Resource-Conscious 3D U-Net Models for Brain Tumor Segmentation: An Ablation Study

Evans Nyeduala Siaw
Matrikel-Nr: 4745660
M.Sc. Biomedical Engineering

Yilin Wu
Matrikel-Nr: 4742492
M.Sc. Scientific Computing

Abstract

Accurate brain tumor segmentation is critical for diagnosis and treatment planning, but current state-of-the-art methods like nnU-Net are computationally demanding. This work explores lightweight 3D U-Net variants tailored for resource-limited settings. Using the BraTS 2020 dataset, we compared three preprocessing pipelines and conducted ablations studies on both baseline and optimized models, by incorporating residual blocks, attention gates, deep supervision, normalization refinements, dropout tuning, and class-weighted loss. The best model achieved Dice scores of 0.880 (WT), 0.812 (TC), and 0.691 (ET), showing consistent improvements over the baseline while remaining computationally efficient. These results highlight the trade-off between efficiency and accuracy, and point to practical strategies for segmentation under limited GPU resources.

1. Introduction

Brain tumors are among the most severe neurological diseases, often associated with high morbidity and mortality. Accurate assessment of tumor extent is essential for clinical tasks such as surgical planning, radiotherapy design, and monitoring treatment response. Magnetic resonance imaging (MRI) is the preferred modality for brain tumor imaging, as it provides high-resolution anatomical and functional information across multiple contrasts (e.g., T1, T2, FLAIR, T1ce). Despite its central role in neuro-oncology, tumor delineation in MRI is typically performed manually by expert radiologists. This process is labor-intensive, time-consuming, and prone to significant inter-observer variability. Automated segmentation methods are therefore highly desirable to ensure reproducibility, efficiency, and consistency in clinical workflows.

Deep learning, particularly convolutional neural networks (CNNs), has emerged as the state-of-the-art for medical image segmentation. The U-Net architecture and its 3D variants have demonstrated remarkable performance on brain tumor segmentation tasks, enabling accurate localization of complex tumor subregions such as whole tumor (WT), tumor core (TC), and enhancing tumor (ET). Benchmark competitions such as the Brain Tumor Segmentation (BraTS) challenge have further accelerated progress by providing large, annotated datasets for rigorous evaluation. However, many high-performing models require significant computational resources, which limits their applicability in real-world clinical settings with constrained hardware. This motivates the development of lightweight yet accurate segmentation approaches that balance performance with efficiency. In this study, we investigate a 3D U-Net framework for brain tumor segmentation under limited GPU resources, analyzing preprocessing pipelines and network design choices through systematic ablation experiments.

2. Related Work

The U-Net architecture, introduced by Ronneberger et al. [1], is a cornerstone of biomedical image segmentation. Its encoder–decoder design with skip connections preserves spatial detail while enabling deep feature extraction. Extensions to 3D U-Net variants by Çiçek et al. [2] further improved performance on volumetric data such as MRI, which is critical for brain tumor analysis.

The Brain Tumor Segmentation (BraTS) challenge has provided a standardized benchmark for multimodal MRI segmentation, focusing on clinically relevant subregions: whole tumor (WT), tumor core (TC), and enhancing tumor (ET). Over the years, increasingly sophisticated methods have been proposed, including cascaded CNNs, attention-based models, and ensemble approaches. A milestone is the nnU-Net framework by Isensee et al. [3], winner of BraTS 2020. Its automated adaptation of preprocessing, architecture, training, and postprocessing achieves strong performance across datasets. However, nnU-Net's comprehensive design is computationally expensive, often requiring large-memory GPUs, long runtimes, and ensembles. Scaling it down typically reduces segmentation accuracy.

To address the challenge of limited-resource environments, several works have explored lightweight alternatives. These include: reducing input patch sizes to lower memory usage [4], incorporating residual connections for improved gradient flow [5], introducing dropout to enhance regularization [6], leveraging attention gates to selectively filter features [7], and employing deep supervision for stronger gradient signals [8]. While such strategies often fall short of nnU-Net's peak performance, they provide a promising balance between segmentation accuracy and computational feasibility.

Our work builds on these efforts by:

1. **Systematically comparing preprocessing pipelines (P1-P3)** to identify a training setup that balances anatomical fidelity and runtime efficiency.

2. **Implementing a baseline 3D U-Net** and introducing incremental architectural refinements (residual blocks, attention mechanisms, deep supervision, dropout, SE modules) to assess their contribution to performance.

3. **Evaluating models using BraTS-standard metrics (Dice for WT, TC, and ET)**, situating results within the broader efficiency-accuracy trade-off in brain tumor segmentation.

3. Method

We are focusing on brain tumor segmentation through a U-Net architecture, where BraTS 2020 dataset is leveraged for model training as well as evaluation. The Brain Tumor Segmentation (BraTS) 2020 dataset provides pre-operative multimodal MRI scans of gliomas, collected from 19 different institutions using diverse scanner and acquisition protocols. Each case consists of four MRI modalities, namely native T1-weighted (T1), post-contrast T1-weighted (T1Gd), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (T2-FLAIR) (see Figure 1). All scans were co-registered to the same anatomical template, resampled to an isotropic resolution of 1 mm^3, and skull-stripped to remove non-brain tissue. Manual annotations were performed by one to four raters and subsequently verified by experienced neuro-radiologists. The segmentation labels include three tumor sub-regions: enhancing tumor(ET, label 4), the peritumoral edema (ED, label 2), and the necrotic and non-enhancing tumor core (NCR/NET, label 1) [9]. (see Figure 2)

3.1. Preprocessing

All MRI volumes were preprocessed to ensure consistency across cases and facilitate model training. First, voxel intensities were clipped to the 0.5th and 99.5th percentiles [3] in order to remove extreme outliers and reduce

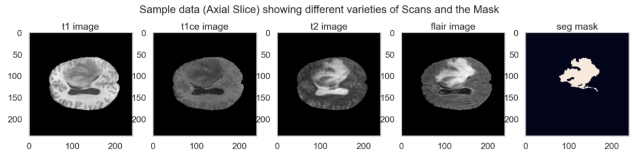

Figure 1. Representative axial slice from the BraTS2020 dataset. Columns show the four input MRI modalities (T1, T1Gd, T2, and FLAIR), alongside the corresponding expert segmentation mask.

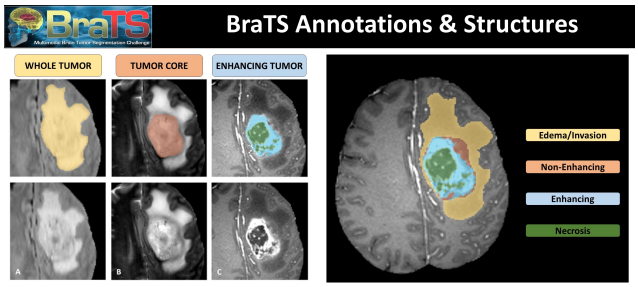

Figure 2. BraTS dataset structure, illustrating expert annotations. The segmentation highlights the clinically relevant subregions: whole tumor (WT), tumor core (TC), and enhancing tumor (ET). Source: https://www.kaggle.com/code/khaledsayedaaaaa/3d-mri-brain-tumor-segmentation-u-net-acc-99

noise. To further reduce redundant background, each volume was cropped to the nonzero bounding box of the corresponding segmentation mask, with a fixed margin to preserve contextual information. The cropped images and corresponding segmentation masks were subsequently resampled to a uniform spatial resolution of $128 \times 128 \times 128$. Image volumes were resampled by trilinear interpolation, while segmentation masks were resampled with nearest-neighbor interpolation to preserve label integrity [3]. Finally, z-score normalization was applied within the brain foreground, with normalized values clipped to the range of [-5,5].

Random 3D patches of size $96 \times 96 \times 96$ were extracted from the preprocessed volumes of $128 \times 128 \times 128$. This patch-based strategy reduces computational and memory requirements, introduced spatial variability, which mitigates overfitting by exposing the network to diverse tumor and background voxels[10]. Segmentation labels were remapped to four categories (0 = background, 1 = non-enhancing core , 2 = edema, 3 = enhancing tumor) and converted into one-hot encoding for multi-class training.

Preprocessing was implemented under three data pipelines: P1 (on-the-fly preprocessing), P2 (cached full resampled volumes with random patching at runtime), and P3 (cached random patches). The caching strategy of P2 and P3 substantially reduced I/O overhead and enabled reproducible experiments with controlled patch diversity. A systematic comparison of these pipelines is presented in Section 4, where we evaluate their trade-offs in efficiency and

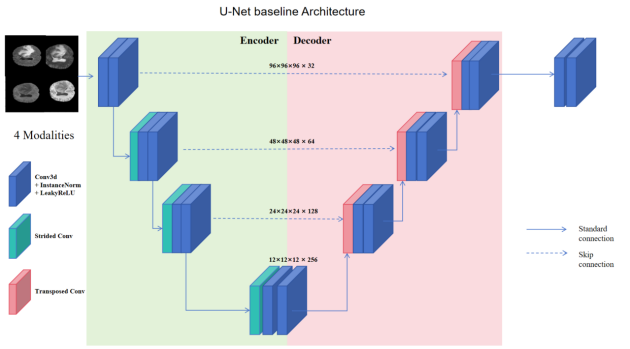

Figure 3. Baseline 3D U-Net Architecture used in our experiments. The model takes four MRI modalities as input and consists of an encoder-decoder structure with skip connections. Each convolutional block includes 3D convolution, instance normalization, and LeakyReLU activation. Downsampling is performed via strided convolutions, and upsampling via transposed convolutions.

segmentation performance.

To further improve robustness and generalization, we employed several data augmentations, including random flips along all spatial axes, 90° in-plane rotations, intensity scaling , and the addition of Gaussian noise [3].

3.2. 3D U-Net

The network architecture, illustrated in Figure 3, follows a 3D U-Net design with an encoder-decoder structure and symmetric skip connections, each consisting of three resolution steps.

In the encoder path, each resolution step contains a convolutional block with two successive 3D convolutions (kernel size $3 \times 3 \times 3$), each followed by instance normalization and a LeakyReLU activation ($\alpha = 0.01$). Downsampling is performed via a $2 \times 2 \times 2$ strided convolution, which doubles the number of feature channels at each step.

In the decoder path, upsampling is performed using a $2 \times 2 \times 2$ transposed convolution, halving the number of feature channels. The upsampled feature maps are concatenated with the corresponding encoder features via skip connections, followed by a convolutional block identical to those in the encoder.

Finally, a $1 \times 1 \times 1$ convolution maps the feature representation to the desired number of output classes.

3.3. Evaluation

3.3.1 Region-based metric

The Dice Similarity Coefficient (DSC) is a region-based metric widely used in medical image segmentation to evaluate the overlap between the predicted segmentation and the ground truth. It ranges from 0 to 1, where 1 indicates perfect overlap and 0 indicates no overlap.

$$D(P,T) = \frac{2|P \cap T|}{|P| + |T|}, \tag{1}$$

where $P \in 0, 1$ denotes the set of voxels predicted as belonging to the structure of interest, $T \in 0, 1$ denotes the ground truth voxels, $|P|$ and $|T|$ denote the number of voxels in each set, and $|P \cap T|$ represents the number of overlapping voxels [9].

The Dice coefficient is maximized to achieve better segmentation performance. For training networks, it is typically reformulated as a loss function, namely the Dice loss, which is defined as

$$\mathcal{L}_{\text{Dice}} = 1 - D. \tag{2}$$

By minimizing Dice loss during training, the model learns to maximize the overlap between predicted and ground truth regions.

3.3.2 Combined Dice and Cross-Entropy Loss

While the Dice loss directly optimizes region overlap, it can be unstable when class imbalance is severe, since small structures such as the enhancing tumor may contribute little to the global overlap. To mitigate this, we combined Dice loss with voxel-wise cross-entropy (CE) loss, which enforces pixel-level classification accuracy. The final loss function is defined as a convex combination:

$$\mathcal{L}_{\text{total}} = \alpha\, \mathcal{L}_{\text{CE}} + (1 - \alpha)\, \mathcal{L}_{\text{Dice}}, \tag{3}$$

where $\alpha \in [0, 1]$ balances the contributions of cross-entropy and Dice loss. In this study, $\alpha = 0.5$ was used, giving equal weight to regional overlap (Dice) and voxel-wise classification (CE). Class weighting was optionally applied to the cross-entropy term to address label imbalance, particularly for the enhancing tumor region, which is underrepresented in the dataset.

4. Experiments

4.1. Preprocessing Pipeline Comparison

To systematically evaluate how different preprocessing strategies impact training dynamics, efficiency, and anatomical consistency, we designed three pipelines: P1 (on-the-fly preprocessing), P2 (cached resampled volumes), and P3 (cached pre-extracted patches). Each pipeline used the same BraTS2020 multimodal MRI data and segmentation masks, but differed in the stage at which computationally heavy operations (clipping, normalization, cropping, resampling, patch extraction) were applied. The goal of this comparison was to balance fidelity of training data with computational feasibility.

4.1.1 Qualitative Assessment

We first examined the visual integrity of patches produced by each pipeline under two complementary visualization modes(Figure 4):

- **Training-like visualization:** Random patch sampling with augmentations enabled (flips, rotations, intensity scaling, noise). This mode replicated the data stream seen during training and confirmed that all pipelines produced anatomically plausible tumor patches with aligned segmentation masks.
- **Deterministic visualization:** Augmentations disabled and seeds fixed to extract comparable subvolumes across pipelines. Under this mode, P1 and P2 produced pixel-identical crops, verifying that caching did not compromise label alignment. P3, however, showed spatial offsets due to pre-sampled patch coordinates, though patches remained anatomically consistent.

Visual inspection also revealed subtle differences in sharpness and positioning: P1 patches were the sharpest, P2 sometimes exhibited mild interpolation artifacts, and P3 occasionally showed reduced diversity due to fixed pre-extracted patches.

4.1.2 Quantitative Consistency

To complement visual inspection, we performed a deterministic patch-level comparison. For each patient and seed, identical coordinates were sampled for P1 and P2, while one representative pre-cached patch was selected for P3. Segmentation masks were collapsed to integer labels, and Dice similarity was computed per class. Results showed:

- P1 vs. P2 $\rightarrow$ Dice = 1.000 across all classes, confirming perfect alignment.
- P1 vs. P3 $\rightarrow$ Dice = 0.000, reflecting decoupled sampling rather than label corruption.
- P2 vs. P3 $\rightarrow$ Dice = 0.000 for the same reason.

Despite low deterministic overlap with P3, its patches remained valid during training due to inherent diversity in the precomputed dataset.

4.1.3 Runtime and Efficiency

Each pipeline was benchmarked on baseline 3D U-Net training to measure computational load. Table 1 summarizes total wall time, average epoch time, and GPU peak memory across seeds.

Table 1. Training summary for pipelines across seeds.

	P1	P2	P3
Total Time (min)	139-151	111-115	111-114
Avg Epoch Time (s)	279-302	223-231	222-229
GPU Peak (MB)	$\sim$2217	$\sim$2191	$\sim$2190

Table 2. Aggregated Dice scores (mean $\pm$ std) of best models across pipelines, averaged over three random seeds. WT = whole tumor, TC = tumor core, ET = enhancing tumor.

	P1	P2	P3
WT	0.850 ± 0.015	$\mathbf{0.856 \pm 0.008}$	0.842 ± 0.010
TC	0.745 ± 0.028	$\mathbf{0.755 \pm 0.005}$	0.737 ± 0.007
ET	0.648 ± 0.018	$\mathbf{0.667 \pm 0.006}$	0.647 ± 0.007

GPU memory usage was nearly identical across pipelines, especially for P2 and P3, but runtime differed significantly. P1 was the slowest compared to P2 or P3. P2 offered a strong balance, reducing wall time while retaining online randomness. P3 achieved the shortest runtimes but sacrificed sampling flexibility, making it less stable across seeds.

4.1.4 Training Dynamics

Loss curves for each pipeline (Figure 5) illustrate differences in convergence. P1 and P2 both showed stable decreases in training and validation loss, with P2 reaching comparable minima in roughly a little over half the runtime of P1. P3 occasionally showed noisier validation curves, consistent with its reduced patch diversity.

4.1.5 Pipeline Selection

Quantitative segmentation performance (on baseline model described in section 3.2) further supported this choice. Table 2 reports the aggregated Dice scores of the best models across three random seeds. While background segmentation was consistently near-perfect ($\approx$ 0.993) for all pipelines, differences emerged in tumor subregions. P2 achieved slightly higher Dice scores for whole tumor (WT: 0.856 ± 0.008) and tumor core (TC: 0.755 ± 0.005) compared to P1 and P3, while enhancing tumor (ET) remained similar across pipelines. Per-class Dice followed the same trend(see demo.ipynb or training results csv), with P2 marginally outperforming the others in edema and enhancing regions. The trade-offs can thus be summarized as:

- **P1:** Maximum fidelity and diversity, but prohibitively slow for large-scale experiments.
- **P2:** Balanced fidelity and efficiency; halved runtime while slightly improving segmentation accuracy.

- **P3:** Fastest pipeline, but reduced patch diversity, noisier validation curves, and marginally lower Dice scores.

Based on both efficiency and segmentation accuracy, **P2 was selected as the default preprocessing pipeline** for all subsequent model ablations and comparisons.

Table 3. Comparison of baseline and optimized U-Net models trained on pipeline P2. Models are denoted as M1–M5 for consistency with the ablation study narrative. WT = whole tumor, TC = tumor core, ET = enhancing tumor.

	M1	M2	M3	M4	M5
WT	0.867	0.878	0.872	0.877	**0.880**
TC	0.770	0.790	0.788	0.806	**0.812**
ET	0.657	0.682	0.673	0.689	**0.691**

5. Ablation Study

To understand how architectural refinements affect segmentation performance, we conducted a series of controlled ablations using pipeline P2. Each model was trained under identical conditions (50 epochs, AdamW optimizer with ReduceLROnPlateau scheduler, early stopping with patience of 10, and learning rate 1×10^{-4} after rigorous smoke tests). For consistency, we report results only on the clinically relevant BraTS regions: Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET).

5.1. Baseline Model (M1)

The baseline U-Net (M1) adopts the classical design, described in section 3.2.(Figure 3). While lightweight and efficient, this design lacks residual connections, attention mechanisms, and deep supervision.

5.2. Optimized Variant 1: Residual + Attention + Deep Supervision (M2)

The first optimized model (M2) introduces:

- Residual convolutional blocks with InstanceNorm and Dropout3D, improving gradient flow and regularization.
- Attention gates on decoder skip connections, filtering irrelevant encoder features.
- Deep supervision via auxiliary heads at intermediate decoder levels, providing additional gradient signals.

This configuration yielded a marked improvement in segmentation accuracy compared to the baseline, particularly for the tumor core and enhancing tumor.

5.3. Optimized Variant 2: GroupNorm + SE + Attention (M3)

The second optimized model (M3) extended this design by replacing InstanceNorm with GroupNorm [11] and integrating Squeeze-and-Excitation (SE) blocks for channel recalibration [12]. This combination aimed to further stabilize training and enhance the sensitivity to small subregions such as ET. While TC improved further compared to the baseline, gains over M2 were more modest.

5.4. Optimized Variant 3: Dropout Removal (M4)

A third experiment disabled dropout in M3, producing M4. The rationale was to evaluate whether stochastic regularization improved or hindered performance under strong data augmentation. Results in Table 3 show that removing dropout further boosted performance, especially for TC (0.806) and ET (0.689), surpassing all previous variants. This suggests that, in this setting, dropout limited convergence rather than preventing overfitting.

5.5. Optimized Variant 4: Class-Weighted Loss (M5)

The final experiment (M5) extended M4 by incorporating class weights into the Dice+CrossEntropy loss to address label imbalance. Results show further improvements across all three regions, achieving the highest Dice scores overall: WT (0.880), TC (0.812), and ET (0.691). This indicates that explicitly rebalancing the loss function can enhance sensitivity to small and clinically relevant regions such as TC and ET.

5.6. Results

Table 3 summarizes the Dice scores for WT, TC, and ET across the baseline and optimized variants. Training and validation dynamics are shown in Figure 6, which overlays training/validation losses and validation Dice across all models. Figure 7 shows sample predictions.

5.7. Training Dynamics and Observation

Figure 6 highlights key trends: M1 converges reliably but saturates at lower Dice for TC and ET. M2 improves ET but remains limited by dropout noise. M3 introduces GroupNorm and SE, with slightly more stable losses but only marginal Dice gains. M4 achieves a better balance, showing smoother validation curves and higher TC/ET Dice. M5 builds on this by leveraging class-weighted loss to further improve all three tumor subregions, particularly TC and ET, while maintaining WT performance. An important observation is that M1 often reaches the lowest absolute loss values, despite lower Dice scores. This apparent mismatch arises because the loss function (Dice + CrossEntropy) is biased toward the dominant background voxels, which M1

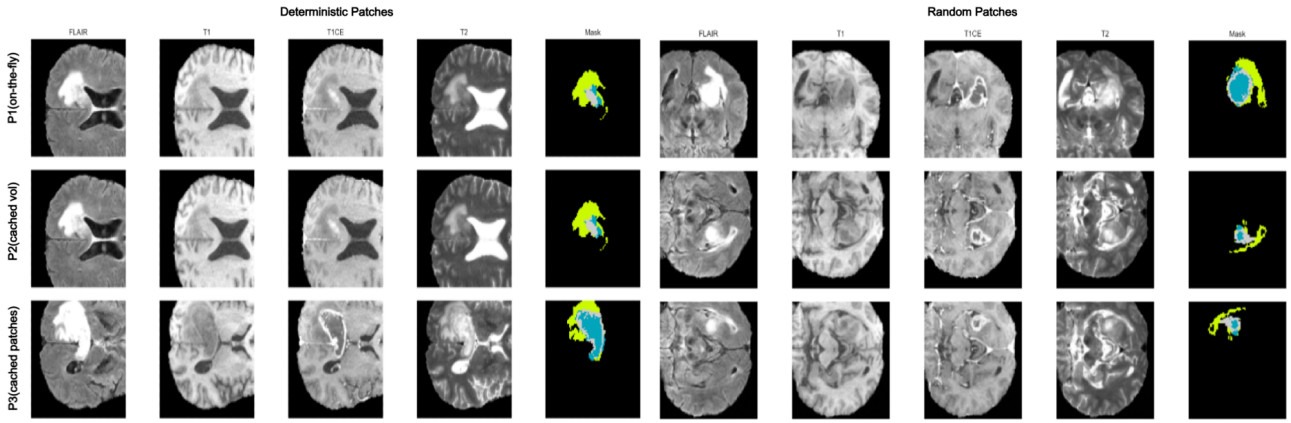

Figure 4. Qualitative Sanity checks: Visualizing patches produced by the 3 pipelines (P1,P2,P3): mainly checking for consistency of anatomical regions (radomly = "what the model sees!") and also the integrity of data deterministically (seeded + no augmentations).

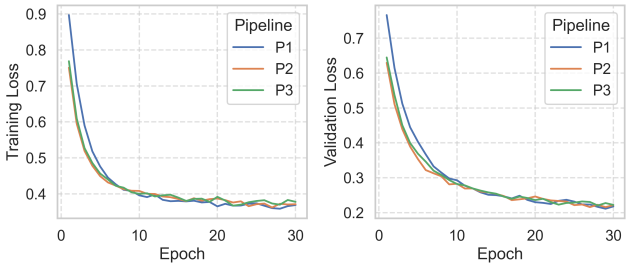

Figure 5. Training and Validation Losses across pipelines P1-P3 (Averaged over 3 seeds). Training was done with the baseline model 3D UNet

learns well. In contrast, optimized models maintain slightly higher losses while improving Dice by focusing more effectively on smaller, clinically critical tumor regions (TC and ET). Thus, Dice, rather than loss convergence alone, provides the more reliable indicator of segmentation quality.

5.8. Summary

Across the ablations, architectural refinements consistently improved segmentation of clinically important tumor subregions (TC and ET), with WT remaining stable across models. M2 provided clear gains, M3 stabilized training, M4 (no dropout) yielded stronger Dice scores, and M5 (class-weighted loss) delivered the best overall performance, demonstrating that imbalance-aware optimization can meaningfully improve sensitivity to critical tumor subregions.

6. Discussion and Conclusions

The ablation study demonstrated that incremental architectural refinements consistently improved segmentation performance, particularly for the clinically relevant tumor core (TC) and enhancing tumor (ET). Among the tested variants, M5 (Optimized U-Net 4 with class-weighted loss) achieved the strongest results, with Dice scores of 0.880 for WT, 0.812 for TC, and 0.691 for ET. These results represent clear improvements over the baseline (M1: 0.867, 0.770, 0.657), confirming the benefit of residual connections, attention gates, deep supervision, normalization refinements, and imbalance-aware training.

Despite these gains, a noticeable gap remains compared to nnU-Net, which reported Dice scores of 88.95 (WT), 85.06 (TC), and 82.03 (ET) on BraTS 2020 [13]. The largest shortfall lies in ET segmentation, where M5 lags by nearly 13 percentage points. Several factors likely contribute: training was limited to 50 epochs with early stopping, whereas nnU-Net employs longer, dataset-adaptive schedules. Extending training or increasing patch size could capture more contextual information, particularly beneficial for ET.

Although class-weighted Dice+CE loss improved balance, the chosen weights ([0.10, 0.20, 0.30, 0.40]) were not fully optimal. Earlier configurations stalled training, and while the final setup improved ET, the gap persisted. Alternative imbalance-aware strategies such as focal loss, dynamic re-weighting, or uncertainty-based losses may prove more effective. Likewise, ensemble training standard in nnU-Net - was not attempted here but could improve robustness.

Another factor is postprocessing. nnU-Net includes dataset-specific refinements such as removing small disconnected predictions, which are particularly effective for ET. Our pipeline omitted such steps, likely contributing to weaker ET scores.

Efficiency analysis underscores the practicality of our approach. M1 completed training in 189 minutes with 2039 MB peak GPU memory, while optimized models (M3–M5) required 217–219 minutes and 3330–3620 MB. These increases are modest compared to nnU-Net's substantially higher computational demands.

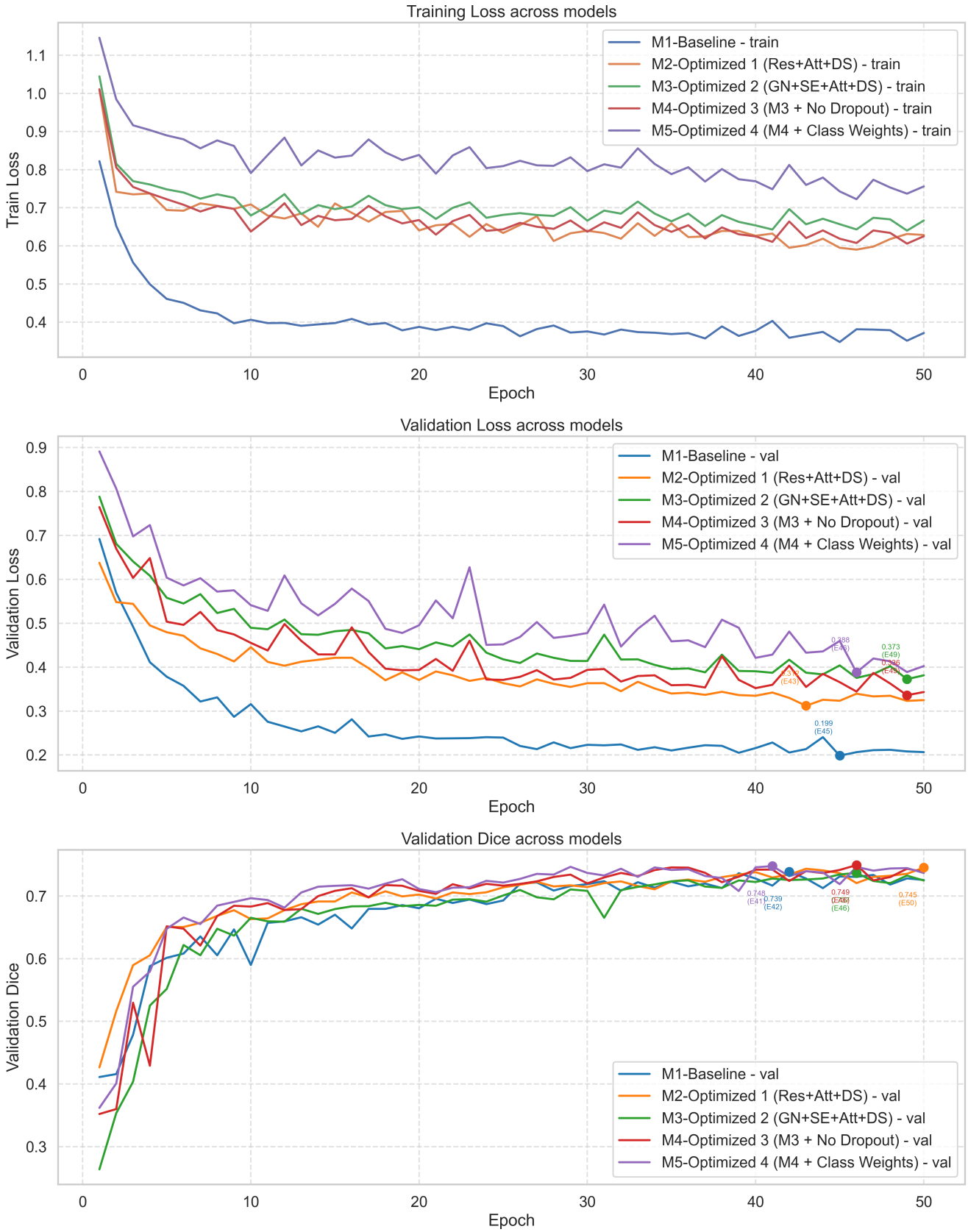

Figure 6. Training dynamics of baseline and optimized U-Net models (M1–M5). Top: training loss trajectories across epochs. Middle: validation loss trajectories, with minima highlighted for each model. Bottom: validation Dice trajectories, with peak values annotated. Architectural refinements (M2–M5) consistently improve Dice performance over the baseline (M1), particularly for tumor core (TC) and enhancing tumor (ET). The class-weighted loss in M5 yields the best overall balance, achieving the highest Dice scores while maintaining stable convergence.

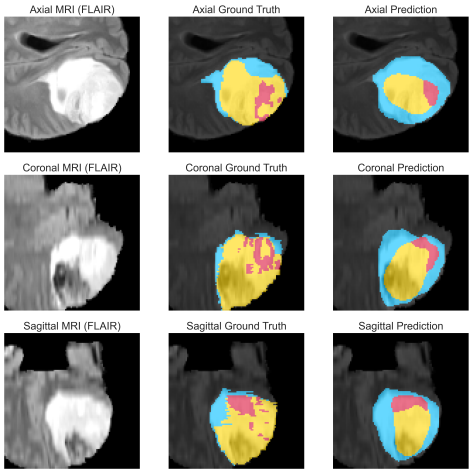

Figure 7. Visualization of model predictions across axial, coronal, and sagittal planes. Each row corresponds to one anatomical view. Columns show the original FLAIR MRI, ground truth segmentation, and model prediction. Segmentation masks highlight background (black), non-enhancing tumor core (yellow), edema (blue), and enhancing tumor (red). Predictions capture the overall tumor extent but show noticeable discrepancies in subregions, particularly in separating non-enhancing (yellow) and enhancing (red) components, consistent with the quantitative Dice results for TC and ET

In conclusion, while nnU-Net remains the state of the art, it requires significant resources. Our optimized U-Net variants show that competitive WT and TC segmentation, and moderately improved ET performance, can be achieved in a lighter framework. Future work should explore extended training, refined loss weighting, and lightweight postprocessing to further close the ET gap while preserving efficiency. This strengthens the case for resource-conscious 3D U-Net designs as viable tools for clinical and research applications under limited hardware constraints.

References

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *arXiv preprint arXiv:1505.04597*, 2015.

[2] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 424–432. Springer, 2016.

[3] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

[4] Mohammad Havaei, Nicolas Guizard, Nicolas Chapados, and Yoshua Bengio. Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, 35:18–31, 2017.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[6] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[7] Ozan Oktay, Jo Schlemper, Loïc Le Folgoc, Matthew Lee, Mattias P. Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[8] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pages 562–570, 2015.

[9] Bastian H. Menze, Ándras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, John Kirby, Yvette Burren, Nicolas Porz, Jens Slotboom, Roland Wiest, and Koen van Leemput. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2015.

[10] Roey Mechrez, Jacob Goldberger, and Hayit Greenspan. Patch-based segmentation with spatial consistency: Application to ms lesions in brain mri. *International Journal of Biomedical Imaging*, 2016(1):7952541, 2016.

[11] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 11217 of *Lecture Notes in Computer Science*, pages 3–19. Springer, Cham, 2018.

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018.

[13] Fabian Isensee, Paul F. Jaeger, Peter M. Full, Philipp Vollmuth, and Klaus H. Maier-Hein. nnu-net for brain tumor segmentation, 2020.