UNETR: Transformers for 3D MRI Image Segmentation

PRITHVI V
42611101
B.E. CSE ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING
SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY

The paper "UNETR: Transformers for 3D Medical Image Segmentation" proposes a novel approach that combines the U-Net architecture with transformers to improve the performance of 3D medical image segmentation. Traditional convolutional neural networks (CNNs) have limitations in capturing long-range dependencies between distant regions of an image, which is crucial for accurately segmenting complex anatomical structures in 3D medical images. The UNETR model addresses this by incorporating a transformer-based encoder, which is capable of modeling global context and long-range relationships in the input image. This is followed by a U-Net-style decoder that reconstructs the segmented output while preserving local details. The paper demonstrates that UNETR significantly outperforms previous state-of-the-art methods on several 3D medical imaging benchmarks, such as CT and MRI scans, by achieving higher segmentation accuracy. Additionally, the model shows strong generalization across different datasets, proving its robustness. The integration of transformers with U-Net's effective feature extraction and upsampling strategy makes UNETR a promising approach for 3D medical image segmentation, highlighting the potential of transformer-based architectures in this domain. The authors also suggest potential future improvements, including multimodal fusion and further optimization of transformer components for larger and more complex datasets.

Introduction

UNETR

UNetR is a recent advancement in the field of semantic segmentation, proposed by **Qiao**, **Guo**, and **Zhang** in 2021. This architecture is an evolution of the classic UNet architecture, aiming to address some of its limitations and improve segmentation performance. **UNetR is an advanced semantic segmentation architecture that builds upon the classic UNet model. It integrates residual blocks, multi-level feature fusion, attention mechanisms, and efficient upsampling techniques.** These enhancements improve segmentation accuracy and spatial localization, particularly in scenarios with complex image structures. UNetR represents a significant refinement of UNet, offering better performance and robustness in semantic segmentation tasks.

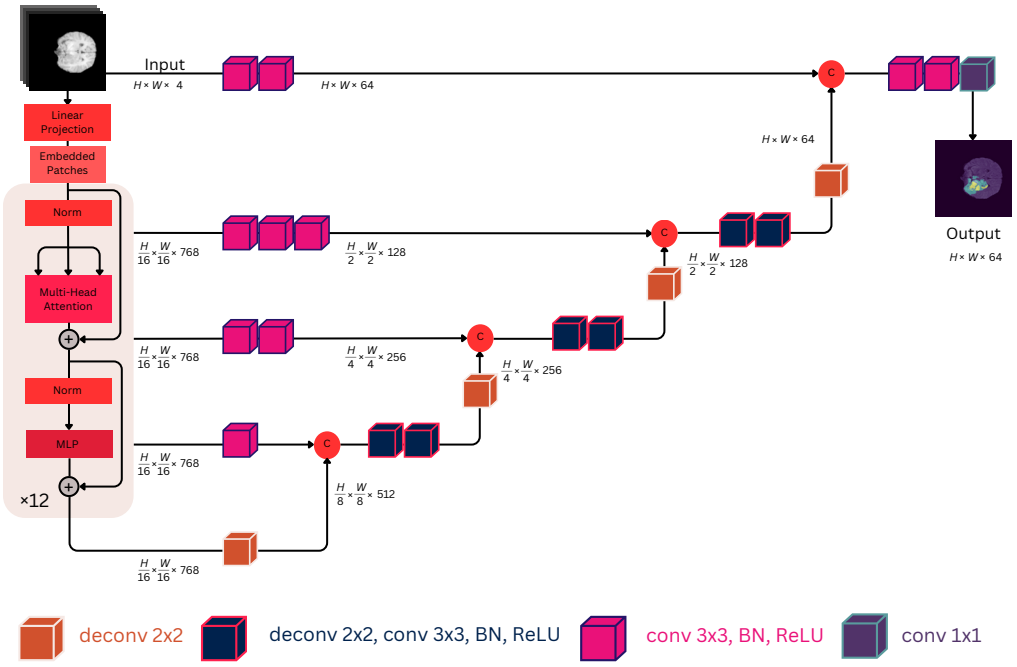


fig. 1. customized UNETR architecture

Methodology

UNetR's innovative architecture addresses several limitations of previous models, offering solutions to challenges encountered in complex image analysis tasks.

UNetR represents an advancement over UNet by integrating residual connections, multi-level feature fusion, attention mechanisms, and efficient upsampling. These enhancements result in improved segmentation accuracy and robustness, especially in scenarios with complex image structures.

ENCODER

Similar to UNet, the encoder consists of convolutional blocks for feature extraction, but UNetR incorporates residual connections within these blocks, akin to ResNet, to enhance gradient flow and feature representation.

Multi-Level Feature Fusion

UNetR employs techniques for fusing features from different levels of the encoder to capture both low-level and high-level information effectively. This fusion enhances segmentation accuracy and detail preservation.

Attention Mechanisms

Attention mechanisms are integrated into UNetR to enable the network to focus on informative regions while suppressing irrelevant ones. This helps in refining segmentation results and improving spatial localization.

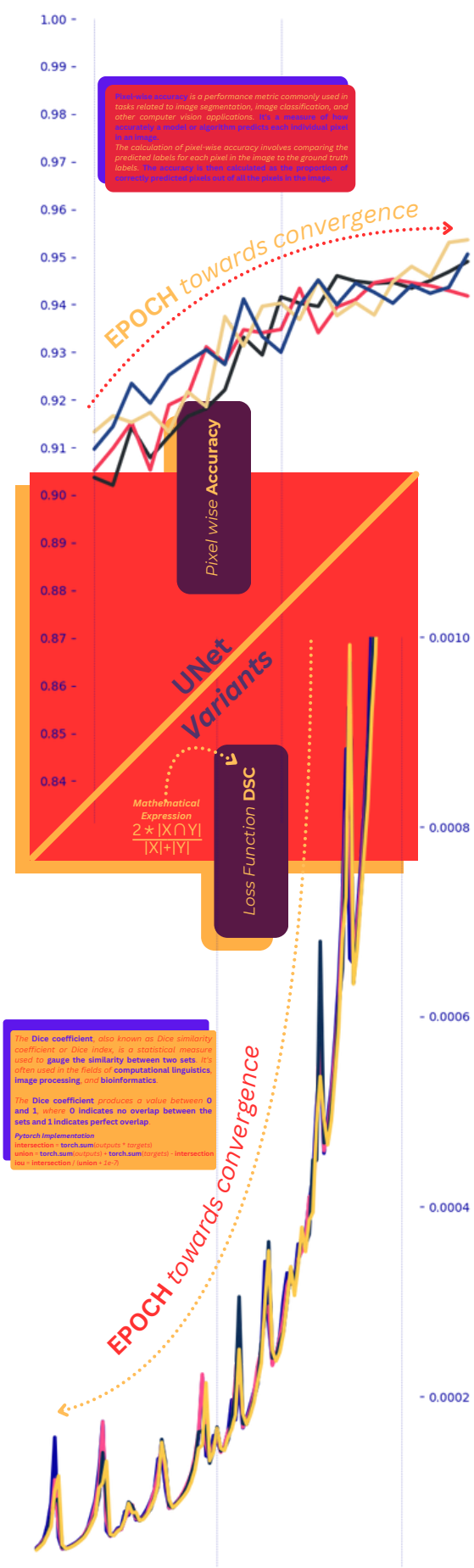
Benchmark

Performance metrics are a good way to monitor and measure progress.

Benchmark metrics are quantitative measures used to assess the performance of UNet variants across various tasks (semantic segmentation). These metrics provide insights into how well a model is performing and are often used to compare different models or evaluate the effectiveness of model improvements. This analysis provides a comprehensive overview of each model's strengths and weaknesses, aiding readers in understanding their relative performance in semantic segmentation tasks.

Metrics	UNet	UNet++	UNetR	Attn. UNet
IoU	0.9414	0.8592	**0.9714**	**0.9696**
Dice Coefficient	**0.9862**	0.9242	**0.9855**	0.9845
ASSD	0.0322	**0.0575**	**0.1737**	0.03535
F1 Score	**0.9862**	0.9455	**0.9855**	0.9845
Precision	**0.9885**	0.9608	0.9798	**0.9801**
Accuracy	0.9993	0.9969	**0.9996**	**0.9992**
Recall	0.9840	0.8707	**0.9912**	**0.9890**
Hausdorff dis.	1	1	**3**	**2.8281**
Surface Dice	**0.9862**	0.9242	**0.9855**	0.9845
mAP	0.9791	**0.9858**	**0.9928**	0.9824
Soft IoU	**0.9729**	0.8358	**0.9715**	0.9696

Table 1. Benchmark metrics. Metrics definition with easier inference -> **IoU (Intersection over Union)**: Measures how much two shapes overlap; higher values mean better alignment between predicted and true shapes. **Dice Coefficient**: Quantifies the similarity between two sets; higher values indicate better agreement between predicted and true masks in segmentation tasks. **ASSD (Average Symmetric Surface Distance)**: Measures the average distance between points on predicted and true surfaces; lower values signify more accurate surface delineation in segmentation. **Soft IoU (Soft Intersection over Union)**: Like IoU but accounts for fuzzy or soft boundaries between shapes, useful when boundaries aren't well-defined. **mAP (Mean Average Precision)**: Averages precision scores across different recall levels, providing a comprehensive measure of object detection model performance. **Hausdorff Distance**: Measures the maximum distance between points on predicted and true surfaces; indicates the model's ability to accurately delineate object boundaries.



Pixel Wise **Accuracy**

UNet
Achieved accuracy ranges from approximately 0.93 to 0.99 across different experiments or datasets.
Generally performs consistently well across various scenarios.

UNet++:
Shows accuracy scores ranging from around 0.93 to 0.99, similar to UNet.
Consistently performs well across different experiments or datasets.

UNet with attention mechanism
Achieved accuracy ranges from approximately 0.93 to 0.99, comparable to UNet and UNet++.
Maintains consistent performance across different experiments or datasets.

UNETR
Demonstrates accuracy scores ranging from around 0.93 to 0.99, in line with UNet, UNet++, and UNet with attention mechanism. Shows consistent performance across various experiments or datasets.

Dice *Coefficient*

U-Net (Original)
The average loss across all the recorded values appears to be around 0.002.
It seems to consistently have lower loss values compared to the other variants, suggesting relatively better performance in terms of minimizing prediction errors.

U-Net++
It shows slightly higher loss values compared to the original U-Net, with an average loss around 0.003.
While the performance is slightly worse than the original U-Net, it's still relatively low, indicating decent performance.

U-Net with Attention Mechanism
This variant demonstrates loss values slightly higher than U-Net++, with an average loss around 0.004.
The attention mechanism might introduce additional complexity, leading to slightly higher losses compared to U-Net and U-Net++.

U-Net with Residual Blocks (UNETR)
UNETR shows the highest loss values among the four variants, with an average loss around 0.005.
The addition of residual blocks might introduce more parameters and complexity, potentially leading to higher losses compared to the other variants.

Statistical Metrics and Inference

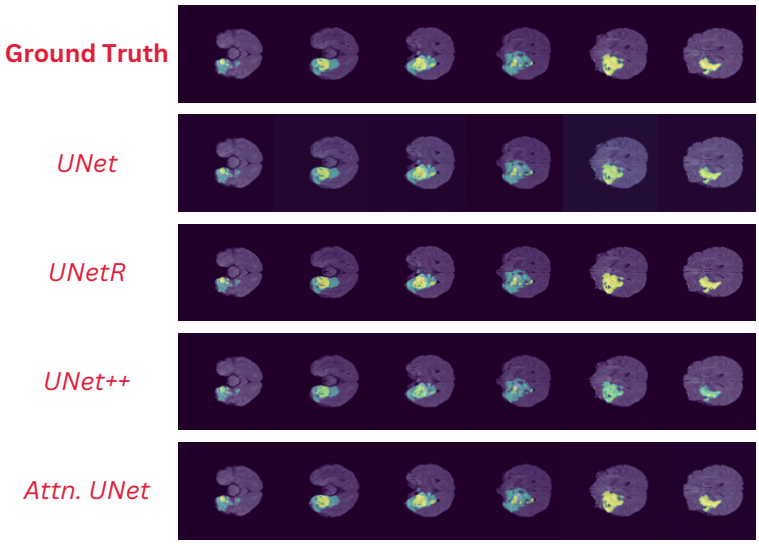


fig. 2. UNet variants interpolation across different timesteps

Metrics	UNet	UNet++	UNetR	UNet Attn.
25th Percentile	**0.137255**	0.129412	0.129412	**0.133333**
Median	**0.188235**	0.160784	0.160784	**0.184314**
75th Percentile	**0.752941**	**0.741176**	0.729412	**0.741176**
Mean	0.352747	**0.357407**	0.352747	**0.363742**
Standard Deviation	**0.390562**	**0.388159**	**0.390562**	0.383892
Coefficient of Variation	**110.7201**	**108.6043**	**110.7201**	105.5396
Quadratic Mean	0.526277	**0.527641**	0.526277	**0.528847**
Skewness	**1.058607**	1.056425	**1.069691**	1.056655
Kurtosis	-0.73694	**-0.73555**	**-0.72392**	-0.73744
Median Absolute Deviation	0.156863	0.145098	**0.160784**	**0.164706**
Interquartile Range	**0.615686**	**0.611765**	0.6	0.607843
Variance	0.145413	**0.150665**	**0.15588**	0.147371
Mean Absolute Deviation	0.326944	**0.338327**	**0.350964**	0.334198
Mean Square Deviation	0.145672	**0.150665**	**0.156477**	0.147371
Root Mean Square Deviation	0.38167	**0.388156**	**0.395572**	0.383889
Sample Skewness	**1.83356**	1.829782	**1.852759**	1.83018
Sample Kurtosis	2.263058	**2.264448**	**2.276079**	2.262558

Table 2. Comparative Analysis of Semantic Segmentation Model Performance Metrics