MIP final project

ID: r12922075, r12922068

Name: 吳韋論, 吳浩平

Data

1. Dataset (Brain Tumor Segmentation 2020)

Brain Tumor Segmentation(BraTS2020) challenge is the semantic segmentation of brain tumors. This is achieved by providing a 3D MRI dataset with voxel-wise ground truth labels annotated by medical professionals.

It contains training dataset comprises 369 subjects, each with four 3D MRI modalities:

- native (T1)
- post-contrast T1-weighted (T1Gd)
- T2-weighted (T2)
- T2 Fluid-attenuated Inversion Recovery (T2-FLAIR)

The input image size is 240 × 240 × 155. Data were collected from various institutions using diverse MRI scanners. Annotations encompass three tumor sub-regions: enhancing tumor, peritumoral edema, and necrotic and non-enhancing tumor core. These annotations are further grouped into three nested sub-regions: Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET).

This dataset size is roughly 40 GB.

Project Objective

Segmentation

Medical volumetric segmentation is a important process in analyzing medical images, focusing on lesion areas pixel by pixel. The accuracy of segmentation can help doctors to diagnose diseases and make well-informed treatment decisions.

Segmentation in 3D medical imaging data has many methods, ranging from traditional 3D-CNN-based models to more recent Transformer-based models. We aim to experiment with different state-of-the-art (SOTA) models, exploring their strengths and weaknesses. Finally, we will compare their performances in terms of WT, TC, ET, and visualize the segmentation results for a comprehensive evaluation.

Methodology

We chose four different types of popular models: 3D-CNN-based (3D-UNet, SegResNet), Transformer-based (Swin-UNETR), and Diffusion-Based (Diff-UNet). Each model has unique characteristics explained below.

1. 3D-UNet (2016 Jun)

- **Network:** Derived from the U-Net architecture, 3D-UNet operates on 3D volumes.
- **Encoder-Decoder Structure:** Utilizes a contracting encoder for global image analysis and an expanding decoder for full-resolution segmentation.
- **Operations:** Processes 3D volumes using dedicated 3D operations, including:
    - 3D convolutions
    - 3D max pooling
    - 3D up-convolutional layers

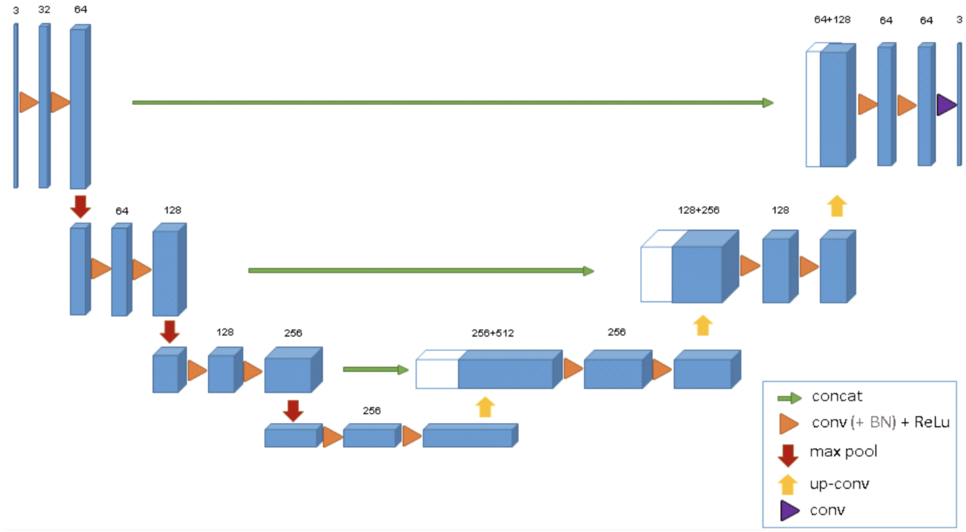


Fig. 2: The 3D u-net architecture. Blue boxes represent feature maps. The number of channels is denoted above each feature map.

2. SegResNet (2018 Nov)

Encoder:

- ResNet blocks, where each block consists of two convolutions with normalization and ReLU, followed by additive identity skip connection.

Decoder:

- Bilinear upsampling is applied, and convolution is performed across three dimensions, utilizing the Sigmoid activation function.

VAE:

- Sampling from a Gaussian distribution with mean and variance, then upsampling with a structure identical to the decoder, but without skip connections.

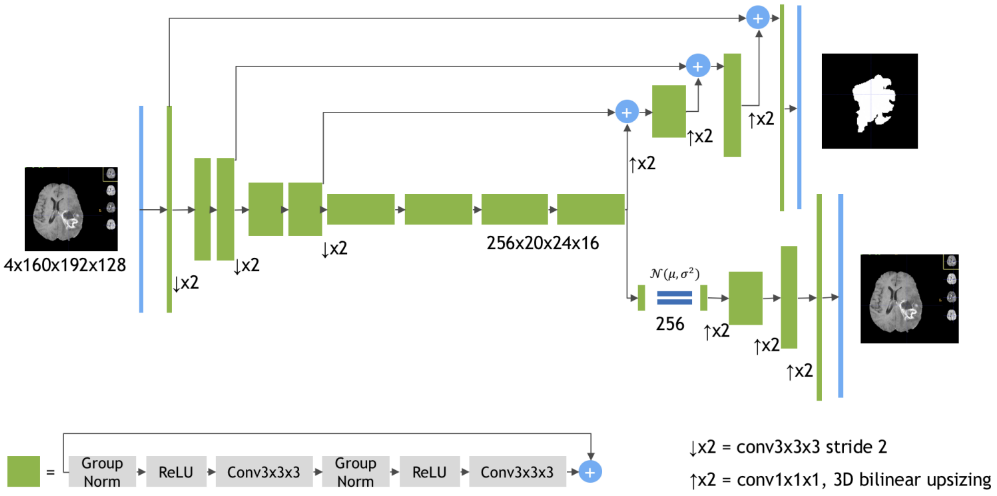


Fig. 1. Schematic visualization of the network architecture. Input is a four channel 3D MRI crop, followed by initial 3x3x3 3D convolution with 32 filters. Each green block is a ResNet-like block with the GroupNorm normalization. The output of the segmentation decoder has three channels (with the same spatial size as the input) followed by a sigmoid for segmentation maps of the three tumor subregions (WT, TC, ET). The VAE branch reconstructs the input image into itself, and is used only during training to regularize the shared encoder.

3. Swin-UNETR (2022 Jan)

Encoder

- Hierarchical Swin transformer responsible for processing the input sequence and extracting features at five different resolutions.

Attention Mechanism:

- Utilizes shifted windows for computing self-attention during the feature extraction process.

Decoder:

- FCNN-based decoder connected to the Swin transformer encoder at each resolution via skip connections.

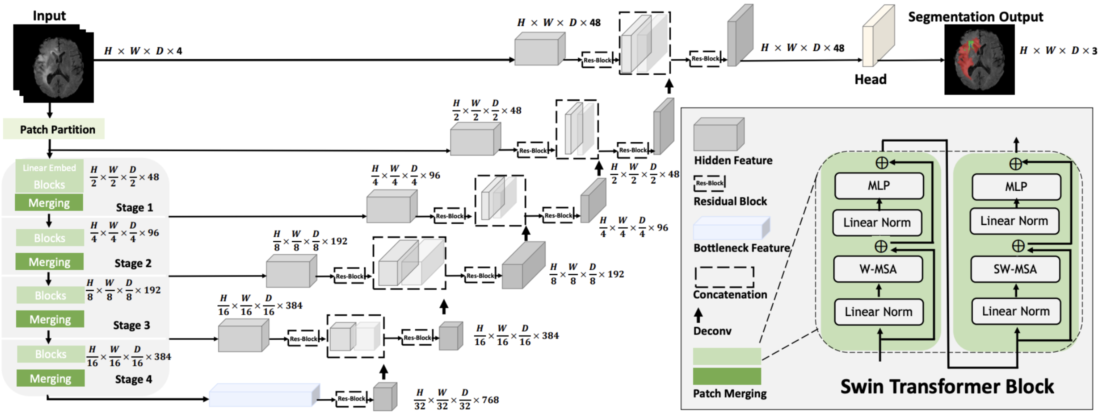


Fig. 1. Overview of the Swin UNETR architecture. The input to our model is 3D multi-modal MRI images with 4 channels. The Swin UNETR creates non-overlapping patches of the input data and uses a patch partition layer to create windows with a desired size for computing the self-attention. The encoded feature representations in the Swin transformer are fed to a CNN-decoder via skip connection at multiple resolutions. Final segmentation output consists of 3 output channels corresponding to ET,WT and TC sub-regions.

4. Diff-UNet (2023 Mar)

Diff-UNet Architecture:

- Diff-UNet uses Denoising Diffusion Models within a U-shaped architecture.

Effective Semantic Information Extraction:

- The U-shaped architecture captures both low-level and high-level features, and increase the model's understanding of complex structures in the 3D-image.

Robustness through SUF Module:

- To enhance the robustness of the diffusion model's predictions, Diff-UNet introduces a StepUncertainty based Fusion (SUF) module during inference.
- The SUF module strategically fuses outputs from the diffusion models at each step, which has more reliable pixel-wise representations.

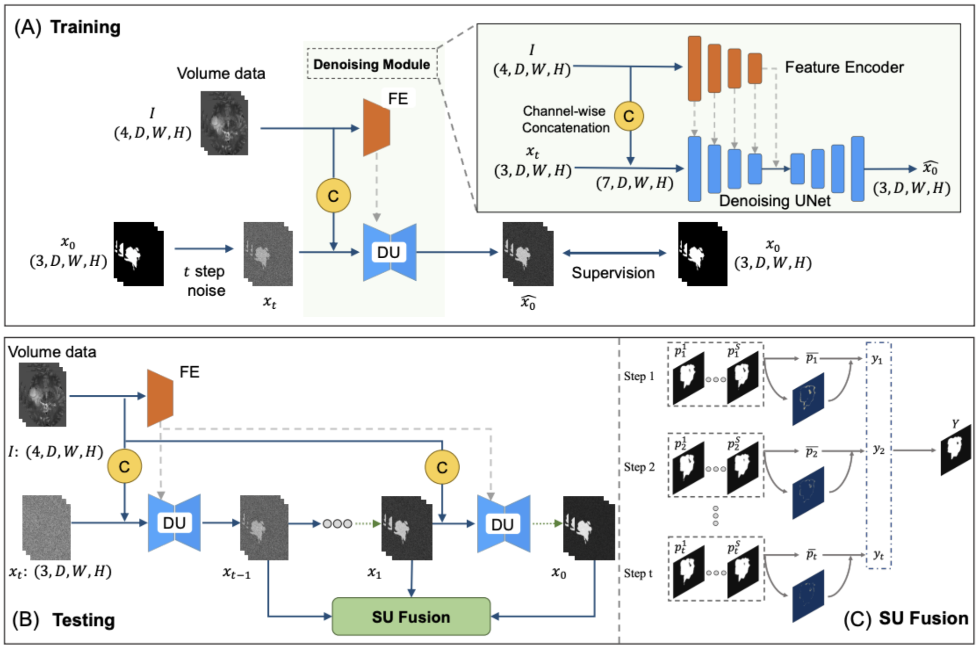


Fig. 1: The overview of proposed Diff-UNet. (A) is the training phase of Diff-UNet to learn a denoising function by the Denoising Module. (B) is the testing phase to generate segmentation results by an iterative way. (C) is the computation process of our SUF module.

Experiments

1. Implementation Details

- Training time
  - 3D UNet: 6 hours without pretrained weights
  - SegResNet: 20 hours without pretrained weights
  - Swin UNETR: 8 hours with pretrained weights
  - Diff-UNet: 18 hours without pretrained weights
- Split ratio is 0.7, 0.1, 0.2 for the training set, the validation set, and the test set.
- Machine Spec
  - GPU: RTX 4080
  - CPU: i7-13700K

  - RAM: 64GB

- Due to the GPU RAM requirements of the model and dataset, the batch size is limited to 1~2.

2. Dice Score

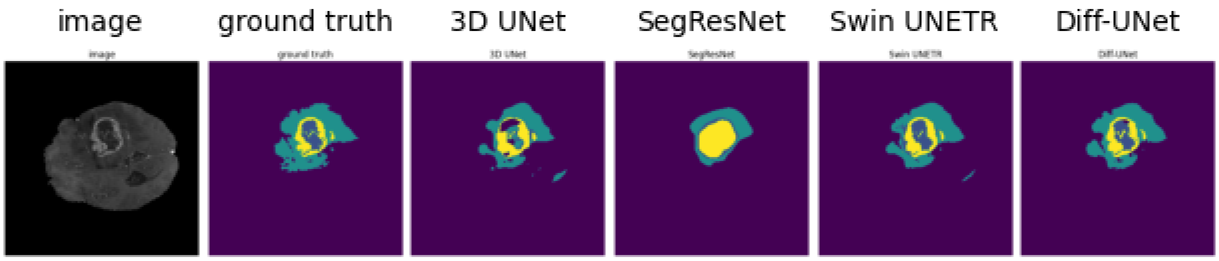
	Whole Tumor(WT)	Tumor Core(TC)	Enhancing Tumor(ET)	Average
3D UNet	0.7249	0.5936	0.5925	0.6370
SegResNet	0.7889	0.6802	0.5685	0.6792
Swin-UNETR	0.8718	0.6981	0.7454	0.7718
Diff-UNet	0.9153	0.8764	0.7772	0.8563

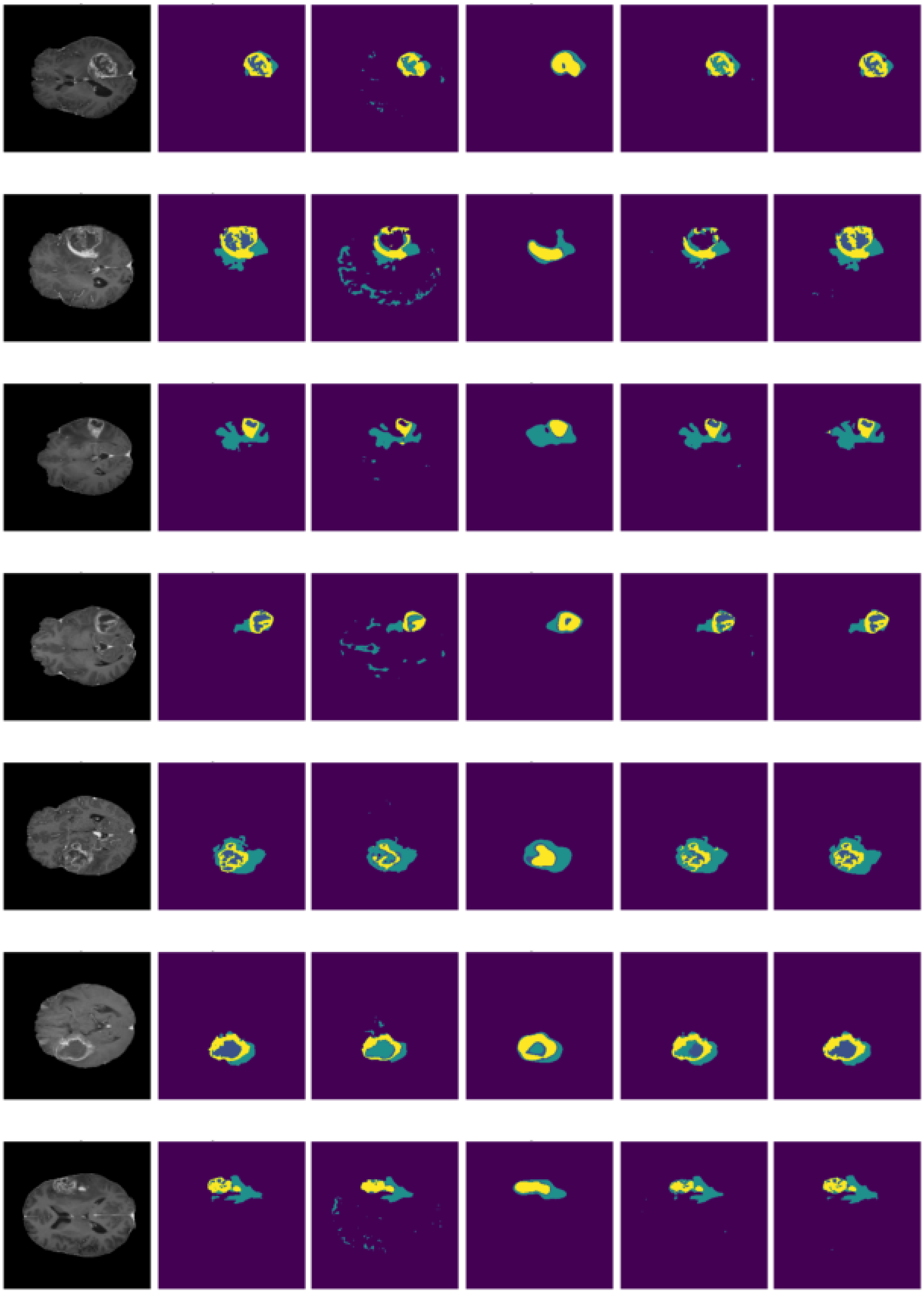
Average Analysis:

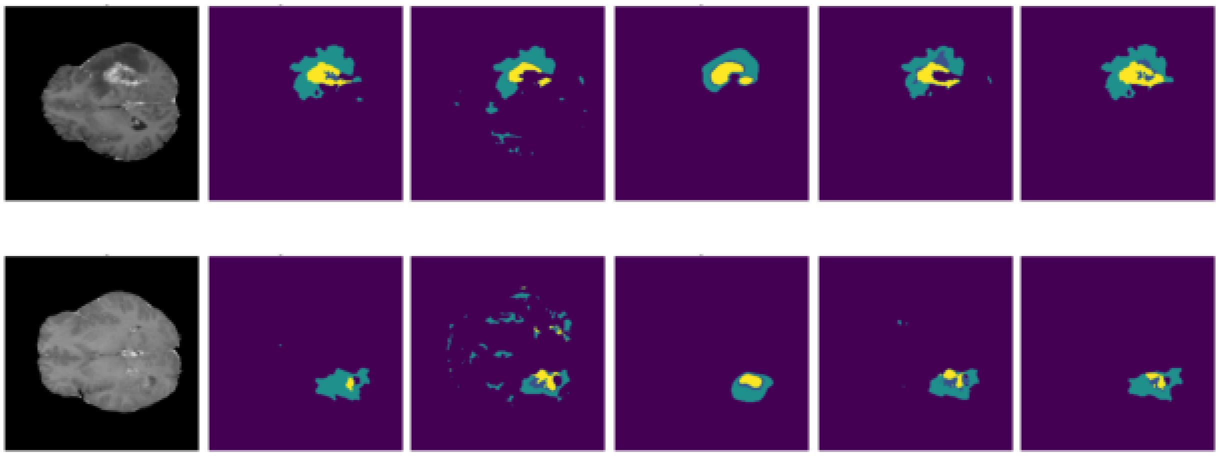
- Diff-UNet has the highest average score, it has robust performance across all regions.
- The scores for transformer-based methods are higher than those for CNN-based methods.
- 3D UNet has a comparatively lower average score, especially in Tumor Core(TC) and Enhancing Tumor(ET).

Visualization & Analysis

Below, we showcase some visual results of the segmentation obtained by each model:







The visualizations presented herein depict the performance characteristics of different models in the context of this segmentation task. Results from the 3D UNet model reveal a prevalence of fragmented misclassification regions, potentially indicative of the model's sensitivity to specific patterns or structures. SegResNet, on the other hand, exhibits less sharpness in capturing shape details, which might be due to its larger Region Of Interest (ROI). In contrast, Swin UNETR demonstrates commendable performance, while Diff-UNet stands out for its precision in delineating segmentation boundaries across diverse details.

It is noteworthy, however, that despite the superior accuracy demonstrated by Diff-UNet, its adoption of a diffusion-based model structure comes at the cost of significantly longer inference times compared to other models, which takes several minutes to perform a single segmentation job. Hence, in specific scenarios where time constraints are a critical factor, Diff-UNet may not necessarily be the optimal choice.

Conclusion

In our final project, we trained four different segmentation models on the BraTS2020 dataset, including 3D UNet, SegResNet, Swin UNETR, and Diff-UNet. Subsequently, we evaluated their performance on three distinct segmentation targets using the Dice score and conducted comparative analyses. Finally, we visually inspected and compared the segmentation results produced by each of the four models to intuitively assess their effectiveness.

Reference

BraTS2020 Dataset

MONAI tutorial for 3D-UNet, SegResNet, Swin-UNETR

Diff-UNet github

3D-UNet paper link

SegResNet paper link

Swin-UNETR paper link

Diff-UNet paper link