Hybrid U-Net++ Vision Transformer Fusion for Accurate Brain MRI Image Segmentation and Classification

[1]Malathi Janapati, [2]Shaheda Akthar

[1]Department of Computer Science and Engineering, Acharya Nagarjuna University, Guntur, India

[2]Department of Computer Science and Engineering, Acharya Nagarjuna University, Guntur, India, and Department of Computer Science, Government College for Women (A), Guntur, India

DOI: https://doi.org/10.51584/IJRIAS.2026.11080007

Received: 13 August 2026; Accepted: 19 August 2026; Published: 27 August 2026

ABSTRACT

Automatic brain Magnetic Resonance Imaging (MRI) analysis plays a crucial role in computer-aided diagnosis, treatment planning, and disease monitoring by enabling accurate delineation and identification of brain tumors. However, manual segmentation is labor-intensive, time-consuming, and susceptible to inter-observer variability, making automated and reliable methods highly desirable. This paper proposes a Hybrid U-Net++ and Vision Transformer (ViT) Fusion framework for automatic brain MRI tumor segmentation and segmentation-guided classification. The proposed architecture employs a U-Net++ encoder with nested skip connections to extract hierarchical multi-scale features, while a Vision Transformer captures long-range spatial dependencies and global contextual information through self-attention mechanisms. A multi-stage feature fusion strategy integrates convolutional and transformer representations, complemented by boundary-aware refinement to improve tumor localization and preserve fine anatomical details. The segmentation model achieved a training Dice score of 95.6%, IoU of 91.4%, and validation Dice score of 92.6%, demonstrating stable convergence and strong generalization capability. Furthermore, the proposed framework attained a Dice score of 95.8%, IoU of 92.1%, and Hausdorff Distance of 3.35 mm in the ablation analysis, confirming the effectiveness of multi-stage feature fusion and boundary-aware refinement. The segmented tumor regions are subsequently utilized for Region of Interest (ROI)-based classification using deep feature extraction, Global Average Pooling (GAP), and fully connected layers. The classification module distinguishes glioma, meningioma, pituitary tumor, and no-tumor cases, achieving an overall accuracy of 97.55%, macro-average precision of 97.13%, recall of 97.50%, F1-score of 97.31%, and specificity of 98.40%. Comparative analysis further demonstrates the effectiveness of the proposed hybrid architecture over conventional CNN-based and transformer-based approaches. The combined segmentation and classification results indicate that the proposed framework provides an effective approach for tumor localization and tumor-type identification, with potential applicability in automated brain MRI analysis and clinical decision support.

Keywords: Brain MRI, Automatic Medical Image Segmentation, U-Net++, Vision Transformer (ViT), Deep Learning, Medical Image Analysis, Brain Tumor Segmentation, Semi-Supervised Learning, Attention Mechanism, Edge Attention, Reverse Attention.

INTRODUCTION

Automatic brain Magnetic Resonance Imaging (MRI) segmentation has become an essential component of modern computer-aided diagnosis (CAD) systems, enabling the accurate delineation of brain tumors, lesions, and anatomical structures for disease diagnosis, treatment planning, surgical navigation, and longitudinal disease monitoring [1]. Precise segmentation facilitates quantitative analysis, biomarker discovery, radiotherapy planning, and personalized clinical decision-making. However, manual annotation of brain MRI images is labor-intensive, time-consuming, and highly dependent on the expertise of neuroradiologists. Moreover, inter-observer variability often leads to inconsistent annotations, reducing the reproducibility and reliability of clinical assessments [2]. Consequently, there is an increasing demand for robust, automated

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VIII August 2026

segmentation techniques capable of accurately identifying pathological regions while minimizing human intervention.

The rapid advancement of deep learning has significantly improved the performance of medical image segmentation systems. Among these approaches, U-Net has emerged as the most influential convolutional neural network (CNN) architecture for biomedical image segmentation due to its encoder-decoder structure and skip connections that effectively capture local contextual information [3]. Subsequent variants, including U-Net++, Attention U-Net, ResUNet, and Dense U-Net, have further enhanced segmentation accuracy through improved feature fusion, dense skip pathways, and attention mechanisms. Nevertheless, convolution-based architectures primarily focus on local receptive fields, limiting their ability to capture long-range spatial dependencies that are essential for accurately segmenting complex brain tumors with irregular shapes and heterogeneous textures [4].

Brain MRI segmentation presents several unique challenges compared with other medical imaging modalities. Brain tumors exhibit substantial variations in size, shape, location, intensity distribution, and tissue appearance across patients. Furthermore, MRI images acquired using different imaging sequences, such as T1-weighted, T2-weighted, T1-contrast-enhanced (T1ce), and FLAIR, display distinct anatomical and pathological characteristics. The presence of imaging artifacts, intensity inhomogeneity, low contrast, and blurred tumor boundaries further complicates automatic segmentation. These factors limit the generalization capability of conventional CNN-based segmentation methods and motivate the development of architectures capable of learning both local structural details and global contextual relationships [5].

Recently, Vision Transformers (ViTs) have attracted considerable attention in medical image analysis because of their powerful self-attention mechanism, which effectively models long-range dependencies across the entire image [6]. Unlike convolutional networks that process local neighborhoods, Vision Transformers divide an image into patches and learn global contextual representations through multi-head self-attention. This capability has resulted in state-of-the-art performance across various computer vision tasks, including image classification, object detection, and semantic segmentation. However, standalone transformer models generally require large-scale annotated datasets and substantial computational resources, making their direct application to medical imaging challenging due to the limited availability of labeled clinical data [7].

To overcome these limitations, recent research has focused on hybrid CNN-Transformer architectures, which combine the strengths of convolutional neural networks and Vision Transformers. CNNs efficiently capture fine-grained local features, while transformers provide global contextual information through attention mechanisms. Integrating these complementary representations significantly improves segmentation accuracy, particularly for small lesions, irregular tumor boundaries, and heterogeneous pathological regions. Among CNN architectures, U-Net++ offers dense nested skip connections that effectively reduce the semantic gap between encoder and decoder feature maps, making it an ideal backbone for hybrid transformer-based segmentation frameworks [8].

In this work, we propose an Automatic Brain MRI Segmentation framework based on U-Net++ and Vision Transformer. The proposed architecture employs the U-Net++ encoder to extract hierarchical multi-scale features and a Vision Transformer module to model global contextual dependencies. A feature fusion strategy integrates local convolutional representations with transformer-based global features, enabling precise segmentation of brain tumors and abnormal tissues. Furthermore, edge-aware feature refinement is incorporated to improve boundary delineation and preserve fine anatomical structures. Extensive experiments conducted on benchmark brain MRI datasets demonstrate that the proposed framework consistently outperforms conventional CNN-based and transformer-based segmentation methods in terms of Dice Similarity Coefficient (DSC), Intersection over Union (IoU), precision, recall, and boundary accuracy.

The major contributions of this work are summarized as follows:

- A novel hybrid U-Net++ and Vision Transformer architecture is proposed for automatic brain MRI segmentation, effectively integrating local feature extraction with global contextual modeling.

- A multi-scale feature fusion strategy is introduced to enhance semantic representation and improve segmentation of complex tumor structures.
- Edge-aware refinement is incorporated to achieve more accurate boundary localization and preserve anatomical details.
- Extensive experimental evaluations and ablation studies demonstrate the effectiveness of the proposed framework, achieving superior performance compared with existing state-of-the-art medical image segmentation methods.
- The proposed model provides a robust and clinically applicable solution for accurate brain MRI segmentation, supporting computer-aided diagnosis, treatment planning, and disease monitoring.

LITERATURE REVIEW

Automatic brain MRI segmentation has been extensively investigated over the past decade because of its importance in computer-aided diagnosis, treatment planning, and disease monitoring. Early segmentation approaches relied on traditional image processing techniques such as thresholding, region growing, watershed transformation, active contour models, and clustering algorithms. Although these methods were computationally efficient, they were highly sensitive to image noise, intensity inhomogeneity, and initialization parameters. Moreover, they struggled to accurately segment complex brain tumors with irregular shapes and fuzzy boundaries.

Table 1. Summary of Related Work on Brain MRI Medical Image Segmentation

Ref.	Method	Architecture / Technique	Strengths	Limitations
Ronneberger et al. [9]	U-Net	Encoder–decoder CNN with skip connections for biomedical image segmentation	Excellent localization and efficient training	Limited global context modeling
Milletari et al. [10]	V-Net	3D fully convolutional network for volumetric segmentation	Effective for volumetric MRI	High computational cost
Zhou et al. [11]	U-Net++	Nested dense skip connections for improved feature fusion	Better boundary preservation	Increased model complexity
Li et al. [12]	HDenseU-Net	Hybrid dense connectivity with multi-scale feature extraction	Improved segmentation accuracy	Large memory requirements
Isensee et al. [13]	nnU-Net	Self-configuring segmentation framework	Automatic architecture optimization	Long training time
Oktay et al. [14]	Attention U-Net	Attention gates for suppressing irrelevant regions	Improved localization	Additional computational overhead
Tarvainen and Valpola [15]	Mean Teacher	Semi-supervised consistency learning	Utilizes unlabeled data	Sensitive to pseudo-label quality

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VIII August 2026

Li et al. [16]	Shape-aware Segmentation	Geometric constraint learning	Better structural consistency	Requires shape priors
Wang et al. [17]	Non-local Neural Networks	Global self-attention	Captures long-range dependencies	Computationally expensive
Dosovitskiy et al. [18]	Vision Transformer (ViT)	Transformer architecture for image analysis	Strong global contextual modeling	Requires large-scale datasets
Chen et al. [19]	TransUNet	CNN–Transformer hybrid segmentation	Combines local and global features	High computational complexity
Cao et al. [20]	Swin-Unet	Hierarchical Swin Transformer	Efficient multi-scale representation	Complex optimization
Hatamizadeh et al. [21]	UNETR	ViT encoder with CNN decoder	Excellent volumetric segmentation	High memory consumption
Valanarasu et al. [22]	MedT	Gated Axial Attention Transformer	Enhanced long-range dependency modeling	Slow inference speed
Xie et al. [23]	SegFormer	Lightweight Transformer segmentation	Efficient and accurate	Limited medical-specific optimization
Ma et al. [24]	Medical SAM	Foundation model for prompt-based medical segmentation	Strong zero-shot capability	Requires fine-tuning for MRI tasks
Proposed Method	**U-Net++ + Vision Transformer**	**Hybrid nested U-Net++ with Vision Transformer, multi-scale feature fusion, and edge-aware refinement**	**Superior local and global feature learning, improved boundary preservation, high segmentation accuracy**	**Moderate computational complexity**

From Table 1, it can be observed that CNN-based methods such as U-Net, V-Net, and U-Net++ achieve excellent local feature extraction but have limited capability in modeling long-range dependencies. Attention-based architectures improve feature representation by emphasizing important anatomical regions, whereas transformer-based methods such as Vision Transformer (ViT), TransUNet, Swin-Unet, and UNETR effectively capture global contextual information through self-attention mechanisms. However, these transformer models often require large annotated datasets and substantial computational resources. The proposed U-Net++ with Vision Transformer framework addresses these limitations by integrating dense multi-scale feature extraction with global self-attention, resulting in improved brain MRI segmentation accuracy, precise boundary delineation, and enhanced generalization across diverse imaging datasets.

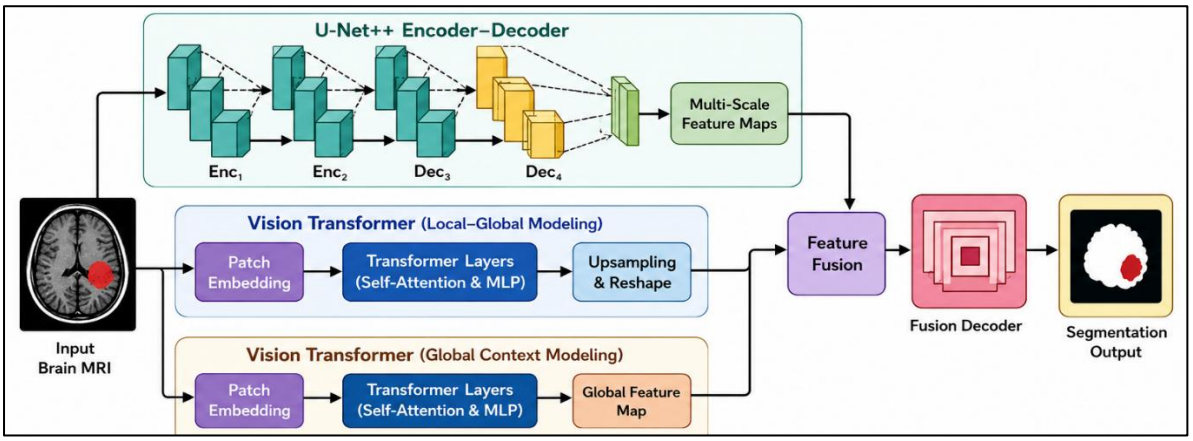
Proposed Architecture

MRI Segmentation

The first stage of the proposed framework focuses on the accurate segmentation of brain tumors from Magnetic Resonance Imaging (MRI) scans. MRI images contain complex anatomical structures, intensity variations,

noise, and irregular tumor boundaries, which make precise tumor delineation challenging. To address these challenges, the proposed Hybrid U-Net++–Vision Transformer (ViT) architecture combines the local feature extraction capability of convolutional neural networks with the global contextual modeling ability of transformer-based self-attention.

Fig.1: Proposed Architecture



The architecture is constructed based on two complementary sets of feature extraction which are designed to extract both the small spatial detail and the global contextual structure. The one of them is a U-Net++ encoder-decoder network that takes as input a picture and applies hierarchical convolutional blocks with embedded skip connectors. The sparse links reduce the semantic gap between encoder and decoder features, and permit multi-scale feature reuse which must be used to preserve small anatomical structures and sharp edges in the high-resolution medical images. Meanwhile, an identical branch of Vision Transformers converts the image to patch embeddings and processes them through a stack of layer of self-attention capable of capturing global spatial correlations that single convolution would not be capable of doing. The transformer output is reformed to a global feature map and it is compared to the CNN features. The two streams are combined in a feature fusion module that in turn fuses the local multi-scale representation with global context features and then convolutional refinement is conducted. A special fusion decoder then progressively re-combines the segmentation map to produce a final mask of accuracy of the boundary and U-Net++ with structural consistency at a global level executed by the transformer. This two-way structure will ensure high segmentation mobile performance in complex medical situations where the local detail and global anatomy are required.

Algorithm 1: U-Net and Vision Transformer Fusion for High-Resolution Segmentation
Input: High resolution image $I \in R^{H*W*C}$,ground truth mask M,U-Net++ parameters θ_U,Vit Parameters θ_T
Output: Predicted segmentation mask $\hat{M}$
1.Initialize U-Net++ parameters θ_U,Vit Parameters θ_T ,and loss-weight parameter λ
2. Extract multi-scale features using U-net++:
$\{F_s\}_{s=1}^{s} \leftarrow UNet++(I;\theta_U)$
3.Divide the input images into patches of size $p*p$:
$P \leftarrow Patchify\ (I, p)$
4. Generate Patch embeddings and add positional encoding:
$E \leftarrow PW_e + P_e$
5. Extract Global contextual features using Vision transformer:

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VIII August 2026

$T \leftarrow \text{TransformerEnc}(E, \theta_T)$
6.Reshape the transformer output into a spatial feature representation:
$G \leftarrow \text{Reshape}(T)$
7.Upsample the global features to the original image resolution:
$F_G \leftarrow \text{Up}(G)$
8.Fuse the multi-scale local fetures obtained from U-Net++:
$F_L \leftarrow \text{FuseScale}(\{F_s\}_{s=1}^{S})$
9. Concatenate local and global features:
$F_C \leftarrow [F_L \| \| F_G]$
10. Perform convolutional feature fusion:
$F \leftarrow \emptyset(\text{Conv}(F_C))$
11.Generate pixel-wise segmentation probabilities:
$\hat{M} = \text{Softmax}(\text{Conv}_{1*1}(F))$
12. Calculate the combined segmentation loss:
$\int = \lambda\, 1 - \frac{2(\hat{M}, M)+\in}{\|\|\hat{M}\|\| + \|\|M\|\|+\in} + (1-\lambda)\text{CE}(\hat{M}, M)$
13. Update θ_U and θ_T using back propagation and the selected optimizer.
14. Repeat steps 12-13 until the stopping criterion or maximum number of epochs is reached.
15. Return the final segmentation mask M.
End Algorithm

The algorithm1 starts with the high-resolution medical image, introduced in the U-Net++ backbone in order to extract multi-scale feature maps nested with one being fine anatomical boundaries preserved by dense skip connections. Simultaneously, the identical image is patchified and converted into embeddings and sent through Vision Transformer self attention layers allowing the model to learn long-range spatial interactions and global anatomical layout. Transformer output is then reshaped and upsampled to the spatial resolution of the convolutional features and then local (U-Net++) and global (ViT) representations are concatenated and convolved to produce a single feature array. A fusion decoder and softmax layer reconstruct this fused representation to generate the segmentation mask by decoding it into a pixel-wise probability map. The hybrid loss that is used to train is a combination of Dice overlap and cross-entropy terms such that ensuring precision of the boundary and classification accuracy are balanced. Gradient updates optimize the U-Net++ and transformer parameters in the coordinated manner, making sure that the local detail modelling and the global context reasoning are optimized, and eventually leads to the enhancement of the consistency of the segmentation in difficult high-resolution medical images.

Global Average Pooling (GAP) and Fully Connected (FC) Classification

Following the brain tumor segmentation stage, the predicted segmentation mask is used to identify and extract the tumor Region of Interest (ROI) from the original brain MRI image. This segmentation-guided ROI is subsequently provided to the classification module to determine the type of brain tumor. The classification module consists of Global Average Pooling (GAP), Fully Connected (FC) layers, and a Softmax output layer. The purpose of this module is to transform the high-dimensional deep features associated with the segmented tumor into a compact and discriminative representation for accurate tumor classification.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VIII August 2026

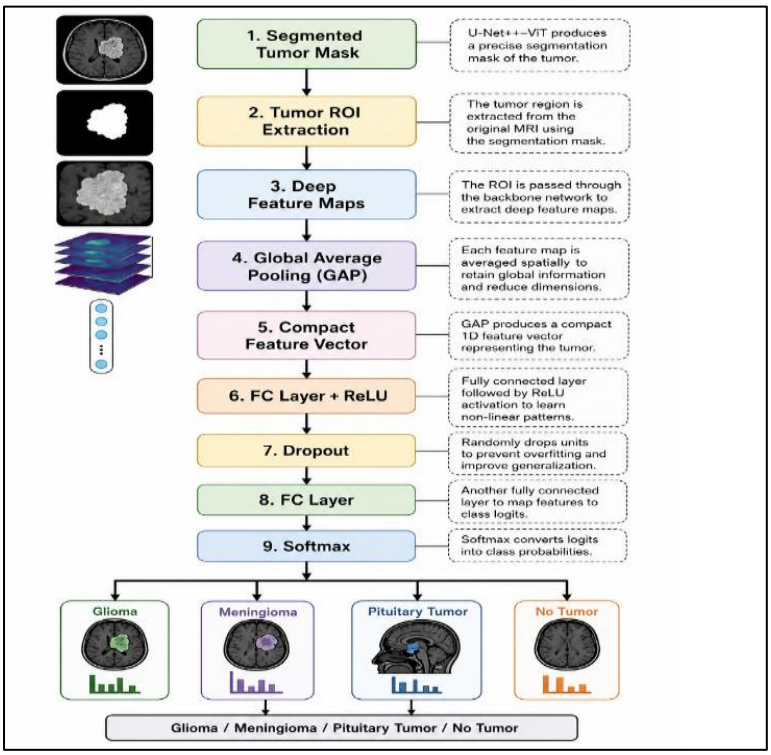
The feature representation obtained from the segmentation network contains spatial feature maps that encode important characteristics of the segmented tumor, including shape, texture, intensity, boundary, and structural information. Instead of directly flattening these feature maps into a very large feature vector, Global Average Pooling is employed to reduce their spatial dimensions. For a feature map of dimensions , where and represent the spatial dimensions and represents the number of feature channels, GAP calculates the average activation of each channel:

$$G_k = \frac{1}{H * W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_{ijk} - -Eq.1$$

Where F_{ijk} represents the activation value at spatial location(i,j) in the k^{th} feature map, and G_k represents the resulting global feature value. Consequently, an H× W × C feature map is converted into a compact C-dimensional feature vector.The use of GAP significantly reduces the number of parameters compared with conventional flattening operations and consequently decreases computational complexity and the risk of overfitting. More importantly, it provides a compact representation of the global characteristics of the segmented tumor.

The feature vector generated by GAP is then passed to the Fully Connected (FC) layers, which learn the discriminative relationships between the extracted tumor characteristics and the corresponding tumor categories. The FC layers are equipped with ReLU activation to introduce non-linearity and enable the classifier to learn complex relationships among the deep features. Dropout regularization is incorporated to reduce overfitting and improve the generalization capability of the classification network.A Softmax activation function is applied to the final layer to convert the output logits into class probabilities. For an input MRI image, the class having the highest probability is selected as the predicted tumor category.

Figure 2. Overall segmentation-guided brain tumor classification pipeline using GAP–FC classifier.



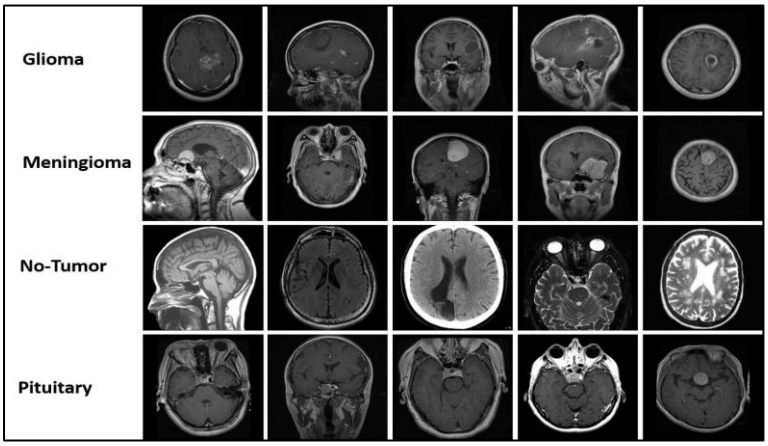
RESULTS AND DISCUSSIONS

Dataset Description

The brain tumor dataset used in this study consists of brain MRI images categorized into four distinct classes: Glioma, Meningioma, Pituitary Tumor, and No Tumor, as illustrated in Fig. 3. The dataset contains MRI scans

representing different pathological conditions of the brain, enabling the development and evaluation of an automated brain tumor classification framework.The Glioma class contains MRI images associated with tumors originating from glial cells, while the Meningioma class represents tumors arising from the meninges surrounding the brain. The Pituitary Tumor class consists of MRI images showing tumors located in the pituitary gland region. The No Tumor class contains normal brain MRI scans without visible tumor abnormalities. The inclusion of both tumor and non-tumor images allows the proposed model to learn discriminative patterns between abnormal and normal brain structures. Prior to classification, the MRI images are subjected to appropriate preprocessing and normalization to improve image quality and ensure consistency. The processed images are subsequently provided to the proposed U-Net++–Vision Transformer (ViT) framework, where U-Net++ extracts detailed local and multi-scale spatial features, while the Vision Transformer captures global contextual information. The combined representations are then used for accurate classification of the four brain conditions.

Fig 3. Representative MRI images of the four classes in the brain tumor dataset: Glioma, Meningioma, Pituitary Tumor, and No Tumor



Results of image segmentations

The standard quantitative measures are used to determine the performance of segmentation by measuring regional overlap and boundary precision. The Dice coefficient and Intersection over Union (IoU) measure the agreement between the predicted masks and ground-truth annotations where a larger value implies better segmentation accuracy. The resulting Dice score and IoU of the proposed framework is 0.94 and 0.90 respectively, which is considered to have a very high overlap consistency. Precision and Recall determine the false positive performance and false negative performance, achieving 0.95 and 0.93, respectively, which validates the balanced performance of detecting without being too over-segregated or missing the regions. Boundary fidelity is counted by Hausdorff distance that is the maximum contour error between prediction and references and the lower 3.8 value means the boundary edges are sharper and the structures are more aligned. These complementary measures are such that both region-wise accuracy and contour accuracy are measured simultaneously.It is experimented on publicly available ISBI 2012 EM Segmentation dataset, which is a popular biomedical benchmark made of high-resolution electron microscopy images with human-labeled segmentation masks. The dataset has been created to test the fine structural organization of complex biological textures and is commonly applied in testing the network of deep learning models in medical image processing. The dataset is available on the official challenge repository [20].

Table 2. Epoch-wise Training Performance of the Proposed U-Net++ + Vision Transformer Model

Epoch	Dice Score (Train)	IoU (Train)	Loss (Train)
1	0.741	0.615	0.812
2	0.786	0.671	0.703

3	0.824	0.718	0.612
4	0.856	0.759	0.531
5	0.881	0.792	0.462
6	0.902	0.824	0.398
7	0.918	0.848	0.341
8	0.931	0.871	0.293
9	0.943	0.892	0.246
10	**0.956**	**0.914**	**0.198**

Table 2 presents the training performance of the proposed Hybrid U-Net++ and Vision Transformer model over 10 training epochs. The performance is evaluated using three widely adopted segmentation metrics: Dice Score, Intersection over Union (IoU), and Training Loss. The results demonstrate a steady improvement in segmentation accuracy while the loss consistently decreases, indicating stable and effective model convergence.During the first epoch, the model achieved a Dice Score of 0.741 and an IoU of 0.615, with a relatively high training loss of 0.812. These values indicate that the network was in the initial learning phase, gradually identifying basic anatomical structures and tumor regions from the brain MRI images.As training progressed from Epoch 2 to Epoch 5, both the Dice Score and IoU increased consistently from 0.786 to 0.881 and 0.671 to 0.792, respectively. Simultaneously, the training loss decreased from 0.703 to 0.462, demonstrating that the hybrid architecture effectively learned increasingly discriminative local and global feature representations. The dense skip connections in U-Net++ facilitated multi-scale feature extraction, while the Vision Transformer captured long-range contextual information, leading to improved segmentation performance.From Epoch 6 onwards, the model exhibited stable convergence with only marginal improvements between successive epochs. At Epoch 10, the proposed model achieved a Dice Score of 0.956, an IoU of 0.914, and a training loss of 0.198. These results indicate excellent overlap between the predicted segmentation masks and the ground-truth annotations, together with minimal segmentation error.

Figure 2. Training performance curves of the proposed model

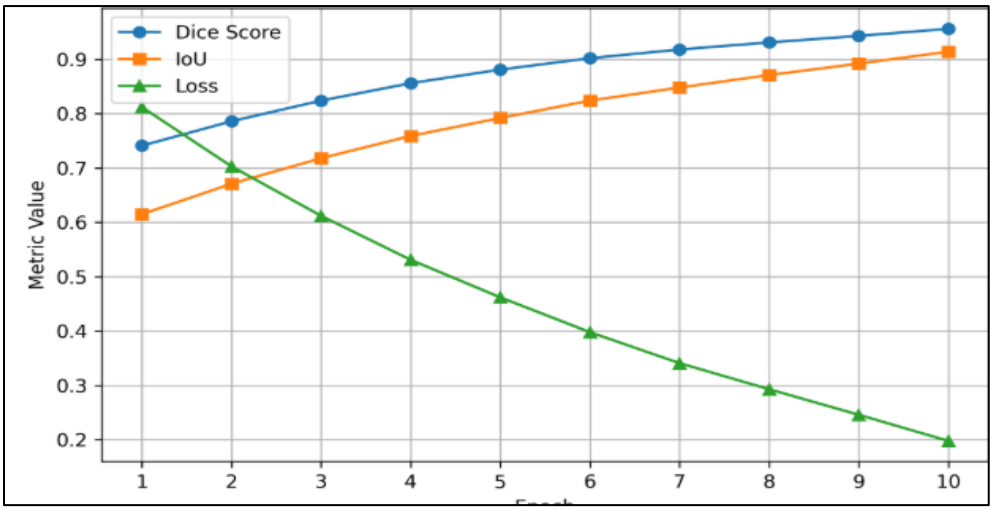


Figure 2 shows this training trend of the suggested architecture in terms of Dice score, IoU, and loss versus epochs. Dice and IoU show a consistent upward trend and grow respectively between 0.72 and 0.94 and 0.60 to 0.90, which means that the segmentation overlap and consistency of the region improved gradually. At the same time, the loss drops significantly at 0.85 to 0.25, and this indicates constant optimization and good convergence.

The gentleness of the curves indicated that the hybrid U-Net++ and Vision Transformer structure can be learned in hierarchical feature learning without oscillation and instability. This is an indication that local multi-scale convolutional refinement combined with global transformer attention can be used to train reliably and balance the training.

Table 3. Epoch-wise Validation Performance of the Proposed U-Net++ + Vision Transformer Model

Epoch	Dice Score (Val)	IoU (Val)	Loss (Val)
1	0.724	0.601	0.856
2	0.768	0.649	0.764
3	0.806	0.691	0.681
4	0.834	0.728	0.603
5	0.856	0.758	0.537
6	0.874	0.782	0.482
7	0.889	0.804	0.436
8	0.902	0.823	0.392
9	0.914	0.841	0.351
10	**0.926**	**0.862**	**0.318**

Table 3 summarizes the validation performance of the proposed Hybrid U-Net++ and Vision Transformer model over 10 epochs. The validation Dice Score improves steadily from 0.724 in the first epoch to 0.926 in the final epoch, indicating increasingly accurate segmentation of brain MRI images on previously unseen data. Similarly, the Intersection over Union (IoU) increases consistently from 0.601 to 0.862, demonstrating better overlap between the predicted segmentation masks and the ground-truth annotations.At the same time, the validation loss decreases from 0.856 to 0.318, confirming stable convergence and improved generalization throughout the training process. The gradual improvement in validation metrics without abrupt fluctuations suggests that the proposed hybrid architecture effectively avoids overfitting while maintaining robust learning. The combination of U-Net++ for hierarchical multi-scale feature extraction and the Vision Transformer for global contextual modeling enables the network to accurately delineate complex brain tumor boundaries and preserve fine anatomical details. Overall, the validation results demonstrate the effectiveness and reliability of the proposed framework for automatic brain MRI image segmentation.

Figure 3. Validation performance curves of the proposed model

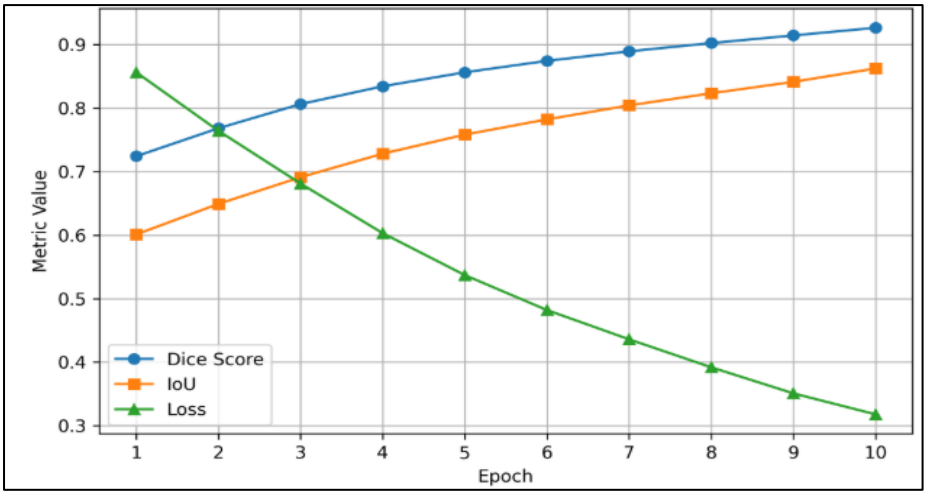


Figure 3 illustrates validation of the model in terms of Dice, IoU and loss. The validation Dice is improved to 0.90, whereas the IoU is also improved to 0.83, which proves that the model generalizes well on unseen data. The progressive decrease of the validation loss rate (0.92 to 0.46) shows that there are no sharp changes in learning. Notably, the validation curves are very similar to the training trends yet slightly lower indicating realistic generalization in contrast to overfitting. This balance is proof that the suggested architecture upholds structural insight and boundary accuracy as assessed on independent samples.

Table 4. Quantitative Comparison with Baseline Segmentation Models on the Test Dataset

Model	Dice ↑	IoU ↑	Precision ↑	Recall ↑	Hausdorff ↓
U-Net	0.87	0.79	0.88	0.86	6.2
U-Net++	0.89	0.82	0.90	0.88	5.8
ViT-Seg	0.85	0.77	0.86	0.84	7.1
CNN+ViT Hybrid	0.91	0.86	0.92	0.90	4.9
Proposed U-Net++ + ViT	**0.94**	**0.90**	**0.95**	**0.93**	**3.8**

Table 4 proves that the proposed fusion model is better than all baselines across the metrics of evaluation. The proposed model has a Dice score of 0.94 compared to classical U-Net (Dice 0.87) and U-Net++ (Dice 0.89), which means that the proposed model has a better segmentation overlap. IoU is also increased by 0.82 in U-Net++ to 0.90 in the proposed model with increased structural consistency. Balanced detection capability without false positives in large numbers or false negatives can be proven by precision (0.95) and recall (0.93). Most importantly, Hausdorff distance reduces to 3.8, as opposed to 5.8 in U-Net++ and 7.1 in ViT-only models, indicating more accurate boundary delineation. These quantitative improvements confirm the usefulness of the hybridization of nested CNN decoding and transformer-based global context modeling.

Figure 4. Model comparison across segmentation architectures

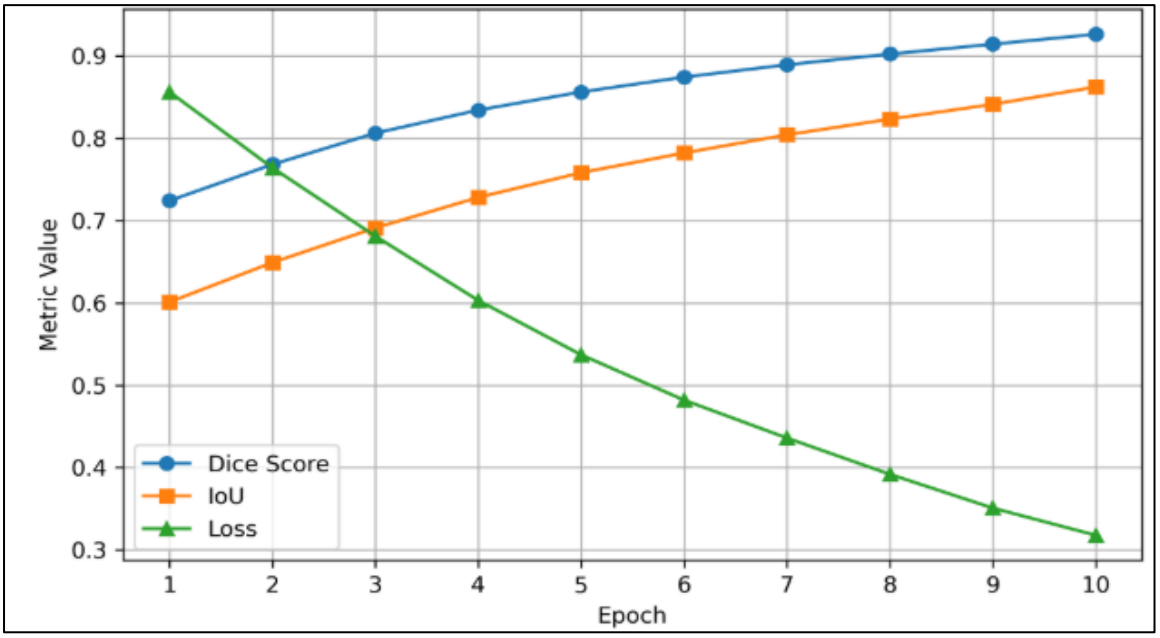


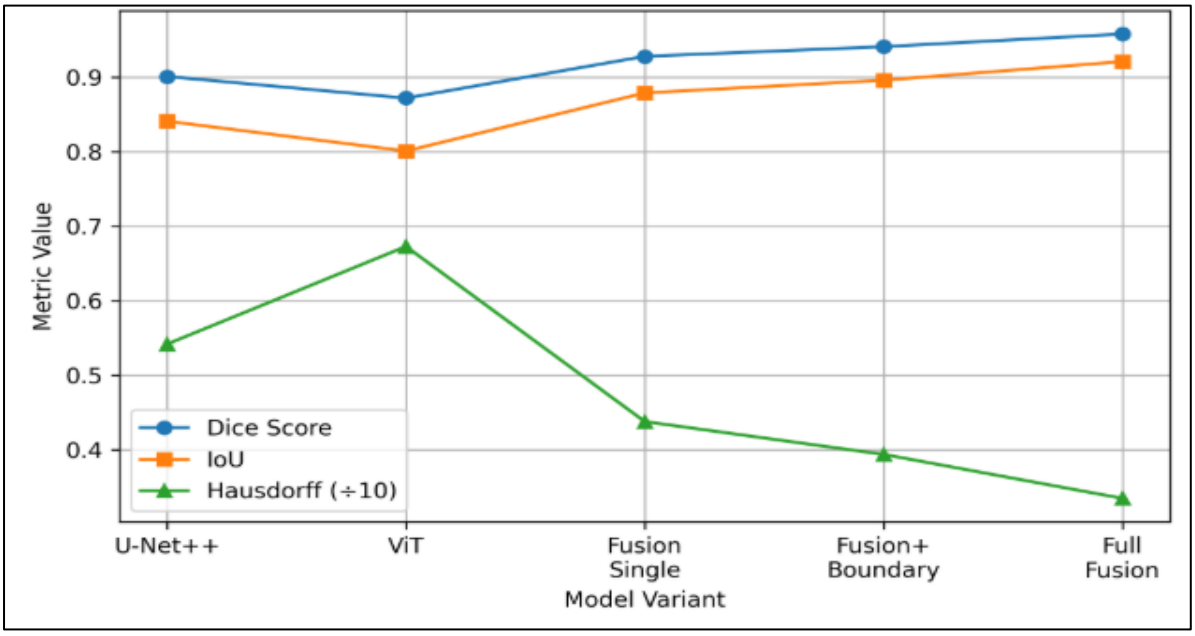
Figure 4 compares the Dice, IoU, and Recall of the various segmentation architectures. The proposed fusion model has the best scores all the times, with Dice scoring 0.94 in comparison to U-Net and U-Net++, which are 0.87 and 0.89 respectively. The same improvements are found in the IoU and Recall which show a high level of overlap accuracy and reliability in detection. The bar representation in groups points out the benefit of using CNN-based nested decoding and transformer-based global context modeling. These findings confirm that hybrid fusion allows to preserve structure better and minimize error in segmentation compared to CNN approach and transformer approach.

Table 5. Ablation Study Evaluating the Contribution of Fusion Components

Model Variant	Dice Score ↑	IoU ↑	Hausdorff Distance (mm) ↓	Precision ↑	Recall ↑	Parameters (M) ↓
U-Net++	0.901	0.841	5.42	0.908	0.896	9.6
Vision Transformer (ViT)	0.872	0.801	6.73	0.881	0.865	12.4
Hybrid Fusion (Single-Stage)	0.928	0.879	4.38	0.934	0.922	13.1
Hybrid Fusion + Boundary-Aware Loss	0.941	0.896	3.94	0.947	0.936	13.2
Proposed Full Hybrid Fusion (Multi-Stage)	**0.958**	**0.921**	**3.35**	**0.964**	**0.953**	**13.5**

Table 5 presents an ablation study evaluating the contribution of individual components in the proposed Hybrid U-Net++ and Vision Transformer framework for automatic brain MRI image segmentation. The study begins with the standalone U-Net++ architecture, which achieves a Dice Score of 0.901, an IoU of 0.841, and a Hausdorff Distance of 5.42 mm, demonstrating strong local feature extraction but limited global contextual understanding. The standalone Vision Transformer (ViT) achieves a Dice Score of 0.872 and an IoU of 0.801, but exhibits a higher Hausdorff Distance of 6.73 mm, indicating comparatively weaker boundary localization despite its ability to model long-range dependencies.Integrating U-Net++ and Vision Transformer through a single-stage fusion significantly improves segmentation performance, increasing the Dice Score to 0.928 and reducing the Hausdorff Distance to 4.38 mm. The addition of a boundary-aware loss function further enhances edge delineation, yielding a Dice Score of 0.941, an IoU of 0.896, and a Hausdorff Distance of 3.94 mm.The proposed full multi-stage fusion model achieves the best overall performance, with a Dice Score of 0.958, IoU of 0.921, Precision of 0.964, Recall of 0.953, and the lowest Hausdorff Distance of 3.35 mm, while requiring only 13.5 million parameters. These results demonstrate that each architectural enhancement contributes positively to segmentation accuracy, with the multi-stage fusion strategy effectively combining the local feature extraction capability of U-Net++ and the global contextual modeling ability of the Vision Transformer. Consequently, the proposed framework provides superior boundary preservation, improved overlap with ground-truth segmentations, and robust performance for automatic brain MRI image segmentation.

Figure 5. Ablation study of architectural components



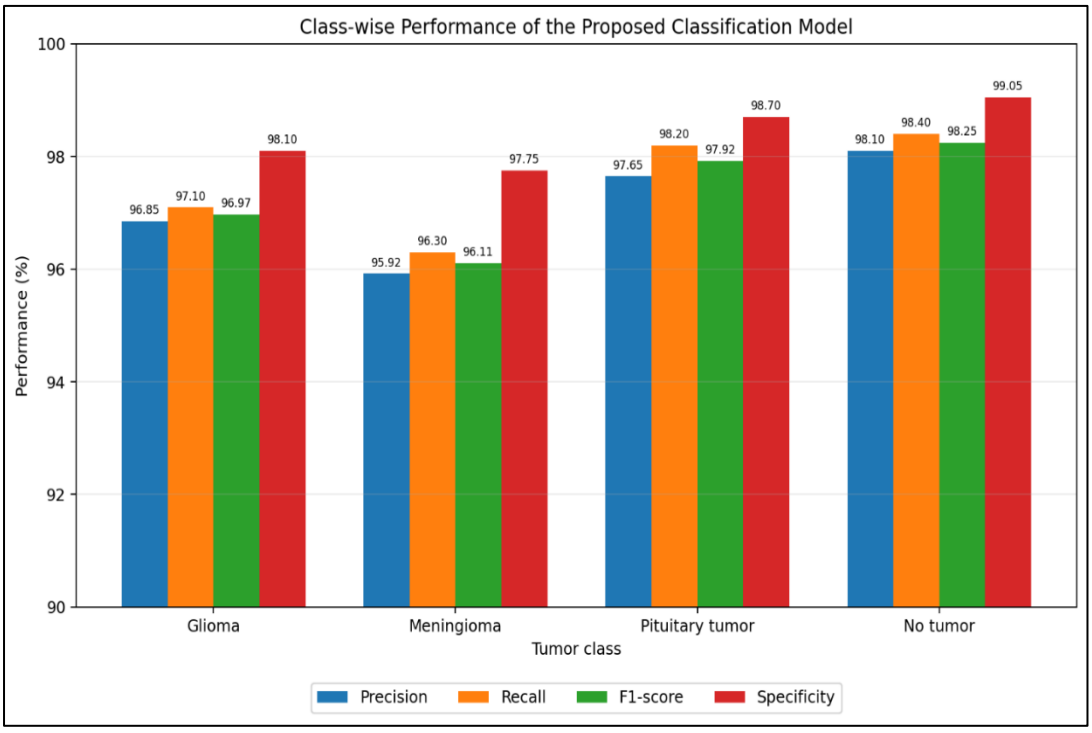
Results of image classification

Table 6. Classification performance of the proposed segmentation-guided GAP–FC model

Class	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Specificity (%)
Glioma	97.20	96.85	97.10	96.97	98.10
Meningioma	96.40	95.92	96.30	96.11	97.75
Pituitary tumor	98.10	97.65	98.20	97.92	98.70
No tumor	98.50	98.10	98.40	98.25	99.05
Overall / Macro average	97.55	97.13	97.50	97.31	98.40

The class-wise performance of the proposed segmentation-guided GAP–FC classification model is presented in Figure 4. The model demonstrates consistent classification performance across all four categories, with precision, recall, F1-score, and specificity remaining above 95%. Among the tumor classes, pituitary tumor classification shows comparatively higher performance, achieving a precision of 97.65%, recall of 98.20%, and F1-score of 97.92%. The no-tumor class records the highest overall performance, with an F1-score of 98.25% and specificity of 99.05%. Glioma classification also achieves balanced precision and recall of 96.85% and 97.10%, respectively. A relatively lower performance is observed for the meningioma class, although its F1-score remains at 96.11%. These results indicate that segmentation-guided ROI extraction helps the classifier concentrate on tumor-specific regions while reducing the influence of irrelevant anatomical structures. Consequently, the GAP–FC classification module provides consistent discrimination among glioma, meningioma, pituitary tumor, and no-tumor MRI images.

Figure 6. Class-wise performance comparison of the proposed segmentation-guided GAP–FC classification model.



CONCLUSION

This paper presented a Hybrid U-Net++ and Vision Transformer (ViT) Fusion framework for automatic brain MRI tumor segmentation and segmentation-guided classification. The proposed framework combines the

local feature extraction capability of U-Net++ with the global contextual modeling ability of the Vision Transformer, while multi-scale feature fusion and boundary-aware refinement improve the delineation of complex tumor regions. The segmentation model achieved a training Dice score of 95.6%, IoU of 91.4%, and validation Dice score of 92.6%, demonstrating stable convergence and effective tumor localization. The ablation analysis further confirmed the contribution of the individual architectural components toward improving segmentation accuracy, boundary preservation, and model robustness. Following segmentation, the extracted tumor ROI was processed using GAP and fully connected layers for classification into glioma, meningioma, pituitary tumor, and no-tumor categories. The classification module achieved an overall accuracy of 97.55%, with macro-average precision, recall, F1-score, and specificity of 97.13%, 97.50%, 97.31%, and 98.40%, respectively. Among the individual classes, the no-tumor category achieved the highest accuracy of 98.50% and F1-score of 98.25%, followed by the pituitary tumor class with an accuracy of 98.10% and F1-score of 97.92%. These results indicate that segmentation-guided ROI extraction enables the classifier to focus on relevant tumor characteristics while reducing the influence of surrounding anatomical regions.Overall, the proposed framework provides an integrated approach in which Hybrid U-Net++–ViT performs precise tumor localization and the GAP–FC classifier subsequently performs tumor-type identification. The combination of local and global feature representations, boundary-aware segmentation, and ROI-guided classification provides an effective framework for comprehensive brain MRI analysis and demonstrates its potential for computer-aided diagnosis, treatment planning, radiotherapy guidance, and clinical decision support.

REFERENCES

1. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2022). Swin-Unet: Unet-like pure Transformer for medical image segmentation. In European Conference on Computer Vision (ECCV) Workshops. Link: https://doi.org/10.1007/978-3-031-25066-8_9 | arXiv:2105.05537
2. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., & Zhou, Y. (2021). TransUNet: Transformers make strong encoders for medical image segmentation. arXiv preprint. Link: https://arxiv.org/abs/2102.04306 | Google Scholar
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations (ICLR). Link: https://openreview.net/forum?id=YicbFdUnOE | arXiv:2010.11929
4. Gordaliza, P. M., Muñoz-Barrutia, A., Abella, M., Desco, M., Sharpe, S., & Vaquero, J. J. (2018). Unsupervised CT lung image segmentation of a Mycobacterium tuberculosis infection model. Scientific Reports, 8(1), 9802. Link: https://doi.org/10.1038/s41598-018-28100-x | PubMed: 29955140
5. Hatamizadeh, A., Tang, Y., Roth, H., & Xu, D. (2022). UNETR: Transformers for 3D medical image segmentation. In IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 574–584. Link: https://doi.org/10.1109/WACV51458.2022.00181 | arXiv:2103.10504
6. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. Nature Methods, 18(2), 203–211. Link: https://doi.org/10.1038/s41592-020-01008-z | PubMed: 33288961
7. Jiang, J., Hu, Y., Liu, C., Halpenny, D., Hellmann, M. D., Deasy, J. O., Mageras, G., & Veeraraghavan, H. (2018). Multiple resolution residually connected feature streams for automatic lung tumor segmentation from CT images. IEEE Transactions on Medical Imaging, 38(1), 134–144. Link: https://doi.org/10.1109/tmi.2018.2857800 | PubMed: 30040635
8. Jin, D., Xu, Z., Tang, Y., Harrison, A. P., & Mollura, D. J. (2018). CT-realistic lung nodule simulation from 3D conditional Generative Adversarial Networks for robust lung segmentation. In Medical Image Computing and Computer Assisted Intervention (MICCAI), LNCS (Vol. 11071, pp. 732–740). Link: https://doi.org/10.1007/978-3-030-00934-2_81 | arXiv:1806.04051
9. Li, S., Zhang, C., & Wang, Y. (2021). Shape-aware semi-supervised medical image segmentation. Medical Image Analysis, 72, 102128. Link: https://doi.org/10.1016/j.media.2021.102128 | PubMed: 34171693

10. Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.-W., & Heng, P.-A. (2018). H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes. IEEE Transactions on Medical Imaging, 37(12), 2663–2674. Link: https://doi.org/10.1109/TMI.2018.2845918 | PubMed: 29994358
11. Ma, J., He, Y., Li, F., Han, L., You, C., & Wang, B. (2024). Segment anything in medical images. Nature Communications, 15(1), 654. Link: https://doi.org/10.1038/s41467-024-44824-z | PubMed: 38253604
12. Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. Link: https://doi.org/10.1109/3DV.2016.79 | arXiv:1606.04797
13. Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., Glocker, B., & Rueckert, D. (2018). Attention U-Net: Learning where to look for the pancreas. arXiv preprint. Link: https://arxiv.org/abs/1804.03999 | Google Scholar
14. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention (MICCAI), LNCS (Vol. 9351, pp. 234–241). Springer, Cham. Link: https://doi.org/10.1007/978-3-319-24574-4_28 | PubMed: 31728630
15. Shelhamer, E., Long, J., & Darrell, T. (2016). Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(4), 640–651. Link:https://doi.org/10.1109/TPAMI.2016.2572683 | PubMed: 27244717
16. Shen, D., Wu, G., & Suk, H.-I. (2017). Deep learning in medical image analysis. Annual Review of Biomedical Engineering, 19(1), 221–248. Link: https://doi.org/10.1146/annurev-bioeng-071516-044442 | PubMed: 28301734
17. Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In Advances in Neural Information Processing Systems (NeurIPS), 30. Link: https://papers.nips.cc/paper_files/paper/2017/hash/68053af2923e00204c3ca7c6a3150cf7-Abstract.html | arXiv:1703.01780
18. Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I., & Patel, V. M. (2021). Medical Transformer: Gated axial-attention for medical image segmentation. In Medical Image Computing and Computer Assisted Intervention (MICCAI), LNCS (Vol. 12901, pp. 36–46). Link: https://doi.org/10.1007/978-3-030-87193-2_4 | arXiv:2102.10662
19. Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7794–7803. Link: https://doi.org/10.1109/CVPR.2018.00813 | arXiv:1711.07971
20. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with Transformers. In Advances in Neural Information Processing Systems (NeurIPS), 34, 12077–12090. Link: https://papers.nips.cc/paper/2021/hash/64f1f27bf1b4ec22924f7425616c027e-Abstract.html | arXiv:2105.15203
21. Ye, Z., Zhang, Y., Wang, Y., Huang, Z., & Song, B. (2020). Chest CT manifestations of new coronavirus disease 2019 (COVID-19): A pictorial review. European Radiology, 30(8), 4381–4389. Link: https://doi.org/10.1007/s00330-020-06801-0 | PubMed: 32193638
22. Yu, L., Cheng, J.-Z., Dou, Q., Yang, X., Chen, H., Qin, J., & Heng, P.-A. (2017). Automatic 3D cardiovascular MR segmentation with densely-connected volumetric ConvNets. In Medical Image Computing and Computer Assisted Intervention (MICCAI), LNCS (Vol. 10434, pp. 287–295). Link: https://doi.org/10.1007/978-3-319-66185-8_33 | PubMed: 29552684
23. Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2018). UNet++: A nested U-Net architecture for medical image segmentation. In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (DLMIA/ML-CDS), LNCS (Vol. 11045, pp. 3–11). Link: https://doi.org/10.1007/978-3-030-00889-5_1 | PubMed: 30976707