2025 World Journal of Innovation and Modern Technology, Vol. 8, Issue 7 (Jul)
ISSN 2682-5910
DOI: 10.53469/wjimt.2025.08(07).13

Advancements in Multimodal Image Fusion and Deep Learning-based Segmentation Techniques for Gliomas: A Comprehensive Review

Lirong Chen[1], Liqiang Wang[1,2], Wei Wang[3]

[1]School of Electronic Engineering, Tianjin University of Technology and Education, Tianjin 300222, China
[2]Tianjin Engineering Research Center of Fieldbus Control Technology, Tianjin 300202, China
[3]Xuanwu Hospital of Capital Medical University, No. 45 Changchun Street, Xicheng District, Beijing 100053, China

Abstract: ***This paper reviews recent developments in deep learning techniques for multimodal image fusion and segmentation of brain tumors. Gliomas, the most common tumors of the central nervous system in adults, require accurate image segmentation to support effective diagnosis and treatment. Multimodal image fusion integrates information from different imaging modalities, offering a more comprehensive and precise characterization of tumors. In this review, we introduce the characteristics of gliomas, outline preprocessing and fusion methods for multimodal images, and summarize commonly used deep learning models for glioma segmentation. We also highlight the benefits of integrating attentional mechanisms and multiscale features into deep learning architectures. In addition, current evaluation metrics and publicly available datasets are discussed. Finally, we address key challenges such as data management, protection of surrounding organs, and model interpretability, aiming to provide researchers with a valuable reference for future studies in multimodal brain tumor segmentation.***

Keywords: Deep learning, Glioma, Medical imaging, Multimodal, Segmentation.

1. INTRODUCTION

Gliomas are the most common primary tumors occurring in the brain and spinal cord of adults, accounting for approximately 30% of all central nervous system (CNS) neoplasms. According to the World Health Organization (WHO) classification, gliomas are divided into grades I to IV based on biological behavior and growth potential. Grades I and II are classified as low-grade gliomas (LGG) with a median survival of approximately 5.5 years, while grades III and IV are considered high-grade gliomas (HGG) with a significantly shorter median survival of around 1.1 years [1,2]. Conventional diagnostic approaches primarily rely on imaging modalities such as magnetic resonance imaging (MRI) and computed tomography (CT), which remain the clinical gold standard alongside histopathological confirmation. However, these methods are often limited by inter-observer variability and suboptimal diagnostic accuracy. Therefore, exploring advanced imaging technologies and automated analysis tools is of great clinical relevance [3].

Different imaging techniques such as CT, MRI, and positron emission tomography (PET) provide complementary information about tumor characteristics, as shown in Figure 1 [4]. CT offers rapid imaging and is suitable for initial screening but exhibits low soft tissue contrast [5]. MRI provides high soft tissue contrast without radiation exposure, making it ideal for detailed brain imaging [6-9]. PET reveals functional and metabolic information but lacks fine anatomical detail [10]. PET/CT combines metabolic and structural imaging, while PET/MR integrates functional imaging with high soft tissue contrast [11]. Moreover, advanced MRI sequences, including magnetic resonance spectroscopy (MRS), magnetic resonance fingerprinting (MRF), chemical exchange saturation transfer (APT), diffusion-weighted imaging (DWI), diffusion tensor imaging (DTI), diffusion kurtosis imaging (DKI), and blood oxygenation level-dependent imaging (BOLD), further enable comprehensive assessment of gliomas [12]. Compared with single-modality imaging, multimodal imaging provides more comprehensive tumor characterization, laying a foundation for subsequent automatic analysis using advanced machine learning techniques.

Recent advances in machine learning, particularly deep learning, have facilitated the automatic delineation and diagnosis of gliomas from multimodal images. However, accurate segmentation of glioma regions remains challenging. Firstly, extensive data preprocessing and alignment are required due to artifacts in the acquired images

[13]. Secondly, the highly variable shapes and sizes of gliomas complicate treatment planning, and even minor deviations during segmentation could risk damaging critical brain structures such as the thalamus [14]. Thus, precise boundary definition and segmentation are critical research directions. Several review articles have summarized the application of deep learning in medical image analysis. Litjens et al. [15] discussed basic deep learning architectures and key techniques but did not explore multimodal image fusion methods. Bernal et al. [11] focused on applying deep convolutional neural networks (CNNs) in brain MRI analysis without emphasizing multimodal fusion strategies. Zhou et al. [16] briefly introduced multimodal fusion but did not investigate the performance and applicability of different strategies in specific segmentation tasks.

In this paper, we conduct a comprehensive literature review using databases such as PubMed, Medline, Embase, and Cochrane Library. We categorize and summarize multimodal fusion strategies, review recent advances, outline available glioma datasets, and briefly evaluate the performance of typical network architectures based on relevant metrics. Finally, we discuss current challenges and propose future research directions.

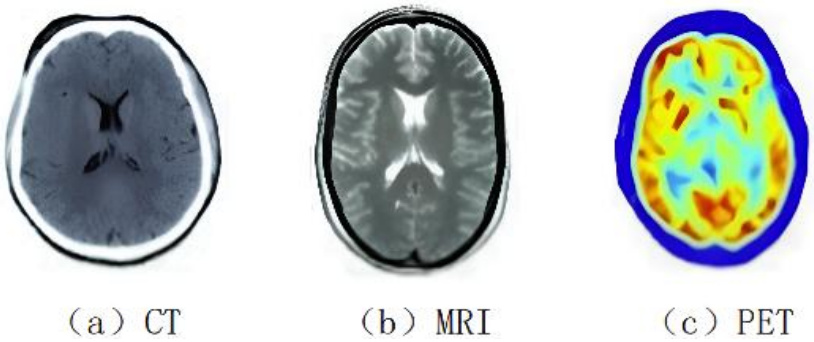

Figure 1: Conventional medical imaging diagram.

2. MULTIMODAL GLIOMA SEGMENTATION

Multimodal medical image segmentation combines data from different imaging modalities to improve the accuracy of segmenting regions of interest (e.g., tumors) in medical images. This technique utilizes the unique advantages of each imaging technique. The multimodal image segmentation process is shown in Figure 2.

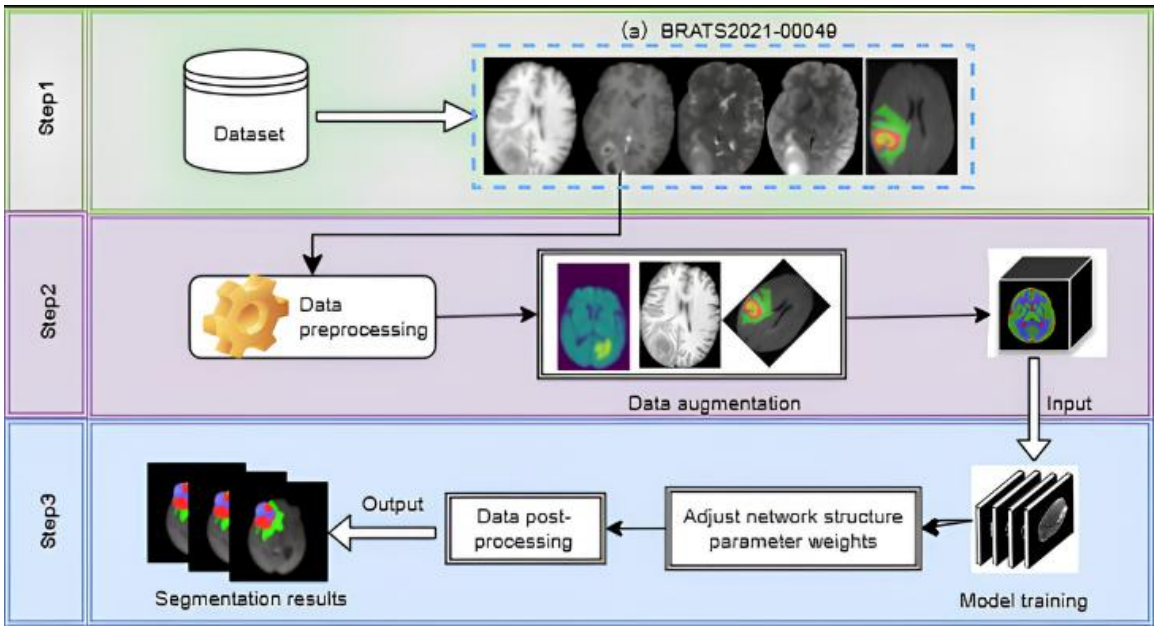

Figure 2: Multimodal glioma image segmentation process

2.1 Public Datasets and Preprocessing

2.1.1 Glioma imaging datasets

To advance multimodal glioma segmentation, publicly available datasets and standardized preprocessing techniques are essential. Datasets used for glioma research typically consist of CT, MRI, and PET images accompanied by corresponding pathological labels. These datasets provide researchers with rich imaging and clinical information to facilitate the development and validation of various algorithms. Among them, the Medical Segmentation Decathlon (MSD) dataset [17] is widely utilized, particularly for brain tumor segmentation tasks. It comprises 484 multi-modal MRI scans, each containing four imaging modalities Fluid Attenuated Inversion Recovery (FLAIR), T1-weighted (T1w), contrast-enhanced T1-weighted (T1gd), and T2-weighted (T2w). The data were collected from 19 different medical institutions and include a subset of the data used in the 2016 and 2017 Brain Tumor Segmentation (BraTS) challenges [18]. The statistically most commonly used BraTS 2013–2024 and MSD multi-modal brain tumor dataset, provided by the International Medical Image Computing and Computer Assisted Intervention Society (MICCAI), is summarized in Table 1.

Table 1: Summary of Glioma Datasets

Dataset name	Data volume			type	Image size
	Training	*Validation*	*Testing*	*MRI*	dataset type
BraTS2024	3600	-	900	T1 T1Gd T2 Flair	240x240 x155 XX.nii.gz
BraTS2023	1251	-	219		
BraTS2022	484	-	266		
BraTS2021	1251	219	570		
BraTS2020	369	125	166		
BraTS2019	335	125	166		
BraTS2018	285	66	191		
BraTS2017	285	46	146		
BraTS2016	200	-	191		
BraTS2015	200	-	74		
BraTS2014	200	-	38		
BraTS2013	35	-	25		
MSD	484	-	266		

Taking the BraTS2020 multimodal brain tumor dataset as an example (datasets from other years follow a similar structure), this dataset consists of 369 training cases and 125 validation cases. The training set includes both HGG and LGG, while the glioma grades in the validation set are not disclosed. Each case includes four MRI modalities: T1w, T2w, T1gd, and FLAIR. The training dataset provides corresponding ground truth segmentation labels, including background (label 0), non-enhancing tumor region (NET, label 1), edema region (ED, label 2), and enhancing tumor region (ET, label 4). For visualization, the segmentation labels are commonly represented using color codes: green for ED (label 2), red for NET (label 1), yellow for ET (label 4), and black or transparent for the background (label 0). The final segmentation task focuses on three tumor subregions: the whole tumor (WT), the tumor core (TC), and the enhancing tumor (ET). The WT region includes all tumor-related components (NET, ED, and ET), corresponding to the union of labels 1, 2, and 4. The TC region comprises NET and ET (labels 1 and 4), and the ET region includes only the enhancing tumor component (label 4). Most of the time, a large number of labels used for training are not available for several reasons. Labeling datasets requires experts in the field, which is both expensive and time-consuming. The overfitting problem needs to be considered when training large neural networks from limited training data [19]. Data augmentation is a way to reduce overfitting and increase the amount and diversity of data. Large training data helps in algorithm generalization. It generates more diverse training samples by transforming the images in the training dataset (rotating, translating, scaling, flipping, distorting, and adding some noise, e.g., Gaussian noise). Both the original and created images are fed into the neural network. Various data enhancement techniques, such as random rotation, random scaling, random elastic deformation, gamma-corrected enhancement, and dynamic mirroring, can be utilized to solve the overfitting problem during the training process [20]. Among the common data enhancement methods are random angle rotation of the image, which helps the model adapt to different image orientations; random horizontal or vertical translation of the image, which enhances the robustness of the model to displacement changes; random scaling of the image, which helps the model deal with tumors at different scales; random horizontal or vertical flipping of the image, which increases the diversity of the training samples; and adding random noise to enhance the noise immunity of the model. Through these data enhancement techniques, a large amount of diverse training data can be generated to reduce the risk of model overfitting and improve the segmentation effect and model generalization ability.

2.1.2 Data preprocessing

Data preprocessing is an important step to ensure the effectiveness of model training and inference, including the following aspects.

1) Normalization: Due to variations in image acquisition, different brain MRI volumes exhibit varying intensity distributions. To mitigate grayscale inconsistencies across images and facilitate subsequent multimodal fusion and processing, intensity normalization is commonly employed in glioma segmentation studies. Among these techniques, z-score normalization is widely used: for training data, normalization is often performed around the tumor region when segmentation masks are unavailable, while for validation data, the entire volume is typically normalized. In such approaches, the mean and standard deviation are calculated over non-zero voxels within the relevant regions, ensuring that the normalized intensities have zero mean and unit variance.

2) Alignment: Alignment processing ensures that images of the same patient at different time points or different modalities are spatially aligned such that each pixel location represents the same anatomical structure in different images. Commonly used alignment methods include rigid alignment (translation and rotation) and non-rigid alignment (deformation).

3) Registration: The multimodal images are spatially and accurately aligned so that each pixel position represents the same anatomical location in different modal images, facilitating fusion and analysis. Commonly used alignment methods include feature point-based alignment, image intensity-based alignment, and machine learning-based alignment. In glioma research, alignment processing can ensure the spatial consistency of MRI, CT, and PET images and improve the accuracy of multimodal image fusion and segmentation [21].

2.2 Fusion Strategies for Multimodal Inputs

Multimodal medical image segmentation plays a crucial role in the segmentation of gliomas, as it combines information from different imaging modalities to improve the accuracy and robustness of tumor delineation. In early fusion strategies, images from multiple modalities are merged at the network's input layer. In contrast, late fusion strategies process each modality independently and combine results at a deeper stage in the network. The categorization of common strategies based on the fusion stage is shown in Figure 3.

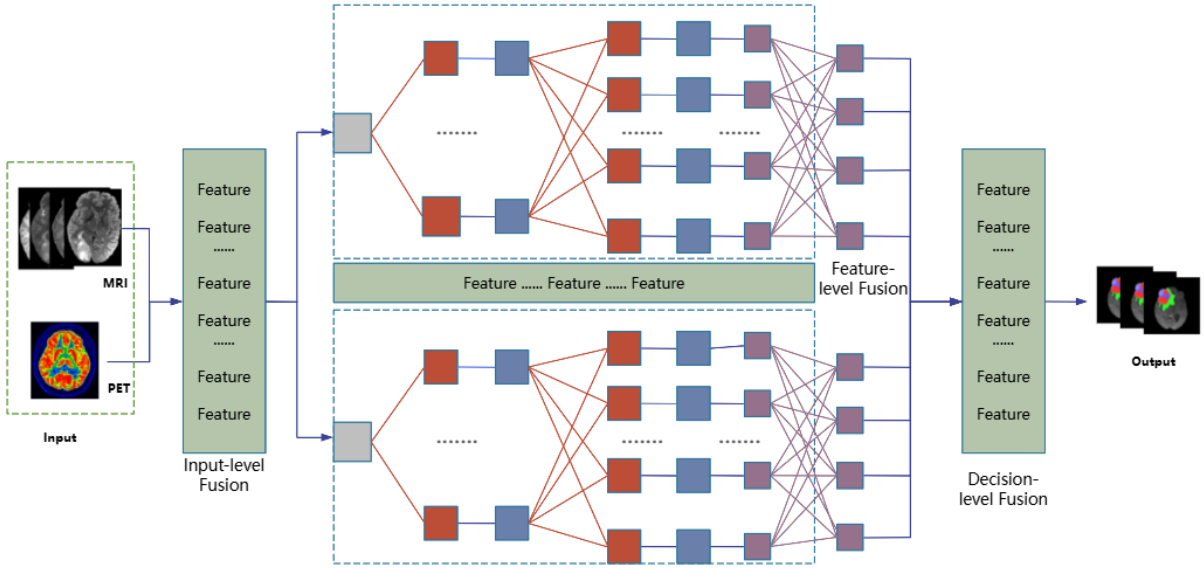

Figure 3: Multi-modal fusion network structure

2.2.1 Input-level fusion

Input-level fusion, also known as early fusion, combines multimodal medical images by stacking them along the channel dimension at the input stage of a convolutional neural network (CNN). This approach enables joint feature learning from the shallowest layers and is widely used for combining different MRI sequences (e.g., T1, T2, T1ce, FLAIR) or hybrid modalities such as PET/CT and PET/MRI. Well-established models like 3D U-Net and nnUNet have successfully adopted early fusion strategies in glioma segmentation tasks such as BraTS [1, 22]. Despite its simplicity and ability to preserve original image information, early fusion tends to ignore modality-specific importance, often leading to overfitting to redundant or noisy inputs. Furthermore, it is highly sensitive to missing or corrupted modalities, which limits its robustness.

2.2.2 Feature-level fusion

Feature-level fusion lies between early and late fusion, offering a compromise by first extracting modality-specific representations through separate branches and then combining them at intermediate stages of the network. This fusion is typically implemented via multi-stream architectures with independent encoders for each modality that merge at certain layers. Common fusion strategies include concatenation, element-wise summation, and attention-based mechanisms. This approach helps preserve unique modality characteristics while enabling effective cross-modality semantic integration. For instance, HyperDenseNet introduced by Dolz et al., integrates densely connected modality-specific pathways to support multilevel fusion [23], while ME-Net, proposed by Zhang et al., performs fusion during decoding to maximize the utility of modality-specific information [24]. Liu et al. further improved this by introducing channel-wise attention, allowing the model to adaptively emphasize more informative modalities [25].

2.2.3 Decision-level integration

Decision-level fusion, also referred to as late fusion, combines segmentation predictions from independently trained models, each focusing on a specific modality. Common techniques include probability map averaging, majority voting, and meta-classification. A key advantage of this strategy is its natural robustness to missing or degraded modalities, as each model operates independently. However, the lack of joint feature learning limits its ability to capture cross-modal interactions. Moreover, training and maintaining multiple networks increases computational and optimization complexity. Although decision-level fusion enhances robustness, it has gradually fallen out of favor as integrated, end-to-end architectures have become more prevalent and effective. For example, Sun et al. [26] proposed an ensemble-based approach in the context of the BraTS challenge, where multiple modality-specific networks were trained and their outputs were fused using a majority voting strategy, showing promising results in brain tumor segmentation.

2.2.4 Fusion under missing modalities

Handling missing modalities is a critical challenge in real-world medical imaging scenarios, where patients may undergo incomplete scanning due to cost, time, or clinical constraints. Existing strategies for dealing with missing MRI sequences can be broadly categorized into three main approaches: hetero-modal segmentation, missing data synthesis, and knowledge distillation. Hetero-modal segmentation methods, such as HeMIS, address this issue by replacing the missing modality with statistical representations (e.g., mean, variance) computed from the available data, thereby enabling model training under incomplete input conditions [27]. However, HeMIS assumes equal informativeness across all modalities, an assumption that may not hold in practice. Recent approaches have employed generative models, such as variational autoencoders (VAEs), to reconstruct missing modalities or infer shared latent representations. For example, Zhu et al. proposed the XLSTM-HVED model [28], which integrates a hetero-modal variational encoder-decoder framework with a Vision XLSTM module and a multi-task learning paradigm. This model enhances feature integration via a Self-Attention Variational Encoder (SAVE) and coordinates segmentation and modality reconstruction through a Squeeze-Fusion-Excitation Cross-Awareness (SFECA) module. It achieved state-of-the-art performance on the BraTS 2024 dataset, particularly under missing modality scenarios, highlighting the effectiveness of joint modeling and adaptive attention mechanisms in addressing such complex challenges. In terms of data synthesis approaches, Sharma et al. [29] explored the potential of generative adversarial networks (GANs) to synthesize missing MRI sequences. Knowledge distillation techniques aim to transfer knowledge from a teacher model trained on complete modality data to a student model operating on partial modalities. This strategy maintains segmentation performance while reducing reliance on complete data [30]. Originally developed for intra-domain CNN knowledge transfer, recent studies have extended this paradigm to inter-domain and cross-modal distillation [31]. Rahimpour et al. [32] proposed a cross-modal distillation method that leverages multi-sequence training data to enhance the segmentation performance of single-sequence CNNs.

2.3 Model Design and Optimization

2.3.1 Representative deep learning models

Commonly used deep learning models have significantly advanced the field of medical image segmentation. AlexNet is a deep convolutional neural network designed to process 2D image data, typically taking RGB images of size 224×224 pixels as input [33]. VGG improves performance by increasing the network depth, allowing for

more complex feature extraction [34]. Fully Convolutional Networks (FCNs) enable pixel-level predictions for images of arbitrary sizes through an end-to-end architecture, making them highly suitable for 2D and 3D medical image segmentation tasks. By removing the fully connected layers found in traditional CNNs, FCNs preserve the spatial resolution of input images, thereby enhancing segmentation accuracy [35]. GoogLeNet extracts features at multiple scales using a parallel architecture with convolution kernels of varying sizes. This design improves computational efficiency and reduces the number of parameters [36]. DenseNet, through its densely connected structure, enables feature reuse throughout the network. This approach not only decreases the number of parameters but also alleviates the vanishing gradient problem, thereby improving both performance and efficiency [37]. The DeepLab family of models further boosts segmentation accuracy by incorporating atrous (dilated) convolutions and Conditional Random Fields (CRFs) for post-processing [38]. Among them, DeepLabv3+ introduces a feature aggregation module to better integrate multi-scale features within an encoder-decoder architecture using dilated convolutions. Building upon FCN, U-Net introduces skip connections to merge high-level semantic information with low-level spatial details in the decoding phase. This design preserves information at every scale and improves feature map utilization, making U-Net particularly effective for biomedical segmentation tasks [39]. V-Net extends the U-Net architecture to 3D, specifically for volumetric medical imaging such as MRI and CT scans. Although this increases computational demands, it also enables the network to capture richer contextual information [40]. Recently, Transformers—originally developed for sequence modeling—have been adopted in medical imaging due to their ability to model long-range dependencies. Utilizing the self-attention mechanism, Transformers capture global contextual information effectively and are particularly well-suited for processing complex and multimodal medical image data [41, 42]. Each of these models offers unique strengths and advanced feature extraction capabilities, and they are widely employed in the development of modern medical image segmentation systems.

2.3.2 Loss functions and evaluation

The overall performance of a segmentation model depends not only on the network structure but also on the loss function [43]. The distribution of brain tumor regions and non-tumor regions makes the segmentation task inherently class-imbalanced, and the choice of an appropriate loss function for a given task has a great impact on the experimental results. The loss function is used to represent the degree of difference between the predicted and labeled values. During the training process, the model continuously fine-tunes the weights and biases of the network to minimize the loss function value and improve the performance of the model. An overview of commonly used loss functions for network models is provided in Table 2. The commonly used loss functions are not suitable for training the network. If these loss functions are used singularly, the training of the convolutional network will be dominated by non-tumor regions with more pixels, and smaller brain tumor regions will have a hard time learning their features, which will reduce the effectiveness of the network and lead to poor segmentation results. Therefore, in most cases, multiple loss functions can be used to adaptively weight the categories according to the specific task, or targeted to use loss functions based on their characteristics to speed up convergence.

2.3.3 Evaluation indicators

The current performance evaluation metrics for assessing model architectures are shown in Table 3.

Table 2: Commonly used loss functions for network models.

Loss Function	Formula	Describe	Proposer
Cross-Entropy Loss	$\text{Loss}_{\text{CEL}}(p_{i,k}, g_{j,k}) = -\sum_{i=1}^{n}\sum_{k=1}^{K} g_{i,k}\, log(p_{i,k})$	Applicable to multi-class classification tasks at the pixel level.	LeCun et al. [44]
Binary Cross-Entropy Loss	$\text{Loss}_{\text{BCEL}}(p_i, g_i) = -\frac{1}{n}\sum_{i=1}^{n}(g_i \log(p_i) + (1-g_i)\, log(1-p_i))$	Performs binary classification for each pixel (e.g., foreground/background).	Shelhamer et al. [35]
Weighted Cross-Entropy Loss	$\text{Loss}_{\text{WCEL}} = -\frac{1}{n}\sum_{i=1}^{n}(w_{g_i} g_i\, log(p_i) + w_{1-g_i}(1-g_i)\, log(1-p_i))$	Solve the class imbalance problem when classifying each pixel.	LeCun et al. [44]
Dice Loss	$\text{Loss}_{\text{DL}}(p_i, g_i) = 1 - \frac{2\sum_{i=1}^{n} p_i g_i + \varepsilon}{\sum_{i=1}^{n} p_i + \sum_{i=1}^{n} g_i + \varepsilon}$	Responsible for global brain tumor segmentation prediction	Milletari et al. [40]
Generalized Dice Loss	$\text{Loss}_{\text{GDL}}(p_{i,k}, g_{j,k}) = 1 - \frac{2\sum_{k=1}^{K} w_k \sum_{i=1}^{n} p_{i,k} g_{i,k}}{\sum_{k=1}^{K} w_k (\sum_{i=1}^{n} p_{i,k} + \sum_{i=1}^{n} g_{i,k})}$	Normalize the contribution of each category.	Sudre et al. [45]

Focal Loss	$\text{Loss}_{\text{FL}} = -\frac{1}{n}\sum_{i=1}^{n}((1-p_i)^{\gamma} g_i \log(p_i) + p_i^{\gamma}(1-g_i)\log(1-p_i))$	Solve the class imbalance problem.	Lin et al. [46]
Fusion Loss	$\text{Loss}_{Fusion} = \frac{1}{N}(\|\|F-S_1\|\|_F^2 + \alpha\|\|F-S_2\|\|_F^2) + \beta[(1-\text{SSIM}(F,S_1)) + (1-\text{SSIM}(F,S_2))]$	Guides multimodal feature fusion via intensity and structure preservation.	Liu et al. [47]
Jensen-Shannon Divergence	$\text{Loss}_{\text{JS}}(p,g) = \frac{1}{2}KL(g\|\|m) + KL(p\|\|m), m = \frac{1}{2}(p+g)$	Measures similarity between two distributions; used in deep supervision.	Englesson et al. [48]

Note: Let n denote the number of samples, P_i and g_i represent the predicted value and the ground truth, respectively. ε is the smoothing term to prevent division by zero, W_k denotes the weight of the k-th class of the i-th sample. γ is an adjustment parameter, typically set to 2. F represents the fused image, while S_1 and S_2 are the input modality images, α and β are weighting coefficients, usually set to 2. SSIM stands for the Structural Similarity Index, and $||\,.\,||_F$ denotes the Frobenius norm. KL(.||.) is the Kullback–Leibler divergence between two distributions.

Table 3: Evaluation indicators for glioma segmentation.

Evaluation index	Formula	Function description
Dice similarity coefficient	$\text{DSC} = \frac{2 \times \text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$	Calculate the similarity between two samples.
Hausdorff distance	$\text{HD}(P,G) = max\{\sup_{p\in P}\inf_{g\in G}\|\|p-g\|\|, \sup_{g\in G}\inf_{p\in P}\|\|g-p\|\|\}$	Measures the maximum distance between the boundary points of two sets
specificity	$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$	Reliability of the model in identifying normal tissue or background
Sensitivity/Recall	$\text{Sensiti}vity(Re\,c\,all) = \frac{TP}{TP+FN}$	The ability to detect actual existing lesions or targets to prevent missed detections
Accuracy	$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$	Measures the overall proportion of correct classification by the model. It is difficult to reflect when the categories are unbalanced.
F1-Score	$\text{F1-Score} = \frac{2 \times \text{Accuracy} \times \text{Recall}}{\text{Accuracy} + \text{Recall}}$	Balance the recall rate and precision rate to evaluate the overall performance of the model.
Intersection of Union	$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}$	Evaluate the overall performance of the segmentation model.
AUC	$\text{AUC} = S$	Evaluate the overall performance of the model under all possible classification thresholds.

Note: p and g represent points in the prediction area and the true value area, respectively. P and G represent two different sets, and sup and inf represent the maximum and minimum values in the set. ||.|| represents the Euclidean distance, and S represents the area under the ROC curve. TP, FP, FN, and TN denote true positive, false positive, false negative, and true negative, respectively.

Currently, deep learning algorithms for glioma and other related medical image segmentation tasks still face several critical challenges, including overfitting, class imbalance, and limited segmentation accuracy. To overcome these issues, researchers have proposed a range of optimization strategies, encompassing data augmentation [33], post-processing techniques, fusion strategies, and loss function design.

To alleviate overfitting, commonly adopted methods include data augmentation, regularization techniques, and early stopping. Data augmentation applies transformations such as rotation, translation, and scaling to training images [33], thereby generating more diverse samples and reducing the model's dependence on specific data distributions. Meanwhile, L2 regularization and Dropout are often used during training to control model complexity. Early stopping helps prevent overfitting by monitoring the performance on a validation set and halting training when improvement stagnates. To address class imbalance, weighted loss functions and resampling techniques are frequently utilized. The former assigns different loss weights to tumor and non-tumor regions, enabling the model to focus more on tumor structures during training. The latter balances class distribution using oversampling or undersampling, thus reducing the negative impact of data imbalance on model performance. To further enhance segmentation precision, researchers have designed more sophisticated network architectures and incorporated diverse fusion strategies to improve the model's feature representation capability.

Widely used deep learning architectures include Convolutional Neural Networks (CNNs), DeepLab, Generative Adversarial Networks (GANs), Transformers, and U-Net. These frameworks demonstrate strong capabilities in modeling complex structures and fine-grained details in medical images, while also offering excellent flexibility and scalability. On this basis, numerous studies have proposed optimized variants tailored to specific segmentation tasks. These models have become mainstream solutions in medical image analysis. A statistical overview of optimization strategies applied to different backbone networks is shown in Figure 4.

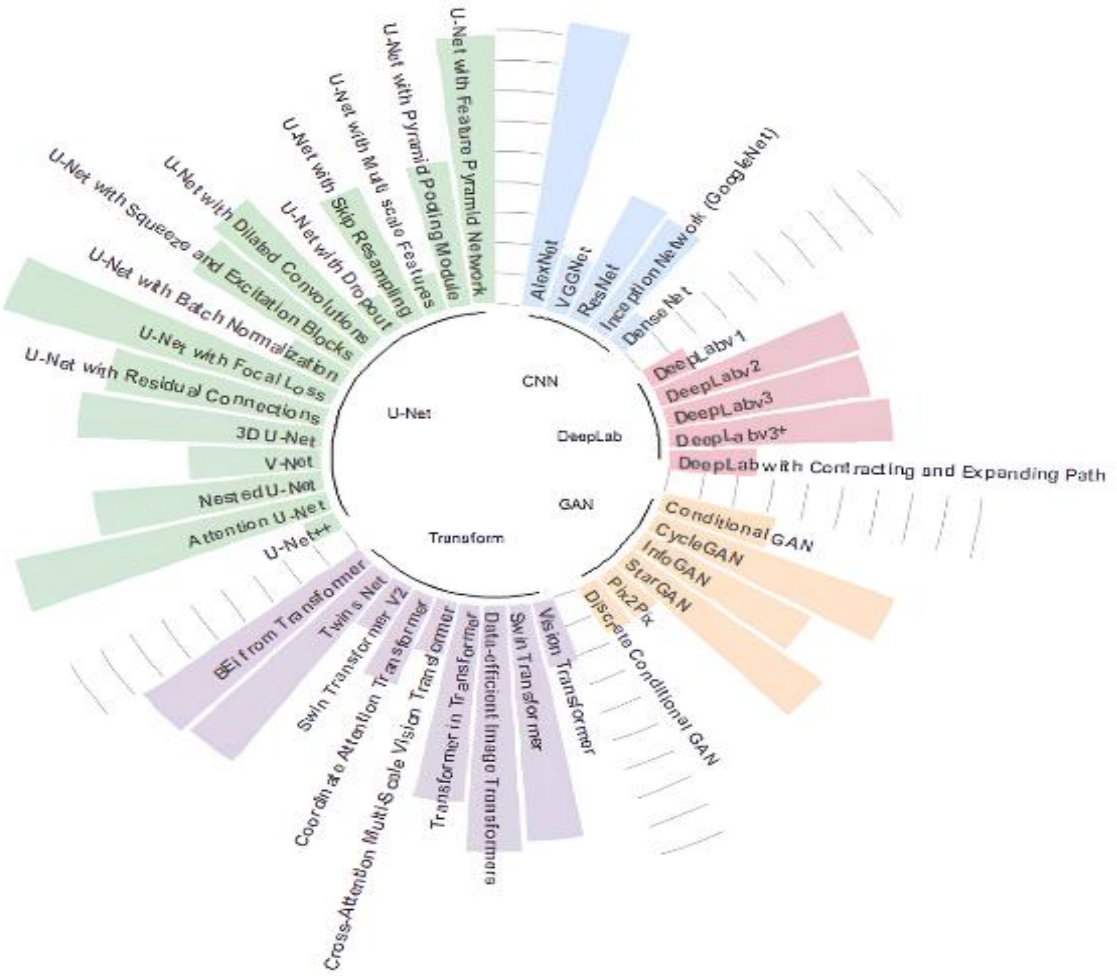

Figure 4: Overview of deep neural network architectures for medical image

Table 5 further categorizes segmentation models based on fusion strategies and evaluates their performance using commonly adopted metrics such as Dice Similarity Coefficient (DSC) and 95% Hausdorff Distance (HD95). Among the publicly available datasets, BraTS 2018–2021 is the most frequently used. Notably, BraTS 2021 has become the most widely adopted dataset due to its comprehensive multi-modality inputs, high-quality annotations, and large volume. In terms of loss function design, the most commonly used combination is Dice loss and Cross-Entropy Loss (CEL). Additionally, in Transformer-based architectures, there is a growing trend toward incorporating deep supervision and alignment-constrained losses to further improve segmentation performance.

Table 5: Comparison of segmentation performance of network models.

Model	Loss Function	Fusion Strategies	Datasets	DSC↑			HD95(mm)↓		
				WT	*TC*	*ET*	*WT*	*TC*	*ET*
FCN [35]	CEL	FLF	BraTs2018	0.711	0.709	0.725	-	-	-
MFD-Net [49]	BCEL+DL	ILF+FLF	BraTS2018	0.908	0.856	0.820	5.99	7.00	2.85
			BraTS2019	0.908	0.857	0.767	3.41	5.83	4.45
			BraTS2021	0.927	0.887	0.854	3.51	5.77	13.98
DPAF-Net [50]	GDL	FLF	BraTS2018	0.895	0.799	0.789	5.82	8.61	5.34
			BraTS2019	0.890	0.812	0.782	8.53	7.43	3.82
			BraTS2020	0.894	0.832	0.781	-	-	-
PIF-Net [51]	DL+BCEL	FlF	BraTS2019	0.894	0.814	0.771	5.35	10.90	5.85
			BraTS2020	0.895	0.819	0.775	5.31	9.43	4.47
AABTS-Net [52]	BCEL+DL+Deep supervision	ILF	BraTS2019	0.911	0.838	0.777	3.99	6.03	3.246
			BraTS2021	0.922	0.861	0.830	4.00	11.18	17.73
MAF-Net [53]	CEL	MF+FLF	BraTS2020	0.880	0.679	0.418	-	-	-
DMFNet [54]	GDL	ILF	BraTS2018	0.906	0.845	0.801	4.66	6.44	3.06
SF-Net [47]	DL+Fusion	FLF	BraTS2020	0.891	0.834	0.810	7.10	6.44	3.89
UDA-GS [55]	DL+CEL	FLF	BraTS2020	0.827	0.704	0.843	6.40	6.40	3.60
MCC-AFFM [56]	DL	ILF+FLF	BraTS2018	0.865	0.870	0.794	4.60	3.60	2.50
AD-Net [57]	CEL+JS	ILF	BraTS2019	0.900	0.810	0.760	4.31	12.40	35.50

			BraTS2020	0.900	0.800	0.760	7.22	15.30	35.20
AMCA-Net [58]	BCEL+DL	FLF	BraTS2018	0.904	0.890	0.802	10.20	7.40	4.30
			BraTS2019	0.910	0.842	0.801	10.70	8.40	4.80
MAFF-ResUNet [7]	BCEL+DL	FLF	BraTs2019	0.912	0.918	0.902	2.16	1.39	1.20
mResU-Net [59]	WCEL+DL	FLF	BraTS2021	0.928	0.929	0.8965	5.12	4.16	2.29
U-Net [60]	WCEL	-	BraTS2018	0.766	0.665	0.561	9.21	10.24	11.12
AttnUnet [61]	DL	FLF	BraTS2018	0.767	0.683	0.543	9.00	10.46	10.45
nnUNet [20]	DL+CEL	ILF	MSD	0.920	0.854	0.810	3.64	4.91	4.06
			BraTS2021	0.926	0.874	0.837	3.55	10.56	22.44
V-Net [40]	DL	-	BraTS2018	0.801	0.528	0.361	-	-	-
			BraTS2019	0.887	0.766	0.739	6.26	8.705	6.13
			BraTS2021	0.840	0.806	0.782	15.69	25.08	20.80
AGSE-VNet [62]	DL	ILF	BraTS2020	0.850	0.690	0.68	8.44	31.60	47.40
3D U-Net [63]	WCEL	ILF	BraTS2018	0.886	0.812	0.764	7.90	7.6	5.60
			BraTS2021	0.905	0.854	0.827	3.78	5.40	2.83
MM-UNet [64]	DL+FL	DLF	BraTS2020	0.850	0.765	0.762	8.24	10.77	6.39
MSFR-NET [65]	DL+CEL	FLF	BraTS2015	0.860	0.740	0.650	-	-	-
			BraTS2018	0.909	0.858	0.807	4.24	6.72	2.73
MultiEncoder UNet [66]	CEL+DL	DLF	BraTS2018	0.910	0.782	0.776	4.37	13.68	29.87
			BraTS2021	0.922	0.880	0.851	5.11	7.16	11.09
Mirror U-Net [67]	DL+CEL	FLF	MSD	0.925	0.858	0.781	-	-	-
DTASUnet [68]	DL+Deep supervision	FLF	BraTS2018	0.905	0.845	0.808	-	-	-
			BraTS2020	0.906	0.844	0.790	-	-	-
TransUNet [69]	DL+CEL	FLF	MSD	0.706	0.684	0.542	14.03	14.5	10.42
UNETR [70]	CEL+DL	ILF	MSD	0.789	0.761	0.585	8.27	8.85	9.35
Swin UNet3D [71]	DL	FLF	BraTS2018	0.874	0.761	0.716	-	-	-
			BraTS2021	0.905	0.866	0.834	-	-	-
Swin UNETR [72]	DL	ILF	BraTS2021	0.926	0.885	0.858	5.83	3.77	6.02
TransBTS [73]	DL	ILF	BraTS2019	0.900	0.819	0.789	5.64	6.049	3.74
			BraTS2020	0.901	0.817	0.787	4.96	9.77	17.95
CKD-TransBTS [74]	DL	FLF	BraTs2021	0.933	0.902	0.885	6.20	6.54	5.93
Transformer-DSUNET [75]	DL+FL	FLF	BraTS2020	0.908	0.923	0.914	-	-	-
nnFormer [76]	CEL+DL+Auxiliary Supervision	FLF	MSD	0.913	0.860	0.818	3.80	4.49	3.87
mmFormer [77]	DL+	MF+FLF	Brats2018	0.896	0.858	0.776	6.82	7.54	7.32
NestedFormer [78]	DL+CEL	FLF	BraTS2020	0.920	0.864	0.800	4.57	5.32	5.27
CMAF-Net [79]	CEL+DL	MF+FLF	Brats2018	0.889	0.846	0.755	4.38	6.59	5.95
			BraTs2020	0.909	0.868	0.778	4.21	5.35	4.02
XLSTM-HVED [28]	DL+CEL	MF+FLF	BraTS2024	0.868	0.779	0.659	11.73	11.30	8.74

Note: ILF: Input-level fusion; FLF: Feature-level fusion; DLF: Decision-level fusion; MF: Missing-modality fusion.

Regarding fusion strategies, each has its own strengths and application scenarios: ILF is suitable for scenarios with complete multimodal inputs. Representative models such as V-Net and 3D U-Net concatenate modality-specific images at the input stage, allowing the network to process them jointly. FLF is the most widely adopted fusion strategy. Models like AttnUnet, UNETR, nnFormer, U-Net++, and nnUNet fall into this category. A common enhancement in FLF involves incorporating attention mechanisms to improve the quality of multi-modal feature fusion. For instance, AGSE-VNet [62] integrates SE (Squeeze-and-Excitation) modules [39] in each encoder to model inter-channel dependencies and adaptively recalibrate feature responses. Another representative module is CBAM (Convolutional Block Attention Module) [80], which effectively enhances the network's attention to salient features. Additionally, many models embed multi-scale feature extraction modules, such as ASPP (Atrous Spatial Pyramid Pooling) [81], which capture contextual information at different receptive field scales, thereby improving the network's ability to delineate tumor boundaries and structures [60]. MF has gained increasing attention in recent years, particularly in clinical scenarios where some input modalities may be unavailable. This strategy enhances model robustness and generalization under incomplete data conditions. DLF is more commonly found in earlier approaches. A notable example is the DLF model proposed by Xie et al. [82], which simulates a multi-atlas segmentation process by constructing a three-stage U-Net architecture with a weighted voting sub-network to perform decision-level fusion of multiple subnetworks. This approach is particularly well-suited for ensemble models with parallel structures.

3. RESEARCH AND PROSPECTS

3.1 Dataset Acquisition and Standardization

Data availability and standardization are fundamental challenges in advancing deep learning applications for glioma segmentation. While deep learning models can automatically extract rich feature representations from medical images, they require a large volume of high-quality, expert-labeled datasets for effective training. However, obtaining such datasets is often constrained by the cost and labor-intensive nature of manual annotation. To mitigate these limitations, dimensionality reduction techniques and tumor slicing methods have been employed, though they often compromise contextual information, leading to inaccurate boundary delineation [83]. Additionally, substantial variations in imaging devices, acquisition protocols, spatial resolution, contrast, and noise levels across institutions complicate the integration and comparison of datasets [84]. The fusion of data from different modalities, such as MRI, CT, and PET, further increases the complexity of standardization due to modality-specific characteristics and formats. Addressing these issues requires the development and adoption of unified standards for data acquisition and preprocessing. International organizations, including the National Cancer Institute (NCI) and the Radiological Society of North America (RSNA), are actively promoting standardized imaging protocols. Furthermore, the application of deep learning-based alignment and normalization techniques has shown promise in reducing inter-institutional data inconsistencies, thereby facilitating more effective multimodal data fusion and model training.

3.2 Outline Accuracy and Related Organ Protection

Precise delineation of target regions is crucial for maximizing the efficacy of glioma treatments while minimizing radiation exposure to healthy tissues. Accurate segmentation ensures that therapeutic doses are concentrated on tumor tissues, thus improving treatment outcomes and reducing the risk of cognitive and functional impairments. For instance, protecting critical structures such as the hippocampus during radiotherapy is essential, as damage to this region may result in significant memory and learning deficits. Given the interpatient variability in tumor characteristics and brain anatomy, personalized radiotherapy planning based on individual imaging profiles and biomarkers is necessary to achieve optimal therapeutic efficacy with minimal adverse effects. This process demands close multidisciplinary collaboration among radiation oncologists, radiologists, physicists, and dosimetrists, supported by rigorous quality control, periodic imaging evaluations, and adaptive treatment planning to account for anatomical changes throughout therapy.

3.3 Model Performance and Interpretability

As deep learning models increasingly influence clinical decision-making, ensuring their high performance and interpretability becomes imperative. While many models achieve impressive segmentation accuracy, the "black box" nature of deep learning remains a significant barrier to clinical acceptance. Improving model transparency through interpretability techniques is essential for gaining clinicians' trust and facilitating the safe deployment of AI systems in practice. Integrating imaging features with pathological and histomorphological data can enhance the biological relevance of model predictions, enabling a deeper understanding of glioma behavior at cellular and tissue levels. In parallel, stringent validation protocols using diverse and independent datasets are necessary to assess model robustness and generalizability, ensuring their reliability across different clinical settings.

3.4 Clinical Translation

Bridging the gap between research and clinical practice requires careful validation, regulatory approval, and integration of multimodal image segmentation technologies into existing healthcare infrastructures. Thorough clinical validation is essential to demonstrate the safety, effectiveness, and generalizability of these models. In addition, clinicians must receive comprehensive training on the operation, interpretation, and limitations of AI-assisted tools to ensure their safe application. Modifications to hospital information systems and workflows are also necessary to accommodate new technologies seamlessly. However, financial and resource constraints often hinder widespread adoption. Therefore, fostering multidisciplinary collaboration among researchers, clinicians, and engineers is critical to facilitate translation efforts. Conducting large-scale clinical trials and validation studies, developing detailed operational guidelines, and promoting supportive policies and funding mechanisms are vital strategies to accelerate the clinical implementation of multimodal segmentation technologies.

4. CONCLUSION

In this paper, we present a comprehensive overview and analysis of the application of deep learning techniques in multimodal image fusion and segmentation for gliomas. Based on an extensive review of recent literature, several key conclusions can be drawn.

Deep learning has demonstrated significant potential for brain tumor segmentation compared to traditional methods. Techniques such as convolutional neural networks (CNNs), DeepLab, generative adversarial networks (GANs), Transformers, and U-Net architectures have enabled accurate semantic segmentation using multimodal images including CT, MRI, and PET. These advances have improved the diagnostic accuracy and treatment planning for gliomas, ultimately contributing to increased patient survival rates and quality of life. With the evolution of medical imaging technology, multimodal fusion approaches have provided considerable advantages by integrating complementary information from different imaging modalities. Strategies such as layer-level fusion and decision-level fusion enhance both segmentation accuracy and the descriptive richness of tumor characteristics, improving the robustness and versatility of brain tumor segmentation. Despite the current challenges related to data acquisition, standardization, and multimodal integration, these limitations are expected to be progressively addressed through the development of public datasets, the establishment of unified imaging standards, and the application of advanced preprocessing techniques. Furthermore, improving model interpretability and accuracy will strengthen clinical practitioners' trust in artificial intelligence technologies, facilitating their adoption in clinical practice. It is important to emphasize that computer-aided diagnostic systems are intended to assist, rather than replace, human expertise.

Future research should focus on the optimization of segmentation algorithms, the effective fusion of incomplete or heterogeneous multimodal data, and the overall enhancement of model performance, aiming to achieve more precise glioma diagnosis and better prognostic outcomes for patients.

REFERENCES

[1] Louis DN, Perry A, Wesseling P, Brat DJ, Cree IA, Figarella-Branger D, et al. The 2021 WHO classification of tumors of the central nervous system: a summary. Neuro Oncol. 2021;23:1231-51.
[2] Valbuena Rubio S, García-Ordás MT, García-Olalla Olivera O, Alaiz-Moretón H, González-Alonso MI, Benítez-Andrades JA, et al. Survival and grade of the glioma prediction using transfer learning. PeerJ Comput Sci. 2023;9:e1723.
[3] Koizumi S, Oishi T, Iwaizumi M, Kurozumi K. Genomic medicine advances for brain tumors. Int J Clin Oncol. 2024;29:1407-16.
[4] Du P, Chen H, Lv K, Geng D. A survey of radiomics in precision diagnosis and treatment of adult gliomas. J Clin Med. 2022;11:3802.
[5] Tang F, Liang S, Zhong T, Huang X, Deng X, Zhang Y, et al. Postoperative glioma segmentation in CT image using deep feature fusion model guided by multi-sequence MRIs. Eur Radiol. 2020;30:823-32.
[6] Thust SC, Heiland S, Falini A, Jäger HR, Waldman AD, Sundgren PC, et al. Glioma imaging in Europe: a survey of 220 centres and recommendations for best clinical practice. Eur Radiol. 2018;28:3306-17.
[7] He X, Xu W, Yang J, Mao J, Chen S, Wang Z. Deep convolutional neural network with a multi-scale attention feature fusion module for segmentation of multimodal brain tumor. Front Neurosci. 2021;15:782968.
[8] Zhang H, Ille S, Sogerer L, Schwendner M, Schröder A, Meyer B, et al. Elucidating the structural-functional connectome of language in glioma-induced aphasia using nTMS and DTI. Hum Brain Mapp. 2022;43:1836-49.
[9] Cirillo S, Battistella G, Castellano A, Sanvito F, Iadanza A, Bailo M, et al. Comparison between the inferior frontal gyrus intrinsic connectivity network and the verb-generation task fMRI network for presurgical language mapping in healthy controls and glioma patients. Brain Imaging Behav. 2022;16:2569-85.
[10] Brindle KM, Izquierdo-García JL, Lewis DY, Mair RJ, Wright AJ. Brain tumor imaging. J Clin Oncol. 2017;35:2432-8.
[11] Bernal J, Kushibar K, Asfaw DS, Valverde S, Oliver A, Martí R, et al. Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review. Artif Intell Med. 2019;95:64-81.
[12] Wei Y, Chen X, Zhu L, Zhang L, Schonlieb CB, Price S, et al. Multi-modal learning for predicting the genotype of glioma. IEEE Trans Med Imaging. 2023;42:3167-78.
[13] Magadza T, Viriri S. Deep learning for brain tumor segmentation: a survey of state-of-the-art. J Imaging. 2021;7:19.
[14] Baeesa S, Maghrabi Y, Moshref R, Al-Maghrabi J. Optic pathway-hypothalamic glioma apoplexy: a report of two cases and systematic review of the literature. Front Surg. 2022;9:891556.
[15] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60-88.
[16] Zhou T, Ruan S, Canu S. A review: deep learning for medical image segmentation using multi-modality fusion. Array. 2019;3-4:100004.

[17] Simpson AL, Antonelli M, Bakas S, Bilello M, Farahani K, Ginneken BV, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv [Preprint]. 2019.
[18] Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, et al. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans Med Imaging. 2015;34(10):1993-2024.
[19] Khodadadi Shoushtari F, Dehkordi ANV, Sina S. Quantitative and visual analysis of data augmentation and hyperparameter optimization in deep learning-based segmentation of low-grade glioma tumors using Grad-CAM. Ann Biomed Eng. 2024;52:1359-77.
[20] Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021;18:203–11.
[21] Azam MA, Khan KB, Salahuddin S, Rehman E, Khan SA, Khan MA, et al. A review on multimodal medical image fusion: compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics. Comput Biol Med. 2022;144:105253.
[22] Bakas S, Reyes M, Jakab A, Bauer S, Rempfler M, Crimi A, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BraTS challenge. arXiv preprint. 2018; arXiv:1811.02629.
[23] Dolz J, Desrosiers C, Ayed IB. HyperDense-Net: a hyper-densely connected CNN for multi-modal image segmentation. IEEE Trans Med Imaging. 2018;38(5):1116–26.
[24] Zhang Y, Wu J, Zhu J, et al. ME-Net: Multiencoder network for brain tumor segmentation. IEEE J Biomed Health Inform. 2020;24(10):2743–52.
[25] Liu L, Zhang Y, Liu Q, et al. CBAM-UNet: Attention-based network for medical image segmentation. Comput Methods Programs Biomed. 2021;208:106304.
[26] Sun L, Zhang S, Chen H, Luo L. Brain tumor segmentation and survival prediction using multimodal MRI scans with deep learning. Front Neurosci. 2019;13:810.
[27] Havaei M, Davy A, Warde-Farley D, Biard A, Courville A, Bengio Y, et al. HeMIS: Hetero-modal image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer; 2016. p. 469–77.
[28] Zhu S, Chen Y, Jiang S, Chen W, Liu C, Wang Y, et al. XLSTM-HVED: cross-modal brain tumor segmentation and MRI reconstruction method using vision XLSTM and heteromodal variational encoder-decoder. arXiv [Preprint]. 2024.
[29] Sharma A, Hamarneh G. Missing MRI pulse sequence synthesis using multi-modal generative adversarial network. IEEE Trans Med Imaging. 2019;39(4):1170-83.
[30] Gupta S, Hoffman J, Malik J. Cross-modal distillation for supervision transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016; Las Vegas, NV, USA. p. 2827-36.
[31] Zhang D, Huang F, Liu S, Wang X, Ge Z. Cross-modality deep feature learning for brain tumor segmentation. Pattern Recognit. 2021;110:107630.
[32] Rahimpour M, Bertels J, Radwan A, Vandermeulen H, Sunaert S, Vandermeulen D, et al. Cross-modal distillation to improve MRI-based brain tumor segmentation with missing MRI sequences. IEEE Trans Biomed Eng. 2022;69(7):2153-64.
[33] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. 2017;60(6):84-90.
[34] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015); 2015 May 7-9; San Diego, CA, USA.
[35] Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Trans Pattern Anal Mach Intell. 2017;39:640-51.
[36] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2015 Jun 7-12; Boston, MA, USA. p. 1-9.
[37] Iandola F, Moskewicz M, Karayev S, Girshick R, Darrell T, Keutzer K. DenseNet: implementing efficient ConvNet descriptor pyramids. arXiv [Preprint] 2014.
[38] Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018; Munich, Germany. p. 801-18.
[39] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2018; Salt Lake City, UT, USA. p. 7132-41.
[40] Milletari F, Navab N, Ahmadi SA. V-net: fully convolutional neural networks for volumetric medical image segmentation. In: Proceedings of the 4th International Conference on 3D Vision (3DV); 2016; Stanford, CA, USA. p. 565-71.

[41] Dosovitskiy A. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv [Preprint] 2020.
[42] Shamshad F, Khan S, Zamir SW, Khan MH, Hayat M, Khan FS, et al. Transformers in medical imaging: a survey. Med Image Anal. 2023;88:10280.
[43] Nguyen QD, Thai HT. Crack segmentation of imbalanced data: the role of loss functions. Eng Struct. 2023;297:116988.
[44] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436-44.
[45] Sudre CH, Li W, Vercauteren T, Ourselin S, Cardoso MJ. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Proceedings of the 3rd International Workshop on Deep Learning in Medical Image Analysis (DLMIA); 2017 Sep 14; Québec City, Canada. p. 240-8.
[46] Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV); 2017; Venice, Italy. p. 2980-8.
[47] Liu Y, Mu F, Shi Y, Chen X. SF-Net: a multi-task model for brain tumor segmentation in multimodal MRI via image fusion. IEEE Signal Process Lett. 2022;PP:1-5.
[48] Englesson E, Azizpour H. Generalized Jensen-Shannon divergence loss for learning with noisy labels. arXiv [Preprint]. 2021.
[49] Hou Q, Peng Y, Wang Z, Wang J, Jiang J. MFD-Net: modality fusion diffractive network for segmentation of multimodal brain tumor image. IEEE J Biomed Health Inform. 2023;27:5958-69.
[50] Chang Y, Zheng Z, Sun Y, Zhao M, Lu Y, Zhang Y. DPAFNet: a residual dual-path attention-fusion convolutional neural network for multimodal brain tumor segmentation. Biomed Signal Process Control. 2023;79:104037.
[51] Liu Y, Mu F, Shi Y, Cheng J, Li C, Chen X. Brain tumor segmentation in multimodal MRI via pixel-level and feature-level image fusion. Front Neurosci. 2022;16:1000587.
[52] Tian W, Li D, Lv M, Huang P. Axial attention convolutional neural network for brain tumor segmentation with multi-modality MRI scans. Brain Sci. 2023;13(1):12.
[53] Huang Z, Lin L, Cheng P, Peng L, Tang X. Multi-modal brain tumor segmentation via missing modality synthesis and modality-level attention fusion. arXiv preprint arXiv:2203.04586. 2022 Mar 9.
[54] Chen C, Liu X, Ding M, Zheng J, Li J. 3D dilated multi-fiber network for real-time brain tumor segmentation in MRI. arXiv [Preprint]. 2019.
[55] Hu Z, Sun Y, Bian L, Luo C, Zhu J, Zhu J, et al. UDA-GS: a cross-center multimodal unsupervised domain adaptation framework for Glioma segmentation. Comput Biol Med. 2025;185:109472.
[56] Zhou T. Modality-level cross-connection and attentional feature fusion based deep neural network for multi-modal brain tumor segmentation. Biomed Signal Process Control. 2023;81:104524.
[57] Peng Y, Sun J. The multimodal MRI brain tumor segmentation based on AD-Net. Biomed Signal Process Control. 2023;80:104336.
[58] Wu S, Cao Y, Li X, Liu Q, Ye Y, Liu X, et al. Attention-guided multi-scale context aggregation network for multi-modal brain glioma segmentation. Med Phys. 2023;50(12):7629-40.
[59] Li P, Li Z, Wang Z, Li C, Wang M. mResU-Net: multi-scale residual U-Net-based brain tumor segmentation from multimodal MRI. Med Biol Eng Comput. 2024;62(3):641-51.
[60] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI); 2015; Munich, Germany. p. 234-41.
[61] Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, et al. Attention U-Net: learning where to look for the pancreas. arXiv [Preprint] 2018.
[62] Guan X, Yang G, Ye J, Yang W, Xu X, Jiang W, et al. 3D AGSE-VNet: an automatic brain tumor MRI data segmentation framework. BMC Med Imaging. 2022;22:6.
[63] Gamal A, Bedda K, Ashraf N, Ayman S, AbdAllah M, Rushdi MA. Brain tumor segmentation using 3D U-Net with hyperparameter optimization. In: Proceedings of the 3rd Novel Intelligent and Leading Emerging Sciences Conference (NILES); 2021; Cairo, Egypt. p. 269-72.
[64] Zhao L, Ma J, Shao Y, Jia C, Zhao J, Yuan H. MM-UNet: a multimodality brain tumor segmentation network in MRI images. Front Oncol. 2022;12:950706.
[65] Li X, Jiang Y, Li M, Zhang J, Yin S, Luo H. MSFR-Net: multi-modality and single-modality feature recalibration network for brain tumor segmentation. Med Phys. 2023;50:2249-62.
[66] Pan Y, Yong H, Lu W, Li G, Cong J. Brain tumor segmentation by combining MultiEncoder UNet with wavelet fusion. J Appl Clin Med Phys. 2024;25(11):e14527.
[67] Marinov Z, Reiß S, Kersting D, Kleesiek J, Stiefelhagen R. Mirror U-Net: marrying multimodal fusion with multi-task learning for semantic segmentation in medical imaging. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW); 2023; Paris, France. p. 2275-85.

[68] Ma B, Sun Q, Ma Z, Li B, Cao Q, Wang Y, et al. DTASUnet: a local and global dual transformer with the attention supervision U-network for brain tumor segmentation. Sci Rep. 2024;14:28379.
[69] Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, et al. TransUNet: Transformers make strong encoders for medical image segmentation. arXiv [Preprint]. 2021.
[70] Hatamizadeh A, Tang Y, Nath V, Yang D, Myronenko A, Landman B, et al. Unetr: Transformers for 3D medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2022; Waikoloa, HI, USA. p. 574-84.
[71] Cai Y, Long Y, Han Z, Liu M, Zheng Y, Yang W, et al. Swin Unet3D: a three-dimensional medical image segmentation network combining vision transformer and convolution. BMC Med Inform Decis Mak. 2023;23:33.
[72] Hatamizadeh A, Nath V, Tang Y, Yang D, Roth H, Xu D, et al. Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. arXiv [Preprint]. 2022.
[73] Wang W, Chen C, Ding M, Yu H, Zha S, Li J. TransBTS: multimodal brain tumor segmentation using transformer. In: de Bruijne M, Cattin PC, Cotin S, et al., editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2021: 24th International Conference; 2021; Strasbourg, France. Cham: Springer; 2021. p. 109-19. (Lecture Notes in Computer Science; vol. 12901)
[74] Lin J, Lin J, Lu C, Chen H, Lin H, Zhao B, et al. CKD-TransBTS: clinical knowledge-driven hybrid transformer with modality-correlated cross-attention for brain tumor segmentation. IEEE Trans Med Imaging. 2023;PP:1.
[75] Zakariah M, Al-Razgan M, Alfakih T. Dual vision Transformer-DSUNET with feature fusion for brain tumor segmentation. Heliyon. 2024;10(18):e37804.
[76] Zhou HY, Guo J, Zhang Y, Yu L, Wang L, Yu Y. nnformer: interleaved transformer for volumetric segmentation. IEEE Trans Image Process. 2021;30:9210-21.
[77] Zhang Y, He N, Yang J, Li Y, Wei D, Huang Y, et al. mmFormer: multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In: Wang L, Dou Q, Fletcher PT, et al., editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2022: 25th International Conference; 2022; Singapore. Cham: Springer Nature Switzerland; 2022. p. 107-17. (Lecture Notes in Computer Science; vol. 13435).
[78] Xing Z, Yu L, Wan L, Han T, Zhu L. NestedFormer: nested modality-aware transformer for brain tumor segmentation. In: Proceedings of the 25th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2022); 2022; Singapore. p. 140-50.
[79] Sun K, Ding J, Li Q, Chen W, Zhang H, Sun J, et al. CMAF-Net: a cross-modal attention fusion-based deep neural network for incomplete multi-modal brain tumor segmentation. Quant Imaging Med Surg. 2024;14:4579-604.
[80] Woo S, Park J, Lee JY, Kweon IS. CBAM: convolutional block attention module. In: Proceedings of the 15th European Conference on Computer Vision (ECCV 2018); 2018 Sep 8-14; Munich, Germany. p. 3-19.
[81] Chen LC, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation. arXiv [Preprint] 2017.
[82] Xie L, Wisse LEM, Wang J, Ravikumar S, Khandelwal P, Glenn T, et al. Deep label fusion: a generalizable hybrid multi-atlas and deep convolutional neural network for medical image segmentation. Med Image Anal. 2022;83:102683.
[83] Zhang R, Wei Y, Wang D, Chen B, Sun H, Lei Y, et al. Deep learning for malignancy risk estimation of incidental sub-centimeter pulmonary nodules on CT images. Eur Radiol. 2024;34(7):4218-29.
[84] Huang W, Tan K, Zhang Z, Hu J, Dong S. A review of fusion methods for omics and imaging data. IEEE/ACM Trans Comput Biol Bioinform. 2022;20:74-93.