# Advancements in Medical Segmentation: Comparative Analysis of MedSegDiff-V2 and Swin UNETR in Brain and Skin Image Processing

Tianxing Ma, Yichen Lu, Dianze Li, Beining Zhu
Department of Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, MI 48105
{tianxinm, nechy, dianzeli, bxz}@umich.edu

## Abstract

*Medical image segmentation is crucial for diagnostic analysis and therapeutic interventions, and this paper introduces MedSegDiff-V2, a novel model that combines diffusion probabilistic models (DPMs) with vision transformers, tailored for medical imaging. Originally validated on brain imaging datasets, we extend its application to dermatological imaging to evaluate its generalization capabilities. Comparing MedSegDiff-V2 with traditional models like UNet, particularly in skin melanoma and brain tumor segmentation, reveals that MedSegDiff-V2 excels in handling ambiguous boundaries and complex patterns, often challenging in medical images. This study demonstrates the model's adaptability across different imaging modalities, showing significant enhancements in segmentation precision, which underscores its potential utility in clinical settings where accurate and reliable image analysis is crucial.*

*Our code is released on: https://github.com/ethan-charles/MedSegDiffsion & https://github.com/ethan-charles/swinUNETR*

## 1. Introduction

Medical image segmentation, which involves dividing a medical image into regions of interest, is pivotal for numerous clinical applications such as diagnostic analysis and image-guided interventions. Recently, diffusion probabilistic models (DPMs) have garnered attention due to their success in various computer vision tasks, including image generation and segmentation. Inspired by this, our project explores the MedSegDiff-V2 model, a novel application of DPM integrated with vision transformers for medical image segmentation, specifically adapted for dermatological imaging.

This study extends the use of MedSegDiff-V2, initially validated on brains, to assess its applicability and performance on skin images, and then comparing it with the Swin UNETR, a UNet-derived model. The project draws its motivation from the potential of DPMs to handle ambiguous boundaries in medical images, which is critical for accurately segmenting skin lesions that often exhibit irregular shapes and blurred edges.

## 2. Background

**Related Works**: Junde Wu et al.'s MedSegDiff and MedSegDiff-v2, which integrate DPMs with transformers for segmentation, and Swin UNETR, which uses Swin Transformers for enhanced spatial processing in medical imaging. These works demonstrate the superiority of transformers in capturing long-range dependencies compared to CNNs.

**Basic Knowledge**: Medical image segmentation and machine learning, including CNNs. An understanding of transformers, which use attention mechanisms for feature extraction, and diffusion models, which iteratively refine noise into structured outputs. In the project, we utilize the performance metrics like the Dice coefficient and Intersection over Union (IoU) for evaluating segmentation methods, which will be mentioned below.

## 3. Methodology

### 3.1. MedSegDiff-v2 [8]

The MedSegDiff-v2 model is an innovative approach to medical image segmentation that combines the Transformer architecture with a diffusion probabilistic model to enhance segmentation performance. This hybrid model is tailored for medical imaging tasks and addresses common challenges such as blurred boundaries in medical images.

MedSegDiff-v2 uses a generative approach to model the segmentation process as a reverse diffusion process. The model starts with the noise distribution and iteratively refines the segmentation mask through a series of steps that

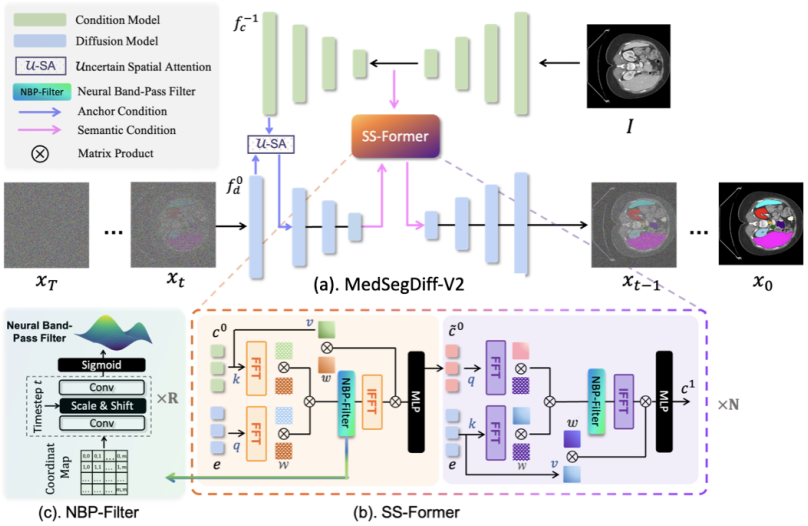



Figure 1. Illustration of the MedSegDiff-V2 Pipeline and Component Models

invert the noise addition. This process is mathematically modeled to gradually reduce noise to the image until clear segmentation is achieved.

At the heart of MedSegDiff-v2 lies the Transformer mechanism, which is used to enhance the feature extraction capabilities of the model. Unlike traditional UNet, which relies heavily on convolutional layers, MedSegDiff-v2 uses a transformer to process features extracted from raw images, allowing it to capture more complex patterns and relationships in the data.

The core innovation of MedSegDiff-v2 lies in its use of two distinct conditioning techniques during the diffusion process, which are termed Anchor Condition and Semantic Condition. The Anchor Condition employs an Uncertainty Spatial Attention (U-SA) mechanism that integrates conditional segmentation features into the diffusion model encoder, reducing variance during the diffusion process. The operation can be expressed as:

$$f_{\text{anc}} = \text{Max}(f_{-1} * k_{\text{Gauss}}, f_{-1}) \tag{1}$$

and

$$f_0' = \text{Sigmoid}(f_{\text{anc}} * k_{1\times1}) \cdot f_0 + f_0 \tag{2}$$

Here, $f_{-1}$ and $f_0$ denote conditional and diffusion features respectively, $k_{\text{Gauss}}$ is a learnable Gaussian kernel, and $k_{1\times1}$ is a convolutional kernel used to modulate the influence of the anchor feature. This method allows the model to generate more stable and accurate predictions by refining the diffusion steps based on input from the raw image and the evolving segmentation map.

The Semantic Condition, on the other hand, incorporates a novel Spectrum-Space Transformer (SS-Former). The SS-Former is designed to bridge the gap between the noise patterns inherent in diffusion models and the semantic features derived from medical images. It achieves this through a cross-attention mechanism that operates in the frequency domain, allowing for a refined integration of noise and semantic features across different stages of the diffusion process. Additionally, the model architecture includes a timestep-adaptive Neural Band-pass Filter (NBP-Filter), which aligns noise and semantic features for each timestep in the diffusion process, further enhancing the model's ability to generate precise segmentations, which can be expressed as:

$$\text{SS-Former} = \text{Cross-Attention}(F(c_0), F(e)) \tag{3}$$

and

$$M = (F(c_0)W_q)(F(e)W_k)^T \tag{4}$$

Here, $F$ denotes a Fourier transform applied to embeddings $c_0$ and $e$, $W_q$ and $W_k$ are learnable weights for the query and key in the Fourier space, representing the condition and noise embeddings, respectively.

The architecture employs a complex encoder-decoder structure in which the encoder utilizes a transformer mechanism to efficiently embed image features. The decoder then employs these transformed features and reconstructs the segmentation map step by step in a backpropagation process. The output of each step is conditioned on the output of the previous step and additional information in the original image, ensuring that each subsequent output is more refined.

Training of MedSegDiff-v2 is controlled by a composite loss function, which includes a standard noise prediction loss and an anchor loss. The anchor loss is a mixture of the soft dice loss and the cross-entropy loss, which helps

to efficiently supervise the conditional model and further refine the model's output. This is formulated as:

$$L_{\text{total}} = L_n + (t \equiv 0 \mod \alpha)(L_{\text{dice}} + \beta L_{\text{ce}}) \quad (5)$$

where $L_n$ is the noise prediction loss, $L_{\text{dice}}$ is the dice loss, $L_{\text{ce}}$ is the cross-entropy loss, $\alpha$ and $\beta$ are hyperparameters that manage the frequency and impact of the anchor loss on training.

MedSegDiff-v2 is designed to function across a wide range of medical imaging modalities, demonstrating versatility and robustness. Its effectiveness is validated by extensive evaluation of several medical image segmentation tasks where it consistently outperforms previous state-of-the-art methods.

### 3.2. Swin UNETR [1]

In our search for an effective tool for brain tumor segmentation, we have incorporated Swin UNETR into our approach. This novel architecture utilizes the strengths of Swin Transformers as its core coding mechanism to revolutionize the semantic segmentation of brain tumors in MRI images. Swin UNETR utilizes a U-shaped network design, which is commonly used for medical image segmentation due to its effective feature extraction and reconstruction capabilities.

At the core of Swin UNETR is the Swin Transformer encoder. It reimagines the segmentation task as a sequence-to-sequence prediction problem. By projecting multimodal input data into a one-dimensional embedded sequence, it can efficiently capture and utilize remote spatial information, which is often a challenge for traditional convolutional neural network (CNN) approaches.

Swin UNETR's encoder operates on a hierarchical basis, utilizing a shift window to compute self-attention:

$$\hat{Z}_{l+1} = \text{SW-MSA}(\text{LN}(Z_l)) + Z_l \quad (6)$$

and

$$Z_{l+1} = \text{MLP}(\text{LN}(\hat{Z}_{l+1})) + \hat{Z}_{l+1} \quad (7)$$

Here, SW-MSA is the shifted window multi-head self-attention, LN is layer normalization, MLP is a multi-layer perceptron, and $Z_l$ and $\hat{Z}_{l+1}$ represent the output feature maps at layers $l$ and $l+1$, respectively. This mechanism allows it to extract features at multiple resolutions, thus capturing a wide range of contextual information. These features are then directed to the FCNN-based decoder at each resolution stage via jump connections. This synergy between the Swin Transformer and the FCNN decoder ensures that the encoder captures detailed features and pinpoints them at the decoding stage, which is crucial for accurate segmentation of brain tumors.

The model is trained using the soft Dice loss function which is given by:

$$L(G, Y) = 1 - \frac{2\sum_{j=1}^{J}\sum_{i=1}^{I} G_{i,j}Y_{i,j}}{\sum_{j=1}^{J}\sum_{i=1}^{I} G_{i,j}^2 + \sum_{j=1}^{J}\sum_{i=1}^{I} Y_{i,j}^2} \quad (8)$$

In this function, $I$ denotes the total number of voxels, $J$ the number of classes, $Y$ the predicted probabilities, and $G$ the ground truth in one-hot encoded form.

## 4. Experiments

### 4.1. Datasets

**Brain Tumor Segmentation from MRI (BraTS):** The datasets provide NIfTI format brain scans including T1, post-contrast T1-weighted, T2-weighted, and T2-FLAIR images from 19 institutions, manually segmented and vetted by neuroradiologists.

**Melanoma Segmentation from Skin Images (ISIC):** The datasets provide dermoscopic lesion images in JPEG format and Associated segmentation binary masks in PNG format.

### 4.2. Implementation Details

Our experiments utilized the PyTorch platform and were conducted on a single NVIDIA RTX 4090 GPU. We standardized the image resolution to 256×256 pixels for uniform processing. The networks were trained end-to-end using the AdamW optimizer [4] with a batch size of 32, and an initial learning rate of $1 \times 10^{-4}$. The models underwent 1000 diffusion steps during inference, a reduction from the 25 repetitions used in MedSegDiff [7]. For ensemble methods, we ran the model four times and employed the STAPLE algorithm [6] to merge the results. We assessed model performance using gradient norm and loss functions.

For data management, we developed a new data loader method, which is shown in Appendix A that processes image lists recorded in a ".json" file, formatting as Appendix B.

We initially targeted the BraTS datasets for our segmentation tasks, but found the process exceedingly time-consuming, with over 12 hours required for every 5000 training steps. Consequently, we opted to test the performance of our MedSegDiff models on the smaller ISIC dataset. Although these trials demonstrated strong generalizability, the performance was suboptimal. Since the model outputs are now limited to only the contours of skin lesions, we developed an algorithm utilizing OpenCV in Python to generate the final segmentation images, whose code is shown in Appendxi C. This approach allows us to transform contour data into comprehensive segmentation maps effectively. This led us to integrate a similar framework, Swin UNETR, to refocus our efforts on the BraTS datasets.

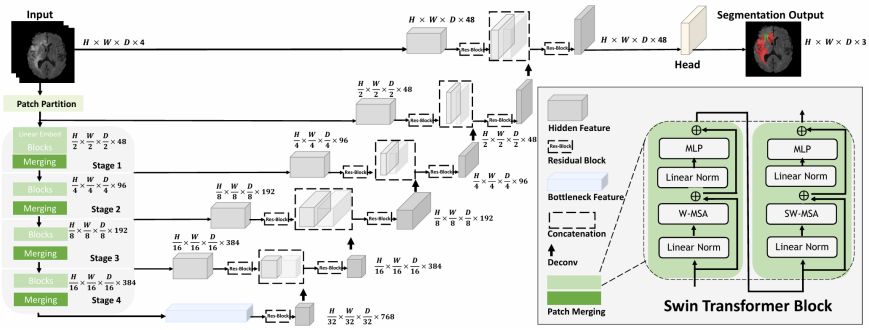



Figure 2. Architecture of the Swin UNETR

### 4.3. Evaluation

In evaluating segmentation performance, we adopted standardized metrics such as the Dice score and Intersection over Union (IoU).

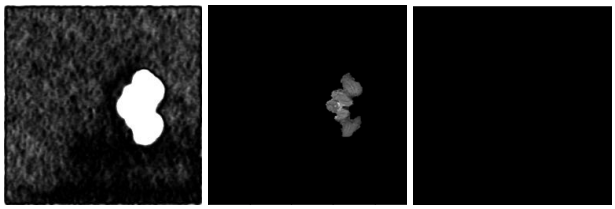
$$Dice = \frac{2 \times Area\ of\ overlap}{Total\ area} \tag{9}$$

$$IoU = \frac{Area\ of\ overlap}{Area\ of\ union} \tag{10}$$

## 5. Results

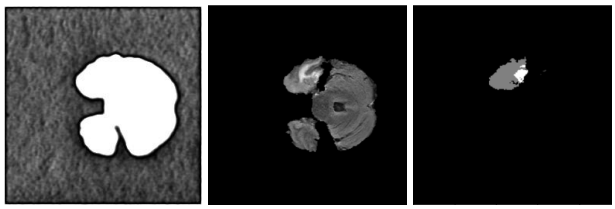
### 5.1. Brain Tumor Segmentation on MedSegDiff V2

Results after training 5000 steps



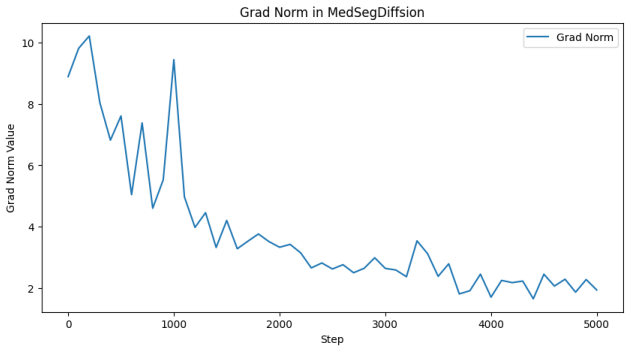
(a) Generated Output (b) Flair Image (c) Ground Truth

Figure 3. Example figures from slice 14 of BraTS20_Training_009

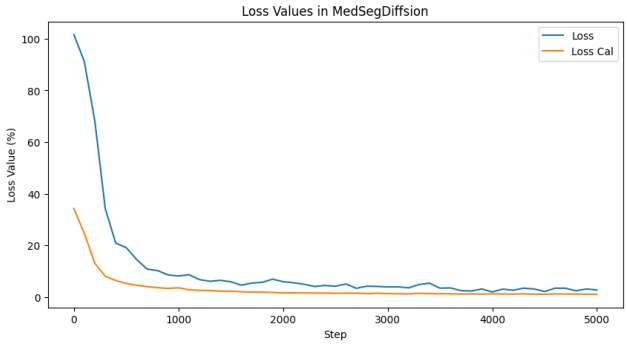


(a) Generated Output (b) Flair Image (c) Ground Truth

Figure 4. Example figures from slice 41 of BraTS20_Training_009




(a) Gradient Norm of the model

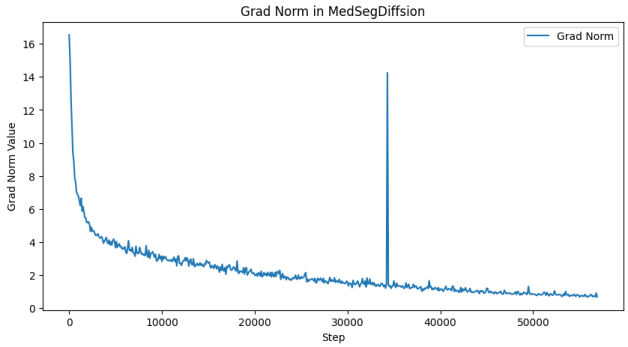



(b) Loss of the model

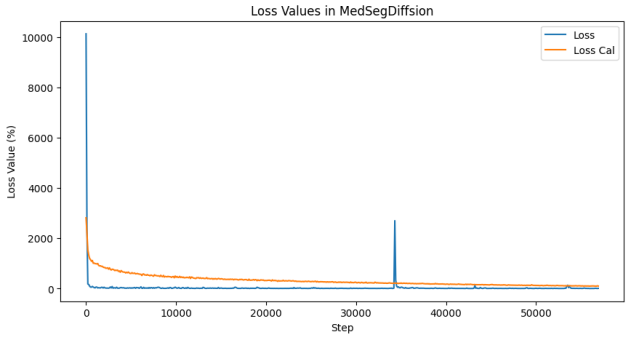
Figure 5. Gradient Norm and Loss of MedSegDiff-v2 on Brain Tumor dataset

According to the figure 5, the grad norm and the loss converge that mean the model has sufficiently trained. However, figure 3 and figure 4 shows the model only learn how to generate the whole brain rather than brain tumor. One possible reason from the author of this paper says is we need train more, which is deagree with the convergence of the training process. Another possible reason is that during the U-SA processing, the model weighting to much on the origianl image rather than segmentation, which needs further modification of original model.

## 5.2. Skin Melanoma Segmentation on MedSegDiff V2

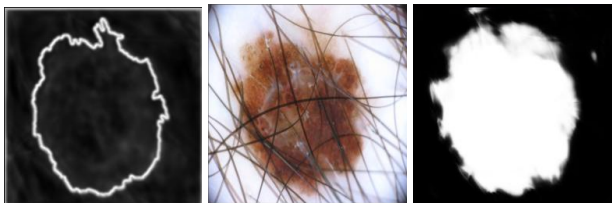



(a) Gradient Norm of the model




(b) Loss of the model

Figure 6. Gradient Norm and Loss of MedSegDiff-v2 on Skin Melanoma dataset

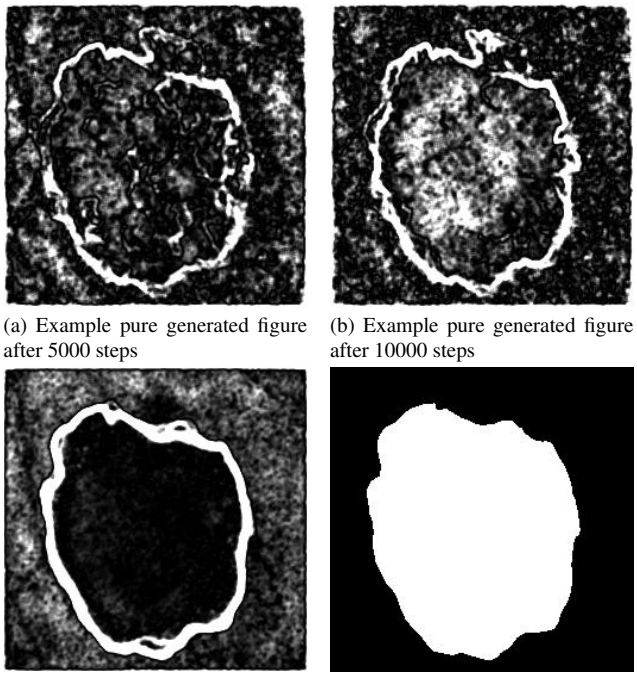


(a) Generated Output (b) Skin Image (c) Ground Truth

Figure 7. Comparation of the pure output of original model and the ground truth on Skin Melanoma dataset

In this study, we observed that the loss curve converged rapidly within the initial 10,000 steps. However, the generated outputs were suboptimal for segmentation tasks, as illustrated in Figure 8. We assessed the ensemble performance of various diffusion-based medical segmentation models over training intervals ranging from 5,000 to 50,000 steps. Our analysis, as depicted in the diagram, suggests that extending training durations significantly improves model performance.

Following noise reduction and filling outlined areas, we processed the contours of skin lesions to produce black-and-white images, enhancing the clarity of the segmentation and facilitating compression.



(a) Example pure generated figure after 5000 steps (b) Example pure generated figure after 10000 steps

(c) Example pure generated figure after 50000 steps (d) Example processed generated figure after processing

Figure 8. Example generated figures in different stages of training MedSegDiff V2 on Skin Melanoma dataset

| Model | dataset | Dice | IoU |
|---|---|---|---|
| MedSegDiffsion | ISIC | 91.3 | 84.1 |
| MedSegDiffsion-v2 | ISIC | 78.2 | 70.5 |

Table 1. Model Performance Comparison on ISIC dataset

## 5.3. Brain Tumor Segmentation on Swin UNETR

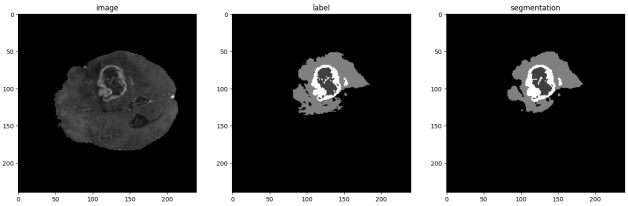



Figure 9. Comparation of the pure output of original model and the ground truth

The chart clearly demonstrates that the Swin Unet Transformer model can generate segmentation for various areas

in different gray levels. This model achieves superior Dice scores compared to the MedSegDiffsion model, although it exhibits lower Intersection over Union (IoU) values. This discrepancy is primarily attributed to the scaling laws affecting our model's performance on the test sets. We will explore this issue in the discussion section.

| Model | dataset | Dice | IoU |
|---|---|---|---|
| MedSegDiffsion | BraTS | 88.9 | 81.2 |
| Swin UNETR | BraTS | 89.2 | 80.5 |

Table 2. Model Performance Comparison on BraTS dataset

### 5.4. Analysis

Scaling laws describe the relationship between enhancements in system performance, such as accuracy or speed, and increases in resources such as data, computational power, or model size. The performance of our models is significantly constrained by the available computing resources and the size of the datasets, as well as the time of fine tuning.

Although the MedSegDiffsion-v2 model was capable of producing closed segmentation contours, the presence of numerous outliers and an insufficient number of training and diffusion steps led to a performance that did not surpass its predecessor. The original codebase was unclear and contained several bugs, particularly problematic given the modifications to the skin datasets among versions 1 to 3B. These issues likely contributed to the difficulty in adjusting parameters and fine-tuning the models, which we believe are fundamental reasons for MedSegDiffsion's suboptimal performance.

In contrast, the Swin Unet model has met our expectations, achieving results very close to those reported for MedSegDiffsion in the literature. We are optimistic about the potential of the Swin model and anticipate that further fine-tuning could yield even higher performance.

## 6. Conclusion

### 6.1. Discussion

In this project, we began an extensive evaluation of the MedSegDiff-v2 model, initially applying it to the brain dataset. Our results show that while MedSegDiff-v2 is promising for medical image segmentation, its performance on the brain data is not ideal, characterized by insufficient segmentation quality and high computational demands, especially due to the 3D nature and size of the datasets involved.

To address these limitations, we shifted our focus to skin datasets, which are inherently two-dimensional and thus require less computational power for processing. In initial experiments, MedSegDiff-v2 produced contour segmentation results that were too simple to transform skin images into grayscale maps. To address this issue, we added a processing technique that resulted in ground truth-like grayscale map results. These adjustments yielded significantly improved outcomes, achieving segmentation results that closely aligned with the ideal.

Building on this foundation, we then applied the Swin UNETR architecture to brain datasets. Swin UNETR integrates a shift-window based self-attention mechanism in U-Net, which proved to be more suitable for capturing complex patterns in brain scans.

Our results show that Swin UNETR not only improves segmentation results but also runs more efficiently compared to our initial experiments using MedSegDiff-v2.

### 6.2. Future Work

During implementation of MedSegDiff V2, performance is not ideal. To further achieve the performance mentioned in the paper, we need to further modify the codes in published version of this model or try to rebuild the model from scratch.

Currently, there is a trend that Mamba is replacing Transformer in generation AI owing to its strong long context retrieve ability. And there is already some work to combine mamba with Unet such as U-Mamba [5], Vmamba [3] and Swin-UMamba [2]. So it is nature to think about apply Mamba on diffusion model to enable model to directly process the whole 3D image rather than slice and process them seperately.

## References


[1] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images, 2022. 3

[2] Jiarun Liu, Hao Yang, Hong-Yu Zhou, Yan Xi, Lequan Yu, Yizhou Yu, Yong Liang, Guangming Shi, Shaoting Zhang, Hairong Zheng, and Shanshan Wang. Swin-umamba: Mamba-based unet with imagenet-based pretraining, 2024. 6

[3] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model, 2024. 6

[4] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 3

[5] Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation, 2024. 6

[6] S.K. Warfield, K.H. Zou, and W.M. Wells. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921, 2004. 3

[7] Junde Wu, Rao Fu, Huihui Fang, Yu Zhang, Yehui Yang, Haoyi Xiong, Huiying Liu, and Yanwu Xu. Medsegdiff: Med-

ical image segmentation with diffusion probabilistic model, 2023. 3

[8] Junde Wu, Wei Ji, Huazhu Fu, Min Xu, Yueming Jin, and Yanwu Xu. Medsegdiff-v2: Diffusion based medical image segmentation with transformer, 2023. 1

# Appendices

## A. Create Standard Format in .json

```python
def create_json(directory, output_file):
    data = {"training": []}
    subdirectories = [os.path.join(directory, d) for d in os.listdir(directory) if os.path.isdir(os.path.join(directory, d))]
    num_directories = len(subdirectories)
    fold_0_limit = int(0.7 * num_directories)  # 70% of the directories

    for idx, subdir in enumerate(subdirectories):
        fold_number = 0 if idx < fold_0_limit else 1
        entry = {
            "fold": fold_number,
            "image": [],
            "label": ""
        }

        for file in os.listdir(subdir):
            filepath = os.path.join(subdir, file)
            if file.endswith(".nii.gz"):
                if "seg" in file:
                    entry["label"] = filepath
                else:
                    entry["image"].append(filepath)

        if entry["image"] and entry["label"]:
            data["training"].append(entry)

    with open(output_file, 'w') as f:
        json.dump(data, f, indent=4)
```

## B. .json file format

```
{
    "training": [
        {
            "fold": 0,
            "image": [
                os.path.join(dir_path, "BraTS20_Training_250/BraTS20_Training_250_flair.nii.gz"),

                os.path.join(dir_path, "BraTS20_Training_250/BraTS20_Training_250_t1.nii.gz"),

                os.path.join(dir_path, "BraTS20_Training_250/BraTS20_Training_250_t1ce.nii.gz"),

                os.path.join(dir_path, "BraTS20_Training_250/BraTS20_Training_250_t2.nii.gz")
            ],
            "label":
                os.path.join(dir_path, "BraTS20_Training_250/BraTS20_Training_250_seg.nii.gz")
        },
        ...
    ]
}
```

## C. Segmentation Image Processing

```python
def pred_process(img):
    border_width = 4
    img[:border_width, :] = 0
    img[-border_width:, :] = 0
    img[:, :border_width] = 0
    img[:, -border_width:] = 0

    blurred_img = cv2.GaussianBlur(img, (5, 5), 0)

    _, threshold = cv2.threshold(blurred_img, 127, 255, cv2.THRESH_BINARY)
    contours, _ = cv2.findContours(threshold, cv2.RETR_EXTERNAL, cv2.CHAIN_APPROX_SIMPLE)
    contour_img = np.zeros_like(img)
    cv2.drawContours(contour_img, contours, -1, (255), thickness=cv2.FILLED)

    labeled_array, num_features = label(contour_img)

    # Find the largest object
    largest_object = np.argmax(np.bincount(labeled_array.flat)[1:]) + 1

    # Create an array where only the largest object is white, and everything else is black
    largest_object_array = np.where(labeled_array == largest_object, 255, 0).astype(np.uint8)

    # Convert back to image
    largest_object_image = Image.fromarray(largest_object_array)

    return largest_object_image
```