UNIFIED MULTIMODAL COHERENT FIELD: SYNCHRONOUS SEMANTIC-SPATIAL-VISION FUSION FOR BRAIN TUMOR SEGMENTATION

Mingda Zhang, Yuyang Zheng, Ruixiang Tang, Jingru Qiu, Haiyan Ding*

School of Software, Yunnan University, Kunming 650500, China
{zhangmingda, zyy0906,tangruixiang, choujingru}@stu.ynu.edu.cn, dinghaiyan@ynu.edu.cn

ABSTRACT

Brain tumor segmentation requires accurate identification of hierarchical regions including whole tumor (WT), tumor core (TC), and enhancing tumor (ET) from multi-sequence Magnetic Resonance Imaging (MRI) images. Due to tumor tissue heterogeneity, ambiguous boundaries, and contrast variations across MRI sequences, methods relying solely on visual information or post-hoc loss constraints show unstable performance in boundary delineation and hierarchy preservation. To address this challenge, we propose the Unified Multimodal Coherent Field (UMCF) method. This method achieves synchronous interactive fusion of visual, semantic, and spatial information within a unified 3D latent space, adaptively adjusting modal contributions through parameter-free uncertainty gating, with medical prior knowledge directly participating in attention computation, avoiding the traditional "process-then-concatenate" separated architecture. On Brain Tumor Segmentation (BraTS) 2020 and 2021 datasets, UMCF+nnU-Net achieves average Dice coefficients of 0.8579 and 0.8977 respectively, with an average 4.18% improvement across mainstream architectures. By deeply integrating clinical knowledge with imaging features, UMCF provides a new technical pathway for multimodal information fusion in precision medicine.

***Index Terms*—** Brain tumor segmentation, multimodal fusion, medical imaging, deep learning, attention mechanism

1. INTRODUCTION

Brain tumor segmentation is one of the most challenging fundamental tasks in neuro-oncology. Its results directly impact clinical decisions including preoperative assessment, treatment monitoring, and radiotherapy planning. While multi-sequence MRI (including T1-weighted (T1), T1-weighted contrast-enhanced (T1ce), T2-weighted (T2), and Fluid-Attenuated Inversion Recovery (FLAIR)) provides complementary contrast information, actual lesions typically exhibit tissue heterogeneity, irregular morphology, and ambiguous boundaries. Additionally, the nested hierarchical relationship between ET, TC, and WT (ET⊂TC⊂WT) requires global consistency from models[1]. These factors collectively cause solutions relying on single-modal information or post-processing corrections to compromise on boundary delineation and hierarchy preservation[2].

Current research explores two main directions: designing powerful visual feature extraction networks and improving multimodal fusion strategies. However, Convolutional Neural Networks (CNNs) struggle with global feature relationships[3, 4], while Transformers incur high computational costs[5, 6]. Most fusion approaches simply stack multimodal images with equal weights, ignoring MRI sequences' differential sensitivity to specific pathological regions[1].

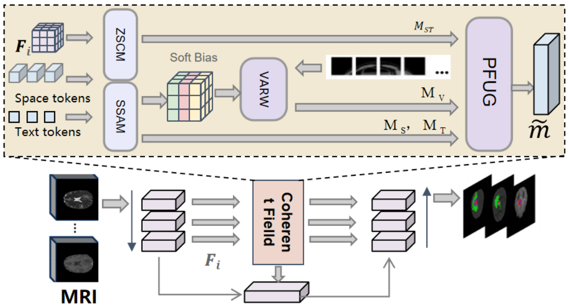

Fig. 1: UMCF overall architecture diagram

Recent attention-based methods still perform multimodal interaction after rather than during feature extraction, failing to leverage modality-specific associations (T1/T1ce for tumor core, FLAIR/T2 for edema) in attention computation[7].

This paper's main contributions are:

1. We propose the UMCF framework for synchronous fusion of visual, semantic, and spatial information within a unified 3D latent space, embedding medical priors directly into attention computation.
2. We design a parameter-free coordination mechanism including Zero-parameter Semantic-Spatial Channel Modulation (ZSCM), Parameter-Free Uncertainty Gating (PFUG), and convex optimization updates, improving cross-modal consistency without additional trainable parameters.

2. METHOD

2.1. Overall Architecture

UMCF changes the traditional "process-then-concatenate" pattern by constructing a unified 3D latent space where visual (V), semantic/text (T), and spatial prior (S) information interact in real-time. This plug-and-play fusion layer, inserted between encoder and decoder of U-Net architectures[3, 4], receives multi-sequence MRI images and clinical text descriptions as input. As shown in Figure 1, multi-sequence MRI images pass through an encoder to extract multi-scale features F_i. Multimodal fusion occurs at the bottleneck layer through coordinated processing by sub-modules: ZSCM, Semantic-Spatial Attention Modulation(SSAM), Visual Attention Read-Write with Medical Priors(VARW), and PFUG, producing segmentation results that satisfy medical hierarchical relationships.

2.2. Data Preprocessing and Feature Extraction

Four MRI sequences (T1, T1ce, T2, FLAIR) undergo registration, intensity normalization, and resampling to form $X \in \mathbb{R}^{H\times W\times D\times C}$. Clinical text is parsed using medical NLP tools and encoded via BiomedCLIP or ClinicalBERT[7, 10]. A U-Net encoder produces multi-scale feature pyramid $\{E_s\}_{s=1}^{L}$. The bottleneck feature $B = E_L$ is projected to d dimensions through $1\times1\times1$ convolution, forming initial latent field $F^{(0)} \in \mathbb{R}^{H_b\times W_b\times D_b\times d}$. All feature vectors are L2-normalized to the unit sphere, using cosine similarity $\mathrm{sim}(a, b) = a^\top b$ with temperature parameter $\tau > 0$.

2.3. Multimodal Information Encoding and Coordinated Fusion Mechanism

Visual Token Construction. Visual tokens aggregate features from local image regions:

$$V_i = \frac{\frac{1}{|P_i|}\sum_{x\in P_i} E(x)}{\left\|\frac{1}{|P_i|}\sum_{x\in P_i} E(x)\right\|_2}\,, \tag{1}$$

where P_i indexes the i-th cubic block (e.g., $8\times8\times8$ voxels) at bottleneck resolution. Average pooling within each block provides noise-robust region representations, while normalization ensures all visual tokens lie on the unit sphere for consistent similarity computation.

Semantic Token Construction. Medical concepts from clinical text are converted to semantic vectors aligned with visual features:

$$T_j = \frac{\frac{1}{|\mathcal{P}_j|}\sum_{w\in\mathcal{P}_j} e(w)}{\left\|\frac{1}{|\mathcal{P}_j|}\sum_{w\in\mathcal{P}_j} e(w)\right\|_2}\,, \tag{2}$$

where each medical phrase $\mathcal{P}_j$ (e.g., "ring enhancement", "central necrosis") obtains word embeddings $e(w)$ through pre-trained medical text encoders[7, 10]. Multiple word vectors within a phrase are merged via average pooling then normalized. The semantic prototype $\bar{T}$, computed as the equal-weight average of all semantic tokens, represents the overall semantic features of the current case.

Spatial Token Construction. Spatial prior information captures tumor position, morphology, and topology from current segmentation probability maps $P_c(\cdot)$ where $c \in \{\mathrm{ET}, \mathrm{TC}, \mathrm{WT}\}$:

$$\mu_c = \frac{\sum_{x\in\Omega} P_c(x)\, x}{\sum_{x\in\Omega} P_c(x)}\,, \tag{3}$$

$$\Sigma_c = \frac{\sum_{x\in\Omega} P_c(x)\,(x-\mu_c)(x-\mu_c)^\top}{\sum_{x\in\Omega} P_c(x)}\,, \tag{4}$$

$$(\lambda_1^c \ge \lambda_2^c \ge \lambda_3^c) = \mathrm{eig}(\Sigma_c), \tag{5}$$

$$\overline{D}_c = \frac{1}{|\Omega|}\sum_{x\in\Omega} \mathrm{SDT}_c(x), \tag{6}$$

$$S_c^{\mathrm{hier}} = \frac{\left[\mu_c^\top,\ \lambda_1^c, \lambda_2^c, \lambda_3^c,\ \overline{D}_c\right]^\top}{\left\|\left[\mu_c^\top,\ \lambda_1^c, \lambda_2^c, \lambda_3^c,\ \overline{D}_c\right]^\top\right\|_2}\,. \tag{7}$$

These statistics jointly describe tumor spatial characteristics: centroid μ_c indicates tumor position, eigenvalues $\lambda_{1,2,3}^c$ of covariance matrix Σ_c reflect tumor extension along principal directions, and the average Signed Distance Transform (SDT) $\overline{D}_c$—which measures each voxel's signed distance to the nearest boundary—characterizes boundary thickness and inside-outside relationships[11, 12]. After concatenating and normalizing these features, we obtain hierarchical structure token S_c^{hier}. Additional topological features (neighborhood smoothness, boundary gradients, surface-to-volume ratio) form the complete spatial token set $\{S_k\}$, with spatial prototype $\bar{S}$ as their average.

With token representations established, we construct their interaction mechanism. First, semantic information requires spatial "grounding" through the semantic field:

$$\phi_T(x) = \sigma\left(\frac{\mathrm{sim}\big(F(x),\ \bar{T}\big)}{\tau}\right)\,. \tag{8}$$

This semantic field ϕ_T acts as a soft spatial attention map, evaluating consistency between each voxel position and overall semantics. High-response regions indicate strong alignment between visual features and clinical descriptions, guiding subsequent attention mechanisms.

Based on this semantic field foundation, UMCF achieves deep multimodal fusion through four coordinated modules (Figure 2). These modules transform tokens from three modalities into four complementary message streams (m_S, m_T, m_V, m_{ST}), ultimately fused into unified voxel representation $\tilde{m}$.

Visual Attention Read-Write with Medical Priors (VARW). Attention incorporates semantic and spatial biases:

$$\alpha_{x,i}^{V} = \frac{\exp\big(\big[\mathrm{sim}(F(x), V_i) + \mu_V(x,i)\big]/\tau\big)}{\sum_p \exp\big(\big[\mathrm{sim}(F(x), V_p) + \mu_V(x,p)\big]/\tau\big)}, \tag{9}$$

$$m_V(x) = \sum_i \alpha_{x,i}^{V}\, V_i, \tag{10}$$

$$\mu_V(x,i) = \log\big(1 + \phi_T(x)\big) \ - \ r_{\mathrm{hier}}(x) \ - \ r_{\mathrm{topo}}(x)\,. \tag{11}$$

The bias μ_V includes: semantic encouragement $\log(1 + \phi_T)$ with logarithmic scaling, hierarchical penalty r_{hier} for ET⊂TC⊂WT violations[1], and topological penalty r_{topo} for discontinuous boundaries. Thus $m_V(x)$ represents medically-constrained visual evidence.

Semantic-Spatial Attention Modulation (SSAM). Modality-specific messages aggregate relevant tokens:

$$m_q(x) = \sum_{z\in\mathcal{I}_q} \mathrm{softmax}_z\left(\frac{\mathrm{sim}\big(F(x),\ z\big)}{\tau}\right) z\,,\ q \in \{T, S\}\,. \tag{12}$$

Semantic messages ($q = T$) soft-select relevant medical concepts, while spatial messages ($q = S$) combine position, scale, and boundary information based on feature similarity.

Zero-parameter Semantic-Spatial Channel Modulation (ZSCM). Cross-modal synergy through element-wise multiplication:

$$m_{TS}(x) = \big(\bar{T} \odot \bar{S}\big) \odot F(x)\,. \tag{13}$$

Dimensions activated by both semantic and spatial prototypes are enhanced, while conflicting responses are suppressed, filtering features endorsed by multiple modalities without trainable parameters.

Parameter-Free Uncertainty Gating (PFUG). Four message streams are adaptively fused based on reliability:

$$\tilde{m}(x) = \sum_{q\in\{V,T,S,TS\}} \frac{\exp\big(-u_q(x)\big)}{\sum\limits_{p\in\{V,T,S,TS\}} \exp\big(-u_p(x)\big)}\, m_q(x)\,. \tag{14}$$

Uncertainty measures: u_V uses prediction entropy (ambiguity indicator), u_T measures text-vision inconsistency, u_S evaluates continu-

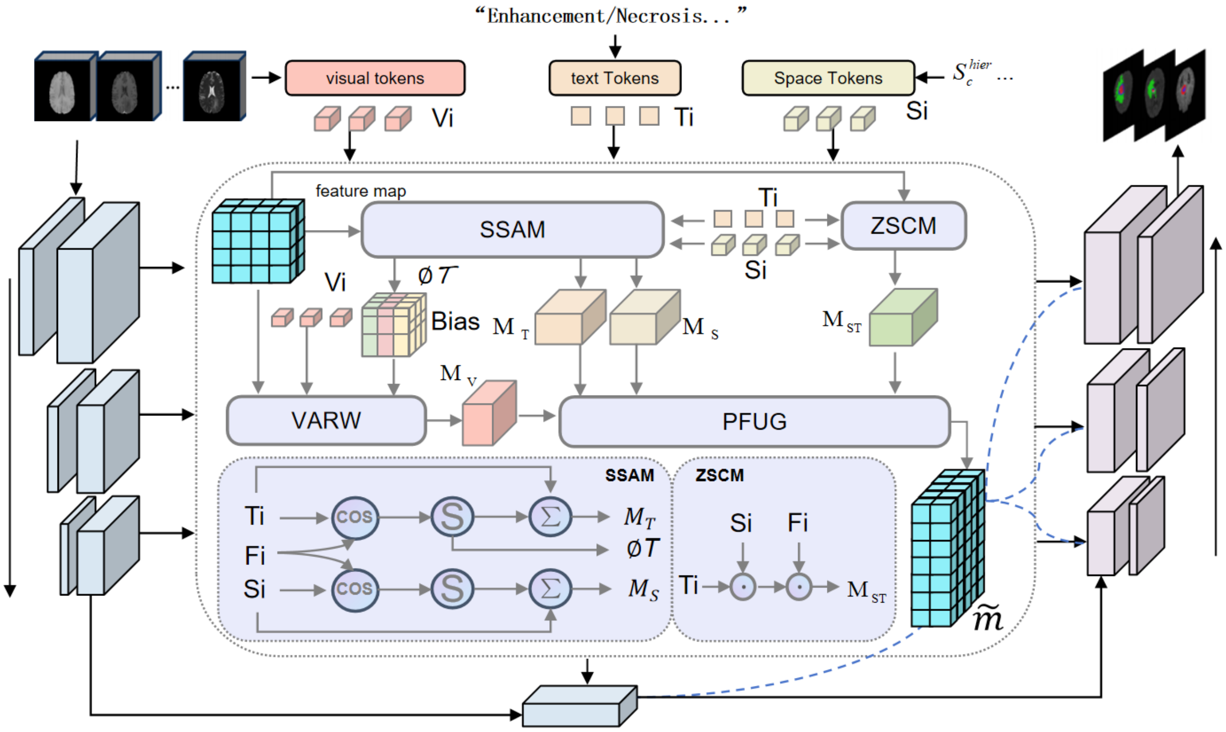

Fig. 2: UMCF core module detail diagram. Shows the semantic-spatial collaboration mechanism of ZSCM/SSAM, visual attention read-write of VARW, and uncertainty-weighted fusion process of PFUG. Visual tokens (Vi), spatial tokens (Si), and text tokens (Ti) participate in attention calculation through soft bias mechanism, producing multi-path messages that ultimately fuse into voxel consensus ($\tilde{m}$).

ity residuals, and u_{TS} averages semantic-spatial uncertainties. Normalized to $[0, 1]$, these weights allow relying on semantic-spatial priors in ambiguous regions while prioritizing visual evidence in clear regions.

2.4. Synchronous Convex Optimization Update

The fused message $\tilde{m}(x)$ integrates with current field representation via:

$$F^{(t+1)}(x) = (1 - \lambda)\, F^{(t)}(x) + \lambda\, \tilde{m}(x)\,. \tag{15}$$

With $\lambda \in (0, 1)$, this convex combination ensures convergence without oscillation. After 2-4 iterations with channel-wise renormalization, the converged field $F^{\star}$ passes to the decoder, producing full-resolution segmentation maps P through layer-wise upsampling.

3. EXPERIMENTS AND RESULTS

3.1. Datasets and Experimental Setup

We evaluate UMCF on BraTS 2020 (369 training, 125 validation) and BraTS 2021 (1251 training, 219 validation) benchmark datasets[13, 9]. Each case includes four co-registered MRI sequences (T1, T1ce, T2, FLAIR) at 1×1×1mm³ resolution with expert annotations for three nested hierarchical structures: enhancing tumor (ET), tumor core (TC), and whole tumor (WT), satisfying ET⊂TC⊂WT[1].

Experiments utilize NVIDIA A100 GPUs with Dice coefficients as the primary metric. The loss combines soft Dice and weighted cross-entropy[14, 15]. UMCF integrates as a plug-and-play module into 3D U-Net and nnU-Net architectures[4, 8] with base channels 32, latent dimension $d = 256$, using AdamW optimizer with OneCycleLR scheduling[16, 17]. Data augmentation includes rotation (±15°), flipping (50%), Gamma correction (0.8-1.2), and modality-specific intensity enhancement. Note that due to differing calculation standards for HD95 distance in different literature (e.g., using surface-distance library or MedPy library) [18, 19], to ensure fairness, this paper does not adopt HD95 metric in method comparisons.

3.2. Comparison Experiments

To evaluate UMCF's performance improvement and verify its generalization ability across different data scales and years, we design comprehensive comparative experiments. We select representative methods from 2020 to 2025, covering BraTS competition winning solutions, recently proposed innovative architectures, and methods introducing multimodal information. Table 1 shows performance comparisons of various methods on BraTS 2020 and 2021 validation set.

The results demonstrate UMCF's architecture-agnostic nature, with consistent 4% performance improvements when integrated with both nnU-Net and simpler 3D U-Net architectures, validating its effectiveness as a plug-and-play module that enhances different backbone networks without architecture-specific modifications. Notably, the most significant gains occur in smaller, ambiguous regions, with TC improving by 6.95% and ET by 7.38%, confirming that multimodal fusion particularly benefits difficult cases where visual information alone is insufficient. Furthermore, while

Table 1: Dice coefficients comparison of different methods on BraTS 2020 and BraTS 2021 datasets

BraTS 2020						BraTS 2021					
Method	Year	Avg	WT	TC	ET	Method	Year	Avg	WT	TC	ET
UMCF + nnU-Net (ours)	2025	**0.8579**	**0.9110**	0.8668	0.7958	UMCF + nnU-Net (ours)	2025	**0.8977**	0.9289	0.9066	0.8577
CLIP-UNet [20]	2025	0.8567	0.8994	**0.8709**	0.8005	Two-branch SR-Net [25]	2025	0.8970	0.9105	0.8930	**0.8861**
nnU-Net (Winner) [8]	2020	0.8535	0.8955	0.8506	**0.8203**	DeepSeg Ensemble (Winner) [26]	2021	0.8960	**0.9294**	0.8788	0.8803
UMCF + 3D U-Net (ours)	2025	0.8505	0.9048	0.8621	0.7846	SegResNet [27]	2025	0.8910	0.9170	0.8960	0.8610
FCFDiff-Net [21]	2025	0.8380	0.8980	0.8300	0.7860	UMCF + 3D U-Net (ours)	2025	0.8874	0.9012	0.9121	0.8488
BU-Net-ASPP [22]	2023	0.8344	0.9073	0.8159	0.7800	BU-Net-ASPP-EVO [22]	2023	0.8740	0.9187	0.8594	0.8434
Modified U-Net [23]	2023	0.8310	0.9050	0.8070	0.7810	3D ResUNet [28]	2022	0.8630	0.8190	**0.9196**	0.8503
LATUP-Net [24]	2024	0.8197	0.8841	0.8382	0.7367	RAL-Net [29]	2022	0.8584	0.8138	0.9076	0.8538
nnU-Net (baseline, ours)	2025	0.8175	0.8784	0.8498	0.7243	nnU-Net (baseline, ours)	2025	0.8522	0.9096	0.8477	0.7993
3D U-Net (baseline, ours)	2025	0.8088	0.8887	0.8099	0.7277	3D U-Net (baseline, ours)	2025	0.8477	0.8601	0.8716	0.8112

Table 2: Ablation study results for UMCF components (based on nnU-Net backbone)

Configuration	BraTS 2020				BraTS 2021			
	Avg	WT	TC	ET	Avg	WT	TC	ET
UMCF	**0.8579**	**0.9110**	**0.8668**	**0.7958**	**0.8977**	**0.9289**	**0.9066**	**0.8577**
w/o m_V	0.8377	0.8989	0.8422	0.7721	0.8835	0.9166	0.8861	0.8478
w/o m_T	0.8421	0.9001	0.8473	0.7789	0.8762	0.8979	0.8842	0.8464
w/o m_{ST}	0.8395	0.8986	0.8418	0.7780	0.8692	0.8943	0.8789	0.8344
w/o m_S	0.8372	0.8987	0.8416	0.7713	0.8646	0.9125	0.8652	0.8161
w/o PFUG	0.8317	0.8977	0.8422	0.7551	0.8638	0.8990	0.8753	0.8171
Pairwise fusion	0.8320	0.8921	0.8430	0.7609	0.8570	0.8761	0.8725	0.8224
Baseline	0.8175	0.8784	0.8498	0.7243	0.8522	0.9096	0.8477	0.7993

CLIP-UNet also incorporates text information for guidance, UMCF achieves superior performance by embedding semantic bias directly into the attention computation process, enabling real-time multimodal interaction throughout the network rather than relying on post-hoc feature concatenation, thus achieving deeper semantic-visual synergy throughout the segmentation process.

3.3. Ablation Study

To understand the specific contributions of each component in the UMCF framework and validate the rationality of design decisions, we conduct systematic ablation studies. Experiments observe performance changes by progressively removing or replacing modules, including four-path message passing mechanism, parameter-free uncertainty gating, and synchronous fusion strategy. Table 2 shows the performance of various configurations.

To analyze component contributions, we conduct ablation studies using nnU-Net backbone. Table 2 presents the quantitative impact of removing individual modules. Results reveal critical insights: removing spatial message m_S causes maximum degradation (average 3.31% decrease on BraTS 2021, with ET decreasing 4.85%), confirming its essential role in capturing tumor morphology and maintaining hierarchical relationships[1]. Semantic message m_T contributes 2.15%, primarily benefiting TC/ET regions where clinical descriptions provide valuable guidance. The parameter-free m_{ST} achieves 2.85% improvement through channel-level coordination. PFUG enables 3.4% gain through uncertainty-adaptive fusion. Most critically, replacing synchronous fusion with pairwise fusion—where modalities are combined two at a time rather than simultaneously yields, demonstrating that simultaneous multimodal interaction within a unified latent space, rather than sequential pairwise processing, is fundamental to UMCF's success.

3.4. Visualization Analysis

Figure 3 illustrates UMCF's segmentation on four representative cases, where blue, green, and orange represent WT, TC, and ET regions respectively. Results demonstrate accurate morphology reconstruction: WT regions (blue) completely envelope lesions without over-segmentation; TC regions (green) precisely capture irregular tumor cores; ET regions (orange), though small and scattered, are correctly identified. Notably, the segmentation boundaries present natural curved morphology without blocky artifacts or discontinuous breaks, with smooth transitions between the three sub-regions while maintaining proper nested relationships (ET⊂TC⊂WT). This confirms that the collaborative action of semantic field ϕ_T and medical bias μ_V successfully achieves spatial modulation of visual attention through medical prior knowledge, enabling UMCF to produce accurate and consistent segmentations across tumors of varying sizes and morphologies.

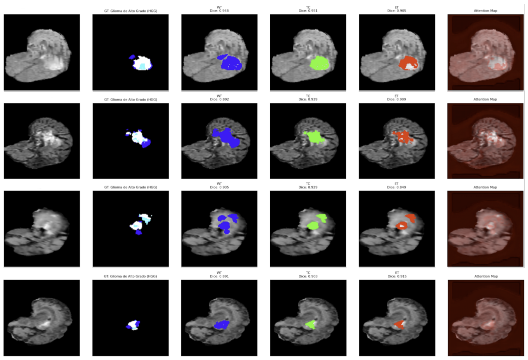

Fig. 3: UMCF segmentation visualization results

4. CONCLUSION

UMCF achieves synchronous fusion of visual, semantic, and spatial information within a unified 3D latent space for brain tumor segmentation. Its core innovations include semantic field-guided localization, medical knowledge-constrained attention, parameter-free coordination, and uncertainty-adaptive fusion. Experimental results show significant boundary quality improvement. As a plug-and-play module, UMCF provides an effective solution for multimodal medical image analysis.

5. REFERENCES

[1] MICCAI BraTS 2018, "MICCAI BraTS 2018: tasks and label definitions," 2018. [Online]. Available: https://www.med.upenn.edu/sbia/brats2018/tasks.html. Accessed: Sep. 16, 2025.

[2] G. Litjens, T. Kooi, B. Ehteshami Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60–88, Dec. 2017.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," in Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI), vol. 9351, Munich, Germany, 2015, pp. 234–241.

[4] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: learning dense volumetric segmentation from sparse annotation," in Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI), Cham, Switzerland, 2016, pp. 424–432.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: transformers for image recognition at scale," in Proc. Int. Conf. Learning Representations (ICLR), 2021.

[6] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. Yuille, and Y. Zhou, "TransUNet: transformers make strong encoders for medical image segmentation," arXiv:2102.04306, 2021.

[7] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong, A. Tupini, Y. Wang, M. Mazzola, S. Shukla, L. Liden, J. Gao, A. Crabtree, B. Piening, C. Bifulco, M. P. Lungren, T. Naumann, S. Wang, and H. Poon, "Large-scale domain-specific pretraining for biomedical vision–language processing (BiomedCLIP)," arXiv:2303.00915, 2023.

[8] F. Isensee, P. F. Jaeger, S.-A. A. Kohl, J. Petersen, and K.-H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," Nature Methods, vol. 18, pp. 203–211, 2021.

[9] J. Chen, A. Zhou, H. Rozycki, et al., "Automatic brain tumor segmentation (BraTS 2021)," arXiv:2111.00742, 2021.

[10] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical BERT embeddings," in Proc. Clinical NLP Workshop (ACL), Minneapolis, MN, USA, 2019, pp. 72–78.

[11] C. R. Maurer Jr., R. Qi, and V. Raghavan, "A linear-time algorithm for computing exact Euclidean distance transforms of binary images in arbitrary dimensions," IEEE Trans. Pattern Anal. Mach. Intell., vol. 25, no. 2, pp. 265–270, Feb. 2003.

[12] P. F. Felzenszwalb and D. P. Huttenlocher, "Distance transforms of sampled functions," Theory of Computing, vol. 8, no. 1, pp. 415–428, 2012.

[13] Z. Zhao, et al., "The brain tumor segmentation challenge 2020," arXiv:2011.03188, 2020.

[14] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: fully convolutional neural networks for volumetric medical image segmentation," in Proc. 4th Int. Conf. 3D Vision (3DV), Stanford, CA, USA, Oct. 2016, pp. 565–571.

[15] J. Lorenzo, A. Gzh, L. Toth, and C. Pal, "On the optimal combination of cross-entropy and soft Dice losses for lesion segmentation," arXiv:2209.06078, 2022.

[16] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in Proc. Int. Conf. Learning Representations (ICLR), 2019.

[17] L. N. Smith and N. Topin, "Super-convergence: very fast training of neural networks using large learning rates," arXiv:1708.07120, 2017.

[18] DeepMind, "surface-distance," GitHub repository. [Online]. Available: https://github.com/google-deepmind/surface-distance. Accessed: Sep. 16, 2025.

[19] MedPy, "Clarification on HD95 implementation," GitHub issue #138, Jul. 2025. [Online]. Available: https://github.com/loli/medpy/issues/138. Accessed: Sep. 16, 2025.

[20] M. Zhang, "A brain tumor segmentation method based on CLIP and 3D U-Net with cross-modal semantic guidance and multi-level feature fusion," arXiv:2507.09966, 2025.

[21] Y. Pang, et al., "FCFDiff-Net: full-conditional feature diffusion embedded network for 3D brain tumor segmentation," Quantitative Imaging in Medicine and Surgery, vol. 15, no. 5, pp. 4217–4234, 2025.

[22] R. Yousef, S. Khan, G. Gupta, B. M. Albahlal, S. A. Alajlan, and A. Ali, "Bridged-U-Net-ASPP-EVO and deep learning optimization for brain tumor segmentation," Diagnostics, vol. 13, no. 16, p. 2633, 2023.

[23] R. Yousef, S. Khan, G. Gupta, T. Siddiqui, B. M. Albahlal, S. A. Alajlan, and M. A. Haq, "U-Net-based models towards optimal MR brain image segmentation," Diagnostics, vol. 13, no. 9, p. 1624, May 2023.

[24] E. J. Alwadee, X. Sun, Y. Qin, and F. C. Langbein, "LATUP-Net: a lightweight 3D attention U-Net with parallel convolutions for brain tumor segmentation," Computers in Biology and Medicine, vol. 184, p. 109353, 2025.

[25] Z. Jia, H. Zhu, J. Zhu, and P. Ma, "Two-branch network for brain tumor segmentation using attention mechanism and super-resolution reconstruction," Computers in Biology and Medicine, vol. 157, p. 106751, May 2023.

[26] U. Baid, et al., "The RSNA–ASNR–MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification," arXiv:2107.02314, 2021.

[27] M. J. Cardoso, et al., "MONAI: an open-source framework for deep learning in healthcare," arXiv:2211.02701, 2022.

[28] L. Pei and Y. Liu, "Multimodal brain tumor segmentation using a 3D ResUNet in BraTS 2021," in Int. MICCAI Brainlesion Workshop, Berlin/Heidelberg, Germany: Springer, 2022, pp. 315–323.

[29] H. Peiris, Z. Chen, G. Egan, and M. Harandi, "Reciprocal adversarial learning for brain tumor segmentation: a solution to BraTS challenge 2021 segmentation task," arXiv:2201.03777, 2022.