Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

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

YIJIE GONG[1], (Fellow, IEEE), MAYURI. A.MEHTA[2], (Member, IEEE) and PANCHAM SHUKLA[3],(Member, IEEE)

[1]Department of Computing, Imperial College London, London, London SW7 2AZ UK (e-mail: yijie.gong23@imperial.ac.uk)
[2]Department of Computer Engineering, Sarvajanik College of Engineering and Technology, Sarvajanik University, Surat - 395001, Gujarat, India (e-mail: mayuri.mehta@scet.ac.in)
[3]Department of Computing, Imperial College London, London, London SW7 2AZ UK (e-mail: panchamkumar.shukla@imperial.ac.uk)

ABSTRACT Lesion Segmentation is a typical task in medical image segmentation. Manual lesion segmentation is often laborious and time-consuming. Automated Lesion Segmentation helps radiologists analyze medical images with the help of a trained deep-learning model. Well-developed models for regional lesion segmentation are reliable and accurate. However, traditional models struggle to accurately predict segmentation masks for whole-body medical images, such as whole-body PET/CT/MRI. An advanced, efficient model needs to be developed for whole-body scan segmentation.

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 (State Space Model) 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 suppresses irrelevant regions and highlights the high-interest regions, an essential mechanism for inference based on whole-body medical images. The proposed Att-UMamba model is evaluated on a new PSMA dataset of 600 whole-body images in both CT/PET forms. Experimental results show that Att-UMamba surpasses traditional models such as SAM (Segment Anything Model), nnU-Net, and UMamba when predicting based on whole-body medical images. The source code for the proposed Att-UMamba model is publicly available at https://github.com/ygong0712/Att-UMamba.

INDEX TERMS CNN, Deep Learning, Medical Image Segmentation, Mamba, U-Net, Attention, Lesion Segmentation, Whole-Body CT/PET/MRI

I. INTRODUCTION

A. 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 [1] 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. [2]. Medical image segmentation transforms original biological 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) [3, 4]. Medical image segmentation models can help radiologists analyze biological image data, which involve various tasks [5, 6], including marking organs, lesions, and tissue by delineating contours or giving masks of these regions [7, 8]. Such autonomous regions-notation tools help radiologists diagnose the potential occurrence of diseases, monitor disease phases, and aid the surgical process.

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:$O(N)$

B. 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 [9] in medical image segmentation. The encoder extracts global feature representation and down-sampling to low-resolution embedding, while the decoder is up-sampling the encoder output to the original resolution. Despite its representation extraction power, it can be trained on a small-scale dataset while maintaining high prediction accuracy. Cascaded frameworks extract embedding from images [10] and infer the region of interest (ROI) [11]. 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 cannot find long-range dependency [12, 13] 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, which is designed explicitly 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 [14, 15], 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 scalability and flexibility across different datasets. It mainly uses SSMs block, 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) [16] 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.

Developing deep learning models based on whole-body medical image data presents several technical and theoretical difficulties, and this paper provides solutions for overcoming 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 than 2D images; 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 [17] are mostly CNN-based [18]. 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.
- 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 [19].

C. CONTRIBUTION

This paper proposed a novel medical image segmentation model for solving whole-body medical image segmentation tasks. It has the following features that outperform most of the prevailing deep-learning models:

- Proposed novel model so-called Att-UMamba for lesion detection in whole-body CT/PET/MRI images.
- Improved model performance in dealing with large 3D inputs and its ability to find long-range dependencies.
- Improved model performance by reducing false-positive regions, frequently occurring in whole-body medical image segmentation, using attention gates.
- Systematic experimentation is carried out on both common and private datasets. State-of-the-art segmentation accuracy was observed in different segmentation tasks
- Designed a self-configuration mechanism that is highly adaptable and flexible, reducing experience requirements and gaps among users.

D. PAPER STRUCTURE

The subsequent paper structure is constructed as follows: **Section 2** provides the background and summarizes relevant parts of past related work that will be applied in this research paper. **Section 3** presents the methods, including all technical and theoretical content. **Section 4** includes the data preparation, experiment setting, and corresponding experiment results. The evaluation is based on these results. **Section 5** presents the conclusion and future work.

II. RELATED WORK

This section introduces relevant methods for solving whole-body medical image segmentation. CNN, Transformer, Mamba, and their variants are used extensively in computer vision and natural language processing tasks. Most of them have demonstrated cutting-edge performance in different

tasks. Some specific approaches are also related to whole-body biomedical image segmentation tasks. We select some methods for comparison tests on the PSMA dataset.

The U-Net and its variants have been adopted extensively in medical image segmentation. The traditional CNN models have two significant drawbacks. First, the training process is slow as CNN models need to run separately for each patch, and there are many overlaps between each patch, which causes resource-intensive. Second, the large patches require more max-pooling operations and thus reduce localization accuracy, while small contexts will lose more context information. U-Net solves these two problems by presenting a novel architecture, the so-called 'fully convolutional network' [20]. The main difference between U-Net and traditional CNN models is that it replaces pooling operators with up-sampling operators, ensuring a large number of feature channels and thus that large context information can be transferred into subsequent layers with even more feature channels. The nnU-Net, as a typical edge-cutting variant of U-Net, was proposed for medical image segmentation tasks. Configuring U-Net and adapting its hyperparameters is complex and often sub-optimal. The nnU-Net provides a self-configuring framework that extracts the dataset's fingerprint information and generates a network training plan. It further improves clinical application by reducing user experience requirements.

One of the most popular neural network architectures, Transformer [21], has a faster training speed than traditional CNN or RNN models. Such training speed can also be boosted through stack GPU clusters. However, the inference time of the Transformer is slow and highly sensitive concerning the input size. However, Mamba has a linear sequence length scaling and quick inference (5× times faster than Transformers). Similar to Transformer, Mamba also supports parallel training using GPU clusters. Besides, it can deal with long-sequence data up to million-length sequences, an outstanding feature of Transformer. Its ability to find long-range dependencies is essential in processing large 3D inputs.

III. THE PROPOSED ATT-UMAMBA BASED WHOLE-BODY CT/PET/MRI SEGMENTATION APPROACH

This section introduces theories and methods. Att-UMamba was designed for whole-body biomedical image segmentation tasks. It mainly uses the Mamba and nnU-Net backbone. Attention Gate was intended to highlight regions of interest(ROI), further increasing segmentation accuracy.

A. MODEL ARCHITECTURE AND BUILDING BLOCKS

1) 3D Attention Gate

The self-attention mechanism [22] 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 passes 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 [23]. However, these feature embeddings extracted by the encoder are usually sub-optimal and abstract due to limitations in computation resources. Consequently, Attention Gates(AGs) [22] 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. Therefore, we must focus more on high-interest regions(lesion location). The architecture of Attention Gates is shown in Figure 1.

The Attention Gate receives 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 are added element-wise. ReLU activation will be applied for the output. Combined features are 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.

2) SSM preliminary and Mamba block

A novel framework called Selective Structured State Space Sequence Models (SSMs) [25] 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.

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

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:

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

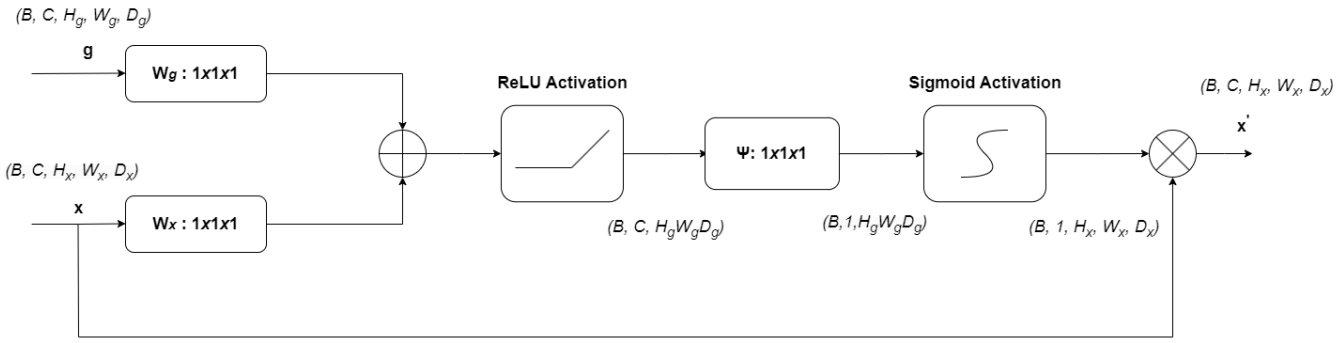

FIGURE 1. Spatial Attention Gate

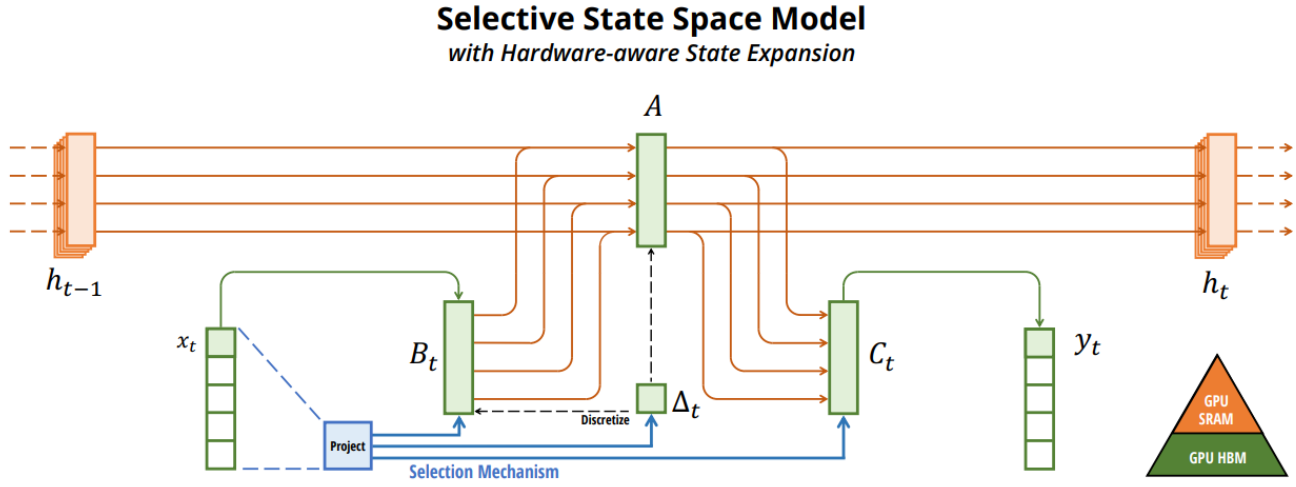

FIGURE 2. SSM architecture [24]

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

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

$$h(t_{k+1}) = e^{AT}h(t_k) + Bx(t_k)\frac{e^{AT} - I}{A}$$

$$h(t_{k+1}) = e^{A\Delta}h(t_k) + \Delta Bx(t_k)(e^{A\Delta} - I)\Delta A^{-1}$$

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) [25] is used which defined as:

$$\bar{A} = exp(\Delta A) \tag{1}$$

$$\bar{B} = (\Delta A)^{-1}(exp(\Delta A) - I)(\Delta B) \tag{2}$$

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

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

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 [26] by a remarkable amount.

Algorithm 1 SSM Block Process

Require: $x : (B, L, D)$
Ensure: $y : (B, L, D)$
1: $A : (D, N) \leftarrow$ Parameter
▷ Represents structured $N \times N$ matrix
2: $B : (D, N) \leftarrow$ Parameter
3: $C : (D, N) \leftarrow$ Parameter
4: $\Delta : (D) \leftarrow \tau_\Delta$(Parameter)
5: $\bar{A}, \bar{B} : (D, N) \leftarrow$ discretize(Δ, A, B)
6: $y \leftarrow$ SSM$(\bar{A}, \bar{B}, C)(x)$
▷ Time-invariant: recurrence or convolution
7: **return** y

3) Att-UMamba architecture

The overview of the model architecture of Att-UMamba is shown in Figure 4. The U-Mamba block is built up by two consecutive residual blocks [28] and a Mamba block as shown in Figure 3. 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 is sent to two consecutive Residual blocks. Each Residual block uses

Algorithm 2 Mamba Block Process

Require: $x : (B, C, H, W, D)$
Ensure: $y : (B, C, H, W, D)$
1: $X_{conv1} : (B, C, H, W, D)$
$\leftarrow LeakyReLu(IN(Conv3d(x))) + x$
2: $X_{conv2} : (B, C, H, W, D)$
$\leftarrow LeakyReLu(IN(Conv3d(X_{conv1}))) + X_{conv1}$
3: $X_{flatten} : (B, L, C) \leftarrow LayerNorm(Flatten(X_{conv2}))$
4: $X_{branch1} : (B, 2L, C)$
$\leftarrow SSM(SiLU(Conv1d(W_{b1}X_{flatten})))$
5: $X_{branch2} : (B, 2L, C) \leftarrow SiLU(Conv1d(W_{b2}X_{flatten}))$
6: $X_{out} : (B, L, C) \leftarrow W_{out}(X_{branch1}X_{branch2})$
7: $y \leftarrow Reshape(X_{out})$
8: **return** y

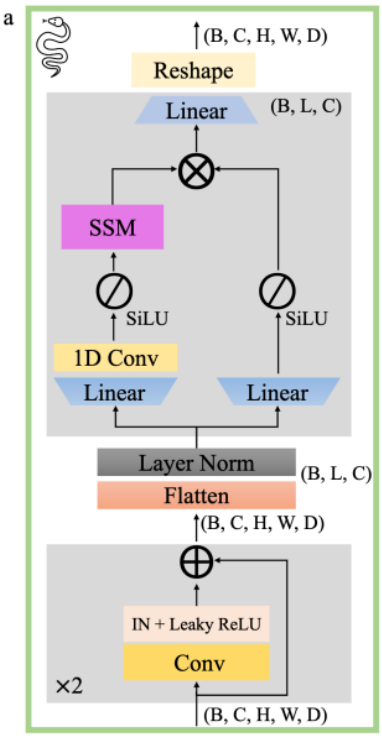

FIGURE 3. U-Mamba Block [27]

a convolution layer followed by the instance normalization layer [29] and Leaky ReLU activation [30]. 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 are 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 [24, 25]). The input is transformed into shapes (B, 2L, C) for the second branch

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 to find 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 recovers the resolution. 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. Each stage has one attention gate, 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.

B. AUTOMATED SEGMENTATION 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 5 [31]. Before training, the model extracts data fingerprints, including image modality, intensity distribution, spacing distribution, and median shape from the training data. The data fingerprints are used to choose rule-based parameters. The preprocessing strategy is 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 an image resampling strategy that ensures images are brought to a common resolution or 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 medical images' pixel or voxel intensity values to a standardized scale. For the PSMA dataset as an example, we use the Z-score normalization [32], 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 automatically configures batch size, patch size, GPU RAM, and network architecture, as discussed in the Experiment section.

There are some fixed parameters derived empirically. We use the polynomial learning rate scheduler [33] 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 learning rate decay and achieves better convergence as training progresses. The Loss function of the optimizer is explained in the Experiment section. The architecture template is also fixed. Here, we use Att-UMamba architecture with a full-resolution Conv3d setting. Typical data augmentation strategies such as rotation, Gaussian blur, and resize are 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.

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

IV. EXPERIMENTS AND RESULTS

The experiment is mainly conducted on Dataset_600 PSMA, Dataset_501 Glioblastoma, and common datasets. The training and inference are run on the NVIDIA A100 GPU with 40 GB RAM to speed up the training process. The results show that Att-UMamba outperforms other models on all four tasks.

A. DATASETS

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

1) Abdomen CT [34] 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 [35] for testing and evaluation.
2) Abdomen MRI dataset [36] 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 668 multi-parametric magnetic resonance imaging (mpMRI) scans for Glioblastoma (GBM) patients from the University of Pennsylvania Health System [37]. It contains well-annotated whole brain data in NIfTI format through computer-aided and manual correction from professional radiologists.
4) Prostate-Specific Membrane Antigen (PSMA) dataset [38] includes 600 PET/CT images. The PSMA Munich dataset was acquired using three different scanner

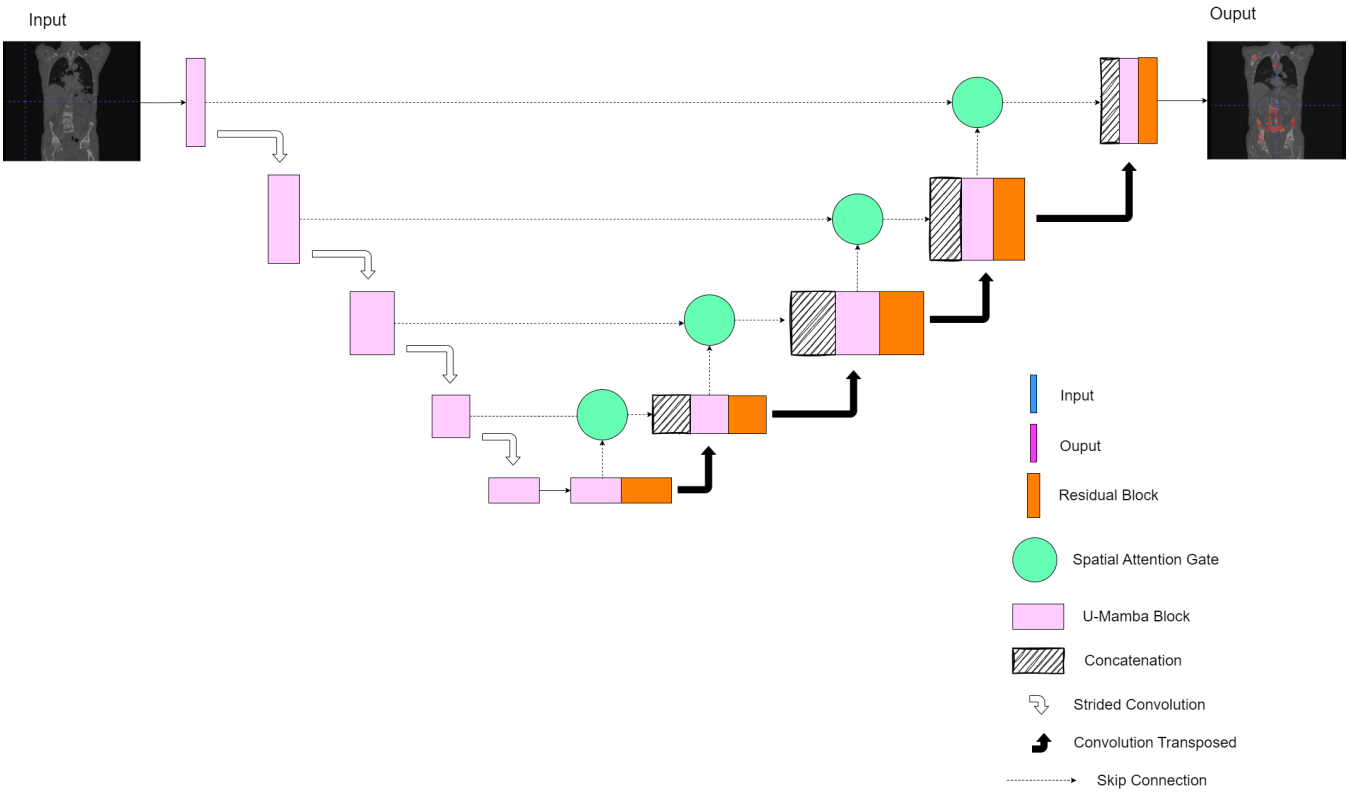

FIGURE 4. **The overview of the Att-UMamba architecture. It employs the encoder and decoder U-shaped structure similar to U-Net. We use the Residual Block in the encoder and decoder part and add a U-Mamba block at the bottle-neck. NnU-Net decides the number of stages depending on the dataset fingerprint. The detailed network configuration parameters are shown in Table 3**

TABLE 3. **Hyperparameter setting**

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

types (Siemens Biograph 64-4R TruePoint, Siemens Biograph mCT Flow 20, and GE Discovery 690).

B. 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 datasets converge around 160 epochs. The configuration generated by the self-configuration pipeline is shown in Table 3.

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 [39]. 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.

C. LOSS FUNCTION

The default loss function combines Dice loss [40] and cross-entropy loss [41] for multi-class segmentation task or Dice loss and binary cross-entropy loss for single-class segmentation task. Focal loss [42] 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:

$$\text{Dice Loss} = 1 - \frac{2\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} g_i} \tag{3}$$

The cross-entropy loss calculates the difference between the predicted probabilities and the actual labels. It is com-

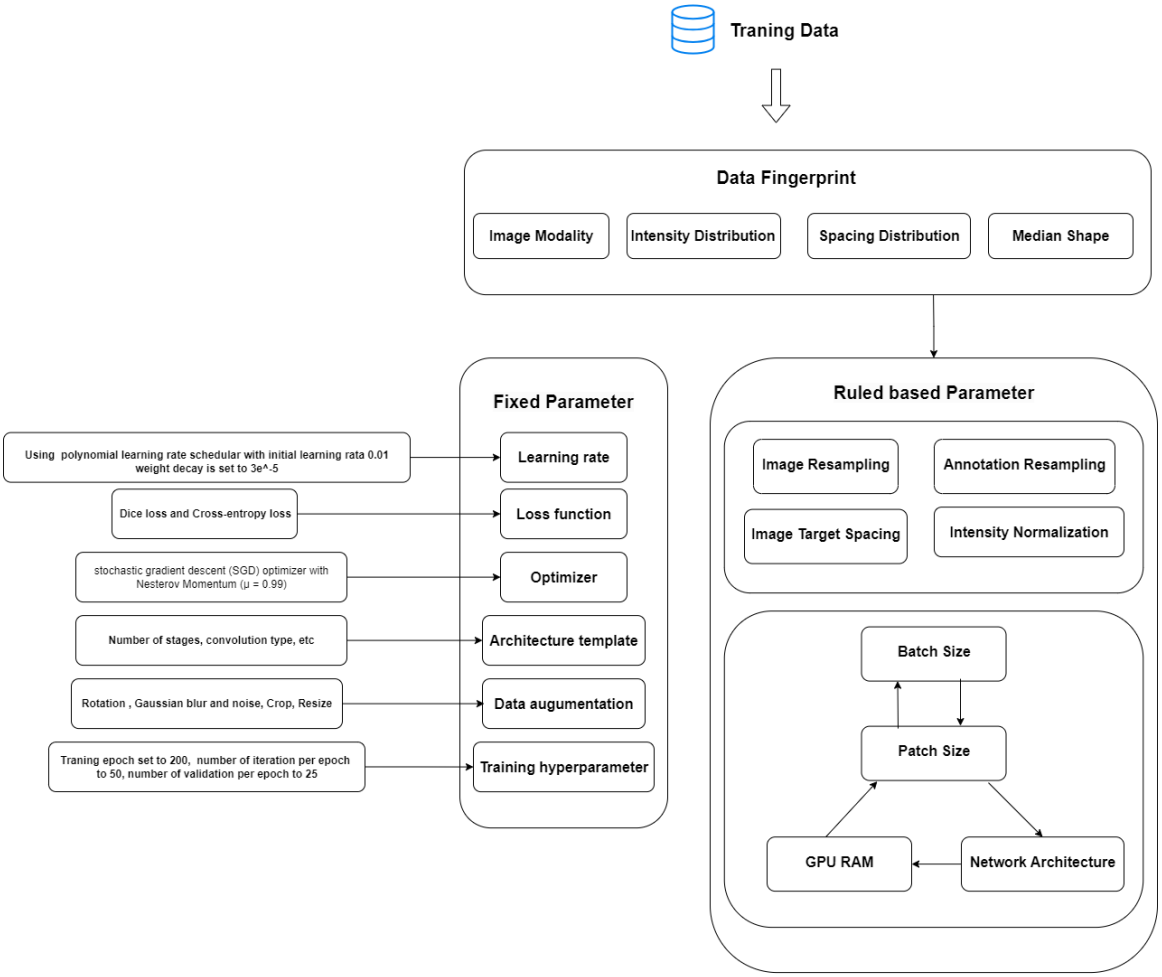

FIGURE 5. Self-configuration mechanism pipeline

monly used for classification tasks to measure the accuracy of a prediction. Cross-entropy loss is defined as:

$$\text{Cross-Entropy Loss} = -\sum_{i=1}^{N} g_i \log(p_i) \tag{4}$$

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

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

D. 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

$$DSC(G,S) = \frac{2|G \cap S|}{|G| + |S|} \tag{5}$$

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 (5), $|G|$ 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) [43] is also used to evaluate 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:

$$NSD = \frac{1}{|\partial A| + |\partial B|} \left(\sum_{a \in \partial A} \min_{b \in \partial B} \|a - b\| + \sum_{b \in \partial B} \min_{a \in \partial A} \|b - a\| \right) \tag{6}$$

where

1) ∂A represents the set of surface points of the ground truth segmentation.
2) ∂B represents the predicted segmentation's surface points.
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.

E. RESULTS AND EVALUATION

The results show that the Att-UMamba performs well on common and novel PSMA datasets. 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 [44]), 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, such as batch size and learning rate scheduler, are used. 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. 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.

Dataset501 Glioblastoma is a regionally based lesion segmentation dataset. It only involves 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

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	$\mathbf{0.866 \pm 0.074}$	$\mathbf{0.903 \pm 0.068}$

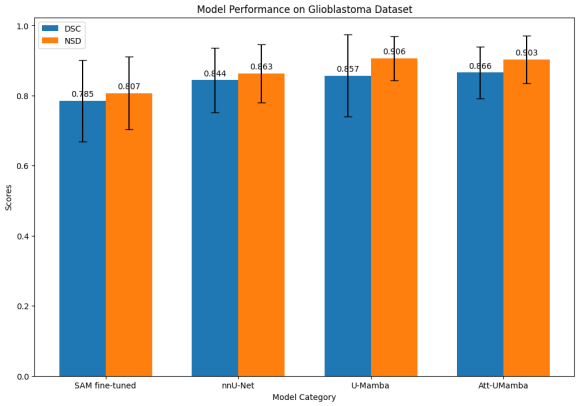

FIGURE 6. Model Performance on the Glioblastoma dataset

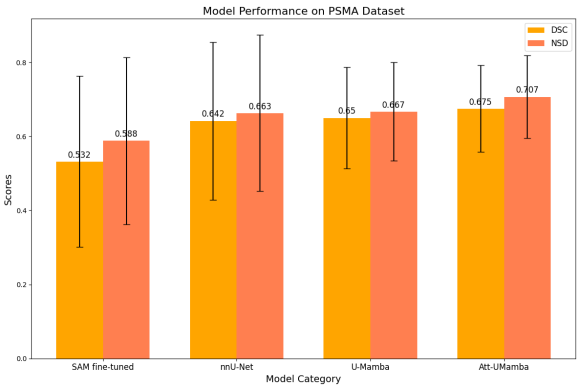

FIGURE 7. Model Performance on the PSMA dataset

extract high-interest regions improve boundary delineation. Unlike 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 misleads the clinical diagnosis process. Att-UMamba shows great potential for suppressing such noise.

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	$\mathbf{0.675 \pm 0.117}$	$\mathbf{0.707 \pm 0.112}$

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, further improving the segmentation accuracy by around 2.5 percent from results generated by U-Mamba.

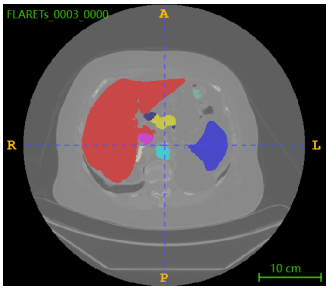

FIGURE 8. Prediction mask of Dataset701 AbdomenCT

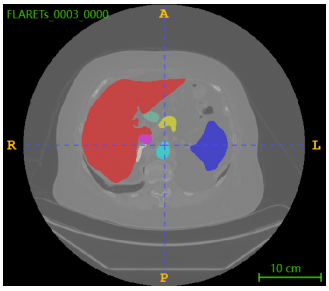

FIGURE 9. Ground truth mask of Dataset701 AbdomenCT

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 outlining more precise ROIs.

The Visualized segmentation example of the T1GD scan of patient No.240 in Dataset501 Glioblastoma is shown in Figure 13 using the Att-UMamba. Att-UMamba is more robust to heterogeneous appearances and has fewer segmentation outliers than nnU-Net and U-Mamba, which also achieve high accuracy in Dataset 501. Adding Attention Gates to the Att-UMamba network highlights the salient feature and thus reduces the number of outliers in segmentation masks.

Figure 14 shows three directional views of ground truth and

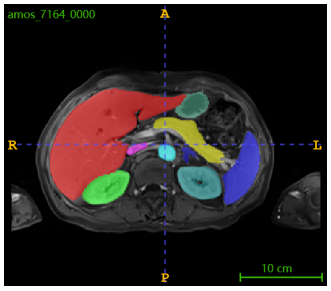

FIGURE 10. Prediction mask of Dataset701 AbdomenMR

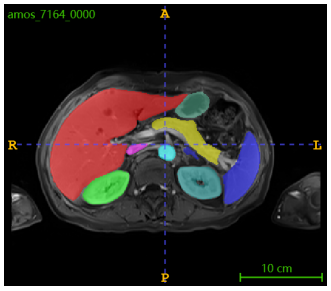

FIGURE 11. Ground truth mask of Dataset702 AbdomenMR

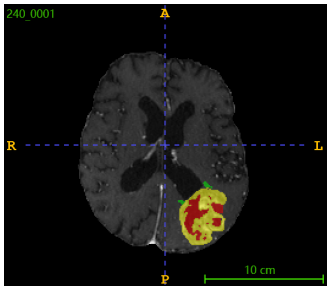

FIGURE 12. Ground truth mask of No.240 patient in Dataset501

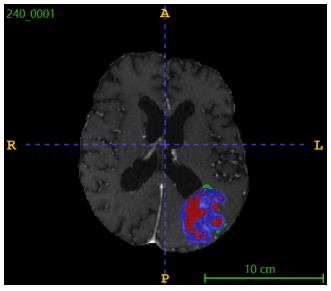

FIGURE 13. Prediction mask of No.240 patient in Dataset501

prediction masks generated by SAM, nnU-Net, U-Mamba, and Att-UMamba of Case ID No.4c9d9614d81f3005. 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 decreases the accuracy of the segmentation. Such a phenomenon probably proves that adding the Attention Gates to skip 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 image slice, 3D segmentation receives a larger input size than 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 than SAM, U-Mamba, and Att-UMamba. Its segmentation boundary is less accurate compared to the other three models. These observations are also shown in DSC and NSD values and 3D visualization shown below.

Figure 15 shows the same case's 3D ground truth and prediction masks generated by SAM, nnU-Net, U-Mamba, and Att-UMamba. From the 3D view, we find that prediction is 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, we have several sparsely distributed lesions in the upper part of the body section, and in the lower part, we have large, aggregated lesions. The prediction for the figure's upper part is less accurate than the lower part. This problem may be attributed to the design of the Att-UMamba architecture. Highlighting ROIs using attention gates may cause some true positive regions to be missing. Nevertheless, the overall prediction accuracy is the highest among all four models. It generates fewer false-positive regions, whereas SAM generates many small ones.

V. CONCLUSION

This paper 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 and other prevailing medical image segmentation tasks or universal lesion segmentation. The outstanding performance is attributed to the architecture design, 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 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 clearer boundaries than the results generated by other models, reducing the false-positive

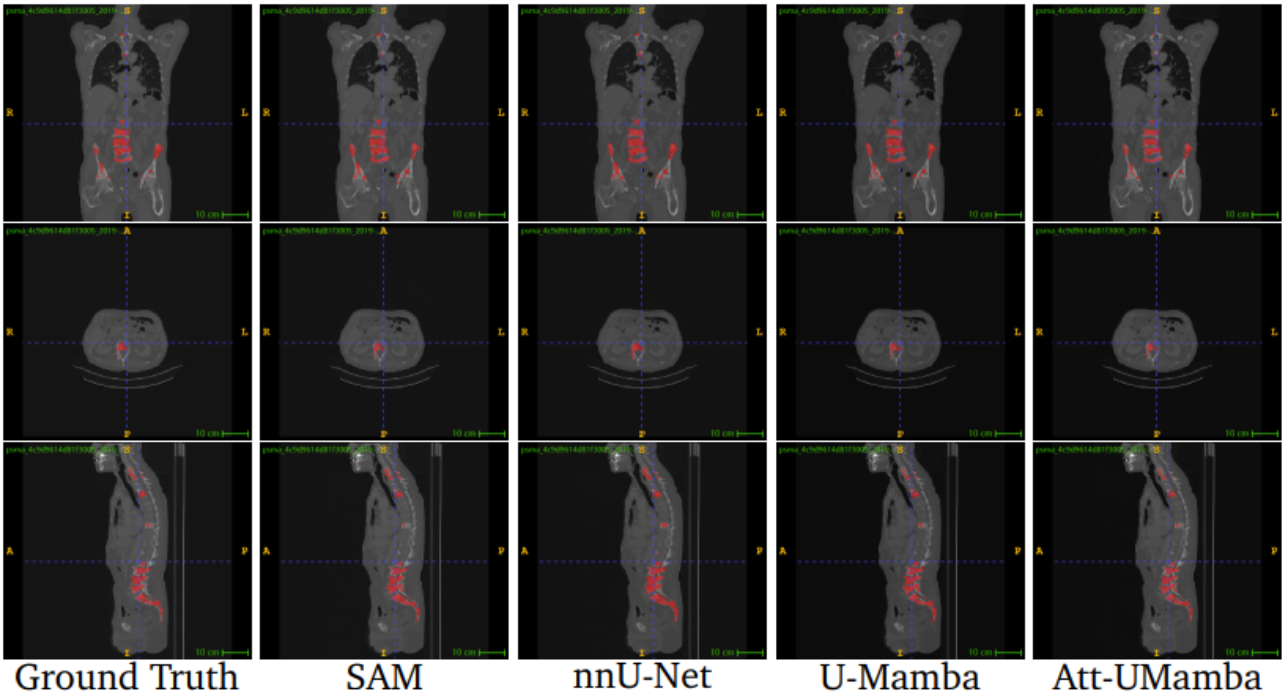

FIGURE 14. Visualized segmentation examples of coronal view (1st row), axial view(2nd row), and sagittal(3rd row) from PSMA dataset. Cursor position (x,y,z) is set to (100, 90, 190) for sliced visualization

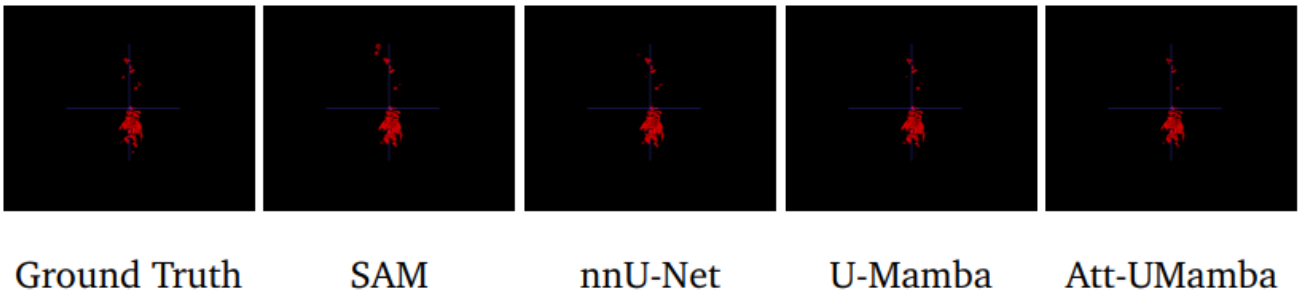

FIGURE 15. Visualized 3D PSMA segmentation masks

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] Siddhartha Jetley, Nicholas A. Lord, Namhoon Lee, and Philip H. S. Torr. Learn to pay attention. In *International Conference on Learning Representations*, 2018.

[2] Stanford Taylor, James M. Brown, Kishan Gupta, J. Peter Campbell, Susan Ostmo, R. V. Paul Chan, Jennifer Dy, Deniz Erdogmus, Stratis Ioannidis, Sang J. Kim, Jayashree Kalpathy-Cramer, Michael F. Chiang, for the Imaging, and Informatics in Retinopathy of Prematurity Consortium. Monitoring Disease Progression With a Quantitative Severity Scale for Retinopathy of Prematurity Using Deep Learning. *JAMA Ophthalmology*, 137(9):1022–1028, 09 2019.

[3] J. De Fauw et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24:1342–1350, 2018.

[4] D. Ouyang et al. Video-based ai for beat-to-beat assessment of cardiac function. *Nature*, 580:252–256, 2020.

[5] Atsushi Saito, Shohei Nawano, and Akifumi Shimizu. Joint optimization of segmentation and shape prior from level-set-based statistical shape model, and its application to the automated segmentation of abdominal or-

gans. *Medical Image Analysis*, 28:46–65, 2016.

[6] Holger R. Roth, Hirohisa Oda, Yuta Hayashi, Masahiro Oda, Noriyuki Shimizu, Masahiro Fujiwara, Keiichi Misawa, and Kensaku Mori. Hierarchical 3d fully convolutional networks for multi-organ segmentation. *arXiv preprint arXiv:1704.06382*, 2017.

[7] Heng Guo, Jianfeng Zhang, Jiaxing Huang, Tony C. W. Mok, Dazhou Guo, Ke Yan, Le Lu, Dakai Jin, and Minfeng Xu. Towards a comprehensive, efficient and promptable anatomic structure segmentation model using 3d whole-body ct scans, 2024.

[8] Pierre-Henri Conze, Gustavo Andrade-Miranda, Vivek Kumar Singh, Vincent Jaouen, and Dimitris Visvikis. Current and emerging trends in medical image segmentation with deep learning. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 7(6):545–569, 2023.

[9] Lars Heinrich, Jan Funke, Constantin Pape, Juan Nunez-Iglesias, and Stephan Saalfeld. Synaptic cleft segmentation in non-isotropic volume electron microscopy of the complete drosophila brain. In Alejandro F. Frangi et al., editors, *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 317–325. Springer, 2018.

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

[11] Russell A. Poldrack. Region of interest analysis for fMRI. *Social Cognitive and Affective Neuroscience*, 2(1):67–70, 03 2007.

[12] Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation, 2024.

[13] Xiongxiao Xu, Yueqing Liang, Baixiang Huang, Zhiling Lan, and Kai Shu. Integrating mamba and transformer for long-short range time series forecasting, 2024.

[14] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[15] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2):353–367, 2011.

[16] AutoPET III Challenge. Autopet iii – grand challenge, 2023. Accessed on: 2024-04-26.

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241. Springer, 2015.

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

[19] RW Pettit, BB Marlatt, SJ Corr, J Havelka, and A Rana. nnu-net deep learning method for segmenting parenchyma and determining liver volume from computed tomography images. *Ann Surg Open*, 3(2):e155, Jun 2022. Epub 2022 Mar 30. PMID: 36275876; PMCID: PMC9585534.

[20] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation, 2014.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[22] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas, 2018.

[23] Michal Drozdzal, Eugene Vorontsov, Gabriel Chartrand, Samuel Kadoury, and Chris Pal. The importance of skip connections in biomedical image segmentation. In Gustavo Carneiro, Diana Mateus, Loïc Peter, Andrew Bradley, João Manuel R. S. Tavares, Vasileios Belagiannis, João Paulo Papa, Jacinto C. Nascimento, Marco Loog, Zhi Lu, Jaime S. Cardoso, and Julien Cornebise, editors, *Deep Learning and Data Labeling for Medical Applications*, pages 179–187, Cham, 2016. Springer International Publishing.

[24] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2023.

[25] Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. Selective structured state space sequence models. *arXiv preprint arXiv:2301.11751*, 2023.

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

[27] Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024.

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

[29] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[30] A.L. Maas, A.Y. Hannun, and A.Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In

International Conference on Machine Learning, volume 28, 2013.

[31] D. Duprez, C. Trauernicht, H. Simonds, and O. Williams. 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):e13988, Aug 2023. Epub 2023 Apr 12.

[32] Nanyi Fei, Yizhao Gao, Zhiwu Lu, and Tao Xiang. Z-score normalization, hubness, and few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 142–151, October 2021.

[33] Purnendu Mishra and Kishor Sarawadekar. Polynomial learning rate policy with warm restart for deep neural network. In *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, pages 2087–2092, 2019.

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

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

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

[37] Spyridon Bakas, Claire Sako, Hamed Akbari, Michel Bilello, Aristeidis Sotiras, Gaurav Shukla, Jeffrey D. Rudie, Nancy Flores Santamaria, Amir Fathi Kazerooni, Saurabh Pati, Sarthak Rathore, Alexander Mamourian, Sung Min Ha, William Parker, Jimit Doshi, Ujjwal Baid, Michael Bergman, Zachary A. Binder, Ragini Verma, 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.

[38] Stefanie Gatidis and Thomas Kuestner. A whole-body fdg-pet/ct dataset with manually annotated tumor lesions (fdg-pet-ct-lesions) [dataset]. The Cancer Imaging Archive, 2022.

[39] Aitor Lewkowycz. How to decay your learning rate, 2021.

[40] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016.

[41] Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[42] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

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

[44] A. Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. In *International MICCAI Brainlesion Workshop*, volume 11384 of *Lecture Notes in Computer Science*, pages 311–320, 2018.

FIRST.AUTHOR.YIJIE.GONG received the B.E degrees in Computer Science and Mathematics from Georgia Institute of Technology, Atlanta, GA, USA in 2023. He is taking the M.E degree in Computing at Imperial College London, London, UK. During his undergraduate curriculum, he received the highest honor and was selected by the Dean's list in 2023. His area of research includes Computer Vision/Deep learning, Medical image segmentation, and Robotics.

SECOND.AUTHOR.MAYURI.A.MEHTA received the B.E. and M.E. degrees in Computer Engineering from Sardar Patel University, Vallabh Vidhyanagar, India in 2000 and 2005 respectively, and a Ph.D. degree in Computer Engineering from Sardar Vallabhbhai National Institute of Technology (SVNIT), Surat, India in 2014. She is a Professor of the Department of Computer Engineering, Sarvajanik College of Engineering and Technology, Surat, India. She is also the International Relations External Affairs Officer of her institute. Her 24 years of professional experience includes academic and research achievements along with administrative and organizational capabilities. Her areas of research include Machine Learning/Deep Learning, Data Science, Medical Image Analysis, Health Informatics, and Computer Algorithms.

THIRD.AUTHOR.PANCHAM.U.SHUKLA