# HBA-VAN: Hierarchical Boundary-Aware Volumetric Attention Network for Multi-Modal Glioma Subregion Segmentation

Jingxuan Zhang*
2023533003
zhangjx12023@shanghaitech.edu.cn

Yatu Zhang*
2023533005
zhangyt2023@shanghaitech.edu.cn

Yiwen Cai*
2023533145
caiyw2023@shanghaitech.edu.cn

Zidong Song*
2024533002
songzd2024@shanghaitech.edu.cn

ShanghaiTech University
*Equal contribution

**Abstract**

Accurate glioma subregion segmentation from multi-modal magnetic resonance imaging (MRI) requires models that capture volumetric anatomy while remaining sensitive to fine tumor boundaries. This is challenging because BraTS subregions are nested, strongly imbalanced, and modality dependent; enhancing tumor (ET), in particular, often occupies only a small fraction of the brain volume. We study this problem through a controlled progression from slice-wise prediction to volumetric reasoning. A strong 2D U-Net establishes the reference point, a 2.5D model evaluates the value of local inter-slice context, and the final model, HBA-VAN, performs full 3D segmentation with residual volumetric encoding, attention-gated skip fusion, boundary-aware supervision, and ET-specific refinement. On the held-out test set, HBA-VANachieves a case-level mean Dice of 0.9167 and a mean HD95 of 2.8165. The largest gain is observed for ET, while the reduction in HD95 indicates improved boundary localization. Worst-case analysis further shows that residual errors concentrate in extremely small ET regions and ambiguous tumor-core boundaries.

## 1 Introduction

Accurate glioma segmentation from multi-modal magnetic resonance imaging (MRI) is a core problem in computational neuro-oncology. Reliable tumor delineation supports volumetric assessment, surgical and radiotherapy planning, and longitudinal monitoring of disease progression. The BraTS benchmark [3, 4, 26] has made this problem measurable by defining three clinically meaningful regions: whole tumor (WT), tumor core (TC), and enhancing tumor (ET). These regions are not ordinary mutually exclusive labels. They form a nested anatomical structure with heterogeneous imaging signatures: WT captures the broad abnormal extent, TC is more compact and boundary-sensitive, and ET is often small, fragmented, and strongly dependent on contrast-enhanced appearance.

This hierarchy creates a tension between efficient learning and anatomically faithful prediction. Slice-wise 2D U-Nets remain appealing because they are stable, memory efficient, and benefit from many supervised axial slices. Yet the underlying pathology is volumetric. A local pattern that appears ambiguous on one slice may become clear only when its continuity across neighboring slices is considered. 2.5D models provide a pragmatic compromise by exposing the network to adjacent slices, but they still treat depth as shallow input context rather than as a spatial dimension represented throughout the feature hierarchy. For BraTS segmentation, the central question is therefore not simply whether increasing model capacity improves performance. The more important question is how much through-plane reasoning is needed to segment nested, imbalanced tumor subregions with reliable boundaries.

We study this question through a controlled progression from 2D prediction to full 3D volumetric modeling. A strong 2D U-Net is used as a meaningful reference rather than a weak baseline. A 2.5D variant then tests whether local inter-slice evidence improves segmentation when the prediction remains center-slice based. Building on the observed value

of context, we propose HBA-VAN, a hierarchical boundary-aware volumetric attention network. The architecture combines residual 3D encoding, attention-gated skip fusion, deep supervision, boundary-aware auxiliary learning, and ET-specific refinement. Its design follows the BraTS containment relation ET $\subset$ TC $\subset$ WT: predictions are modeled as overlapping binary regions while the objective encourages anatomically plausible consistency.

Our experiments report both overlap and boundary-sensitive metrics under patient-level splits. The final model achieves a case-level mean Dice of 0.9167 and mean HD95 of 2.8165 on the held-out test set, with the clearest gain on ET. Beyond aggregate performance, we analyze the transition from 2D to 2.5D to 3D, isolate the effect of ET refinement, and examine the worst-performing cases. This analysis shows that volumetric attention and boundary-aware learning improve the overall error profile, while extremely small ET regions and ambiguous tumor-core boundaries remain the dominant residual failure modes.

## 2 Related Work

### 2.1 Encoder-Decoder Segmentation and Volumetric Learning

Encoder-decoder networks remain the dominant family of architectures for biomedical segmentation. U-Net [32] introduced the combination of a contracting semantic path and an expansive localization path connected by high-resolution skips, while residual learning [13] made deeper convolutional backbones easier to optimize. Subsequent medical segmentation systems extended this template to volumetric data. 3D U-Net [7] and V-Net [27] replaced 2D operators with 3D convolutions, allowing representations to evolve over all anatomical axes. DeepMedic [18], VoxResNet [5], deeply supervised 3D networks [9], and autoencoder-regularized 3D segmentation [29] further established that through-plane context is central for lesion and tumor segmentation. More recent ConvNet families, including nnU-Net [16] and MedNeXt [34], show that strong convolutional models remain competitive when preprocessing, patch sampling, normalization, loss design, and inference are configured carefully.

These architectures also expose two practical constraints that are particularly relevant for BraTS. First, full 3D learning is memory intensive, which motivates patch-based training and sliding-window inference. Second, the apparent simplicity of a U-Net baseline can be misleading: strong performance often depends on foreground-aware sampling, modality normalization, deep supervision, and robust evaluation. Our study therefore treats the Task2 2D U-Net as a strong empirical anchor, uses 2.5D learning as an intermediate probe of local context, and reserves full 3D modeling for the final architecture.

### 2.2 Glioma Segmentation, Imbalance, and Boundary-Sensitive Objectives

The BraTS benchmark standardized multi-modal glioma segmentation and made WT, TC, and ET the canonical evaluation regions [3, 4, 26]. Earlier CNN systems demonstrated the value of multi-modal MRI and dense prediction for tumor delineation, including patch-based CNNs [31], cascaded and multi-scale networks [12, 39, 45], and task-specific BraTS pipelines [15, 17]. These works also highlight the difficulty of ET: it is small, contrast-dependent, and often sparse relative to edema and normal anatomy. The nested region structure ET $\subset$ TC $\subset$ WT further means that the three outputs should not be treated as mutually exclusive semantic classes.

Loss design is therefore as important as architecture. Dice-style optimization [27, 37] directly addresses foreground-background imbalance, focal loss [23] and Tversky variants [1, 35] emphasize hard positives and asymmetric errors, and boundary-aware objectives [19, 20] target surface localization. Deep supervision [9, 22] can make intermediate decoder features region-discriminative, while surface-sensitive metrics such as HD95 reveal boundary failures that Dice alone may hide. HBA-VANcombines these lessons by using region-aware sampling, weighted overlap and calibration terms, auxiliary boundary supervision, and a soft hierarchy penalty aligned with the BraTS containment relation.

### 2.3 Attention, Transformers, and Emerging Context Models

Attention mechanisms improve segmentation by selecting task-relevant features before fusion. Squeeze-and-excitation [14], CBAM [42], and spatial-channel excitation for medical segmentation [33] modulate features channel-wise or spatially, while Attention U-Net [30] gates encoder skips with decoder context. This is especially useful for brain tumor

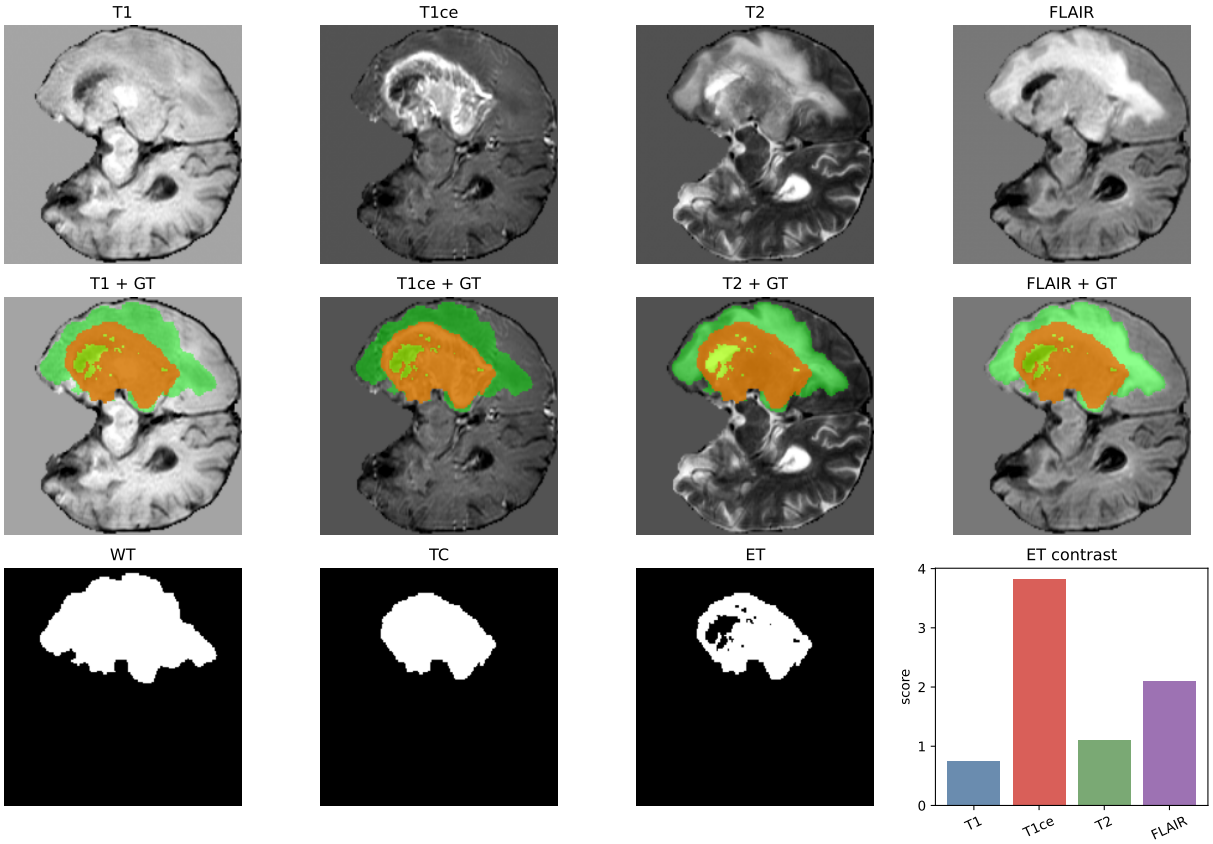


Figure 1: Processed BraTS example showing multi-modal contrast, nested WT/TC/ET masks, and ET visibility on a representative tumor slice.

segmentation: high-resolution MRI contains many normal anatomical structures, and unconditional skip concatenation can reintroduce distractors into the decoder. Our attention gates follow this principle, using high-level decoder evidence to regulate which encoder details are transferred during 3D reconstruction.

Transformer-based models provide a complementary route to context modeling. The general self-attention framework [38] and vision transformers [8] motivated medical architectures such as TransUNet [6], CoTr [43], TransBTS [41], UNETR [11], Swin UNETR [10], nnFormer [46], and UNETR++ [36]. Recent state-space and foundation-model directions, including U-Mamba [25], SegMamba [44], Segment Anything [21], MedSAM [24], SAM-Med3D [40], and medical image foundation-model surveys [2, 28], suggest that long-range context and large-scale pretraining are increasingly important. In this project, however, we deliberately adopt a lighter design: convolutional 3D residual blocks provide the main volumetric representation, while attention is used as a targeted skip-fusion mechanism and ET refinement is introduced as an anatomically motivated module rather than a generic increase in model size.

## 3 Dataset and Preprocessing

Each subject is provided as co-registered NIfTI volumes containing four MRI modalities and one segmentation mask. We use T1n, contrast-enhanced T1c, T2w, and T2f/FLAIR as the input channels. All volumes are first converted into NumPy arrays, stacked into a four-channel tensor, and cropped to the nonzero brain foreground. This foreground cropping removes redundant black background, reduces memory consumption, and keeps the subsequent 2D and 3D training units focused on anatomically meaningful regions. Within the cropped brain mask, each modality is normalized independently by z-score normalization,

$$\hat{X}_m(v) = \frac{X_m(v) - \mu_m}{\sigma_m + \epsilon}, \tag{1}$$

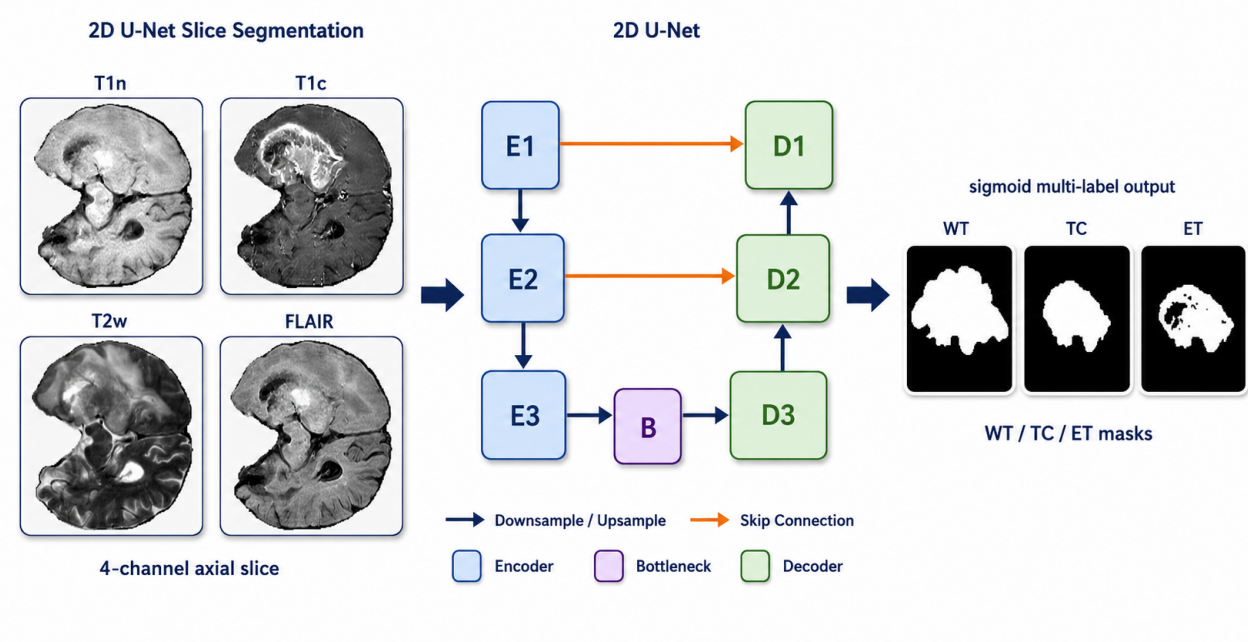


Figure 2: Task2 2D U-Net baseline. Four MRI modalities from one axial slice are processed by a slice-wise encoder-decoder with skip connections and independent sigmoid heads for WT, TC, and ET.

where $\mu_m$ and $\sigma_m$ are computed from nonzero brain voxels of modality $m$.

Raw tumor labels are remapped into the three overlapping regions specified by the BraTS task. Under the conventional BraTS encoding used in the project description, the mapping is

$$\mathrm{WT} = \{1, 2, 4\}, \qquad \mathrm{TC} = \{1, 4\}, \qquad \mathrm{ET} = \{4\}. \tag{2}$$

The BraTS-GLI release used in our experiments may additionally encode enhancing or residual core components with label 3; the implementation therefore uses the generalized mapping $\mathrm{WT} = \{1, 2, 3, 4\}$, $\mathrm{TC} = \{1, 3, 4\}$, and $\mathrm{ET} = \{3, 4\}$ when such labels are present. In both cases, the resulting WT, TC, and ET masks form a multi-label hierarchy rather than mutually exclusive semantic classes.

All cases are split at the patient level into training, validation, and test partitions to avoid slice leakage. Task2 uses the processed 2D slice dataset: axial slices are resized to a fixed spatial size and stored with four-channel images, three-channel region masks, case identifiers, and slice indices. Task3 stores full 3D case arrays and samples volumetric patches during training, while evaluation is performed by sliding-window inference over the full volume.

The preprocessing pipeline is also used for exploratory visualization and contrast analysis, as required by the project specification. For a representative tumor-containing slice, Figure 1 visualizes the four modalities at the same anatomical position, overlays the ground-truth masks, and displays WT/TC/ET separately. We additionally compare the mean intensity of ET against surrounding non-tumor brain tissue across modalities. In the selected example, T1c provides the strongest ET contrast, which is consistent with the role of contrast-enhanced T1 imaging in highlighting enhancing tumor.

## 4 Methodology

### 4.1 Problem Setup and Strong 2D Baseline

Let $\Omega = \{1, \ldots, D\} \times \{1, \ldots, H\} \times \{1, \ldots, W\}$ denote the voxel lattice of a subject. Each case is represented by a four-modality MRI volume $X \in \mathbb{R}^{4 \times D \times H \times W}$ and a multi-label target $Y \in \{0, 1\}^{3 \times D \times H \times W}$ over the region set $\mathcal{R} = \{\mathrm{WT}, \mathrm{TC}, \mathrm{ET}\}$. For voxel $v \in \Omega$, $Y_r(v) = 1$ indicates that $v$ belongs to region $r$. Unlike standard semantic

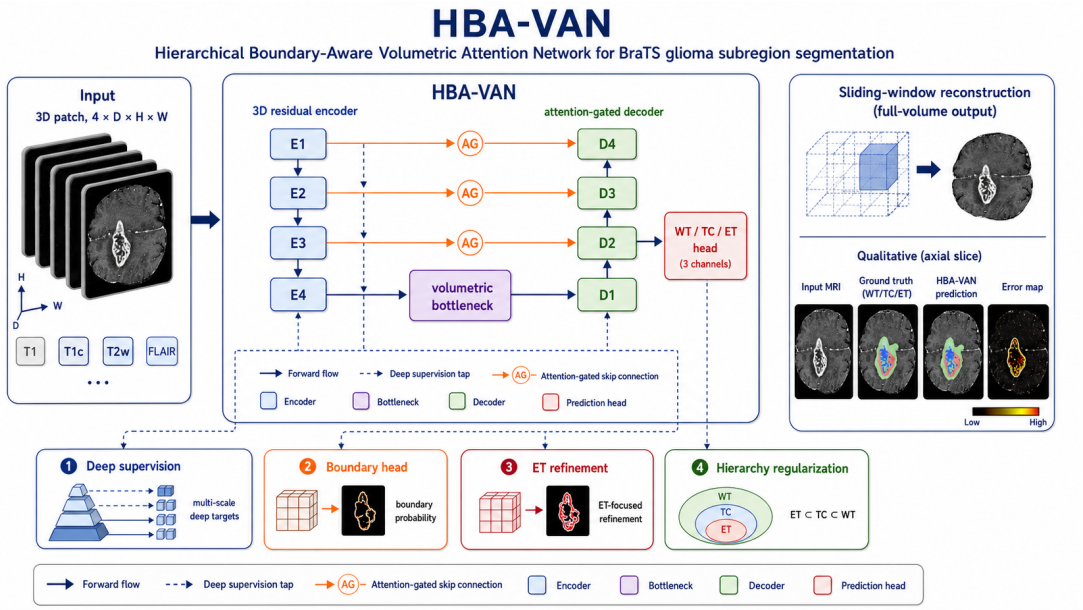


Figure 3: HBA-VAN architecture. A 3D residual encoder-decoder with attention-gated skip connections is augmented with deep supervision, boundary prediction, ET refinement, hierarchy regularization, and sliding-window full-volume reconstruction.

segmentation, the three BraTS outputs are not mutually exclusive classes. Their definitions imply ET $\subset$ TC $\subset$ WT, so a softmax classifier would impose an incorrect competition between clinically nested regions. We instead estimate independent region probabilities:

$$Z = f_\theta(X), \qquad P_r(v) = \sigma(Z_r(v)), \quad r \in \mathcal{R}. \tag{3}$$

The final binary mask for each region is obtained by thresholding $P_r(v)$. This multi-label formulation preserves the overlapping region structure while leaving room for soft anatomical regularization during training.

This formulation also clarifies the role of the Task2 baseline, which serves as the empirical anchor for the rest of the study. For an axial slice $x_z = X_{:,z,:,:} \in \mathbb{R}^{4\times H\times W}$, the 2D U-Net estimates

$$P_{r,z}(u) = \sigma(f_{2D}(x_z)_r(u)), \qquad u \in \{1,\dots,H\} \times \{1,\dots,W\}. \tag{4}$$

The baseline is intentionally strong rather than minimal. It uses all four MRI modalities, skip-connected decoding for high-resolution localization, region-aware loss weighting for WT/TC/ET imbalance, and tumor-biased sampling so that rare positive slices are not overwhelmed by normal anatomy. Figure 2 summarizes the resulting 2D pipeline. This makes the 2D model a meaningful reference: improvements over it are less likely to come from basic preprocessing or optimization choices, and more likely to reflect additional spatial modeling.

At the same time, the 2D baseline exposes a precise representational limitation. Since $P_{r,z}$ depends only on $x_z$, the model has no direct mechanism for checking whether a candidate tumor pattern persists, expands, or disappears across depth. This is especially consequential for TC and ET, whose boundaries and contrast patterns can change rapidly between adjacent slices. The subsequent 2.5D and 3D models are therefore framed as increasingly expressive ways to augment this strong slice-wise baseline with through-plane evidence, rather than as unrelated architectures.

The 2.5D model is introduced as a controlled way to test whether this missing through-plane evidence is useful. Rather than changing the prediction target, it changes only the input evidence available for a center slice. For a context radius $k$, the model receives the local stack

$$\mathcal{X}_z^{2.5D} = \{x_{z-k}, \dots, x_z, \dots, x_{z+k}\}, \tag{5}$$

and predicts the center-slice target $Y_{:,z,:,:}$. The training signal therefore remains slice-level, but the decision is conditioned on local anatomical continuity. This makes 2.5D a useful intermediate formulation: it retains much of the efficiency of 2D training while testing whether adjacent slices resolve ambiguities in the center slice.

Table 1: Design summary of HBA-VANand the BraTS-specific difficulty addressed by each component.

| Design element | Methodological role |
|---|---|
| Volumetric residual backbone | Models through-plane continuity and multi-scale tumor morphology. |
| Attention-gated skip fusion | Selectively transfers high-resolution localization features into the decoder. |
| Deep supervision | Encourages intermediate decoder representations to be region-discriminative. |
| Boundary auxiliary learning | Provides surface-aware supervision for boundary-sensitive evaluation. |
| ET refinement | Uses tumor-core context to improve the smallest and most imbalanced subregion. |
| Hierarchy regularization | Penalizes predictions inconsistent with ET $\subset$ TC $\subset$ WT. |

However, 2.5D still treats depth as input context rather than as a spatial axis of representation. Features are not propagated through a dense 3D lattice, and the network predicts only the center slice instead of a volumetric field. The positive role of 2.5D context motivates the final architecture, while its limitation motivates moving from input-level depth context to feature-level volumetric reasoning.

### 4.2 HBA-VAN Architecture

HBA-VAN implements this transition by predicting dense labels over volumetric patches. Given $X_p \in \mathbb{R}^{4\times d\times h\times w}$, the network produces logits $Z_p \in \mathbb{R}^{3\times d\times h\times w}$ for all voxels in the patch. The full pipeline is shown in Figure 3. The backbone is a residual 3D encoder-decoder: the encoder progressively increases semantic abstraction over the volumetric field, while the decoder restores spatial resolution for voxel-wise prediction. Residual blocks let each stage model refinements of the incoming representation rather than relearning it from scratch, which is helpful when MRI intensity contrast and tumor morphology must be preserved simultaneously.

The architectural motivation is anatomical. Edema, tumor core, and enhancement are not isolated axial patterns but spatially coherent structures. Full 3D convolution allows the model to represent through-plane continuity of WT, compactness of TC, and the spatial relation between ET and surrounding tumor tissue throughout the feature hierarchy. This is the main distinction from 2.5D: depth is no longer a shallow input cue, but a dimension over which intermediate representations are learned.

High-resolution skip connections are essential for localization, but unfiltered skips can also pass normal anatomical texture into the decoder. HBA-VAN therefore uses attention-gated skip fusion so that semantic decoder context regulates which encoder features are reused. For an encoder skip tensor $S$ and decoder gating tensor $G$, the filtered skip is

$$\tilde{S} = S \odot \sigma\left(\psi\left(\phi_s(S) + \phi_g(G)\right)\right), \tag{6}$$

where $\phi_s$ and $\phi_g$ project encoder and decoder features into a shared compatibility space, $\psi$ maps the compatibility representation to an attention mask, and $\odot$ denotes element-wise modulation. The gate therefore preserves high-resolution detail when it is compatible with decoder-level tumor evidence, avoiding unconditional concatenation of all encoder activations.

Table 1 summarizes how the major architectural components map to the BraTS-specific difficulties discussed above.

The decoder is supervised not only at the final output but also through intermediate prediction heads. Deep supervision makes coarse and fine decoder stages directly discriminative for the tumor regions. This is useful in volumetric segmentation because coarse stages capture global tumor extent, whereas high-resolution stages recover local boundaries. The final prediction can therefore rely on a decoder hierarchy whose intermediate features are already aligned with segmentation.

Boundary supervision is introduced for a complementary reason. Dice is an overlap metric and can be relatively insensitive to small surface shifts, particularly for large regions. HD95, however, directly reflects boundary error. To bias shared decoder features toward surface localization, we derive a boundary band from each binary region target:

$$B(Y^r) = \mathrm{dilate}(Y^r) - \mathrm{erode}(Y^r). \tag{7}$$

The boundary head predicts this contour-like target as an auxiliary task. It does not replace region supervision; instead, it injects a surface-aware signal into the shared decoder representation. This is particularly relevant around diffuse

edema and irregular tumor-core margins, where similar Dice values can correspond to substantially different boundary distances.

The ET refinement branch addresses the most imbalanced subregion. ET is small, contrast-dependent, and often responsible for the lowest case-level scores. Predicting it as an isolated channel ignores the anatomical fact that ET should lie inside the tumor core. HBA-VANtherefore refines ET using both the final decoder feature $F_{dec}$ and the main TC logit $Z_{\mathrm{TC}}^{main}$. The branch predicts an additive correction:

$$Z_{\mathrm{ET}}^{final} = Z_{\mathrm{ET}}^{main} + \alpha R_{\mathrm{ET}}\left(F_{dec}, \sigma(Z_{\mathrm{TC}}^{main})\right). \tag{8}$$

The refinement is residual: the main ET prediction remains available, while the correction uses tumor-core evidence to recalibrate ET. A soft hierarchy penalty further discourages probability maps that violate the expected containment relation:

$$\mathcal{L}_{nest} = [\sigma(Z_{\mathrm{TC}}) - \sigma(Z_{\mathrm{WT}})]_+ + [\sigma(Z_{\mathrm{ET}}) - \sigma(Z_{\mathrm{TC}})]_+ . \tag{9}$$

The penalty does not hard-code the masks during learning. Instead, it discourages anatomically implausible probability configurations while allowing the network to resolve uncertain voxels from image evidence.

### 4.3 Optimization and Full-Volume Inference

Training is performed on volumetric patches. This is necessary for memory efficiency, but it also creates a sampling problem: uniform patches would be dominated by healthy tissue and provide little signal for TC or ET. We therefore use region-aware patch sampling that increases the frequency of tumor-centered patches, gives additional emphasis to ET-containing regions, and still retains non-tumor context to control false positives. Spatial and intensity augmentations are applied only in ways that preserve the clinical meaning of the MRI modalities.

The main segmentation objective combines overlap, calibration, and hard-example terms:

$$\mathcal{L}_{main} = \lambda_D \mathcal{L}_{Dice} + \lambda_B \mathcal{L}_{BCE} + \lambda_F \mathcal{L}_{Focal}. \tag{10}$$

Dice emphasizes region overlap under foreground-background imbalance, BCE stabilizes voxel-wise probability calibration, and focal BCE increases the contribution of difficult voxels and rare positives. The full objective ties the region, scale, boundary, ET, and hierarchy terms into one training criterion:

$$\mathcal{L} = \mathcal{L}_{main} + \sum_k \lambda_k \mathcal{L}_{DS}^{(k)} + \lambda_{bd}\mathcal{L}_{bd} + \lambda_{et}\mathcal{L}_{ET} + \lambda_{nest}\mathcal{L}_{nest}. \tag{11}$$

Here $\mathcal{L}_{DS}^{(k)}$ supervises decoder scale $k$, $\mathcal{L}_{bd}$ supervises boundary prediction, $\mathcal{L}_{ET}$ supervises the refinement branch, and $\mathcal{L}_{nest}$ enforces soft anatomical consistency.

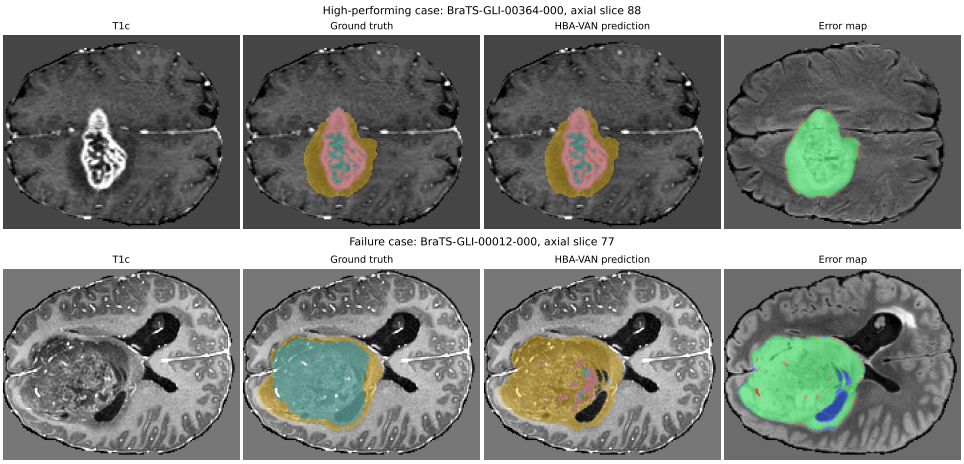


Figure 5: Qualitative test examples for HBA-VAN, illustrating typical behavior and a difficult failure case.

At inference time, the patch model is applied to the full MRI volume with overlapping sliding windows. Patch probabilities are accumulated into a case-level probability volume before thresholding. A final hierarchy-aware post-processing step enforces ET $\subset$ TC $\subset$ WT and removes implausibly small isolated ET components. All reported metrics are computed on these reconstructed full-volume predictions, so the evaluation reflects case-level segmentation rather than isolated patch accuracy.

# 5 Experiments

## 5.1 Evaluation Protocol and Main Comparison

We evaluate segmentation quality using Dice coefficient and the 95th percentile Hausdorff distance (HD95). Dice measures volumetric overlap,

$$\mathrm{Dice}(A, B) = \frac{2|A \cap B|}{|A| + |B|}, \tag{12}$$

whereas HD95 measures boundary discrepancy while reducing sensitivity to extreme outlier points. Higher Dice and lower HD95 indicate better segmentation quality. Unless otherwise stated, all values are case-level metrics computed on reconstructed full-volume predictions from the held-out test set.

Table 2 compares the principal modeling stages. The Task2 2D U-Net obtains a strong mean Dice of 0.8962, which makes it a meaningful slice-wise reference. The best 2.5D model improves mean Dice to 0.9091, supporting the hypothesis that local inter-slice context is useful. HBA-VANfurther improves mean Dice to 0.9167 and yields the strongest ET Dice, increasing ET from 0.8544 in the 2D baseline to 0.8954. Its mean HD95 of 2.8165 indicates that the gain is not restricted to region volume but also improves boundary-sensitive evaluation.

Figure 4 complements the aggregate results in Table 2 with a case-level comparison on the same axial slice. The row-wise layout makes the progression in spatial modeling directly visible: the slice-wise baseline recovers the dominant tumor extent, the intermediate context model improves subregion delineation, and HBA-VAN produces the closest agreement with the reference mask. The corresponding error maps further expose residual false-positive and false-negative regions that are obscured by a single overlap score.

Figure 5 visualizes representative full-volume predictions from the final model. The overlays are designed to show both region agreement and spatial error structure. True-positive regions are shown in green in the error map, false positives in red, and false negatives in blue.

## 5.2 Context and Ablation Studies

The preliminary 2.5D study separates two questions that are otherwise coupled in a full 3D model: whether neighboring slices are useful at all, and whether full volumetric feature learning is necessary beyond local slice context. Table 3 summarizes the available variants. The basic 2.5D reproduction already improves over the 2D baseline, while additional fusion and loss variants provide small but consistent gains. The best 2.5D model reaches a test case-level mean Dice of 0.9091, improving over the 2D baseline by 1.29 points. This result motivates volumetric modeling, but its remaining gap to HBA-VANsuggests that local input context alone is insufficient to fully capture tumor topology.

Table 4 isolates the effect of ET-specific refinement in the full-volume 3D setting. Both models use volumetric attention, deep supervision, and boundary supervision; HBA-VANadditionally includes the ET refinement branch. The refinement branch improves ET Dice by 1.29 points and reduces mean HD95 by 0.6911. This suggests that the refinement branch improves not merely average overlap, but also the spatial reliability of the predicted surfaces. Figure 6

Table 2: Held-out test performance with case-level Dice and HD95.

| Method | Spatial | WT | TC | ET | Mean | HD95 |
|---|---|---|---|---|---|---|
| Task2 U-Net | 2D | 0.9253 | 0.9090 | 0.8544 | 0.8962 | 6.6879 |
| Best 2.5D | 2.5D | 0.9381 | **0.9224** | 0.8667 | 0.9091 | 4.0447 |
| HBA-VAN | 3D | **0.9403** | 0.9144 | **0.8954** | **0.9167** | **2.8165** |

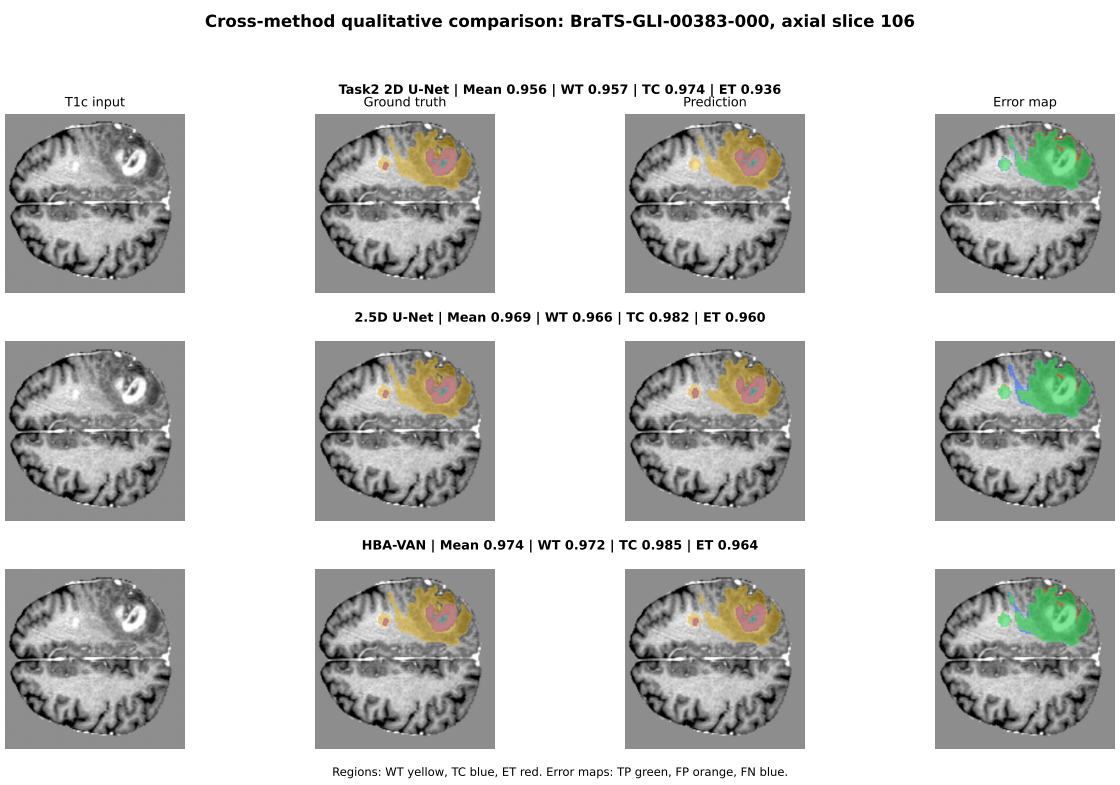


Figure 4: Cross-method qualitative comparison on BraTS-GLI-00383-000. Each row shows one modeling stage using the same T1c slice and ground truth, followed by its segmentation and spatial error map. The reported case-level mean Dice increases from 0.9560 for Task2 to 0.9694 for the intermediate context model and 0.9736 for HBA-VAN. Region overlays denote WT, TC, and ET; error maps show true positives in green, false positives in orange, and false negatives in blue.

complements the table by comparing spatial error patterns before and after ET refinement.

### 5.3 Error Analysis

Table 5 reports the five worst test cases under HBA-VAN. The most severe case contains only 224 target ET voxels, which is also reflected qualitatively in Figure 5. In such cases, Dice becomes extremely sensitive to small localization errors, and ET/TC disagreement can dominate the mean score even when WT is accurate. Other cases show boundary and shape errors, including ET over-segmentation or tumor-core under-segmentation. These results indicate that the remaining failure modes are not uniformly distributed across regions: ET and TC boundaries are the primary bottlenecks.

Table 3: Preliminary 2.5D study with case-level mean Dice on validation and test sets.

| Variant | Design | Val Mean | Test Mean |
|---|---|---|---|
| G1 reproduction | local context | 0.8998 | 0.9057 |
| Fusion-aware | feature fusion | 0.8976 | 0.9068 |
| Modality-aware | modality fusion | 0.8978 | 0.9082 |
| WT-aware loss | weighted objective | 0.8958 | **0.9091** |

Table 4: 3D ablation showing the effect of ET refinement on overlap and boundary accuracy.

| Variant | Attn. | DS | Bdry | ET ref. | ET | HD95 |
|---|---|---|---|---|---|---|
| 3D Attn.+Bdry | ✓ | ✓ | ✓ | – | 0.8825 | 3.5076 |
| HBA-VAN | ✓ | ✓ | ✓ | ✓ | **0.8954** | **2.8165** |

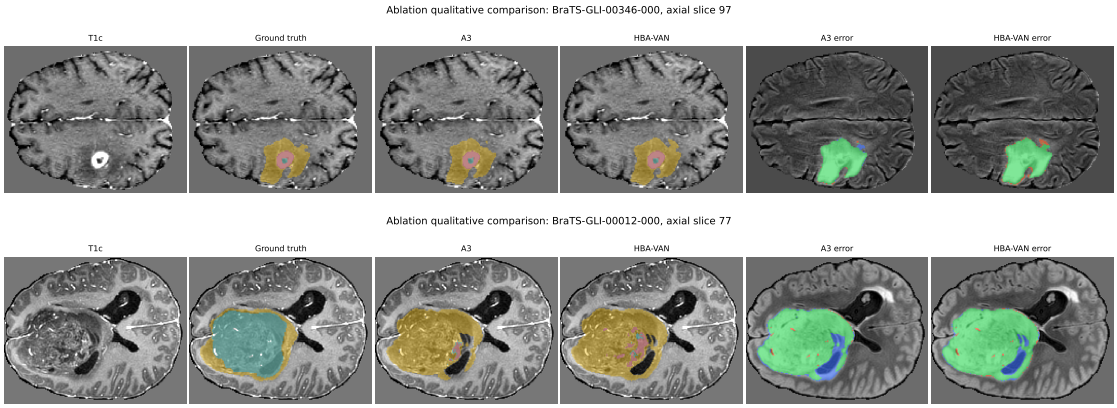


Figure 6: Qualitative ablation comparing spatial errors before and after ET refinement.

## 6 Discussion

The experimental trajectory supports the central design choice of moving from slice-wise prediction to volumetric modeling. The 2D U-Net baseline is strong, but its lack of explicit through-plane context limits its ability to preserve volumetric coherence. The 2.5D model partially mitigates this issue, improving mean Dice by 1.29 points over the 2D baseline. However, the final 3D model provides the strongest result, indicating that true volumetric feature learning remains valuable even when the 2D baseline is well optimized.

The ET refinement branch yields a targeted improvement. ET Dice increases from 0.8825 to 0.8954 relative to the 3D boundary model, while mean HD95 decreases from 3.5076 to 2.8165. This pattern is important: a Dice-only gain could reflect better thresholding or region volume calibration, whereas the HD95 reduction suggests improved boundary localization. The result is consistent with the design of HBA-VAN, which combines attention-gated feature selection, boundary supervision, and ET-specific refinement.

The worst-case analysis also clarifies the limitations of the current model. Extremely small ET lesions remain difficult because both patch sampling and loss weighting have limited signal when the target occupies only a few hundred voxels. Boundary ambiguity is another persistent source of error, especially for TC. Future work could explore uncertainty-aware ensembling, stronger topology-constrained decoding, lesion-size-adaptive focal/Tversky losses, and additional modality-aware fusion for low-contrast ET cases.

Table 5: Worst test cases for HBA-VAN, dominated by small ET regions and ambiguous tumor-core boundaries.

| Case ID | Mean | WT | TC | ET | Failure mode |
|---|---|---|---|---|---|
| BraTS-GLI-00012-000 | 0.3573 | 0.9650 | 0.0690 | 0.0379 | small ET |
| BraTS-GLI-00610-001 | 0.4308 | 0.5615 | 0.4341 | 0.2968 | shape/boundary |
| BraTS-GLI-00366-000 | 0.6160 | 0.9536 | 0.0518 | 0.8425 | TC boundary |
| BraTS-GLI-00149-000 | 0.7442 | 0.8641 | 0.6742 | 0.6941 | shape/boundary |
| BraTS-GLI-00124-000 | 0.7879 | 0.8616 | 0.9552 | 0.5468 | ET over-seg. |

## 7 Conclusion

We presented a staged study for BraTS glioma subregion segmentation. Task2 establishes a strong 2D U-Net baseline, while Task3 moves toward volumetric segmentation. The proposed HBA-VANmodel integrates 3D residual encoding, attention-gated skip fusion, deep supervision, boundary-aware auxiliary learning, and ET-specific refinement. On the held-out test set, HBA-VANachieves the best mean Dice and mean HD95 among the evaluated models, with the largest gain on ET. The results indicate that volumetric context and clinically motivated refinement are effective for improving glioma subregion segmentation beyond strong slice-wise baselines.

## References


[1] Nabila Abraham and Naimul Mefraz Khan. A novel focal Tversky loss function with improved attention U-Net for lesion segmentation. In *IEEE International Symposium on Biomedical Imaging*, pages 683–687, 2019.

[2] Reza Azad, Ehsan Khodapanah Aghdam, Anne Rauland, Yiwei Jia, Amirhossein Bozorgpour Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen, Ehsan Adeli, and Dorit Merhof. Medical image segmentation review: The success of U-Net. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[3] Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C. Kitamura, Sarthak Pati, et al. The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

[4] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S. Kirby, John B. Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific Data*, 4:170117, 2017.

[5] Hao Chen, Qi Dou, Lequan Yu, Jing Qin, and Pheng-Ann Heng. VoxResNet: Deep voxelwise residual networks for brain segmentation from 3d MR images. *NeuroImage*, 170:446–455, 2018.

[6] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[7] 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*, pages 424–432. Springer, 2016.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[9] Qi Dou, Hao Chen, Lequan Yu, Jing Qin, and Pheng-Ann Heng. 3d deeply supervised network for automated segmentation of volumetric medical images. *Medical Image Analysis*, 41:40–54, 2017.

[10] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R. Roth, and Daguang Xu. Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In *International MICCAI Brainlesion Workshop*, pages 272–284. Springer, 2022.

[11] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R. Roth, and Daguang Xu. UNETR: Transformers for 3d medical image segmentation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1748–1758, 2022.

[12] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, 35:18–31, 2017.

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

[14] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[15] Fabian Isensee, Philipp Kickingereder, Wolfgang Wick, Martin Bendszus, and Klaus H. Maier-Hein. No New-Net. In *International MICCAI Brainlesion Workshop*, pages 234–244. Springer, 2018.

[16] 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*, 18:203–211, 2021.

[17] Zeyu Jiang, Changxing Ding, Min Liu, and Dacheng Tao. A two-stage cascaded U-Net for brain tumor segmentation. *IEEE Access*, 8:119382–119391, 2020.

[18] Konstantinos Kamnitsas, Christian Ledig, Virginia F. J. Newcombe, Joanna P. Simpson, Andrew D. Kane, David K. Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis*, 36:61–78, 2017.

[19] Davood Karimi and Septimiu E. Salcudean. Reducing the hausdorff distance in medical image segmentation with convolutional neural networks. *IEEE Transactions on Medical Imaging*, 39(2):499–513, 2020.

[20] Hoel Kervadec, Jihene Bouchtiba, Christian Desrosiers, Eric Granger, Jose Dolz, and Ismail Ben Ayed. Boundary loss for highly unbalanced segmentation. In *Medical Imaging with Deep Learning*, pages 285–296, 2019.

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, et al. Segment anything. In *IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

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

[23] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

[24] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15:654, 2024.

[25] Jun Ma, Feifei Li, and Bo Wang. U-Mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024.

[26] Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2015.

[27] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *International Conference on 3D Vision*, pages 565–571, 2016.

[28] Michael Moor, Oishi Banerjee, Zaid S. H. Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616:259–265, 2023.

[29] Andriy Myronenko. 3d MRI brain tumor segmentation using autoencoder regularization. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 311–320. Springer, 2018.

[30] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias 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.

[31] Sérgio Pereira, Adriano Pinto, Victor Alves, and Carlos A. Silva. Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Transactions on Medical Imaging*, 35(5):1240–1251, 2016.

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[33] Abhijit Guha Roy, Nassir Navab, and Christian Wachinger. Concurrent spatial and channel squeeze and excitation in fully convolutional networks. In *Medical Image Computing and Computer-Assisted Intervention*, pages 421–429. Springer, 2018.

[34] Saikat Roy, Gregor Koehler, Constantin Ulrich, Michael Baumgartner, Jens Petersen, Fabian Isensee, Paul F. Jaeger, and Klaus H. Maier-Hein. MedNeXt: Transformer-driven scaling of convnets for medical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 405–415. Springer, 2023.

[35] Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In *International Workshop on Machine Learning in Medical Imaging*, pages 379–387. Springer, 2017.

[36] Abdelrahman Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. UNETR++: Delving into efficient and accurate 3d medical image segmentation. In *International Workshop on Machine Learning in Medical Imaging*, pages 263–274. Springer, 2022.

[37] Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sébastien Ourselin, and M. Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 240–248. Springer, 2017.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[39] Guotai Wang, Wenqi Li, Sébastien Ourselin, and Tom Vercauteren. Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. In *International MICCAI Brainlesion Workshop*, pages 178–190. Springer, 2017.

[40] Haoyu Wang, Sheng Guo, Jin Ye, Zhongying Deng, Junlong Cheng, Tianze Li, Jian Chen, Yanzhou Su, Ziyan Huang, Yuxin Shen, et al. SAM-Med3D: Towards general-purpose segmentation models for volumetric medical images. *arXiv preprint arXiv:2310.15161*, 2024.

[41] Wenxuan Wang, Chen Chen, Meng Ding, Hong Li, Sen Yu, and Sen Zha. TransBTS: Multimodal brain tumor segmentation using transformer. In *Medical Image Computing and Computer-Assisted Intervention*, pages 109–119. Springer, 2021.

[42] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *European Conference on Computer Vision*, pages 3–19. Springer, 2018.

[43] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. CoTr: Efficiently bridging CNN and transformer for 3d medical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 171–180. Springer, 2021.

[44] Zhaohu Xing, Tianchen Ye, Yijun Yang, Guang Liu, and Lei Zhu. SegMamba: Long-range sequential modeling mamba for 3d medical image segmentation. *arXiv preprint arXiv:2401.13560*, 2024.

[45] Xiaomei Zhao, Yihong Wu, Guanghui Song, Zhenye Li, Yazhuo Zhang, and Yong Fan. Brain tumor segmentation using a fully convolutional neural network with conditional random fields. *Medical Image Analysis*, 43:98–111, 2018.

[46] Hong-Yu Zhou, J. Guo, Yizhe Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnFormer: Volumetric medical image segmentation via a 3d transformer. *IEEE Transactions on Image Processing*, 32:4036–4045, 2023.