Journal of Advanced Health Informatics Research (JAHIR)
Vol. 4, No. 1, (2023), pp. 23-38
ISSN: 2985-6124, DOI: 10.59247/jahir.v4i1.369

Knowledge Distillation in Lightweight U-Net Transformer Architectures for Brain Tumor Segmentation

Toat Tuloh [1], Purwono Purwono [1], Iis Setiawan Mangkunegara [1]
[1] *Departement of Informatics, Universitas Harapan Bangsa, Indonesia*

ARTICLE INFO

Article history:

Received February 05, 2026
Revised March 24, 2026
Published July 31, 2026

Keywords:

U-Net 3D;
Knowledge Distillation;
Dice Score;
IoU;
GFLOPS.

ABSTRACT

Brain tumor segmentation from Magnetic Resonance Imaging was an essential step for therapy planning, prognosis evaluation, and treatment monitoring in glioma patients. Manual delineation required substantial time and was prone to inter-observer variability. Although deep learning models achieved high segmentation accuracy, performance improvements were often accompanied by increased computational complexity, limiting their applicability in resource-constrained clinical environments. To address this issue, an efficiency-oriented segmentation framework was developed based on a lightweight three-dimensional U-Net enhanced with a shallow Transformer module and guided by knowledge distillation. The main contribution of this study was the integration of logit-level and feature-level distillation to improve segmentation capability while maintaining low inference complexity. The framework emphasized a balanced trade-off between segmentation accuracy and computational efficiency rather than benchmark maximization. Experiments were conducted using the BraTS 2020 and BraTS 2021 datasets. The official training sets were internally split into eighty percent for training and twenty percent for validation. The student network was trained using a hybrid segmentation loss combined with temperature-scaled logit distillation and bottleneck feature alignment from a higher-capacity teacher model. Model performance was evaluated using Dice score, Intersection over Union, and computational complexity measured in floating-point operations. On the BraTS 2021 dataset, the proposed model achieved Dice scores of 0.6538 for Whole Tumor, 0.5382 for Tumor Core, and 0.5304 for Enhancing Tumor. Per-class Dice values were 0.9706 for background, 0.3028 for necrotic or non-enhancing tumor core, 0.4938 for edema, and 0.4936 for enhancing tumor. The corresponding Intersection over Union values followed similar trends. The model maintained an inference complexity of approximately 54.38 gigafloating-point operations for input patches of size 64 × 64 × 64. These findings indicated that the proposed framework achieved a stable balance between segmentation performance and computational efficiency, supporting practical deployment under limited computational resources.

Corresponding Author:

Toat Tuloh, Departement of Informatics, Universitas Harapan Bangsa, Indonesia
Email: toatuloh@student.uhb.ac.id

1. INTRODUCTION

Brain tumors, particularly gliomas, are among the most lethal neurological diseases because of their biological complexity and substantial inter-patient heterogeneity [1]. On magnetic resonance imaging (MRI), gliomas are commonly divided into three clinically relevant subregions: the enhancing tumor (ET), the necrotic and non-enhancing tumor core (NCR/NET), and the peritumoral edema (ED), each of which has distinct clinical significance [2]. Accurate segmentation of these subregions plays a critical role in distinguishing tumor tissue from healthy tissue and provides essential information for treatment planning and therapeutic response assessment [3].

Manual segmentation remains widely used in clinical practice. The process is labor-intensive because it requires slice-by-slice delineation with a high level of precision [4]. Segmentation quality also depends on the expertise of individual annotators, resulting in inter-observer variability caused by differences in clinical interpretation [5]. These limitations have driven the development of automated segmentation methods that provide faster, more consistent, and reproducible results while improving clinical workflow efficiency [6].

Advances in deep learning, particularly Convolutional Neural Networks (CNNs), have transformed medical image analysis through hierarchical feature extraction capabilities [7]. Among the available architectures, U-Net has become one of the most widely adopted models for biomedical image segmentation because of its encoder-decoder architecture and skip connections, which preserve spatial information throughout feature reconstruction [8]. The increasing demand for volumetric image analysis has also promoted the adoption of 3D U-Net, which exploits three-dimensional spatial information more effectively than slice-based two-dimensional approaches [9].

Higher segmentation accuracy is often accompanied by increased network complexity, larger numbers of model parameters, and greater computational demands [10]. Such computational requirements present challenges for deployment in clinical environments with limited hardware resources [11]. The development of computationally efficient segmentation models remains essential to facilitate the broader adoption of deep learning in clinical practice [12].

In addition to complexity, brain tumor segmentation also faces class imbalance challenges, especially in small sub-regions such as ET, which are clinically important but have a lower proportion of pixels/voxels compared to other classes [13]. Attention mechanisms have been widely explored to improve feature representation, including the Bottleneck Attention Module (BAM), Convolutional Block Attention Module (CBAM), and Squeeze-and-Excitation (SE) module. Among these approaches, the SE module has demonstrated favorable computational efficiency by recalibrating channel-wise feature responses with minimal computational overhead (He & Gopalakrishnan, 2024).

On the other hand, Transformers are increasingly adopted to capture global dependencies that are difficult for conventional CNNs to learn [14]. Full Transformers tend to be computationally expensive and therefore unsuitable for resource-constrained systems [15]. Consequently, lightweight Transformers can be positioned as enhancement modules to improve global context without excessively increasing GFLOPs [14]. To further reduce complexity, this study applies knowledge distillation as a strategy to transfer knowledge from a more complex teacher model to a lighter student model [16]. The main focus of this study is to develop an efficiency-oriented brain tumor segmentation framework that balances segmentation performance and computational cost, with evaluation on the BraTS 2020 and BraTS 2021 benchmark datasets [17].

The main contributions of this study are summarized as follows [18]. This work proposes a lightweight 3D U-Net segmentation architecture that integrates a shallow Transformer module to jointly capture local anatomical details and global contextual information while maintaining low computational complexity [19]. A multi-level knowledge distillation strategy is introduced, enabling effective knowledge transfer from a high-capacity teacher network to a compact student model, thereby improving segmentation performance, particularly on clinically important sub-regions, without significantly increasing model parameters or GFLOPs [20]. Extensive evaluations are conducted on the BraTS 2020 and BraTS 2021 benchmark datasets, demonstrating that the proposed method achieves stable and consistent Dice and IoU performance under controlled computational complexity, emphasizing deployment feasibility in resource-constrained clinical environments [21].

Unlike many recent segmentation frameworks that prioritize state-of-the-art accuracy through increasingly complex Transformer-based architectures, this study adopts an efficiency-oriented design philosophy. The proposed framework is designed to achieve a balanced trade-off between segmentation performance and computational efficiency rather than maximizing Dice scores at the expense of computational cost. By integrating a lightweight Transformer module into a 3D U-Net backbone and incorporating multi-level knowledge distillation, the model improves representational capacity while maintaining computational

efficiency with controlled GFLOPs. This design distinguishes the proposed framework from accuracy-oriented approaches and supports its potential applicability in resource-constrained clinical environments.

2. METHODS

We used the official training subsets of BraTS 2020 (369 subjects) and BraTS 2021 (1,251 subjects), each containing four MRI modalities (T1, T1ce, T2, and FLAIR) with a spatial resolution of 240 × 240 × 155 voxels. In our experiments, we followed a fixed subject-wise 80/20 split to form training and held-out validation sets (see Section 2.7). For each study, the intensity of each modality was normalized using z-score normalization within the brain mask. The volumes were then resampled and cropped into 128 × 128 × 128 cubes. The 3D data augmentation scheme included random flipping along the axial, coronal, and sagittal axes, 90° rotations, mild elastic deformation, intensity shifting, gamma augmentation (0.8–1.2), Gaussian noise, and channel dropout to simulate missing modalities [22]. The segmentation targets follow the standard BraTS voxel-wise four-class labeling scheme: background (BG), necrotic/non-enhancing tumor core (NCR/NET), edema (ED), and enhancing tumor (ET). In addition, region-based metrics (WT and TC) are computed following the BraTS protocol by aggregating the predicted sub-regions.

2.1. Research Design

This study employed a computational experimental design to develop, train, and evaluate deep learning models for brain tumor segmentation. Experiments were conducted using the BraTS 2020 and BraTS 2021 benchmark datasets. Model performance was evaluated using four segmentation classes: background (BG), necrotic and non-enhancing tumor core (NCR/NET), peritumoral edema (ED), and enhancing tumor (ET).

2.2. Datasets

The BraTS 2020 and BraTS 2021 benchmark datasets were used in this study. Both datasets provide multimodal magnetic resonance imaging (MRI) scans with expert-annotated brain tumor segmentation masks. The BraTS benchmark has been widely adopted in brain tumor segmentation research because it provides curated multi-institutional data and standardized evaluation protocols.

Region-based evaluation followed the standard BraTS protocol. The Whole Tumor (WT) region was defined as the union of the necrotic and non-enhancing tumor core (NCR/NET), peritumoral edema (ED), and enhancing tumor (ET). The Tumor Core (TC) region was defined as the union of NCR/NET and ET, whereas the Enhancing Tumor (ET) region corresponded to the ET label alone.

2.3. Preprocessing Data

Pre-processing stages are performed to improve model training stability. The process includes intensity normalization to make the distribution of each subject more consistent and volumetric input formation to support 3D modeling. Normalization strategies are considered a common approach in medical image data processing to reduce inter-patient intensity variation [23].

2.4. Proposed Model Architecture

The proposed framework adopts a lightweight 3D U-Net architecture as the student segmentation network [26]. The student model follows an encoder–decoder structure with skip connections to preserve spatial details across multiple feature scales. Each encoder stage consists of two consecutive 3D convolution layers followed by instance normalization and LeakyReLU activation. Downsampling is performed using max pooling with stride 2, while upsampling in the decoder is implemented using transposed 3D convolutions.

To enhance global contextual representation, a lightweight Transformer module is inserted at the bottleneck of the network. Unlike full self-attention mechanisms, the proposed Transformer is implemented using 1×1×1 convolutional attention and feed-forward projection blocks with residual connections. This design enables contextual refinement without introducing quadratic complexity typical of standard multi-head self-attention [3]. Consequently, the model maintains low computational overhead while still benefiting from global feature interaction.

The student network uses an initial channel width of 32 and expands hierarchically up to 256 channels at the bottleneck. The final prediction head applies a 1×1×1 convolution to produce four output channels corresponding to background (BG), NCR/NET, edema (ED), and enhancing tumor (ET). WT and TC are not directly predicted as separate output channels; instead, they are derived from the predicted four-class mask via region aggregation during evaluation.

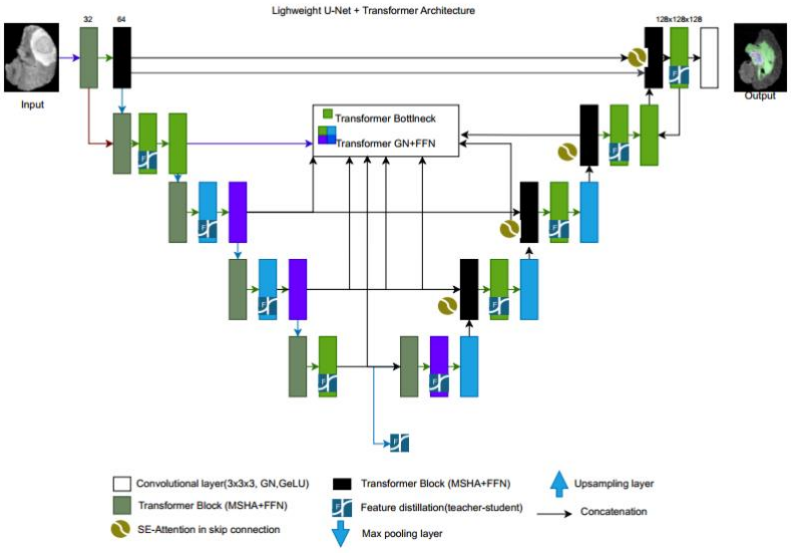

Fig. 1. Architecture of the 3D U-Net Model with Lightweight Transformer Enhancement

Figure 1 illustrates the encoder-decoder architecture of the proposed 3D U-Net with multi-level skip connections. A lightweight Transformer module is positioned at the bottleneck to capture global contextual information before feature reconstruction in the decoder.

A high-capacity teacher network is employed during training to guide the lightweight student model through knowledge distillation. The teacher network shares the same encoder-decoder architecture as the student but uses a wider channel configuration and greater representational capacity. Specifically, it adopts a larger base number of channels and incorporates a full convolutional bottleneck without structural compression, enabling the extraction of richer hierarchical features and more stable decision boundaries.

The teacher network receives the same multimodal MRI volumes as the student and produces identical four-class segmentation outputs. Only the student model is retained during inference, ensuring that the additional complexity of the teacher network does not increase computational requirements during deployment.

The student model is optimized for computational efficiency, with a reduced number of parameters and an inference complexity of approximately 54.38 GFLOPs using input patches of 64 × 64 × 64 voxels. The teacher network has substantially higher representational capacity because of its wider channel configuration and increased model complexity.

This asymmetric teacher-student design enables efficient knowledge transfer during training while preserving the lightweight characteristics of the student model for deployment in resource-constrained environments. The teacher network is removed after training, ensuring that inference is performed without additional computational overhead.

2.5. Knowledge Distillation Scheme

Knowledge distillation is applied to transfer representational knowledge from the high-capacity teacher network to the lightweight student network without increasing inference complexity [4]. The distillation process is performed exclusively during training.

Output-based distillation is applied at the prediction level. The soft probability distribution generated by the teacher network serves as a supervisory signal for the student network. A temperature-scaled softmax is used to produce softened targets that provide richer class relationship information compared to one-hot labels [4]. The student is optimized to minimize the divergence between its predictions and the teacher's softened output.

Feature-based distillation is applied at intermediate representation levels [6]. Feature maps from the teacher bottleneck are aligned with those of the student using a feature-matching loss. This encourages the student to mimic high-level semantic structures learned by the teacher while maintaining its compact architecture. The overall loss function is defined as equation 1:

$$L_{total} = L_{seg} + \lambda 1 L_{logit} + \lambda 2 L_{feature} \quad (1)$$

where L_seg denotes the segmentation loss (Dice-based or hybrid loss), L_logit represents the soft-target distillation loss, and $L_feature$ corresponds to the intermediate feature alignment loss. The weighting coefficients λ1 and λ2 balance segmentation supervision and distillation guidance.

Because distillation is only active during training, the inference computational cost remains identical to that of the lightweight student network.

2.6. Loss Function

The training objective combines a segmentation loss with a knowledge distillation loss to improve the student model performance while maintaining low inference complexity. Considering the severe class imbalance in brain tumor sub-regions, a Dice-based loss is adopted due to its robustness in imbalanced scenarios and its compatibility with probabilistic outputs [7]. In this study, the segmentation loss is defined as a hybrid objective consisting of voxel-wise Cross-Entropy loss and Dice loss as equation 2:

$$L_{seg} = L_CE + Ldice \tag{2}$$

Where L_CE enforces voxel-level class discrimination, while L_dice directly optimizes overlap between prediction and ground truth.

For output-level distillation, the student is guided using softened class probability distributions produced by the teacher. Let zt denote the teacher logits and zs denote the student logits [8]. The softened probabilities are computed using a temperature parameter T as follows equation (3):

$$\begin{aligned} p_t^{(T)} &= softmax(z_t/T) \\ p_s^{(T)} &= softmax(z_s/T) \end{aligned} \tag{3}$$

The logit distillation loss is defined using Kullback–Leibler divergence as equation (4):

$$L\_logit \; = \; T\text{^}2 * KL(p_t^{(T)}) \; ||p_s^{(T)})) \tag{4}$$

The factor T2 is included to preserve gradient magnitude under temperature scaling [9]. The overall prediction supervision is then balanced using a weighting coefficient α, where the student learns from both the hard ground-truth labels and the teacher soft targets [10].

In addition to output supervision, feature-level distillation is applied to transfer higher-level representations learned by the teacher [7]. Since the proposed student architecture places a lightweight Transformer block at the bottleneck, feature distillation is conducted at the bottleneck feature map to encourage alignment of global contextual representations.

Let ft denote the teacher bottleneck feature and fs the student bottleneck feature. In the proposed network, fs corresponds to the output of the encoder stage immediately before upsampling (i.e., the input/output of the Transformer bottleneck) [11]. A mean squared error (MSE) feature matching loss is used equation (5):

$$L_feat = \left|\left|f_t - f_s\right|\right|_2^2 \tag{5}$$

A weighting coefficient β is used to control the contribution of feature distillation. The final training objective is expressed as equation (6):

$$L_total = (1 - \alpha)L_seg + \alpha L_logit + \beta L_feat \tag{6}$$

where α controls the tradeoff between segmentation supervision and logit based distillation, while β determines the strength of feature alignment. Importantly, both L_logit and L_feat are only applied during training; the inference computational cost (GFLOPs) remains identical to the lightweight student network.

2.7. Training Procedure

Training is performed using the Adam optimizer for stability and convergence in deep network training [12]. Data splitting follows the BraTS protocol according to the available training and validation subsets [12]. Validation is conducted separately on each dataset to ensure consistent performance on BraTS 2020 and BraTS 2021.

Training was conducted using input patches of size 64 × 64 × 64 voxels extracted from preprocessed MRI volumes. The student network used a base channel width of 32 and a sliding window size of 8 × 8 × 8 within the lightweight Transformer bottleneck. The batch size was set to 1 due to GPU memory constraints associated

with 3D volumetric inputs. The model was trained for 15 epochs using the Adam optimizer. The initial learning rate was set to 1×10^{-4} without warm-up scheduling. The number of epochs was determined empirically based on validation convergence behavior under computational constraints.

For knowledge distillation, the temperature parameter was set to T = 4 to produce softened probability distributions. The balancing coefficients were empirically chosen as $\alpha = 0.5$ for logit-based distillation and $\beta = 0.1$ for feature-level alignment. An optional additional input channel encoding delta edema information was enabled, resulting in 6 input channels instead of 5 standard MRI-derived channels.

All experiments were executed using Google Colab with NVIDIA GPU acceleration. Mixed-precision training (Automatic Mixed Precision, AMP) was employed to reduce memory consumption and accelerate training.

For each dataset, the samples were randomly split into 80% for training and 20% for held-out validation following the same protocol for both BraTS 2020 and BraTS 2021. The split configuration was fixed using the predefined split file (splits_fixed.json) with FOLD=0 to ensure reproducibility. The split was performed at the subject level (patient-wise) to avoid data leakage between training and validation sets. During inference, a simple post-processing step was applied to reduce spurious predictions by removing small isolated 3D components using connected-component filtering.

2.8. Evaluation Metrics

Model performance is assessed using three main indicators. Dice score is used to measure the overlap agreement between predictions and ground truth. IoU is used as a stricter alternative overlap measure. In addition to accuracy, efficiency is calculated using GFLOPs to evaluate inference computational load so that lightweight models can be objectively assessed. IoU AVG denotes the mean IoU averaged across the four predicted classes (BG, NCR/NET, ED, and ET).

3. RESULTS AND DISCUSSION

Performance evaluation was conducted on the BraTS 2020 and BraTS 2021 datasets using the standard tumor sub-regions, namely the BG, NCR/NET, ED, and ET classes. Additionally, the BraTS region metrics WT and TC are reported by aggregating the predicted sub-regions (WT = NCR/NET ∪ ED ∪ ET; TC = NCR/NET ∪ ET). The primary metrics employed were the Dice Similarity Coefficient (Dice) and the Intersection over Union (IoU) score, which are widely used in the brain tumor segmentation literature to assess overlap accuracy and boundary precision.

3.1. Experimental Results on the BraTS 2021 Dataset

The official BraTS 2021 training dataset was internally split into 80% for training and 20% for held-out validation (Section 2.7). The model was evaluated on the held-out validation set using BG, NCR/NET, ED, and ET classes. Based on validation results, the per-class Dice scores on BraTS 2021 are as follows: BG = 0.9706; NCR/NET = 0.3028; ED = 0.4938; ET = 0.4936. Meanwhile, the per-class IoU values are BG = 0.9613; NCR/NET = 0.2802; ED = 0.4575; ET = 0.4586. Aggregate Dice scores for ET, TC, and WT regions are 0.5304; 0.5382; and 0.6538, respectively.

Table 1. Segmentation evaluation results on BraTS 2021

Model Variant	BraTS 2021					
	Dice BG(%)	Dice WT(%)	Dice TC(%)	Dice ET(%)	IoU AVG(%)	GFLOPs
Student baseline (U-Net only)	96.08	63.78	51.46	52.06	51.73	55.65
Student + Transformer (no KD)	96.43	64.67	52.83	52.48	52.86	55.21
Student + KD (no Transformer)	96.15	64.18	52.62	52.39	52.76	54.65
Proposed (Student + Transformer + KD)	97.06	65.38	53.82	53.04	53.94	54.38

Table 1 presents the ablation study results on the BraTS 2021 dataset, comparing different student model configurations in terms of segmentation performance and computational complexity. The evaluation includes Dice scores for Background (BG), Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET), along with the average IoU and inference complexity measured in GFLOPs.

The baseline student model using a standard 3D U-Net achieves a Dice score of 0.6378 for WT, 0.5146 for TC, and 0.5206 for ET, with a computational cost of 55.65 GFLOPs. Introducing the lightweight Transformer module without knowledge distillation improves segmentation accuracy across all tumor regions, increasing the Dice score to 0.6467 (WT), 0.5283 (TC), and 0.5248 (ET), while slightly reducing GFLOPs to 55.21. This indicates that incorporating global contextual modeling enhances region-level segmentation performance with minimal computational overhead.

When knowledge distillation is applied without the Transformer module, performance also improves compared to the baseline, achieving Dice scores of 0.6418 (WT), 0.5262 (TC), and 0.5239 (ET), with reduced computational complexity (54.65 GFLOPs). This suggests that transferring knowledge from the teacher model enhances the representation capability of the lightweight student network without increasing inference cost.

The proposed model demonstrates the best trade-off between segmentation accuracy and computational efficiency among the evaluated student variants. It obtains Dice scores of 0.6538 for WT, 0.5382 for TC, and 0.5304 for ET, along with the highest average IoU (0.5394) and the lowest computational complexity (54.38 GFLOPs). These results demonstrate that combining contextual modeling and distillation provides complementary benefits, leading to improved segmentation accuracy while simultaneously reducing inference complexity.

The ablation study confirms that each component contributes positively to performance, and their integration yields the most balanced trade-off between segmentation accuracy and computational efficiency. The precision-recall performance of the proposed model across the four segmentation classes is presented in Figure 2.

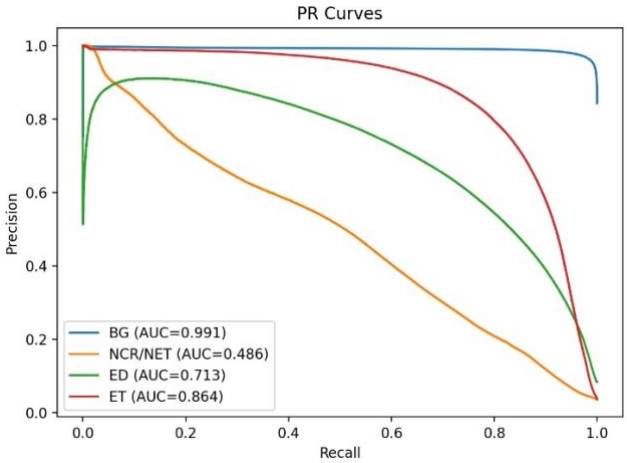

Fig. 2. PR Curve on BraTS 2021

The Precision Recall curve shows prediction quality for each class. BG exhibits the most stable performance with a high curve area, while the NCR/NET class shows a precision drop at high recall, indicating model difficulty on minority classes. The receiver operating characteristic (ROC) curves for all segmentation classes are shown in Figure 3.

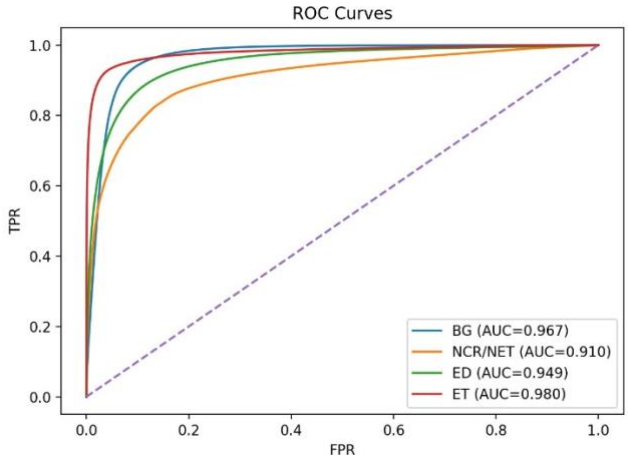

Fig. 3. ROC Curve on BraTS 2021

The ROC curve illustrates the relationship between True Positive Rate (TPR) and False Positive Rate (FPR) for each class. High AUC values for BG and ET indicate strong discrimination capability, whereas the NCR/NET class shows greater challenges. The learning behavior of the proposed model during training is illustrated in Figure 4.

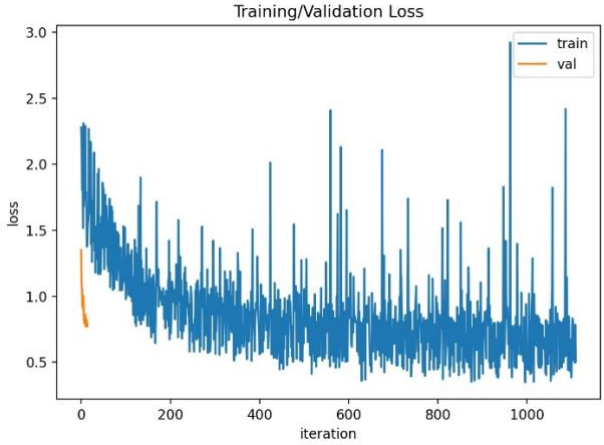

Fig. 4. Training and Validation Loss Graphs on BraTS 2021

The graph shows loss trends during training iterations. Training loss generally decreases but still exhibits fluctuations, while validation loss shows early stability, indicating that the learning process proceeds adequately despite batch variations. Representative segmentation results produced by the proposed model are presented in Figure 5.

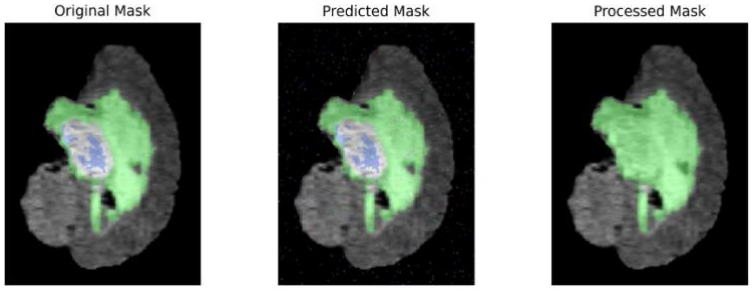

Fig. 5. Visualization of the segmentation mask in BraTS 2021

Visualization shows comparisons between ground truth, model predictions, and post-processing results. The main tumor areas follow the ground truth structure despite prediction noise before processing. Representative best- and worst-case segmentation results are presented in Figure 6.

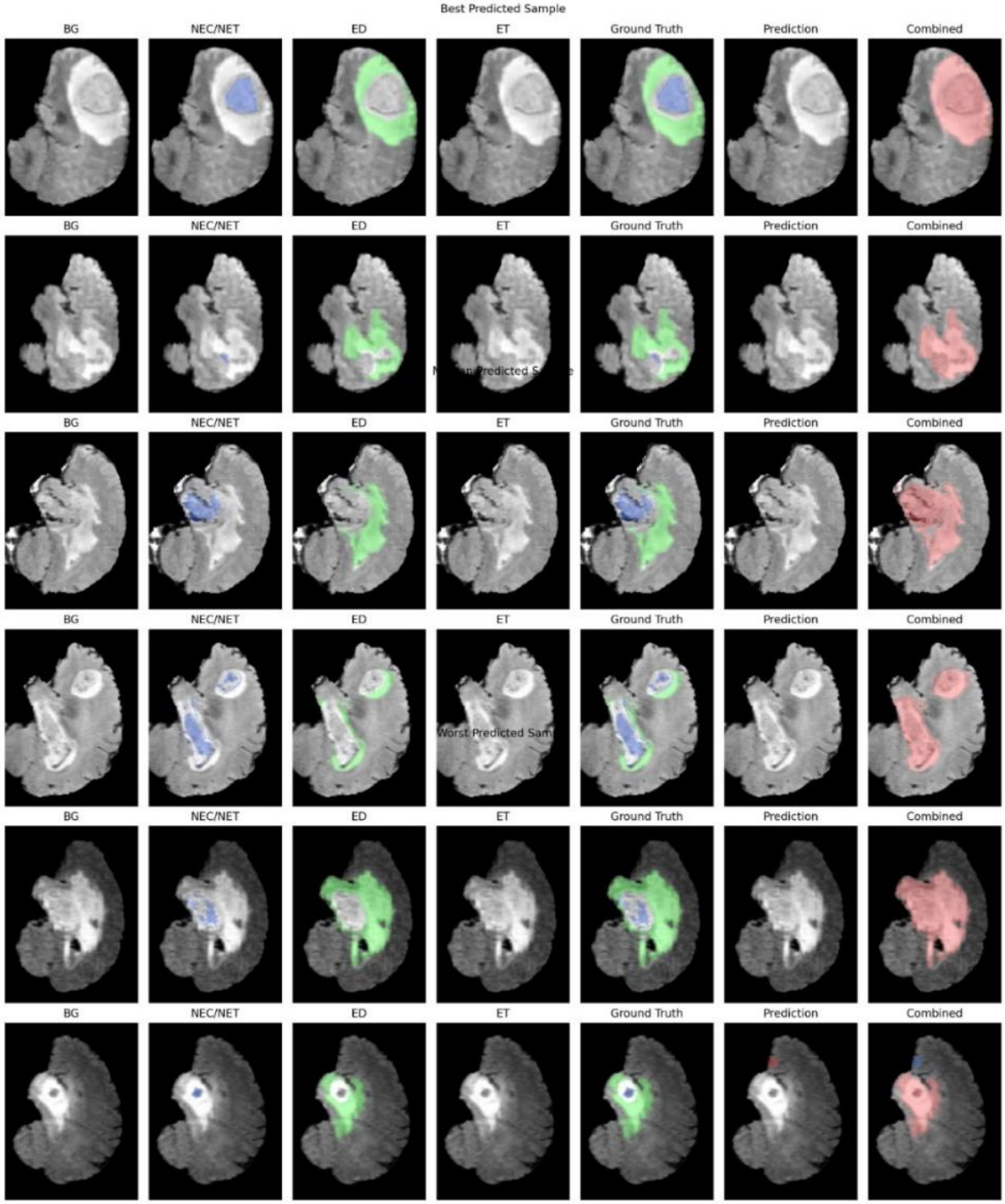

Fig. 6. Visualization of The Best and Worst Samples in BraTS 2021

The figure presents examples of the best and worst predictions for each tumor class along with combined final results. This comparison shows model stability on large tumors and challenges on small sub-regions. The overall classification performance is summarized by the confusion matrix shown in Figure 7.

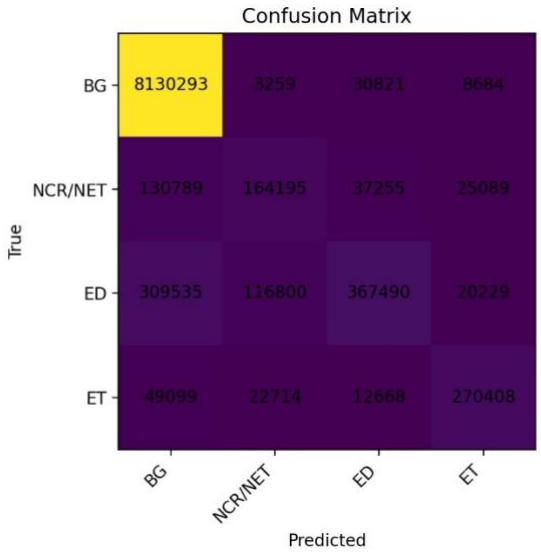

Fig. 7. BraTS 2021 Confusion Matrix

The confusion matrix illustrates the distribution of correct and incorrect predictions across four classes. The BG class has the highest correct predictions, while dominant errors occur between NCR/NET and ED due to intensity similarity and ambiguous boundaries [29].

3.2. Experimental Results on the BraTS 2020 Dataset

For BraTS 2020, the official training data were internally split into 80% for training and 20% for held-out validation (Section 2.7). The model was evaluated on the held-out validation set using BG, NCR/NET, ED, and ET classes. The evaluation on the BraTS 2020 dataset yielded the following per-class Dice scores: BG = 0.9520, NCR/NET = 0.2166, ED = 0.3165, and ET = 0.4212. The corresponding per-class IoU scores were BG = 0.9343, NCR/NET = 0.1967, ED = 0.2502, and ET = 0.3944. The aggregated Dice scores for the ET, TC, and WT regions on BraTS 2020 were 0.4696, 0.4996, and 0.5332, respectively.

Table 2. Segmentation evaluation results on BraTS 2020

Model Variant	BraTS 2020					
	Dice BG(%)	Dice WT(%)	Dice TC(%)	Dice ET(%)	IoU AVG(%)	GFLOPs
Student baseline (U-Net only)	94.15	51.24	47.28	45.21	43.36	55.65
Student + Transformer (no KD)	94.62	52.17	48.63	45.88	44.19	55.21
Student + KD (no Transformer)	94.48	51.96	48.32	45.67	43.94	54.65
Proposed (Student + Transformer + KD)	95.20	53.32	49.96	46.96	44.39	54.38

Table 2 presents the ablation study results on the BraTS 2020 dataset, reporting Dice scores for Background (BG), Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET), along with the average IoU and computational complexity measured in GFLOPs.

The baseline 3D U-Net achieves Dice scores of 0.5124 (WT), 0.4728 (TC), and 0.4521 (ET), with a computational cost of 55.65 GFLOPs. Incorporating the lightweight Transformer without knowledge distillation improves segmentation performance across all tumor regions, particularly WT and TC, indicating that global contextual modeling contributes to better structural delineation of tumor subregions. This configuration also slightly reduces the computational complexity to 55.21 GFLOPs.

Applying knowledge distillation without the Transformer further enhances feature representation while reducing the inference cost to 54.65 GFLOPs. The Dice scores increase compared with the baseline, suggesting that knowledge transfer from a high-capacity teacher model strengthens the discriminative capability of the lightweight student network.

The proposed model, which integrates both the lightweight Transformer and knowledge distillation, achieves the highest segmentation performance across all evaluated tumor regions, with Dice scores of 0.5332 (WT), 0.4996 (TC), and 0.4696 (ET), while maintaining the lowest computational complexity of 54.38 GFLOPs. The improvement, although moderate, demonstrates that contextual modeling and knowledge distillation provide complementary benefits in enhancing tumor boundary delineation and inter-class discrimination.

The performance on BraTS 2020 is lower than that on BraTS 2021. This discrepancy may be attributed to inter-dataset variability, differences in tumor morphology, and case complexity. The consistent improvement across all ablation variants confirms the robustness of the proposed approach and its ability to achieve a balanced trade-off between segmentation accuracy and computational efficiency.

The performance trend across all ablation variants consistently shows incremental improvements when either the Transformer module or knowledge distillation is introduced, and the combined configuration yields the most stable results across the WT, TC, and ET regions. Although repeated multi-run statistical validation was not conducted, the consistent improvement pattern across two independent datasets (BraTS 2020 and BraTS 2021) supports the reliability of the proposed framework. Figure 8 presents the Precision–Recall curves of the proposed model on the BraTS 2020 dataset.

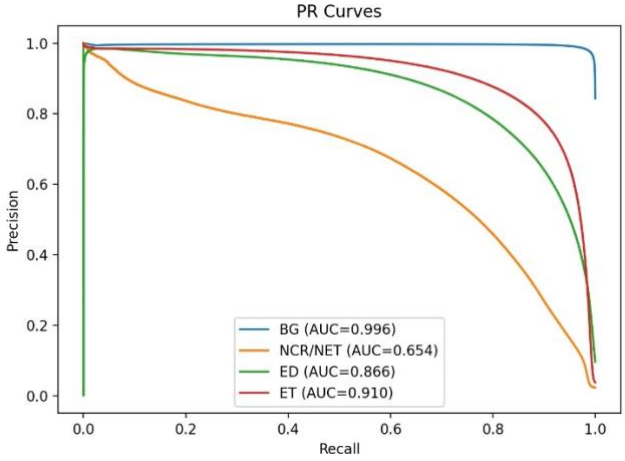

Fig. 8. PR Curve on BraTS 2020

The Precision–Recall curve on BraTS 2020 shows high performance for the BG class, while the NCR/NET class exhibits a lower curve area, indicating difficulty in recognizing relatively small tumor core regions. The receiver operating characteristic (ROC) curves for the BraTS 2020 dataset are shown in Figure 9.

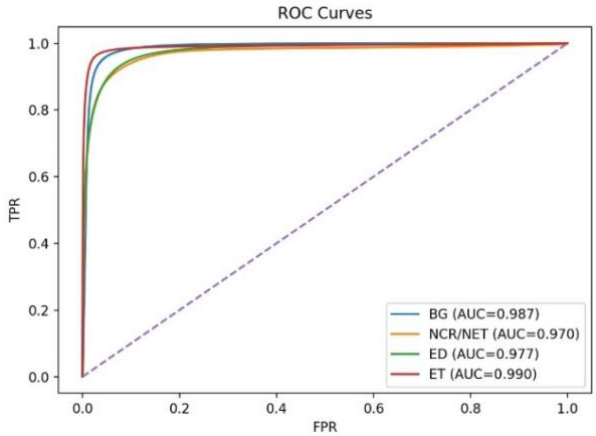

Fig. 9. ROC Curve on BraTS 2020

The ROC curve illustrates the model's ability to discriminate each class. AUC values for ET are relatively high compared to other classes, whereas NCR/NET still shows a performance gap. The training and validation loss curves on the BraTS 2020 dataset are illustrated in Figure 10.

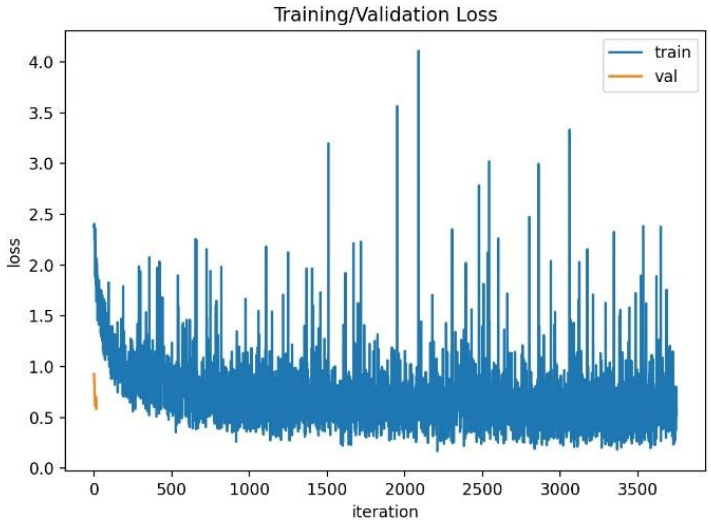

Fig. 10. Training and validation loss graphs on BraTS 2020

The graph shows loss changes during training and validation on BraTS 2020. Loss fluctuations may arise due to volumetric tumor variation and class imbalance distribution. The confusion matrix summarizing the segmentation performance on the BraTS 2020 dataset is presented in Figure 11.

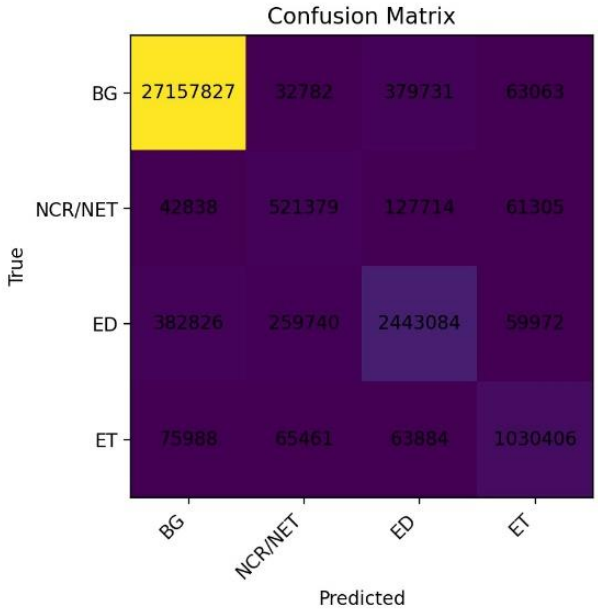

Fig. 11. BraTS 2020 confusion matrix

The confusion matrix shows dominance of correct predictions for the BG class and significant misclassification errors in ED and NCR/NET classes. This is consistent with class imbalance challenges and intensity feature similarity among sub-regions.

3.3. Comparison with Previous Studies

To position the proposed approach within the landscape of recent brain tumor segmentation methods, we compare our results with representative prior studies reported on the BraTS 2020 and BraTS 2021 benchmarks. The comparison focuses on Dice performance for WT, TC, and ET, as well as computational complexity (GFLOPs) when available. For studies that did not report GFLOPs or parameter counts, the corresponding entries are left blank to avoid misleading assumptions.

Table 3. BraTS 2020 comparison

Study	Dice WT(%)	Dice TC(%)	Dice(ET)	Parameters	GFLOPs
Isensee et al.	88.95	85.06	82.03	181.516M	-
Tang et al.	89.29	78.97	70.30	-	-
Ballestar et al.	84.21	75.03	61.75	-	-
Wang et al.	90.09	81.73	78.73	32.99 M	333
Messaoudi et al.	80.68	75.20	69.59	-	-
Raza et al.	86.60	83.57	80.04	30.47 M	374.04
Zhu et al.	90.22	89.20	82.48	-	-
Proposed (Student + Transformer + KD)	53.32	49.96	46.96	-	54.38

Table 4. BraTS 2021 comparison

Study	Dice WT(%)	Dice TC(%)	Dice(ET)	Parameters	GFLOPs
Peiris et al.	90.77	85.39	81.38	-	-
Akbar et al.	89.07	80.73	78.02	-	-
Jia et al.	92.53	87.96	84.80	17.91M	449.79
Li et al.	90.18	81.61	76.89	-	-
Ma et al.	92.59	87.86	82.17	-	-
Hatamizadeh et al.	92.60	88.50	85.80	61.98 M	394.84
Roth et al.	90.60	83.50	79.20	-	-
Proposed (Student + Transformer + KD)	65.38	53.82	53.04	-	54.38

From a deployment perspective, maintaining a computational complexity of 54.38 GFLOPs makes the model significantly more feasible for real-world implementation compared to heavy Transformer-based models that require substantially higher computational budgets. This efficiency-oriented design is particularly relevant for institutions with limited GPU infrastructure. Differences in reported Dice values may also be influenced by variations in data splits, preprocessing pipelines, and evaluation protocols [11].

Although the absolute Dice scores of the proposed method are lower than large-scale Transformer-based architectures reported in prior studies, it is important to note that many of these methods rely on substantially higher computational budgets, with GFLOPs exceeding 300–400 and parameter counts above tens of millions. The proposed framework is intentionally designed under strict computational constraints to emphasize deployability rather than benchmark maximization. Therefore, the comparison should be interpreted in the context of accuracy–efficiency trade-offs rather than absolute performance dominance. The proposed model reduces computational complexity by approximately six to eight times compared to several Transformer-heavy architectures reported in Tables 3 and 4. The proposed method should be interpreted as an efficiency-constrained segmentation framework rather than a benchmark-maximizing architecture, emphasizing deployability under strict computational budgets.

Experimental results show that the 3D U-Net model with a lightweight Transformer is able to perform multi-class segmentation on BraTS 2020 and BraTS 2021 with relatively consistent performance on dominant classes, particularly background [33]. minority classes such as NCR/NET remain a major challenge with relatively low Dice scores, which aligns with the class imbalance problem in brain tumor segmentation [34].

The application of knowledge distillation is expected to contribute to improved student model generalization through the transfer of smoother decision distributions from the teacher model [35]. Thus, the student model can potentially acquire richer feature representations without increasing inference complexity [36]. This approach is particularly relevant when models need to be deployed in clinical systems that prioritize efficiency and limited resources [37].

The complementary effect observed between the lightweight Transformer and knowledge distillation suggests that contextual modeling and representation smoothing address different aspects of segmentation learning. While the Transformer enhances spatial dependency modeling, distillation improves class boundary calibration by transferring softer decision distributions from the teacher network. The combined effect results in improved Dice scores without increasing inference complexity.

From a global context perspective, the use of a lightweight Transformer at the bottleneck strengthens the model's ability to understand long-range spatial dependencies, which are often suboptimal in pure CNNs [10]. Nevertheless, this strategy must still consider the trade-off between global context enhancement and computational efficiency. In this study, stable GFLOPs across both datasets indicate that the lightweight Transformer module does not cause extreme increases in computational burden [38].

Evaluation of PR and ROC curves indicates that the BG class exhibits near-ideal performance, whereas NCR/NET and ED classes are more difficult to learn due to ambiguous boundaries and complex tumor morphology characteristics [39]. The confusion matrix shows a tendency toward misclassification in non-dominant tumor areas, suggesting that future development can focus on more adaptive loss strategies and stronger distillation for minority classes [40].

The morphological similarity and intensity overlap between NCR/NET and edema (ED) regions introduce additional segmentation difficulty. The boundaries between these two sub-regions are often indistinct in MRI modalities, leading to confusion patterns that are also reflected in the confusion matrix. Such boundary ambiguity limits the discriminative capacity of standard convolutional features and challenges accurate sub-region delineation.

Another contributing factor is the fixed input patch size of 64^3 used during training and inference. While this patch size offers a favorable trade-off between computational efficiency and spatial context, it may restrict the availability of broader contextual cues necessary for precisely identifying small and irregular tumor core regions. Increasing patch resolution could potentially improve NCR/NET representation, but at the cost of substantially higher GFLOPs and memory consumption.

Despite these challenges, the ablation study indicates that knowledge distillation provides measurable improvements for minority tumor regions. Compared to the student baseline, the introduction of knowledge distillation increases Dice scores for TC and ET, suggesting enhanced boundary calibration and smoother class probability distributions. The transfer of softened logits from the teacher model appears to reduce overconfident predictions on ambiguous voxels, thereby improving inter-class discrimination without increasing inference complexity.

Further enhancement of minority class segmentation may require more specialized strategies. Future work could explore boundary-aware loss functions, such as focal-Tversky or region-sensitive Dice variants, to better penalize false negatives in small tumor cores. In addition, ROI-focused sampling or adaptive patch selection strategies may help increase exposure to underrepresented regions during training. Finally, boundary-guided or feature-level distillation mechanisms could be investigated to strengthen knowledge transfer specifically for challenging sub-regions such as NCR/NET.

The persistent performance gap in NCR/NET can be attributed to severe voxel imbalance and boundary ambiguity with surrounding edema regions. This indicates that additional class-sensitive optimization strategies, such as boundary-aware losses or region-focused sampling, may further enhance minority class performance in future work.

3.4. Limitations

Despite the promising computational efficiency demonstrated by the proposed framework, several limitations should be acknowledged. Segmentation performance on minority tumor subregions, particularly the necrotic and non-enhancing tumor core (NCR/NET), remained moderate, indicating the need for more effective class imbalance mitigation strategies. Performance variability was not assessed through repeated multi-run experiments, which limits the evaluation of model stability across different training runs. The teacher–student configuration was evaluated using a single teacher architecture, and additional investigations involving teacher models with different capacities may provide further insights into knowledge transfer effectiveness.

4. CONCLUSION

This study proposed a knowledge distillation framework for brain tumor segmentation based on a 3D U-Net enhanced with a lightweight Transformer module. The proposed method was evaluated on the BraTS 2020 and BraTS 2021 benchmark datasets. On BraTS 2021, the model achieved Dice scores of 0.6538 for the whole tumor (WT), 0.5382 for the tumor core (TC), and 0.5304 for the enhancing tumor (ET), while maintaining a computational cost of approximately 54.38 GFLOPs for input patches of $64 \times 64 \times 64$ voxels. These findings demonstrate that the proposed framework provides a balanced trade-off between segmentation performance and computational efficiency, supporting its potential applicability in resource-constrained clinical environments. Future research should focus on improving the segmentation of minority tumor subregions and evaluating model robustness through multi-fold cross-validation.

REFERENCES

[1] Y. R. Park *et al.*, "Expedited safety reporting through an alert system for clinical trial management at an academic medical center: Retrospective design study," *JMIR Medical Informatics*, vol. 8, no. 2, 2020, doi: 10.2196/14379.

[2] X. Feng, H. Bai, D. Kim, G. Maragkos, J. Machaj, and R. Kellogg, "Brain Tumor Segmentation with Patch-Based 3D Attention UNet from Multi-parametric MRI," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, vol. 12963, A. Crimi and S. Bakas, Eds., in Lecture Notes in Computer Science, vol. 12963. , Cham: Springer International Publishing, 2022, pp. 90–96. doi: 10.1007/978-3-031-09002-8_8.

[3] S. Baig *et al.*, "Segmentation of pre- and posttreatment diffuse glioma tissue subregions including resection cavities," *Neuro-Oncology Advances*, vol. 6, no. 1, p. vdae140, Jan. 2024, doi: 10.1093/noajnl/vdae140.

[4] F. Yepes-Calderon and J. Gordon McComb, "Manual Segmentation Errors in Medical Imaging. Proposing a Reliable Gold Standard," in *Applied Informatics*, vol. 1051, H. Florez, M. Leon, J. M. Diaz-Nafria, and S. Belli, Eds., in Communications in Computer and Information Science, vol. 1051. , Cham: Springer International Publishing, 2019, pp. 230–241. doi: 10.1007/978-3-030-32475-9_17.

[5] K. A. Wahid *et al.*, "Large scale crowdsourced radiotherapy segmentations across a variety of cancer anatomic sites," *Sci Data*, vol. 10, no. 1, p. 161, Mar. 2023, doi: 10.1038/s41597-023-02062-w.

[6] T. Stein *et al.*, "Efficient Web-Based Review for Automatic Segmentation of Volumetric DICOM Images," in *Bildverarbeitung für die Medizin 2019*, H. Handels, T. M. Deserno, A. Maier, K. H. Maier-Hein, C. Palm, and T. Tolxdorff, Eds., in Informatik aktuell. , Wiesbaden: Springer Fachmedien Wiesbaden, 2019, pp. 158–163. doi: 10.1007/978-3-658-25326-4_33.

[7] N. Rasool and J. Iqbal Bhat, "Unveiling the Complexity of Medical Imaging through Deep Learning Approaches," *Chaos Theory and Applications*, vol. 5, no. 4, pp. 267–280, Dec. 2023, doi: 10.51537/chaos.1326790.

[8] B. Rajesh, M. Satyanarayana, and P. S. Vasarao, "Automated Image Segmentation for Complex Scenes Using U-Net Architecture," *IJRIAS*, vol. X, no. VII, pp. 77–83, 2025, doi: 10.51584/IJRIAS.2025.100700006.

[9] A. A. Novikov, D. Major, M. Wimmer, D. Lenis, and K. Buhler, "Deep Sequential Segmentation of Organs in Volumetric Medical Scans," *IEEE Trans. Med. Imaging*, vol. 38, no. 5, pp. 1207–1215, May 2019, doi: 10.1109/TMI.2018.2881678.

[10] L. Zhang, M. Li, P. Zhang, and P. Liu, "EfficientSegNet: Lightweight Semantic Segmentation with Multi-Scale Feature Fusion and Boundary Enhancement," *Sensors*, vol. 25, no. 19, p. 5934, Sep. 2025, doi: 10.3390/s25195934.

[11] M. Poddar, J. S. Marwaha, W. Yuan, S. Romero-Brufau, and G. A. Brat, "An operational guide to translational clinical machine learning in academic medical centers," *npj Digit. Med.*, vol. 7, no. 1, p. 129, May 2024, doi: 10.1038/s41746-024-01094-9.

[12] E. Y. Zhu, C. Zhao, H. Yang, J. Li, Y. Wu, and R. Ding, "A Comprehensive Review of Knowledge Distillation-Methods, Applications, and Future Directions," *IJIRCST*, vol. 12, no. 3, pp. 106–112, May 2024, doi: 10.55524/ijircst.2024.12.3.17.

[13] X. Qi, M. Hou, and P. Gao, "Dual-View Brain Tumor MRI Segmentation with Embedding Coordinate Attention Mechanism," in *2023 IEEE 5th International Conference on Power, Intelligent Computing and Systems (ICPICS)*, Shenyang, China: IEEE, Jul. 2023, pp. 140–146. doi: 10.1109/ICPICS58376.2023.10235355.

[14] J. Wei, J. Chen, Y. Wang, H. Luo, and W. Li, "Improved deep learning image classification algorithm based on Swin Transformer V2," *PeerJ Computer Science*, vol. 9, p. e1665, Oct. 2023, doi: 10.7717/peerj-cs.1665.

[15] H. Tabani, A. Balasubramaniam, S. Marzban, E. Arani, and B. Zonooz, "Improving the Efficiency of Transformers for Resource-Constrained Devices," Jun. 30, 2021, *arXiv*: arXiv:2106.16006. doi: 10.48550/arXiv.2106.16006.

[16] L. Zhou, J. Zhao, and R. Shi, "Classifier - Centric Knowledge Distillation Based on Class Activation Weight Sharing," in *2025 6th International Conference on Computer Vision, Image and Deep Learning (CVIDL)*, Ningbo, China: IEEE, May 2025, pp. 1205–1212. doi: 10.1109/CVIDL65390.2025.11085768.

[17] Z. Liu, C. Zhang, H. Peng, Q. Xu, and Y. Gao, "Drug distribution management system based on IoT," *KSII Transactions on Internet and Information Systems*, vol. 16, no. 2, pp. 424–444, 2022, doi: 10.3837/tiis.2022.02.004.

[18] A. Galkin, Yu. Davidich, and H. Samchuk, "Development Of Mathematical Models For Evaluating Demand Parameter Elasticity In Retail Networks Under Consumer-Driven Logistics," *MEC*, vol. 4, no. 185, pp. 262–266, Sep. 2024, doi: 10.33042/2522-1809-2024-4-185-262-266.

[19] K. Alrfou and T. Zhao, "GCtx-UNet: Efficient Network for Medical Image Segmentation," 2024, *arXiv*. doi: 10.48550/ARXIV.2406.05891.

[20] L. Zhao, X. Qian, Y. Guo, J. Song, J. Hou, and J. Gong, "MSKD: Structured knowledge distillation for efficient medical image segmentation," *Computers in Biology and Medicine*, vol. 164, p. 107284, Sep. 2023, doi: 10.1016/j.compbiomed.2023.107284.

[21] M. M. Danesh Pajouh, "Efficient Brain Tumor Segmentation Using a Dual-Decoder 3D U-Net with Attention Gates (DDUNet)," Apr. 14, 2025, *Open Science Framework*. doi: 10.31219/osf.io/ws5xf_v1.

[22] Anonymous, "Data augmentation." Zenodo, Jul. 07, 2023. doi: 10.5281/ZENODO.8123405.

[23] J. Panic, A. Defeudis, G. Balestra, V. Giannini, and S. Rosati, "Normalization Strategies in Multi-Center Radiomics Abdominal MRI: Systematic Review and Meta-Analyses," *IEEE Open J. Eng. Med. Biol.*, vol. 4, pp. 67–76, 2023, doi: 10.1109/OJEMB.2023.3271455.

[24] S. Ali, N. Ali, F. Mohamed, T. Kamal, and M. Salih, "Region segmentation for lung cancer CT image using 3D U-Net model," *Turkish Journal of Internal Medicine*, vol. 7, no. 3, pp. 98–108, Jul. 2025, doi: 10.46310/tjim.1580929.

[25] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Global Context Networks," Dec. 24, 2020, *arXiv*: arXiv:2012.13375. doi: 10.48550/arXiv.2012.13375.

[26] B. Choudhary, D. Singh, and D. S. Sisodia, "Hierarchical Feature and Attention-Based Distillation to Improve Student Model Performance," in *2025 IEEE International Conference on Computer, Electronics, Electrical*

Engineering & their Applications (IC2E3), Srinagar Garhwal, India: IEEE, May 2025, pp. 1–6. doi: 10.1109/IC2E365635.2025.11167035.

[27] D. Chen, J.-P. Mei, C. Wang, Y. Feng, and C. Chen, "Online Knowledge Distillation with Diverse Peers," *AAAI*, vol. 34, no. 04, pp. 3430–3437, Apr. 2020, doi: 10.1609/aaai.v34i04.5746.

[28] D. Liu, M. Kan, S. Shan, and X. Chen, "Function-Consistent Feature Distillation," 2023, *arXiv*. doi: 10.48550/ARXIV.2304.11832.

[29] A. Kheterpal and K. Singh Gill, "The EfficientNetB5 Solution to Multi-Class Bone Marrow Classification Challenges," in *2024 12th International Conference on Internet of Everything, Microwave, Embedded, Communication and Networks (IEMECON)*, Jaipur, India: IEEE, Oct. 2024, pp. 1–5. doi: 10.1109/IEMECON62401.2024.10846120.

[30] A. Tragakis *et al.*, "GLFNET: Global-Local (frequency) Filter Networks for efficient medical image segmentation," 2024, doi: 10.48550/ARXIV.2403.00396.

[31] U. Baid *et al.*, "The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification," Sep. 12, 2021, *arXiv*: arXiv:2107.02314. doi: 10.48550/arXiv.2107.02314.

[32] A. Burrello, M. Risso, B. A. Motetti, E. Macii, L. Benini, and D. J. Pagliari, "Enhancing Neural Architecture Search With Multiple Hardware Constraints for Deep Learning Model Deployment on Tiny IoT Devices," *IEEE Trans. Emerg. Topics Comput.*, vol. 12, no. 3, pp. 780–794, Jul. 2024, doi: 10.1109/TETC.2023.3322033.

[33] N. E. Varghese, A. John, U. D. A. C, and M. J. Pillai, "Transformer-augmented lightweight U-Net (UAAC-Net) for accurate MRI brain tumor segmentation," *Neurological Research*, vol. 47, no. 12, pp. 1212–1227, Dec. 2025, doi: 10.1080/01616412.2025.2517312.

[34] S. Banerjee and S. Mitra, "Novel Volumetric Sub-region Segmentation in Brain Tumors," *Front. Comput. Neurosci.*, vol. 14, p. 3, Jan. 2020, doi: 10.3389/fncom.2020.00003.

[35] X. Gu, R. Jin, and M. Li, "Progressively Relaxed Knowledge Distillation," in *2024 International Joint Conference on Neural Networks (IJCNN)*, Yokohama, Japan: IEEE, Jun. 2024, pp. 1–8. doi: 10.1109/IJCNN60899.2024.10650544.

[36] S. Li *et al.*, "Distilling a Powerful Student Model via Online Knowledge Distillation," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 34, no. 11, pp. 8743–8752, Nov. 2023, doi: 10.1109/TNNLS.2022.3152732.

[37] J. Zhu, Z. Tang, P. Ma, Z. Liang, and C. Wang, "DLKUNET : A Lightweight and Efficient Network With Depthwise Large Kernel for Medical Image Segmentation," *Int J Imaging Syst Tech*, vol. 35, no. 1, p. e70035, Jan. 2025, doi: 10.1002/ima.70035.

[38] D.-H. Le, "Integrating multiple microRNA functional similarity networks for improved disease–microRNA association prediction," *Biology Methods and Protocols*, vol. 10, no. 1, p. bpaf065, Jan. 2025, doi: 10.1093/biomethods/bpaf065.

[39] E. Radiya-Dixit, D. Zhu, and A. H. Beck, "Automated Classification of Benign and Malignant Proliferative Breast Lesions," *Sci Rep*, vol. 7, no. 1, p. 9900, Aug. 2017, doi: 10.1038/s41598-017-10324-y.

[40] Prathipati Silpa Chaitanya and Susanta Kumar Satpathy, "Advancing Brain Tumour Detection and Classification: Knowledge Distilled ResNeXt Model for Multi-Class MRI Analysis," *IJCESEN*, vol. 10, no. 4, Dec. 2024, doi: 10.22399/ijcesen.730.