Mamba vs SOTA for Vision Tasks

Clay Crews†
Department of Computer Science
University of South Carolina
Columbia, SC, United States
jccrews@email.sc.edu

Lexington Whalen†
Department of Computer Science
University of South Carolina
Columbia, SC, United States
LAWHALEN@email.sc.edu

***Abstract*—Abstract This paper explores the performance of the Mamba architecture, a variant of Structured State Space Sequence Models (SSMs), in comparison to state-of-the-art models such as Transformers, U-Net, and ResNet for vision tasks. We focus on image segmentation and classification, with the goal of maintaining high accuracy while keeping parameter counts low. Developing compact models with fewer parameters is crucial for the future of AI, not only to reduce energy consumption but also to enable deployment on resource-constrained edge devices. By leveraging the selective state spaces in Mamba blocks, we aim to achieve efficient and effective segmentation and classification performance while maintaining the linear scalability of SSMs. We implement Mamba-based architectures and compare their results to popular models like U-Net, ResNet, and Transformer-based approaches on standard vision benchmarks. In addition to accuracy, we place a strong emphasis on model size, targeting compact architectures suitable for resource-constrained environments. Through careful design choices and optimizations, we strive to develop Mamba-based models that achieve competitive accuracy with significantly fewer parameters compared to existing state-of-the-art models. This work highlights the potential of Mamba and SSMs for efficient vision tasks, contributing to the development of compact yet accurate models in image segmentation and classification. We open-source our code to facilitate further research and exploration of these architectures: https://github.com/lxaw/mamba-vs-else-vision**

***Index Terms*—Mamba, TinyML, Image Segmentation, Image Classification**

I. Introduction

MRI imaging to identify brain tumors has naturally been a target for accurate image segmentation through the use of machine learning (ML) architecture. The use of computer-aided diagnosis of tumors can be crucial to quickly identify tumors and determine a course of action for the patient. Gliomas are the most common primary brain malignancy with varying degrees of aggressiveness in the brain. The protocol for MRI image annotation of these tumors consists of the following labels: the whole tumor extent, the tumor core, the non-enhancing/necrotic tumor region, and regions of low grade gliomas [5]. These annotations of an MRI are used to determine the size and severity of a tumor.

To accurately segment these images, picking out the small features of the image becomes the focus. The use of the U-Net architecture, presented by Ronneberger, Fischer, and Brox [1], has had success in MRI image segmentation. U-Net builds on the approach of fully convolutional networks by improving the limitation on the large amount of data needed to accurately train the model. Very few training images are needed due to the contracting and expansive paths in this architecture. In the contracting path, images are downsampled and pooled along with increasing the number of feature channels. Each downsampled resolution produces a multi-channel feature map. The expansive path up samples and combines with each segmentation feature map from the contracting layer to localize features in the image. A 1x1 convolution layer is applied to the final image to map the large feature vector to a selected number of classes. All of this essentially creates a feature map where each pixel's relevance is taken into account and evaluated in the end result. An approach using the U-Net architecture would provide the detailed analysis needed for tumor annotation.

An arising issue in the complex recurrent or convolutional neural networks used for image segmentation is the length of the sequences given as input. Maintaining relevance of the current area of an image in relation to the rest of the image is key for more accurate segmentation. An approach with the Transformer architecture has a large emphasis on learning specific features in an image, outlined by Vaswani et al. 2017 [2]. This architecture follows an encoder-decoder pattern, connected through an attention mechanism, a transformer block. Transformer blocks show capability of learning long-distance dependency throughout an input image. In the self-attention mechanism for this block, a single element in a given sequence is compared to all of the elements in the sequence. However, this mechanism becomes computationally very expensive when it comes to long sequences.

Selective State Spaces (SSMs) pose an improvement in computational efficiency over Transformers for this long dependency range. SSMs represent the relevance and context of different parts of the input sequence giving selective attention. This model is widely used in control theory for time variant systems and in fields such as computational neuroscience. Selective attention is very efficient in handling long sequences over time and presents a more localized context of the data. Sections of sequences that deserve attention are represented by the model and will dynamically update these values to reflect the contextual relevance.

Making use of Structured State Spaces, presented by Gu, Goel, and Ré [3], for their superiority in long range dependency tasks, a specific type of SSM, the Mamba architecture,

† Equal contributions

by Gu and Dao [4], proposed a foundational model to operate on arbitrary sequences from a variety of inputs in the domain of sequence modeling. This model achieves the power of Transformers while scaling linearly in sequencing length to be computationally efficient. Additionally, the selection mechanism of Mamba improves on an SSMs ability to focus on or ignore sections in an input sequence by parameterizing the parameters of an SSM to reflect the current context of the input. The Mamba model and its application to brain tumor MRI imaging segmentation will be evaluated in this project.

II. Proposed Methodology

In this paper, we strive to maintain high accuracy in our image segmentation while also keeping low parameter counts. We primarily seek to use pyramidal pooling modules [13] to improve accuracy, and use a variant of the Mamba [4] architecture inspired by [12] to maintain small parameter sizes.

III. Input Data

We shall be using a Brain MRI segmentation dataset found in [14]. This dataset has been used in several papers regarding classification of the shape and severity of tumors, and provides an adequate dataset for research purposes in the field of medical image analysis, particularly in the area of brain MRI segmentation. Researchers have utilized this dataset in various studies focusing on the classification of tumor shapes and the assessment of tumor severity. With its comprehensive collection of brain MRI scans, along with corresponding segmentation masks, the dataset offers valuable resources for developing and evaluating algorithms aimed at automating tumor detection and analysis.

The dataset contains brain MRI images along with their segmentation masks for 110 patients. There were a total of 3143 train images, 393 validation images, and 393 test images. The distribution of the tumor and non-tumor images is shown in Figure 1.

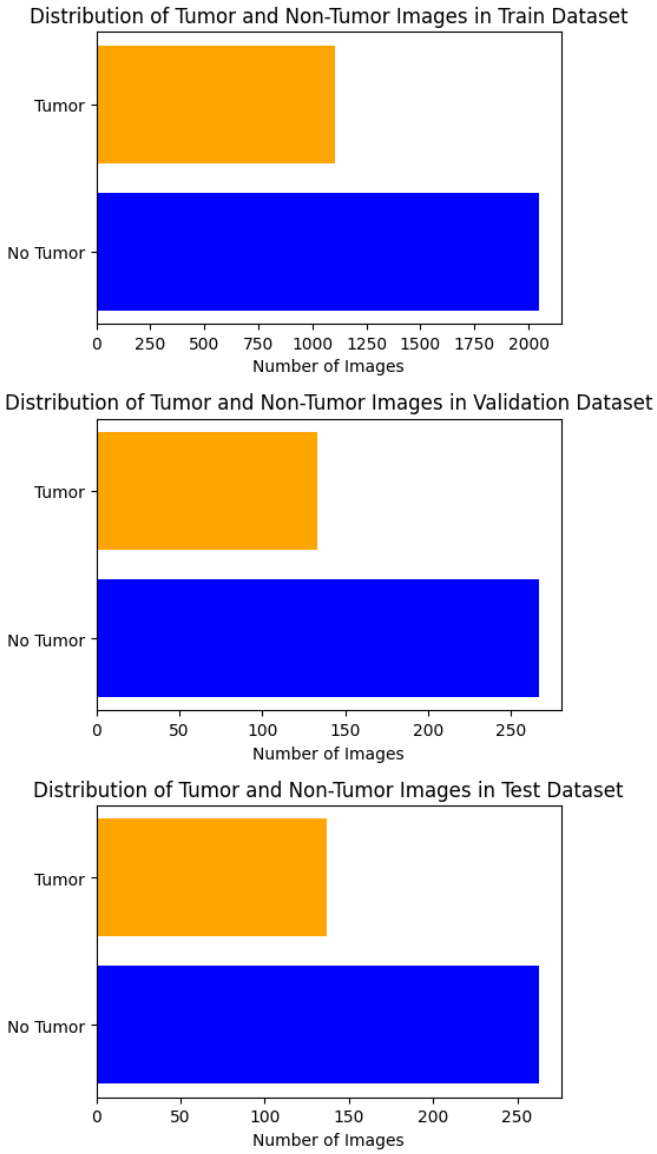

Fig. 1. Distributions of tumor and non-tumor images in the train, validation, and test datasets.

IV. Output Data

The output of this model shall be segmentation masks for new, never before seen images.

V. Mamba

Mamba [4] is a recently proposed architecture that combines the modeling power of Transformers with the linear scaling efficiency of state space models (SSMs). The key idea is to augment SSMs with a selection mechanism that allows the model parameters to vary based on the input, enabling it to perform content-aware reasoning. Standard SSMs map an input sequence $x(t)$ to an output $y(t)$ via a latent state $h(t)$ using a linear time-invariant system:

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t)$$
$$y(t) = \mathbf{C}h(t) + \mathbf{D}x(t)$$

where the parameter matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$ are fixed. In the original paper, the $\mathbf{D}$ matrix is considered only as a skip connection, and thus as it does not directly play a role in the differential equation, was ignored. This allows SSMs to be computed efficiently as a convolution. However, the time-invariance means the model cannot change its behavior for specific inputs, making it difficult to solve tasks requiring selective focus, such as selective copying or induction heads [4]. Mamba introduces a selection mechanism where the SSM parameters $\mathbf{B}$, $\mathbf{C}$, Δ (the state step size) vary for each input token x_t. This allows the model to selectively propagate or forget information based on the current token. The selective SSM is computed using a parallel scan operation to maintain linear complexity. The Mamba block interleaves the selective SSM with linear projections, convolutions, and activations:

1) Linear projection to expand the embedding dimension
2) Convolution to mix information between dimensions
3) Selective SSM via parallel scan to efficiently propagate

information
4) Linear projection to get back to the target embedding size

These blocks are stacked to form the Mamba architecture. Compared to Transformers, Mamba can achieve comparable modeling performance while scaling linearly in sequence length rather than quadratically [4]. Overall, the selective SSM leverages the strengths of RNNs, CNNs and Transformers - it maintains an unbounded context window like RNNs, can be parallelized like CNNs, and achieves content-aware reasoning like Transformers, while being more efficient than all three. This makes Mamba a promising foundation model for processing long sequences across various domains.

VI. Compared Models

Prior to an analysis of how the Pyramidal U-Mamba compares, we shall explain what models we chose for comparison and why.

For this our analysis, we compare four models against our own developed ones. The models are U-Net [1], ResNet18 [9], ResNet50, U-Mamba [6], and SegViT [11]. Below we explain our choices.

A. U-Net

The core idea behind U-Net is to complement a traditional contracting network with successive layers that replace pooling operations with upsampling operators. These upsampling layers aim to increase the resolution of the output. Subsequent convolutional layers can then learn to assemble a precise output based on this high-resolution information [1]. A key modification in the U-Net architecture is the inclusion of a large number of feature channels in the upsampling part. This allows the network to effectively propagate contextual information to higher resolution layers. As a result, the expansive path of the network becomes more or less symmetric to the contracting part, leading to a distinctive U-shaped architecture. To predict pixels at the border regions of the image, the missing context is extrapolated by mirroring the input image. This tiling strategy is crucial for applying the network to large images, as it circumvents resolution limitations imposed by GPU memory constraints [1]. Due to its encoder-decoder architecture, skip connections, multi-scale feature extraction, and its efficiency, U-Net has been used in many denoising and diffusion models. For instance, DDPMs (Denoising Diffusion Probabilistic Models) use an architecture similar to U-Net for denoising and sample generation [8] while the popular Stable Diffusion architecture uses a U-Net based architecture for the diffusion process [7]. We have chosen to compare our novel segmentation architecture against the U-Net architecture due to its well-established reputation and widespread adoption in the field of medical image segmentation.

B. ResNet

We have also selected ResNet18 and ResNet50 as additional benchmarks. These architectures, introduced by He et al. in their seminal work "Deep Residual Learning for Image Recognition" [9], have revolutionized the field of deep learning by addressing the problem of vanishing gradients in deep neural networks. The key innovation in ResNets is the introduction of residual connections, which allow the network to learn residual functions with reference to the input layer, thereby facilitating the training of much deeper networks. ResNet18 and ResNet50, with 18 and 50 layers respectively, have been widely adopted in various computer vision tasks, including image classification, object detection, and segmentation. These models have demonstrated exceptional performance and generalization ability across diverse datasets. By comparing our proposed architecture against ResNet18 and ResNet50, we aim to assess its effectiveness in relation to these well-established and highly influential architectures. This comparison will provide valuable insights into the capabilities of our model and its potential to advance the state-of-the-art in image segmentation tasks.

C. U-Mamba

As our model takes much inspiration from the recently developed U-Mamba design [6], we also choose to incorporate it in our comparison. U-Mamba addresses these limitations by introducing a novel hybrid CNN-SSM block that leverages the strengths of both architectures. The convolutional layers in the block are responsible for local feature extraction, while the State Space Sequence Models (SSMs) [3], a new family of deep sequence models, are known for their strong capability in handling long sequences and capturing long-range dependencies. By integrating these two components, U-Mamba achieves a balance between local and global information processing, enabling it to effectively handle long-range dependencies in biomedical image segmentation tasks. Moreover, U-Mamba incorporates a self-configuring mechanism that allows it to automatically adapt to various datasets without manual intervention, enhancing its versatility and usability.

D. SegFormer

SegFormer [15] is a transformer-based semantic segmentation model that employs a hierarchical structure and a novel attention mechanism called Efficient Self-Attention (ESA). The SegFormer architecture consists of a transformer encoder for capturing long-range dependencies and a lightweight All-MLP decoder for generating high-resolution segmentation masks. The ESA module reduces the computational complexity of self-attention by performing attention operations in a local window and aggregating global information through a depth-wise convolution. This allows SegFormer to efficiently process high-resolution images while capturing both local and global context. Moreover, SegFormer introduces a position-sensitive embedding scheme that encodes both spatial and channel-wise information, enhancing the model's ability to capture fine-grained details. By comparing U-Mamba against SegFormer, we aim to evaluate the effectiveness of our hybrid CNN-SSM approach in relation to the transformer-based segmentation mechanism employed by SegFormer. This comparison will

provide insights into the strengths and weaknesses of both architectures and their ability to handle complex biomedical image segmentation tasks. Furthermore, it will help us understand the potential of transformer-based approaches and hybrid approaches in advancing the state-of-the-art in biomedical image segmentation.

E. UltraLight VM-UNet

The UltraLight VM-UNet [12] is a lightweight neural network architecture designed for skin lesion segmentation tasks. It is built upon the Vision Mamba module, which is a state-space model (SSM) that can efficiently handle long-range dependencies in sequences, making it well-suited for image segmentation tasks.

The key innovation of the UltraLight VM-UNet is the proposed Parallel Vision Mamba Layer (PVM Layer), which processes deep features in parallel using multiple Vision Mamba blocks. Specifically, the input feature map is split into multiple sub-feature maps, each processed by a separate Vision Mamba block with a reduced channel count. This parallel processing approach allows the UltraLight VM-UNet to maintain high segmentation performance while significantly reducing the number of parameters and computational complexity.

The UltraLight VM-UNet is reported to have only 0.049 million parameters and a computational cost of 0.060 GFLOPs, which is significantly lower than traditional convolutional neural networks and transformers used for image segmentation tasks. Despite its lightweight nature, the authors demonstrate that the UltraLight VM-UNet achieves competitive performance on three publicly available skin lesion segmentation datasets, outperforming several state-of-the-art lightweight models.

The success of the UltraLight VM-UNet can be attributed to the authors' in-depth analysis of the key factors influencing the parameters of the Vision Mamba module. By identifying the number of input channels as a critical factor affecting the parameter count, they were able to design the PVM Layer to process features in parallel while keeping the overall channel count constant, leading to a significant reduction in parameters without compromising performance.

VII. Our Models

Inspired by [6] and [12], we sought to implement both models on a different problem: tumor segmentation.

Our goal was to 1) maintain dice scores comparable to state-of-the-art (SOTA) methods such as those listed above, while 2) being smaller than those above.

We now go into what we did for each of our models.

A. UMambaBot_PPBot

This model incorporates pyramidal pooling, which is a strategy used in convolutional neural networks (CNNs) to capture context and incorporate multi-scale information [13]. Pyramidal pooling involves parallel pooling operations at different scales, followed by concatenation of the resulting feature maps. This approach has been shown to improve the performance of CNNs in various computer vision tasks, including segmentation, by enabling the model to capture both local and global context.

In our UMambaBot_PP model, we integrated pyramidal pooling into the U-Mamba architecture, aiming to leverage the strengths of both the Mamba module and multi-scale feature extraction.

B. UL-VM-UNet_v1

For ULMUNet_v1, the initial channel list was modified from [8, 16, 24, 32, 48, 64] to [16, 32, 64, 128, 256], increasing the number of channels at each step of the encoder and decoder. Additionally, the depth of the U-Net was reduced by one layer to keep the parameter count low while increasing the number of parameters in each step. The intention was to improve performance by increasing the capacity of the model while maintaining a reasonable parameter count.

C. UL-VM-UNet_v2

In ULMUNet_v2, the number of parallel branches in the Parallel Vision Mamba (PVM) block was increased from 4 to 8. This modification directly decreases the parameter count, further proving the idea proposed in the UltraLight VM-UNet paper [12] that processing features in parallel with reduced channel counts can significantly reduce parameters while maintaining performance.

D. UL-VM-UNet_v3

ULMUNet_v3 combines the modifications from ULMUNet_v1 and ULMUNet_v2. It increases the number of parallel branches in the PVM block to 8 and modifies the channel list to [16, 32, 64, 128, 256]. Additionally, the depth of the U-Net was reduced to 5 layers. This approach aims to balance the trade-off between model capacity and parameter efficiency.

E. UL-VM-UNet_v4

Pyramidal pooling was used in each of the PVM blocks in the encoder and decoder networks. If we want to do similar to UMambaBot_PP we can do it just in the bottleneck. Let me know if we want this can I can do it fast. This model incorporates pyramidal pooling, similar to UMambaBot_PP, but within the UltraLight VM-UNet architecture. The goal was to leverage the benefits of multi-scale feature extraction while maintaining the parameter efficiency of the UltraLight VM-UNet.

F. UL-VM-UNet_v5

What is UL-PP? v5 is v4 with no PP. If any, v4 is UL-PP. The UL-PP (Ultra Light Pyramidal Pooling) model trained and performed well with 1.1M parameters. Incorporating pyramidal pooling on the UL network scored slightly worse than without pyramidal pooling but required approximately 400,000 more parameters.

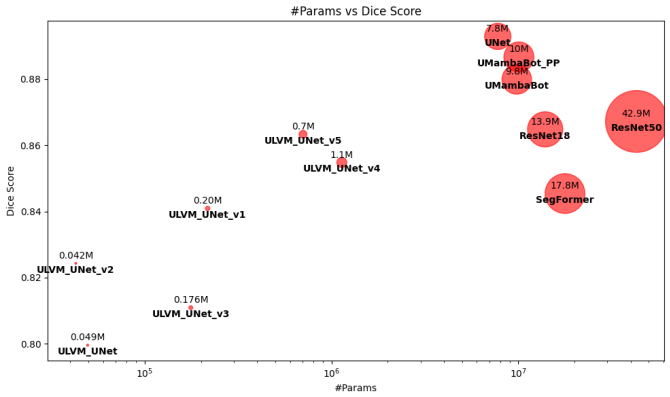

Fig. 2. A scatter plot of Dice Score vs model parameter count, where "M" means "millions. The points are scaled to help represent size.

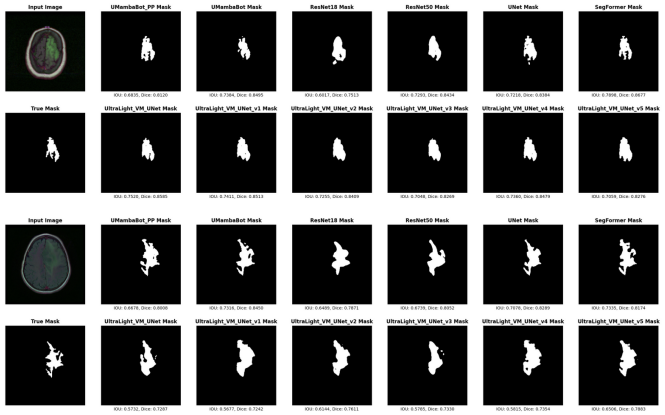

Fig. 3. Example segmentations.

VIII. Results

We now compare our models against ResNet18, ResNet50, standard U-Mamba bottleneck, UNet, and UltraLight VM-UNet. We train for 100 epochs on all models, and use dice-loss with Adam optimizer. We train on roughly 3000 images, and validate on roughly 400. we then test on roughly 400. We show the stats of our data in Figure 1. The results are shown in Figures **??**,2,3.

Model Name	Parameter Count (Millions)	Dice Score
UMambaBot_PP	10.0	0.8867
UMambaBot	9.8	0.8799
ResNet18	13.9	0.8648
ResNet50	42.9	0.8672
SegFormer	17.8	0.8454
UNet	7.8	0.8929
UL-VM-UNet	0.049	0.79955
UL-VM-UNet_v1	0.20	0.8409
UL-VM-UNet_v2	0.042	0.8243
UL-VM-UNet_v3	0.176	0.8109
UL-VM-UNet_v4	1.1	0.8548
UL-VM-UNet_v5	0.7	0.8633

TABLE I
COMPARISON OF DIFFERENT SEGMENTATION MODELS.

do we need a table of the params and scores? May be helpful.. Maybe stick one in the appendix.

Would appendix really need to be in the page limit? Appendix is to include figures that wouldn't fit in paper?

IX. Conclusion

In this work, we explored the application of Structured State Space Sequence Models (SSMs), particularly the Mamba architecture, for tumor segmentation from brain MRI scans. The Mamba architecture leverages selective state spaces, enabling efficient and effective segmentation while maintaining linear scalability.

Our experiments and comparisons with state-of-the-art (SOTA) models demonstrate the potential of the Mamba-based approach in both model size and segmentation accuracy. The smallest model, ULVM_Net_v2, achieved a compelling dice score of 0.8243 with only 0.042M parameters, achieving comparable performance to the larger ResNet50 model (42.9M params) with a dice score of 0.8672, while also outperforming the original UltraLight VM-UNet in terms of both parameter count and dice score.

While the traditional U-Net model achieved the highest dice score of 0.8929, our UMambaBot_PP model closely followed with a dice score of 0.8867 note the raise of 300k params in pp?. Notably, our Mamba-based UMambaBot_PPBot and UMambaBot models outperformed the Segformer transformer model, indicating the potential of Mamba-based approaches to surpass transformer-based models for this task.

We were able to make the ultra-light model smaller and more accurate on our task, further proving the potential of Mamba-based systems that prioritize compactness and efficiency. The selective state spaces and parallel processing strategies employed in our models enabled us to strike a balance between model size and segmentation accuracy.

Overall, this work contributes to the exploration of Structured State Space Sequence Models for medical image segmentation tasks. The Mamba architecture's ability to capture long-range dependencies while maintaining linear scalability positions it as a promising alternative to traditional convolutional and transformer-based approaches, especially in resource-constrained environments.

Future work could involve further refinements and optimizations of the Mamba-based models, as well as their application to other medical imaging modalities and segmentation tasks. Additionally, investigating the interpretability and robustness of these models could provide valuable insights for their deployment in real-world clinical settings.

Appendix

A. Training Pipeline

The training pipeline for our model consists of the following steps:

1. Data preprocessing: The dataset is split into training, validation, and test sets. The images are resized to a uniform size and normalized.

2. Model architecture: We employ a convolutional neural network (CNN) with the following layers: [Describe the layers and their configurations].

3. Training: The model is trained using the training set for [X] epochs with a batch size of [Y]. We use the [Z] optimizer with a learning rate of [L].

4. Evaluation: The trained model is evaluated on the validation set to monitor its performance and prevent overfitting. We use metrics such as accuracy, precision, recall, and F1-score.

5. Testing: Once the model achieves satisfactory performance on the validation set, it is tested on the held-out test set to assess its generalization ability.

B. Supplementary Figures

REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," arXiv:1505.04597 [cs.CV], 2015.

[2] A. Vaswani et al., "Attention Is All You Need," arXiv:1706.03762 [cs.CL], 2023.

[3] A. Gu, K. Goel, and C. Ré, "Efficiently Modeling Long Sequences with Structured State Spaces," arXiv:2111.00396 [cs.LG], 2022.

[4] A. Gu and T. Dao, "Mamba: Linear-Time Sequence Modeling with Selective State Spaces," arXiv:2312.00752 [cs.LG], 2023.

[5] S. Bakas et al., "Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge," arXiv:1811.02629 [cs.CV], 2019.

[6] J. Ma, F. Li, and B. Wang, "U-Mamba: Enhancing Long-range Dependency for Biomedical Image Segmentation," arXiv:2401.04722 [cs.CV], 2024.

[7] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," arXiv:2112.10752 [cs.CV], 2021.

[8] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," CoRR, vol. abs/2006.11239, 2020. [Online]. Available: https://arxiv.org/abs/2006.11239

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," CoRR, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[10] B. Zhang et al., "SegViT: Semantic Segmentation with Plain Vision Transformers," arXiv:2210.05844 [cs.CV], 2022.

[11] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv:2010.11929 [cs.CV], 2021.

[12] R. Wu, Y. Liu, P. Liang, and Q. Chang, "UltraLight VM-UNet: Parallel Vision Mamba Significantly Reduces Parameters for Skin Lesion Segmentation." Accessed: Apr. 20, 2024. [Online]. Available: https://arxiv.org/pdf/2403.20035.pdf

[13] H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia, "Pyramid Scene Parsing Network," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6230-6239, doi: 10.1109/CVPR.2017.660.

[14] "Brain MRI segmentation," www.kaggle.com. https://www.kaggle.com/datasets/mateuszbuda/lgg-mri-segmentation

[15] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. Alvarez, and P. Luo, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers." Available: https://arxiv.org/pdf/2105.15203.pdf