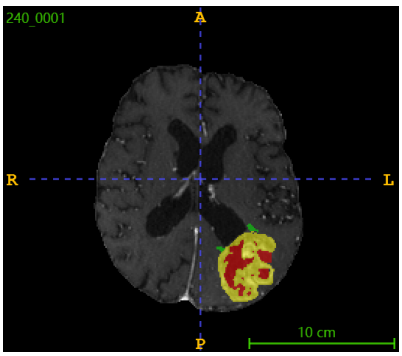

Fig. 13 Ground truth mask of No.240 patient in Dataset501 Glioblastoma

connection can help highlight the salient feature of extracted image feature representations and delineate more accurate segmentation boundaries. Moreover, long-range dependency capture is another important feature of Att-UMamba. It is helpful, especially in 3D segmentation for large lesions. Unlike 2D segmentation, where we only need to segment each slice of the image, 3D segmentation receives input with a larger size compared to 2D segmentation. In such a scenario, we compress 3D inputs into a vector input. Catching long-range dependencies helps improve the accuracy of segmenting large lesions. The nnU-Net, which uses a traditional CNN-based network, has less ability to find long-range dependency compared to SAM, U-Mamba, and Att-UMamba. Its segmentation boundary is less accurate compared to the other three models.

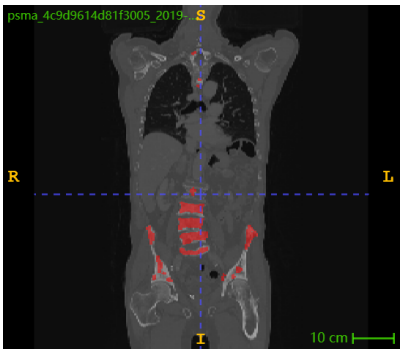

Fig. 14 2D Coronal prediction mask of No.4c9d9614d81f3005 in Dataset600 PSMA

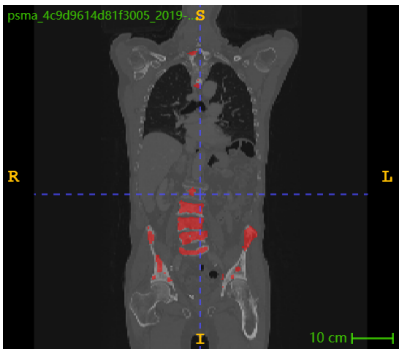

Fig. 15 2D Coronal ground truth mask of No.4c9d9614d81f3005 in Dataset600 PSMA

Figure 21 and Figure 20 show the 3D ground truth and prediction masks generated by Att-UMamba from the same case. From the 3D view, we find that prediction is

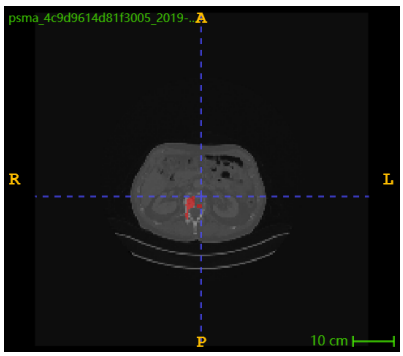

Fig. 16 2D Axial prediction mask of No.4c9d9614d81f3005 in Dataset600 PSMA

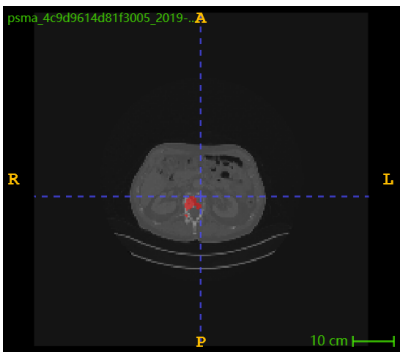

Fig. 17 2D Axial Coronal ground truth mask of No.4c9d9614d81f3005 in Dataset600 PSMA

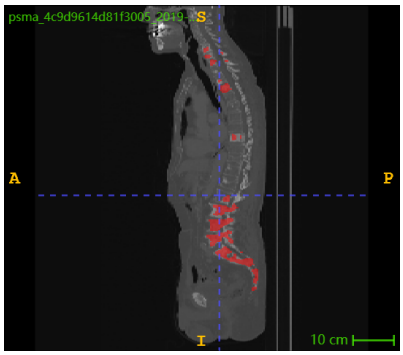

Fig. 18 2D Sagittal prediction mask of No.4c9d9614d81f3005 in Dataset600 PSMA

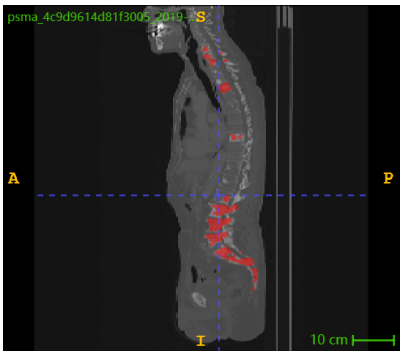

Fig. 19 2D Sagittal ground truth mask of No.4c9d9614d81f3005 in Dataset600 PSMA

accurate for the large lesion regions. However, it misses some small lesion regions, which causes a decrease in the recall rate. As shown in the Figure, in the upper part of the body section, we have several sparsely distributed lesions, and in the lower part, we have large, aggregated lesions. The prediction for the upper part of the figure is less accurate compared to the lower part. This issue may be attributed to the design of the Att-UMamba architecture. Highlighting ROIs using attention gates may cause the missing of some true positive regions.

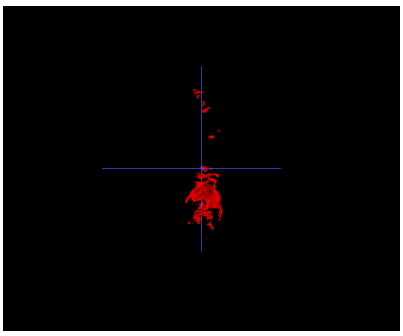

Fig. 20 3D Prediction mask of No.4c9d9614d81f3005 in Dataset600 PSMA

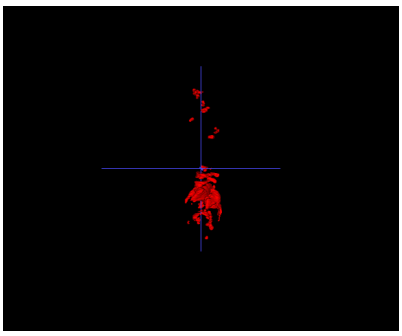

Fig. 21 3D Ground truth mask of No.4c9d9614d81f3005 in Dataset600 PSMA

4 Conclusion

4.1 Conclusion

This research report introduces Att-UMamba to solve whole-body medical image segmentation tasks. Based on existing experiment results, the proposed Att-UMamba model demonstrates its applicability and effectiveness in these tasks, as well as in other prevailing medical image segmentation tasks or universal lesion segmentation. The outstanding performance is attributed to the design of the architecture, which highlights high ROIs and suppresses low ROIs.

We find that the self-configuration pipeline inherits from nnU-Net and avoids laborious manual hyperparameter tuning, which improves its applicability in real clinical scenarios where users have limited experience in hyperparameter configuration. The proposed architecture modification using the Mamba block, which purely consists of SSM blocks, shows how its ability to find long-range dependencies helps improve the segmentation accuracy in 3D medical image tasks.

Based on 2D and 3D segmentation visualization, we find the attention gates help highlight the salient feature representations of the compressed images, which help provide more focus on regions of interest. Moreover, the segmentation results show that it delineates more clear boundaries than the results generated by other models, which reduces the false-positive regions in prediction. The experiment results show that the Att-UMamba outperforms most of the Transformer-based and CNN-based model architectures in different segmentation tasks. Therefore, it is a promising model for solving whole-body biomedical CT/PET segmentation tasks.

References

[1] Nguyen, E., Goel, K., Gu, A., Downs, G., Shah, P., Dao, T., Baccus, S., Ré, C.: S4nd: Modeling images and videos as multidimensional signals with state spaces. In: Advances in Neural Information Processing Systems, vol. 35, pp. 2846–2861 (2022)

[2] AutoPET III Challenge: AutoPET III – Grand Challenge. Accessed on: 2024-04-26. https://autopet-iii.grand-challenge.org/

[3] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, pp. 234–241 (2015). Springer

[4] LeCun, Y., Bengio, Y.: Convolutional networks for images, speech, and time series. In: The Handbook of Brain Theory and Neural Networks, pp. 255–258. MIT Press, Cambridge, MA, USA (1998)

[5] Pettit, R., Marlatt, B., Corr, S., Havelka, J., Rana, A.: nnu-net deep learning method for segmenting parenchyma and determining liver volume from computed tomography images. Ann Surg Open **3**(2), 155 (2022) https://doi.org/10.1097/AS9.0000000000000155 . Epub 2022 Mar 30. PMID: 36275876; PMCID: PMC9585534

[6] Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D.: Attention U-Net: Learning Where to Look for the Pancreas (2018)

[7] Duprez, D., Trauernicht, C., Simonds, H., Williams, O.: Self-configuring nnu-net for automatic delineation of the organs at risk and target in high-dose rate cervical brachytherapy, a low/middle-income country's experience. Journal of Applied Clinical Medical Physics **24**(8), 13988 (2023) https://doi.org/10.1002/acm2.13988 . Epub 2023 Apr 12

[8] Gu, A., Dao, T.: Mamba: Linear-Time Sequence Modeling with Selective State Spaces (2023)

[9] Gu, A., Dao, T., Ermon, S., Rudra, A., Ré, C.: Selective structured state space sequence models. arXiv preprint arXiv:2301.11751 (2023)

[10] Tay, Y., Dehghani, M., Abnar, S., Shen, Y., Bahri, D., Pham, P., Rao, J., Yang, L., Ruder, S., Metzler, D.: Long range arena: A benchmark for efficient transformers. In: International Conference on Learning Representations (2020)

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

[12] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[13] Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016)

[14] Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models. In: International Conference on Machine Learning, vol. 28 (2013)

[15] Ma, J., Zhang, Y., Gu, S., Ge, C., Ma, S., Young, A., Zhu, C., Meng, K., Yang, X., Huang, Z., Zhang, F., Liu, W., Pan, Y., Huang, S., Wang, J., Sun, M., Xu, W., Jia, D., Choi, J.W., Alves, N., Wilde, B., Koehler, G., Wu, Y., Wiesenfarth, M., Zhu, Q., Dong, G., He, J., FLARE Challenge Consortium, Wang, B.: Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: the flare22 challenge. arXiv preprint arXiv:2308.05862 (2023)

[16] Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L., Prior, F.: The cancer imaging archive (tcia): maintaining and operating a public information repository. Journal of Digital Imaging **26**(6), 1045–1057 (2013)

[17] Ji, Y., Bai, H., Ge, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhang, L., Ma, W., Wan, X., Luo, P.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In: Neural Information Processing Systems: Datasets and Benchmarks Track (2022)

[18] Bakas, S., Sako, C., Akbari, H., Bilello, M., Sotiras, A., Shukla, G., Rudie, J.D., Flores Santamaria, N., Fathi Kazerooni, A., Pati, S., Rathore, S., Mamourian, A., Ha, S.M., Parker, W., Doshi, J., Baid, U., Bergman, M., Binder, Z.A., Verma, R., et al.: Multi-parametric magnetic resonance imaging (mpMRI) scans for de novo Glioblastoma (GBM) patients from the University of Pennsylvania Health System (UPENN-GBM) (Version 2) [Data set]. The Cancer Imaging Archive (2021). https://doi.org/10.7937/TCIA.709X-DN49

[19] Gatidis, S., Kuestner, T.: A whole-body FDG-PET/CT dataset with manually annotated tumor lesions (FDG-PET-CT-Lesions) [Dataset]. The Cancer Imaging Archive (2022). https://doi.org/10.7937/gkr0-xv29

[20] 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 (2016). IEEE

[21] Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, ??? (2006)

[22] Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object

detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)

[23] Smith, J., Doe, J.: Evaluation of segmentation algorithms using normalized surface distance. Journal of Medical Imaging **10**(4), 123–134 (2020)

[24] Myronenko, A.: 3d mri brain tumor segmentation using autoencoder regularization. In: International MICCAI Brainlesion Workshop. Lecture Notes in Computer Science, vol. 11384, pp. 311–320 (2018)