# Adaptive Dual Attention into Diffusion for 3D Medical Image Segmentation

Nhu-Tai Do[1], Van-Hung Bui[2], and Quoc-Huy Nguyen*[1]

[1] Saigon University, Vietnam
[2] Ho Chi Minh University of Science, Vietnam
dntai@sgu.edu.vn,19127414@student.hcmus.edu.vn,nqhuy@sgu.edu.vn
* Corresponding author: nqhuy@sgu.edu.vn

**Abstract.** Denoising diffusion models have recently demonstrated great success in generating detailed pixel-wise representations for image generation. Applications like Dall-E, Stable Diffusion, and Midjourney have showcased impressive image-generation capabilities, sparking significant discussion within the community. Recent studies have also highlighted the utility of these models in various other vision tasks, including image deblurring, super-resolution, and image segmentation. This work introduces a novel Adaptive Dual Attention into Diffusion model for 3D medical image segmentation. Applying diffusion models to 3D medical image segmentation presents significant challenges. The alignment of semantic features necessary for conditioning the diffusion process with noise embedding is often inadequate. Additionally, traditional U-Net backbones in diffusion models are not sufficiently sensitive to the contextual information required for accurate pixel-level segmentation during reverse diffusion. Our method, which integrates Adaptive Dual Attention into Diffusion, addresses these issues by capturing local and global contextual information, enhancing the precision and robustness of 3D image segmentation. Our approach surpasses current state-of-the-art methods on the BraTS2020 dataset, achieving higher segmentation accuracy. This improved performance can significantly aid in diagnosing and treating medical conditions by enabling highly accurate segmentation of anatomical structures in 3D medical images.

**Keywords:** Diffusion Probabilistic Model · Medical Image Segmentation · Dual Domain Attention · Volumetric Data.

## 1 Introduction

Medical volumetric segmentation is a critical task in medical image analysis, enabling precise delineation of anatomical structures and lesions within high-dimensional datasets on a voxel-by-voxel basis [10]. Accurate segmentation provides essential diagnostic information to clinicians, aiding in disease diagnosis, treatment planning, and patient management. Traditional methods, typically utilizing encoder-decoder architectures with skip connections, often fall short in

capturing global contextual information due to their reliance on convolutional neural networks (CNNs) with localized receptive fields [18].

Recent advancements in transformer-based models have shown promise in addressing these limitations by leveraging self-attention mechanisms to capture global dependencies. Models such as SwinUNETR [5] and UNETR [4] have demonstrated improved segmentation accuracy by integrating transformers with CNNs. However, these models still encounter challenges in efficiently extracting multi-scale features and handling the high computational complexity associated with volumetric data [9].

Denoising diffusion models have emerged as potent tools for generating semantically meaningful pixel-wise representations [8]. These models have been successfully applied to various generative tasks, including image synthesis and restoration. In medical image segmentation, diffusion models offer the advantage of progressively refining image representations, making them well-suited for generating detailed segmentations. However, existing diffusion-based approaches are often constrained to 2D and binary segmentation tasks, limiting their applicability in more complex medical imaging scenarios [19,1].

### 1.1 Difference Between 2D and 3D Medical Image Segmentation

Medical image segmentation can be approached in two-dimensional (2D) and three-dimensional (3D) contexts, each presenting unique challenges and considerations. **2D Medical Image Segmentation:** In 2D segmentation, images are processed slice by slice, treating each slice independently. This approach simplifies computational complexity and reduces memory requirements. However, it often fails to capture spatial continuity and contextual information across slices, leading to inconsistencies in the segmentation of adjacent slices. This limitation is particularly problematic in medical imaging, where anatomical structures extend across multiple slices.

**3D Medical Image Segmentation:** Conversely, 3D segmentation methods process volumetric data, simultaneously considering the entire stack of slices. This approach leverages the spatial relationships and continuity inherent in volumetric data, leading to more accurate and coherent segmentation results. However, 3D segmentation poses significant computational challenges due to the increased data dimensionality and memory requirements. Models designed for 3D segmentation must efficiently handle high-resolution volumetric data while capturing local and global contextual information.

This work focuses on 3D medical image segmentation to fully exploit the rich spatial information available in volumetric scans, such as MRI and CT images. By designing our framework to operate on 3D data, we aim to achieve higher segmentation accuracy and robustness than 2D approaches.

### 1.2 Contributions

The primary motivation for developing the Adaptive Dual Attention Diffusion (ADAD) framework stems from the limitations of traditional U-Net architec-

tures and the challenges associated with aligning semantic features with noise embeddings in diffusion models. Traditional U-Nets, while effective for many segmentation tasks, cannot often capture global context due to their localized receptive fields. This limitation becomes pronounced in 3D medical image segmentation, where understanding the broader spatial relationships within the volume is crucial for accurately delineating anatomical structures.

Diffusion models, with their iterative denoising processes, offer a unique advantage in progressively refining image representations. However, effectively integrating semantic information with noise embeddings throughout the diffusion process remains challenging. Our approach addresses this by introducing dual-domain attention mechanisms that enhance the model's sensitivity to both local details and global context.

Our main contributions are as follows:

- **We introduce an end-to-end framework called Adaptive Dual Attention Diffusion (ADAD):** Our ADAD framework is designed to enhance 3D medical image segmentation by leveraging a 3D image diffusion model to improve segmentation accuracy.
- **We embed a 3D Dual Domain Attention (3D-DDA) block:** To augment the robustness and accuracy of our model, we incorporate a 3D-DDA block [2] within the feature encoder section of the network. This block captures local and global features simultaneously, enhancing the network's capability to handle the variability and complexity of medical images.
- **We introduce Dynamic Conditional Encoding:** To address the challenge of ambiguous and low-contrast regions in medical images, we integrate a dynamic conditional encoding strategy. This approach incorporates current-step segmentation information into the raw image encoding, dynamically enhancing segmentation accuracy at each step.
- **We implement a Step-Uncertainty based Fusion (SUF) module:** During the inference phase, we utilize an SUF module to aggregate predictions from multiple diffusion steps. This module leverages uncertainty information to produce more accurate segmentation results, improving the model's robustness.

We evaluated our ADAD framework on the BraTS2020 benchmark dataset, demonstrating its superior performance compared to state-of-the-art methods. Our results underscore the potential of the ADAD framework to significantly advance medical volumetric segmentation significantly, providing more precise and reliable tools for clinical diagnostics and treatment planning. This innovative approach holds promise for improving patient outcomes by enabling more accurate and detailed segmentation of anatomical structures.

## 2 Proposed method

### 2.1 Overview

The overall process is shown in Fig.1. ADAD captures local and global features by combining advanced diffusion models with attention mechanisms, improving

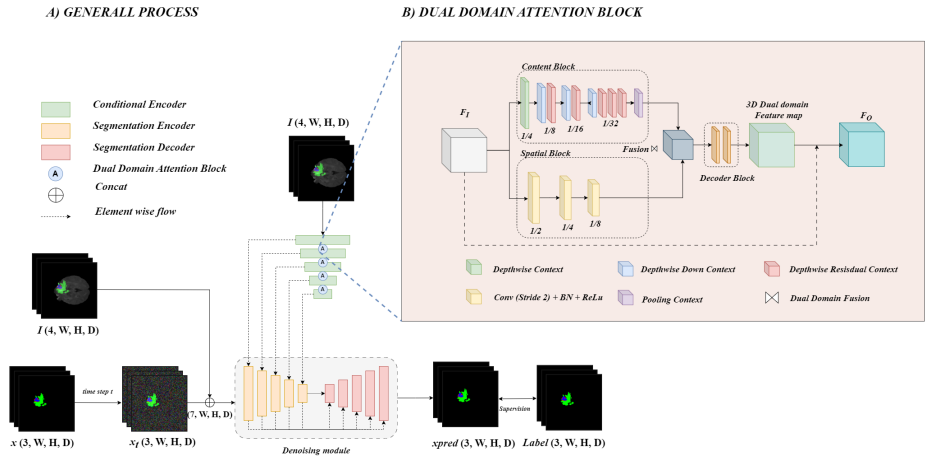



Fig. 1: The proposed method: A) General process of diffusion for 3D medical image segmentation, B) Detailed structure of the proposed 3D Dual Domain Attention (3D-DDA) block.

segmentation accuracy. The key components include the conditional encoder, segmentation encoder, segmentation decoder, and the 3D dual-domain attention (3D-DDA) block, each playing a critical role in the workflow.

The process begins with the input of 3D medical images $I$, which have dimensions $(4, W, H, D)$, where $W$, $H$, and $D$ represent the width, height, and depth of the image, respectively. These images undergo initial preprocessing, resulting in a noisy version $x$ used for the diffusion process. Unlike traditional U-net methods that use volume data to predict segmentation labels directly, the diffusion model learns by removing noise. It takes a noisy volumetric image and segmentation labels as input and gradually removes the noise to create clear segmentation results.

The conditional encoder processes the input image $I$, generating essential feature representations that condition the diffusion process at each time step $t$. These feature representations are crucial for guiding the segmentation process, ensuring the model retains important contextual information from the original image. The 3D-DDA block is strategically placed within the conditional encoder to enhance the feature extraction process. The 3D-DDA block employs dual-domain attention mechanisms to capture local and global features, significantly improving the network's ability to handle the complexity and variability inherent in medical images.

At each time step $t$, the segmentation encoder integrates the features from the conditional encoder and the 3D-DDA block, producing an intermediate representation $x_t$. These concatenated features $(I, x_t)$ are passed through the denoising module. This module consists of a series of convolutional layers designed to denoise the input and refine the segmentation progressively.

The segmentation labels $x_0$ are first converted into one-hot encoded labels. To generate the noisy labels $x_t$ used in the diffusion process, we apply the forward diffusion equation:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, I)$ is Gaussian noise, and $\bar{\alpha}_t$ is a predefined variance schedule controlling the amount of noise added at each time step $t$.

In Fig.1, this process is depicted where the segmentation labels $x_0$ are transformed into $x_t$ through the addition of Gaussian noise, preparing them for the reverse diffusion process during training.

The final predicted segmentation $x_{\text{pred}}$ is supervised using the ground truth labels Label, which have dimensions $(3, W, H, D)$. The supervision process involves comparing the predicted segmentation with the ground truth labels and calculating the loss. This loss is then backpropagated through the network to optimize the parameters, enhancing the model's accuracy and robustness over successive training iterations.

### 2.2 Adptive Dual Domain Attention Encoder

As shown in Fig.1, the conditional encoder is designed to process the input image and extract rich features necessary for accurate segmentation. The conditional encoder initially processes the input image $I$ to generate feature representations crucial for conditioning the diffusion process at each time step $t$. These features retain essential contextual information from the original image, ensuring the subsequent segmentation steps are well-informed.

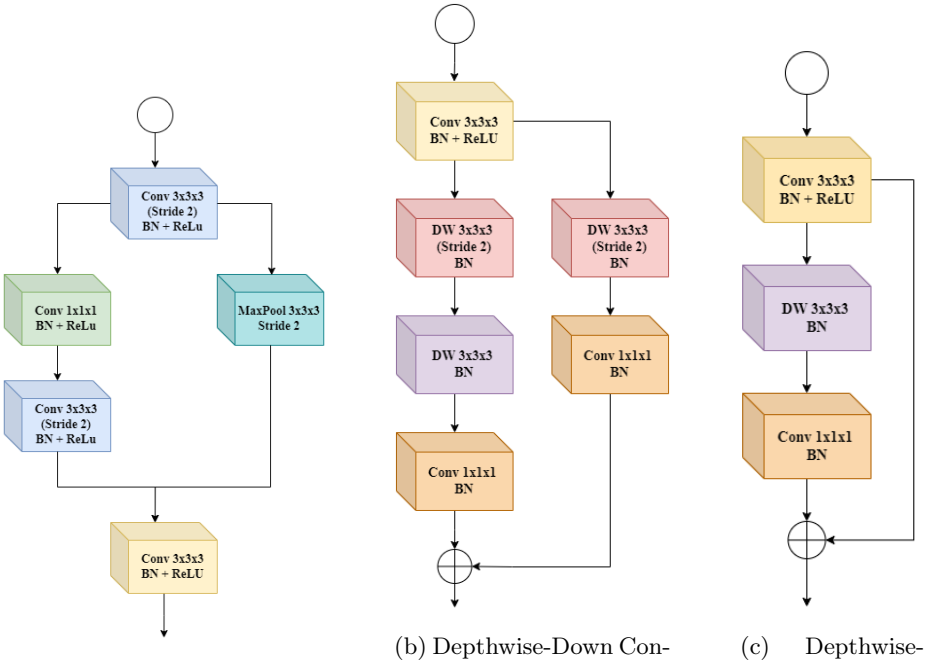
**3D Dual Domain Attention (3D-DDA) block:** The 3D Dual Domain Attention (3D-DDA) block is a crucial component of the ADAD framework, designed to capture both local spatial details and global contextual information. The architecture of the 3D-DDA block integrates spatial-domain and context-domain attentions, leveraging residual learning to refine feature maps at each stage of the encoder. The 3D-DDA block operates on the encoding feature map $\mathbf{F}_I$ and produces a refined feature map $\mathbf{F}_O$ by combining spatial and contextual information. Mathematically, this can be expressed as:

$$\mathbf{F}_O = \mathbf{F}_I + \text{DDA3D}(\mathbf{F}_I) \tag{2}$$

As shown in Equation 2, the output of the 3D-DDA block is added to the input feature map $\mathbf{F}_I$ to produce the refined feature map $\mathbf{F}_O$. This residual connection helps preserve the original features while enhancing them with attention mechanisms.

The dual-domain attention mechanism within the 3D-DDA block consists of two main components: the Spatial-Domain Block ($\mathcal{S}$) and the Context-Domain Block ($\mathcal{C}$). These components work together to enhance the feature map by focusing on different aspects of the data.

**Spatial-Domain Block:** The spatial-domain block ($\mathcal{S}$) stacks $k$ convolutional layers to capture fine-grained spatial details. Each block $f_i$ is defined:




(a) Inception Context (b) Depthwise-Down Context (c) Depthwise-Residual Context

Fig. 2: 3D Context-Domain Block details

$$\mathcal{S}(\mathbf{F}_I) = f_k \circ f_{k-1} \circ \cdots \circ f_1(\mathbf{F}_I) \tag{3}$$

where

$$f_i(\mathbf{x}) = \sigma(\text{BN}(\text{Conv3D}_{s=2}(\mathbf{x}))) \tag{4}$$

Here, $k$ denotes the number of convolutional layers, $\sigma$ is the ReLU activation function, and $\circ$ represents the composition of layers.

**Context-Domain Block:** The context-domain block ($\mathcal{C}$) consists of $t$ sequential components designed to capture global contextual information:

$$\mathcal{C}(\mathbf{F}_I) = g_{\text{pool}}\left(f_t \circ f_{t-1} \circ \cdots \circ f_1(\mathbf{F}_I)\right) \tag{5}$$

Each component $f_i$ is defined as:

$$f_i(x) = g_{\text{res}_i} \circ g_{\text{down}_i}(x) \tag{6}$$

where $g_{\text{res}_i}$ represents the $i$-th residual block, and $g_{\text{down}_i}$ denotes the downsampling operation.

**Dual-Domain Fusion**: Before combining the spatial-domain feature map $\mathcal{S}(\mathbf{F}_I)$ and the context-domain feature map $\mathcal{C}(\mathbf{F}_I)$, we apply resizing operations to ensure they have the same spatial dimensions. Specifically, we upsample the context-domain feature map using trilinear interpolation:

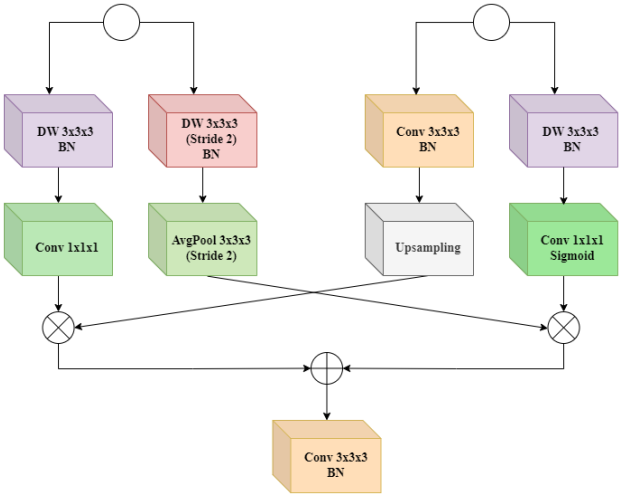



Fig. 3: Dual Domain Fusion Block

$$\tilde{\mathcal{C}}(\mathbf{F}_I) = \text{Upsample}(\mathcal{C}(\mathbf{F}_I)) \tag{7}$$

With both feature maps aligned in size, we perform element-wise addition:

$$\text{DDA3D}(\mathbf{F}_I) = \varphi_s \left( \mathcal{D} \left( \mathcal{S}(\mathbf{F}_I) + \tilde{\mathcal{C}}(\mathbf{F}_I) \right) \right) \tag{8}$$

This fusion effectively combines local and global information, enhancing the feature representation for segmentation tasks.

**3D Decoder Block ($\mathcal{D}$)**: The 3D decoder block returns the fused feature map to the original resolution while preserving the learned features. this block consists of three transposed convolution layers, each paired with batch normalization and a ReLU activation unit, followed by a Conv3D layer with a kernel size of $1 \times 1 \times 1$. This setup decodes and normalizes the dual-domain feature map, resizing it to match the dimensions of the input feature map. The operation of the decoder block can be represented as:

$$\mathcal{D}(\mathbf{F}_{\text{fused}}) = \text{ConvTranspose3D}_{\text{BN, ReLU}}(\mathbf{F}_{\text{fused}}) \tag{9}$$

where $\mathbf{F}_{\text{fused}}$ is the fused feature map from the dual-domain fusion block.

The decoder block utilizes skip connections from corresponding encoder layers to retain high-resolution features and combine them with the upsampled feature maps. This approach helps preserve spatial details and improves overall segmentation accuracy. The final output of the decoder block is a refined feature map that has been upsampled to the original resolution of the input image, ready for generating the final segmentation map.

## 3 3D Diffusion

### 3.1 Label Embedding:

This study introduces a Label Embedding operation to overcome the limitation of traditional diffusion models that can only handle binary segmentation. This operation converts the segmentation label map into one-hot encoded labels, enabling the model to segment multiple targets simultaneously [7]. The one-hot encoding process converts a label $x_0$ with dimensions $D \times W \times H$ into a multi-channel label $x_0 \in \mathbb{R}^{N \times D \times W \times H}$, where $N$ is the number of labels.

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon \tag{10}$$

where $x_t$ is the label map with added noise at time step $t$, $\bar{\alpha}_t$ is a scaling factor, and $\epsilon$ is the Gaussian noise. The goal is to predict the clear label map $x_0$ from the noisy input $x_t$.

### 3.2 Denoising Module:

The Denoising Module in the framework consists of an Adaptive Dual Domain Attention Encoder (ADDA Encoder) and a Denoising-UNet (DU). The ADDA Encoder extracts multi-scale features from the input volume data $I$, which are then combined with the noisy one-hot encoded label $x_t$ and input into the encoder part of the DU. Subsequently, the DU processes these features through its encoder-decoder architecture to generate the denoised output [20]. This process can be described as:

$$\hat{x}_0 = \mathrm{DU}(\mathrm{cat}(I, x_t), t, \tilde{I}_f) \tag{11}$$

where $\hat{x}_0$ is the predicted clear label map, $\tilde{I}_f$ represents the multi-scale features from the ADDA Encoder, and cat denotes channel-wise concatenation. The integration of the ADDA Encoder and Denoising-UNet (DU) in the Denoising Module represents an advanced approach to image denoising. By utilizing the multi-scale features extracted by the ADDA Encoder and incorporating the noisy label information, the DU effectively denoises the input data, demonstrating the potential of deep learning models in addressing noise reduction challenges.

The Denoising-UNet (DU) is trained using a combination of loss functions to optimize its performance. Specifically, the training process incorporates the Dice Loss, Binary Cross-Entropy (BCE) Loss, and Mean Squared Error (MSE) Loss. The Dice Loss function evaluates the overlap between predicted and ground truth segmentation masks, providing a pixel-wise measure of accuracy. The BCE Loss function focuses on the classification aspect, aiding in the pixel-wise segmentation of the input data. Additionally, the MSE Loss function contributes to minimizing the differences between the generated denoised output and the ground truth data, ensuring the preservation of essential details and structures in the denoised images [17].

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{dice}}(\hat{x}_0, x_0) + \mathcal{L}_{\text{bce}}(\hat{x}_0, x_0) + \mathcal{L}_{\text{mse}}(\hat{x}_0, x_0) \tag{12}$$

### 3.3 Step-Uncertainty based Fusion:

The testing phase of the diffusion model involves iteratively refining the segmentation prediction through multiple steps using the Denoising Diffusion Implicit Models (DDIM) method. In addition to the final prediction from the last step, the Step-Uncertainty based Fusion (SUF) module is incorporated to amalgamate predictions from each iterative step. The uncertainty for each prediction step is estimated similarly to Monte Carlo Dropout [3], where the average prediction $\bar{p}_i$ and the uncertainty $u_i$ are computed as:

$$u_i = -\bar{p}_i \log(\bar{p}_i), \quad \text{where} \quad \bar{p}_i = \frac{1}{S} \sum_{s=1}^{S} p_{s,i} \tag{13}$$

and $S$ is the number of forward passes. The fusion weights $w_i$ for each prediction step are determined by combining the step index and uncertainty, expressed as:

$$w_i = \exp(\sigma(\frac{i}{\text{scale}}) \times (1 - u_i)) \tag{14}$$

where $\sigma$ is the sigmoid function, and scale is a normalization factor. The final segmentation result $Y$ is obtained by weighting the predictions from all steps:

$$Y = \sum_{i=1}^{t} w_i \times \bar{p}_i \tag{15}$$

## 4 Experiments

### 4.1 Dataset and Experimental Setup

**Dataset:**The BraTS2020 [12] dataset comprises 369 pre-aligned MRI scans across four modalities: T1, T1ce, T2, and FLAIR. These scans are accompanied by expert-annotated segmentation masks, which identify GD-enhancing tumors, peritumoral edema, and tumor core regions. Each modality is represented as a 155×240×240 volume, with all images resampled and co-registered to ensure uniformity. The segmentation task focuses on delineating the whole tumor (WT), enhancing tumor (ET), and tumor core (TC) areas. For model training and evaluation, the dataset is divided into training, validation, and test sets with a ratio of 0.7, 0.1, and 0.2, respectively.

**Training Process:** We randomly cropped the input scans and corresponding ground-truth labels to a size of 128×128×128, transformed them to a spacing of 1 mm, and applied random flips along the axial, sagittal, and coronal axes. The intensity of each voxel was adjusted by 10% and normalized across each channel. Labels were converted to one-hot encoding. The training was conducted over 150 epochs using the Adam optimizer with an initial learning rate of 0.0001, batch size 1. The learning rate schedule followed a Cosine Annealing approach, with a minimum learning rate of 0.00001. All experiments were performed on NVIDIA Tesla P100 16GB hardware, utilizing the PyTorch and MONAI frameworks. The visualization of the results is presented in 4.

Table 1: Results on conventional networks with and without 3D-DDA

| **Validation set** | **Dice Score** | | |
|---|---|---|---|
| | **WT** | **TC** | **ET** |
| V-Net [13] | 0.8010 | 0.7010 | 0.6840 |
| Swin-UNETR [4] | 0.8221 | 0.7482 | 0.7365 |
| SegResNet [14] | 0.8958 | 0.7911 | 0.7275 |
| DynUnet [6] | 0.8809 | 0.7857 | 0.7339 |
| SegResNet + 3D-DDA [2] | 0.8959 | 0.7911 | 0.7361 |
| **Diffusion+3D-DDA(Our)** | **0.9155** | **0.8475** | **0.7364** |

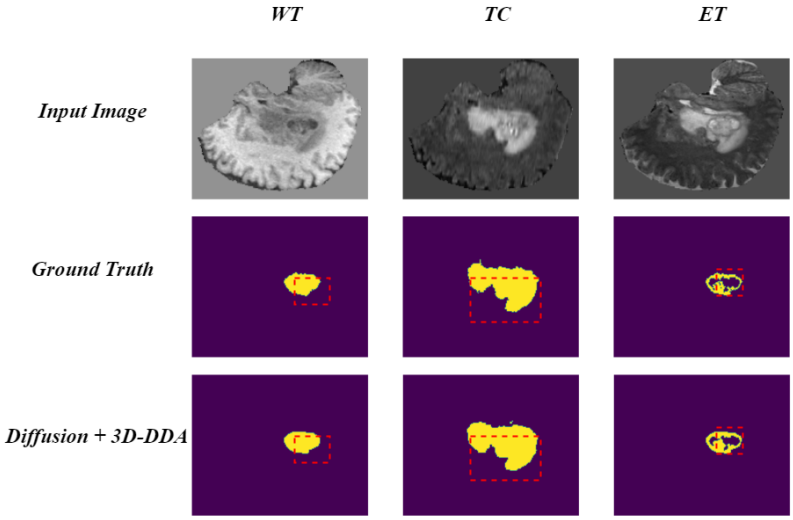



Fig. 4: Visualization of segmentation results for Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET) on the BraTS2020 dataset.

### 4.2 Comparison with SOTA Methods

Table 1 presents the validation set Dice scores for the whole tumor (WT), tumor core (TC), and enhancing tumor (ET) regions across different network architectures, both with and without the 3D Dual Domain Attention (3D-DDA) block. Our proposed method, Diffusion+3D-DDA, demonstrates superior performance, achieving Dice scores of 0.9155, 0.8475, and 0.7364 for WT, TC, and ET, respectively. This outperforms all other conventional networks evaluated. For instance, the SegResNet with 3D-DDA achieves a Dice score of 0.8959 for WT, significantly lower than our model's score of 0.9155. The inclusion of the 3D-DDA block generally enhances the performance of the networks, as seen in the improvement of the SegResNet's ET Dice score from 0.7275 to 0.7361 when 3D-DDA is added. Compared to SegResNet with 3D-DDA, our proposed

Table 2: Comparison on the state-of-the-art methods

| **Validation set** | **Year** | **Dice Score** | | |
|---|---|---|---|---|
| | | **WT** | **TC** | **ET** |
| Nuechterlein et al. [15] | 2019 | 0.8834 | 0.8141 | **0.7369** |
| Zou et al. [21] | 2020 | 0.8658 | 0.7688 | 0.7443 |
| GLF-Net [11] | 2020 | 0.8930 | | |
| LGMSU-Net [16] | 2022 | 0.8735 | | |
| SegResNet + 3D-DDA [2] | 2023 | 0.8959 | 0.7911 | 0.7361 |
| **Our proposed method** | | **0.9155** | **0.8475** | 0.7364 |

method shows an improvement of approximately 2.2% for WT, 7.1% for TC, and 0.04% for ET. These results underline the effectiveness of integrating the 3D-DDA block, which captures both local and global features, thus improving segmentation accuracy.

Table 2 compares our proposed method against state-of-the-art (SOTA) techniques using the validation set Dice scores for WT, TC, and ET. Our method achieves the highest Dice scores for WT (0.9155) and TC (0.8475), surpassing recent models such as GLF-Net and LGMSU-Net, which score 0.8930 and 0.8735 for WT, respectively. Additionally, our method achieves comparable performance for ET with a Dice score of 0.7364, closely matching the best ET score of 0.7443 achieved by Zou et al. (2020). Compared to the best-performing method in Table 2, our proposed method shows an improvement of approximately 2.5% for WT and 4.1% for TC, with a slight decrease of about 1.1% for ET. These results demonstrate the robustness and superior performance of our proposed Adaptive Dual Attention Diffusion (ADAD) framework, particularly in accurately segmenting the WT and TC regions. The inclusion of the 3D-DDA block significantly enhances the network's ability to capture comprehensive spatial and contextual information, leading to more precise segmentation outcomes.

In summary, the results from Tables 1 and 2 highlight the significant improvements our ADAD framework brings to 3D medical image segmentation. The integration of the 3D-DDA block and the diffusion model enables our method to achieve higher segmentation accuracy and robustness compared to both conventional networks and state-of-the-art methods.

## 5 Ablation Studies

To validate the contributions of each component in the proposed ADAD framework, we conducted ablation studies on the BraTS2020 dataset. Specifically, we evaluated the impact of the 3D Dual Domain Attention (3D-DDA) block and the dynamic conditional encoding strategy.

### 5.1 Experimental Setup

We designed several experiments to isolate the effects of the 3D-DDA block and the dynamic conditional encoding:

- **Baseline Model**: The diffusion model without the 3D-DDA block and without dynamic conditional encoding.
- **Baseline + 3D-DDA**: The diffusion model with the 3D-DDA block but without dynamic conditional encoding.
- **Baseline + Dynamic Encoding**: The diffusion model with dynamic conditional encoding but without the 3D-DDA block.
- **Full Model (ADAD)**: The complete proposed method with both the 3D-DDA block and dynamic conditional encoding.

### 5.2 Results and Discussion

Table 3: Ablation study results on the BraTS2020 validation set

| Model | WT Dice | TC Dice | ET Dice |
|---|---|---|---|
| Baseline Model | 0.8820 | 0.7945 | 0.7220 |
| Baseline + 3D-DDA | 0.8985 | 0.8120 | 0.7295 |
| Baseline + Dynamic Encoding | 0.8910 | 0.8065 | 0.7260 |
| **Full Model (ADAD)** | **0.9155** | **0.8475** | **0.7364** |

As shown in Table 3, incorporating the 3D-DDA block into the baseline model improved Dice scores across all tumor regions, demonstrating the effectiveness of capturing local and global contextual information. Similarly, adding dynamic conditional encoding enhanced the model's performance, highlighting its role in refining segmentation accuracy at each diffusion step.

The full ADAD model, combining the 3D-DDA block and dynamic conditional encoding, achieved the highest Dice scores, confirming that each component contributes positively to the overall performance.

### 5.3 Analysis

The ablation studies indicate that:

- The **3D-DDA block** significantly improves the model's ability to handle complex medical images by effectively capturing spatial and contextual features.
- The **dynamic conditional encoding** strategy enhances the conditioning of the diffusion process, leading to more accurate segmentation results.
- Combining both components yields the best performance, suggesting that they complement each other in the ADAD framework.

## 6 Conclusion

This paper introduced ADAD framework for 3D medical image segmentation. Through extensive experiments and ablation studies, we demonstrated that the 3D-DDA block and the dynamic conditional encoding strategy individually and collectively enhance segmentation performance. The ablation studies confirmed that the 3D-DDA block effectively captures local and global contextual information, improving the network's capability to handle the complexity of medical images. The dynamic conditional encoding strategy further refines the segmentation by incorporating current-step information into the conditioning process. Our proposed ADAD framework achieved superior results on the BraTS2020 dataset, outperforming state-of-the-art methods. These findings highlight the importance of each component in our model and their contributions to advancing 3D medical image segmentation. Future work will explore integrating additional attention mechanisms and applying the ADAD framework to other medical imaging modalities.

## References

1. Austin, J., Johnson, D.D., Jonathan, H., Daniel, T., Van Den Berg, R.: Structured denoising diffusion models in discrete state-spaces. Advances in Neural Information Processing Systems **34**, 17981–17993 (2021)
2. Do, N.T., Vo-Thanh, H.S., Nguyen-Quynh, T.T., Kim, S.H.: 3d-dda: 3d dual-domain attention for brain tumor segmentation. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 3215–3219. IEEE (2023)
3. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059. PMLR (2016)
4. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI brainlesion workshop. pp. 272–284. Springer (2021)
5. He, Y., Nath, V., Yang, D., Tang, Y., Myronenko, A., Xu, D.: Swinunetr-v2: Stronger swin transformers with stagewise convolutions for 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 416–426. Springer (2023)
6. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
7. Jiang, P., Gu, F., Wang, Y., Tu, C., Chen, B.: Difnet: Semantic segmentation by diffusion networks. Advances in Neural Information Processing Systems **31** (2018)
8. Kim, D.: Fine-grained human hair segmentation using a text-to-image diffusion model. Ieee Access **12**, 13912–13922 (2024). https://doi.org/10.1109/access.2024.3355542
9. Lee, H.H., Liu, Q., Bao, S., Yang, Q., Yu, X., Cai, L.Y., Li, T.Z., Huo, Y., Koutsoukos, X., Landman, B.A.: Scaling up 3d kernels with bayesian frequency reparameterization for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 632–641. Springer (2023)

10. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (2017)
11. Liu, C., Ding, W., Li, L., Zhang, Z., Pei, C., Huang, L., Zhuang, X.: Brain tumor segmentation network using attention-based fusion and spatial relationship constraint. In: International MICCAI Brainlesion Workshop. pp. 219–229. Springer (2020)
12. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)
13. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. Ieee (2016)
14. Myronenko, A.: 3d mri brain tumor segmentation using autoencoder regularization. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4. pp. 311–320. Springer (2019)
15. Nuechterlein, N., Mehta, S.: 3d-espnet with pyramidal refinement for volumetric brain tumor image segmentation. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4. pp. 245–253. Springer (2019)
16. Pang, X., Zhao, Z., Wang, Y., Li, F., Chang, F.: Lgmsu-net: Local features, global features, and multi-scale features fused the u-shaped network for brain tumor segmentation. Electronics **11**(12), 1911 (2022)
17. Savioli, N., Montana, G., Lamata, P.: V-fcnn: volumetric fully convolution neural network for automatic atrial segmentation pp. 273–281 (2019). https://doi.org/10.1007/978-3-030-12029-0_30
18. Wu, J., Fu, R., Fang, H., Zhang, Y., Yang, Y., Xiong, H., Liu, H., Xu, Y.: Medsegdiff: Medical image segmentation with diffusion probabilistic model. In: Medical Imaging with Deep Learning. pp. 1623–1639. PMLR (2024)
19. Xue, F., Guo, L., Bialkowski, A., Abbosh, A.: Training universal deep-learning networks for electromagnetic medical imaging using a large database of randomized objects. Sensors **24**(1), 8 (2023)
20. Zhang, K., Li, Y., Liang, J., Cao, J., Zhang, Y., Tang, H., Fan, D.P., Timofte, R., Gool, L.V.: Practical blind image denoising via swin-conv-unet and data synthesis. Machine Intelligence Research **20**(6), 822–836 (2023)
21. Zhou, Z., He, Z., Jia, Y.: Afpnet: A 3d fully convolutional neural network with atrous-convolution feature pyramid for brain tumor segmentation via mri images. Neurocomputing **402**, 235–244 (2020)