Meta-Radiology 1 (2023) 100004

Contents lists available at ScienceDirect

Meta-Radiology

journal homepage: www.keaipublishing.com/en/journals/meta-radiology

Vision transformers in multi-modal brain tumor MRI segmentation: A review

Pengyu Wang [a], Qiushi Yang [b], Zhibin He [c], Yixuan Yuan [a,*]

[a] Department of Electronic Engineering, The Chinese University of Hong Kong, Kowloon, Hong Kong SAR 999077 China
[b] Department of Electrical Engineering, City University of Hong Kong, Kowloon, Hong Kong SAR 999077 China
[c] School of Automation, Northwestern Polytechnical University, Xi'an 710072, China

ARTICLE INFO

Keywords:
Brain tumor segmentation
Multi-modal MRI
Vision transformer
Deep learning

ABSTRACT

Brain tumors have shown extreme mortality and increasing incidence during recent years, which bring enormous challenges for the timely diagnosis and effective treatment of brain tumors. Concretely, accurate brain tumor segmentation on multi-modal Magnetic Resonance Imaging (MRI) is essential and important since most normal tissues are unresectable in brain tumor surgery. In the past decade, with the explosive development of artificial intelligence technologies, a series of deep learning-based methods are presented for brain tumor segmentation and achieved excellent performance. Among them, vision transformers with non-local receptive fields show superior performance compared with the classical Convolutional Neural Networks (CNNs). In this review, we focus on the representative transformer-based works for brain tumor segmentation proposed in the last three years. Firstly, this review divides these transformer-based methods as the pure transformer methods and the hybrid transformer methods according to their transformer architectures. Then, we summarize the corresponding theoretical innovations, implementation schemes and superiorities to help readers better understand state-of-the-art transformer-based brain tumor segmentation methods. After that, we introduce the most commonly-used Brain Tumor Segmentation (BraTS) datasets, and comprehensively analyze and compare the performance of existing methods through multiple quantitative statistics. Finally, we discuss the current research challenges and describe the future research trends.

1. Introduction

Many studies show that the brain is the most important and complex organ in the human body, containing over more than a hundred billion neurons and each with up to ten thousand synapses. According to the statistics of the World Health Organization (WHO), brain tumor is one of the most fatal cancers at present, and their morbidity is increasing year by year. Concretely, brain tumors can be divided into four classes, namely glioma, meningioma, pituitary adenoma and nerve sheath tumor. Among them, gliomas are the most common primary brain tumors, where the current mention of brain tumors usually refer to gliomas, which originate in the cells that make up the supporting tissue of the brain, known as glial cells.[1–3] Gliomas are caused by the interaction between congenital genetic high-risk factors and environmental carcinogenic factors. The clinical significance of glioma is that it is a dangerous and fatal brain tumor, which is highly malignant and aggressive, resulting in various symptoms, such as seizures, headaches, vision problems, and changes in speech and behavior. In general, the locations, shapes and sizes of brain tumors have significant impacts on the degree and nature of these symptoms diagnosed by doctors, as well as formulating the treatment and surgery plans. Therefore, brain tumor segmentation is conducive to accurate and efficient localization and recognition of gliomas, which can help doctors improve the diagnosis and prognosis in clinical applications (see Table 1).

In the past decades, scholars have conducted extensive fundamental researches on brain tumors.[4–12] The early researches aim to understand the biological properties of glial cells, and how they transform into malignant malignancies. Over time, researchers gain a better understanding of the genetic and molecular changes that occur in gliomas. These researches have promoted the development of new diagnosis and treatment methods for brain tumors, such as determining the brain tumor grading, heredity and target therapy through genomic data. On the other hand, scholars[13,14] have also investigated the use of various imaging techniques, such as Magnetic Resonance Imaging (MRI), to help diagnose brain tumors and

* Corresponding author.
E-mail address: yxyuan@ee.cuhk.edu.hk (Y. Yuan).

https://doi.org/10.1016/j.metrad.2023.100004
Received 31 May 2023; Accepted 1 June 2023
Available online 5 July 2023

Table 1
The correlations between brain tumors and MRI modalities.

Modality	WT	TC	ET
T1	low	low	low
T1ce	medium	high	high
T2	high	medium	low
FLAIR	high	low	low

monitor the corresponding progression. Considering the accuracy and intuition of MRI, MRI-based brain tumor examination has become the most commonly-used scheme. With the assistance of MRI technology, some new surgical techniques such as minimally invasive surgery have developed rapidly, which makes it possible to precisely remove brain tumors and minimize the damage to surrounding healthy tissues.

Specifically, brain tumor segmentation aims to locate multiple tumor regions by changing the representations of MRI, where the segmentation results are utilized for subsequent prognosis and survival prediction of brain tumors. The brain tumor regions can generally be defined as three types of sub-regions, including Enhancing Tumor (ET), Tumor Core (TC: enhancing tumor, necrosis and non-enhancing tumor), Whole Tumor (WT: peritumoral edema, enhancing tumor, non-enhancing tumor and necrosis), as shown in Fig. 1.

1.1. Multi-modal brain MRI

With the development of MRI technology, multi-modal MRI images have been widely applied in brain tumor segmentation, which can provide a more comprehensive understanding of the tumors and surrounding brain tissues. Concretely, MRI contains four modalities in terms of T1, T2, T1ce and FLAIR, which are complementary imaging modalities for the diagnosis and monitoring of brain tumors,[1–3] i.e., different MRI modalities can provide complementary information about tumor appearance and properties (see Table 2). For example, the T1-weighted MRI modality is often used to provide high-resolution images of the brain, and the T2-weighted MRI modality can provide information about the fluid content of tissue, which is particularly useful for differentiating between tumor and normal brain tissue. Additionally, contrast-enhanced T1-weighted MRI modality can provide information about blood vessels and the enhancement pattern of the tumor, which can be used to help diagnose the type of tumor and its aggressiveness. The combination of multiple MRI modalities can provide a more complete and accurate representation of the tumor and surrounding brain tissue, which is essential for effective glioma segmentation. The use of multi-modal MRI data also enables researchers to evaluate the performance of different segmentation algorithms and compare their results, which can be used to advance the development of new techniques and improve the accuracy of brain tumor segmentation.

Table 2
The property of different MRI modalities.

Modality	Property
T1	Highlight the tissue's T1 relaxation (longitudinal) differences. Brighter tissue denotes the shorter relaxation time.
T1ce	Highlight the tumor signal on T1.
T2	Highlight the tissue's T2 relaxation (transverse) differences. Brighter tissue denotes the longer relaxation time.
FLAIR	Suppressing the cerebrospinal fluid bright signal. Highlight the small hyper-intense tumor regions.

The most well-known dataset used for evaluating brain tumor segmentation is the Brain Tumor Segmentation (BraTS) challenge dataset, which contains a large number of matched MRI modalities including T1, T2, T1ce and FLAIR with manually annotated tumor segmentation masks. The BraTS dataset provides a valuable resource for researchers and clinicians working on glioma segmentation and brain tumor diagnosis.

1.2. Brain tumor segmentation using multi-modal MRI

Over the past years, the rapid development of deep learning has effectively improved the performance of computer-aided diagnosis. Many technical advances[4–12] have been made in multi-modal brain tumor segmentation, leading to a growing number of methods that are able to perform this task with varying degrees of accuracy and speed. The earliest and simplest methods for brain tumor segmentation are manual tracing, where an expert clinician draws a contour around the tumor in the images. However, manual tracing is time-consuming and can be subject to inter- and intra-observer variability. With the advent of computer vision and machine learning techniques, many automated methods have been developed to perform brain tumor segmentation. These methods can be broadly classified into two categories: traditional methods and deep learning methods.

Traditional methods[15–17] include region-growing and level-set methods, and atlas-based segmentation. These methods use image features such as intensity, texture, and gradient information to separate the brain tumor from the surrounding tissue. For example, region-growing methods use seed points to grow the tumor region by including voxels that have similar intensity values to the seed points. In recent years, deep learning methods, particularly Convolutional Neural Networks (CNNs), have become increasingly popular for brain tumor segmentation. CNNs can learn complex image features from large amounts of annotated data,

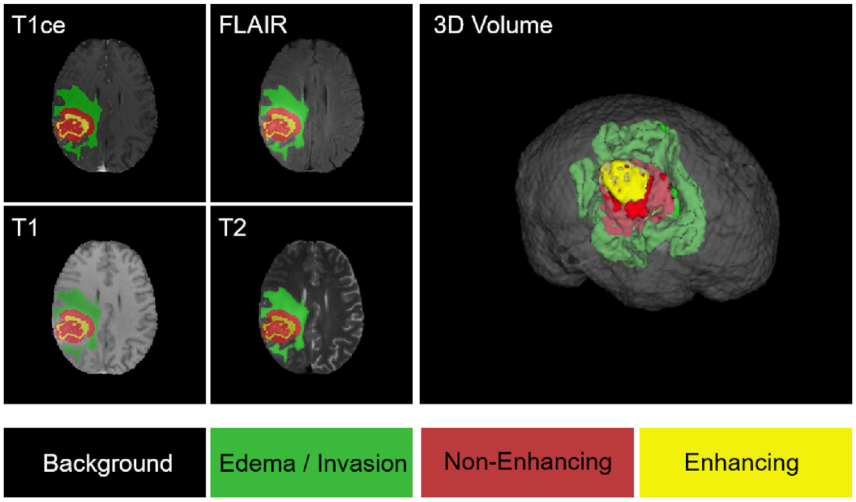

Fig. 1. Illustrations of different sub-regions for brain tumor segmentation.

which leads to improved performance compared to traditional methods. For example, one popular deep learning method for brain tumor segmentation is the U-Net architecture,[18] which uses an encoder-decoder structure to learn both low-level and high-level features of the images. Latest, the vision transformers[19–25] witness remarkable progress and achieve superior performance on brain tumor segmentation. Some works adopt multi-head self-attention layers within transformers to capture more discriminative global features and perform impressive results. Others design transformer-based modality fusion modules to align multi-modal inputs and boost the modality fusion to improve the segmentation performance with multi-modal MRI data. These works aim to learn better and more compatible representations for multi-modal brain tumors and exploit adequate information for superior tumor segmentation.

1.3. Contributions of our review

Different from most previous works, this review mainly focuses on the applications of vision transformers on multi-modal brain tumor MRI segmentation, which currently dominates this task with state-of-the-art segmentation performance. Moreover, we discuss the existing challenges and future research directions from multiple perspectives. In summary, this review contains the following major contributions:

- We review the existing brain tumor segmentation approaches from diverse perspectives including the background, datasets, models and development trends.
- We introduce the vision transformer and sum up the state-of-the-art transformer-based methods on brain tumor segmentation using multi-modal MRI data.
- We conduct a comprehensive statistical analysis on the recently published papers, the widely-used datasets and evaluation metrics, and the comparison of segmentation performance.
- We conclude the existing challenges and discuss future research directions.

In this review, we organize the rest of the paper as following contents: (i) a complete introduction to vision transformer architecture, multi-head self-attention and positional encoding (Section 2); (ii) existing state-of-the-art transformer-based works for brain tumor segmentation including the pure transformer methods and the hybrid transformer methods, which is the main part (Section 3); (iii) a comprehensively statistical analysis on the publications, datasets, evaluation metrics and segmentation performance (Section 4); (iv) current research challenges and future research trends for brain tumor segmentation (Section 5); (v) the discussion and conclusion of this review (Section 6).

2. Vision transformer

Transformer is a typical attention-based deep learning model. In Natural Language Processing (NLP), the transformer is first proposed by Vaswani et al.[25] for machine translation tasks. Different from the locally connected CNNs and RNNs, the transformer can model the long-range dependencies between tokens, resulting in better global feature relation modeling. Recently, transformer-based methods have achieved state-of-the-art performance in multiple NLP tasks, and effectively replace RNNs as the most popular architectures. Inspired by this, Dosovitskiy et al.[26] introduce the classical transformer into computer vision called Vision Transformer (ViT). Specifically, ViT first divides input images into non-overlapping patches, and then models the global relations between such patches through multiple standard transformer layers for image classification. Compared with CNNs and RNNs, transformer-based methods usually have higher computational costs, but break through the performance bottleneck benefits from their non-local receptive fields. In medical image analysis, many transformer-based methods are developed for classification, segmentation and detection tasks, showing promising performance and generalization. This review mainly focuses on transformer-based methods for multi-modal brain tumor MRI segmentation. Referring to the recently proposed ViT variants and relevant review works, we will briefly introduce the core components of transformers in the following contents: (i) transformer encoder-decoder; (ii) Multi-head Self-Attention (MSA); (iii) positional encoding.

2.1. Transformer encoder-decoder

Following the classical implementation of image segmentation networks, most transformer-based segmentation methods adopt the encoder-decoder architecture. Concretely, the transformer encoder has a hierarchical structure, which contains different stages and each stage consists of multiple transformer blocks. To produce hierarchical representations, the down-sampling operations are inserted between adjacent encoder stages, which can reduce the spatial resolution of previous stage features. Next, the transformer decoder has a symmetrical hierarchical structure with the transformer encoder, and the difference is that the up-sampling operations are utilized between adjacent decoder stages. In addition, between the transformer encoder and decoder, the corresponding low-level and high-level features with the same resolution are concatenated through skip connections to ensure effective gradient propagation. In summary, the transformer encoder compresses high-resolution input images as low-resolution encoding features, and the transformer decoder restores low-resolution encoding features to the original resolution semantic maps for producing segmentation results.

2.2. Multi-head self-attention (MSA)

In transformer-based methods, multi-head self-attention is the most important component, which plays a key role in modeling long-distance dependencies between tokens. Specifically, each transformer block contains a MSA, which can learn the correlation between each token and all tokens to generate a self-attention map. Using the self-attention map to adaptively weigh the token channel information, transformers can effectively model the global feature relations. Take ViT as an example, given a 2D input image $I \in \mathbb{R}^{H\times W\times C}$, H and W are the width and height of the image respectively, C is the number of channels. ViT first divides the image I into N non-overlapping patches with $P \times P$ size, where each patch can be regarded as a token. Then, these patches are flattened in channel dimension to produce a token sequence $X \in \mathbb{R}^{N\times(P^2C)}$, N is the total number of patches and P is the length of each patch. As shown in Fig. 2, each transformer block contains two Layer Normalization (LN) layers, a MSA, a Multi-Layer Perception (MLP) and two skip connections. In MSA,

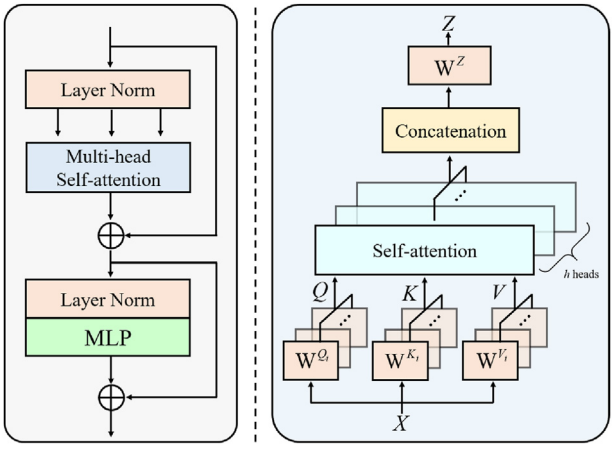

Fig. 2. Architectures of the standard transformer block (left) and the multi-head self-attention (right).

the first LN layer can normalize the input token sequence X. Then, three linear layers with weights W^Q, W^K and W^V project X as query $Q = XW^Q \in \mathbb{R}^{N\times D}$, key $K = XW^K \in \mathbb{R}^{N\times D}$ and value $V = XW^V \in \mathbb{R}^{N\times(P^2C)}$. The self-attention map $A \in^{N\times N}$ is the multiplication result of query Q and key K:

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{D}}\right). \tag{1}$$

Then, the self-attention map A is multiplied with value V to generate the final output of MSA:

$$Z = AV. \tag{2}$$

It is noteworthy that if the number of heads is larger than one in MSA, the above expression can be rewritten as:

$$Z_i = \text{softmax}\left(\frac{(XW^{Q_i})(XW^{K_i})^T}{\sqrt{D/h}}\right)XW^{V_i}, \tag{3}$$

where Z_i denotes the output of the i-th head, W^{Q_i}, $\lceil W^{K_i}$ and W^{V_i} are the weights of the i-th head for producing the query, key and value respectively.

In the above situation, MSA consists of multiple parallel heads, which can capture different global relations between tokens, thereby improving transformer performance. After that, we concatenate the output of each head at channel-level and use a linear layer W^Z to map the concatenation results as the final output of MSA:

$$Z = [Z_1,\ Z_2, \ldots, Z_h]W^Z, \tag{4}$$

where h represents the total number of heads.

In summary, the computation complexity of the self-attention map A is quadratic to the length of the input token sequence X, so that the high-resolution input image will increase huge computations for a transformer. Moreover, although MSA can effectively improve transformer performance, the large number of heads also brings relatively high computations. In order to balance the model performance and complexity, transformer-based methods usually set the number of heads between 2-12.

2.3. Positional encoding

Considering that the position information of tokens is crucial in a sentence, transformer-based methods widely adopt learnable positional encodings, which can adaptively learn the position information of each token to avoid the performance degradation caused by missing position information. In the classical transformer, Vaswani et al.[25] utilize the sine and cosine functions for retaining position information, the corresponding definition is:

$$PE(p, 2j) = \sin\left(\frac{p}{10000^{2j}/d}\right). \tag{5}$$

$$PE(p, 2j+1) = \cos\left(\frac{p}{10000^{2j}/d}\right), \tag{6}$$

where p represents the position of a token in the sequence, d is the word embedding length of a token, and j denotes the index range in 0 to $d-1$.

Similarly, the position information of patches is important for images. Hence, Dosovitskiy et al.[26] also use the learnable positional encoding in ViT, which adds the patch-specific positional encoding to the corresponding patch as the input token sequence to feed the transformer encoder. In addition to the above-mentioned absolute positional encodings, some efficient relative position encoding schemes are developed recently, such as the relative position bias proposed by SwinT,[27] the conditional positional encoding proposed by Chun et al.,[28] and the image relative position encoding proposed by Wu et al.[29]

3. Transformer-based multi-modal brain tumor segmentation methods

Accurate brain tumor segmentation results can provide doctors with rich clinical information, which is essential for brain tumor diagnosis and treatment. In brain tumor segmentation, CNNs are dominating methods for a long time. However, CNN is suboptimal in modeling non-local relations between features, so it usually requires building more complex and deeper network structures. In contrast, the recently proposed vision transformers comprehensively consider all local image patches for establishing global feature relations. Therefore, the correlations between tumor regions and backgrounds can be better explored and modeled in transformers, it is more conducive to improving brain tumor segmentation accuracy. Currently, various transformer-based methods are presented for brain tumor segmentation and show promising performance. In specific, most of them follow a core ideology, that is, based on the transformer with global receptive fields, through novel designs (such as network units or structures) to improve the transformer's ability in modeling local feature relations, so that the transformer can capture both global and local information effectively. According to the transformer architectures as shown in Fig. 3, we introduce two classes of transformer-based methods for brain tumor segmentation: (i) pure transformer methods[30–41]; (ii) hybrid transformer methods.[42–73] Among them, pure transformer methods mainly through designing transformer blocks to improve segmentation performance, while hybrid transformer methods focus on effectively combining transformer and CNN. Next, we will introduce the recent representative works over 2021–2023.

3.1. Pure transformer methods

In this section, pure transformer methods are defined as only adopting transformer-based blocks without combining CNNs. The early proposed pure transformer methods mainly follow the existing effective transformers in computer vision, such as ViT[26] and SwinT.[27] Specifically, Sager et al.[30] present ViTBIS, which expands the classical ViT into a U-shaped encoder-decoder structure. In the encoder, 1D tokens output by each transformer block is first reshaped as 2D features, and then the strided convolutions are used to down-sample features and change the channel dimension. In the decoder, the bilinear interpolations are used to up-sample features. ViTBIS also adopts the convention-based concatenation operation for skip connections. In ViTBIS, the long-range dependencies between features are well established by transformer layers, while the hierarchical U-Net structure reduces the model complexity to a certain extent. Hence, ViTBIS can accurately segment brain tumors as well as synapse multi-organ and spleen. Although ViT can well model long-range dependencies, its computational costs will primarily increase when the input image has a high resolution. Inspired by the dilated convolution, Wu et al.[31] construct an efficient transformer for brain tumor segmentation, namely Dilated Transformer (D-former). In specific, the D-former block contains three groups of transformer blocks that can jointly model global-local feature relations, where each group consists of the continuous Local Scope (LS)-MSA and Global Scope (GS)-MSA. LS-MSA utilizes the local windows similar to SwinT to model the local feature relations, while GS-MSA refers to the dilated convolution to sample tokens at intervals. Compared with classical ViT-style methods, D-former shows a lower model complexity but a more competitive performance. The main reason is that D-former innovatively introduces the ideology of the dilated convolution into the transformer to ensure comprehensive modeling of global-local feature relations.

Unlike the classical ViT, SwinT proposes a shifted windows based MSA to model the global feature relations, which not only improves the performance, but also achieves lower model complexity. Hence, many studies adopt SwinT as the basic transformer backbone for brain tumor segmentation. Typically, Liang et al.[32] propose a 3D U-shaped Symmetrical Swin Transformer-based Network (BTSwin-UNet), which uses 3D SwinT blocks as the basic units to construct the transformer encoder-decoder with skip

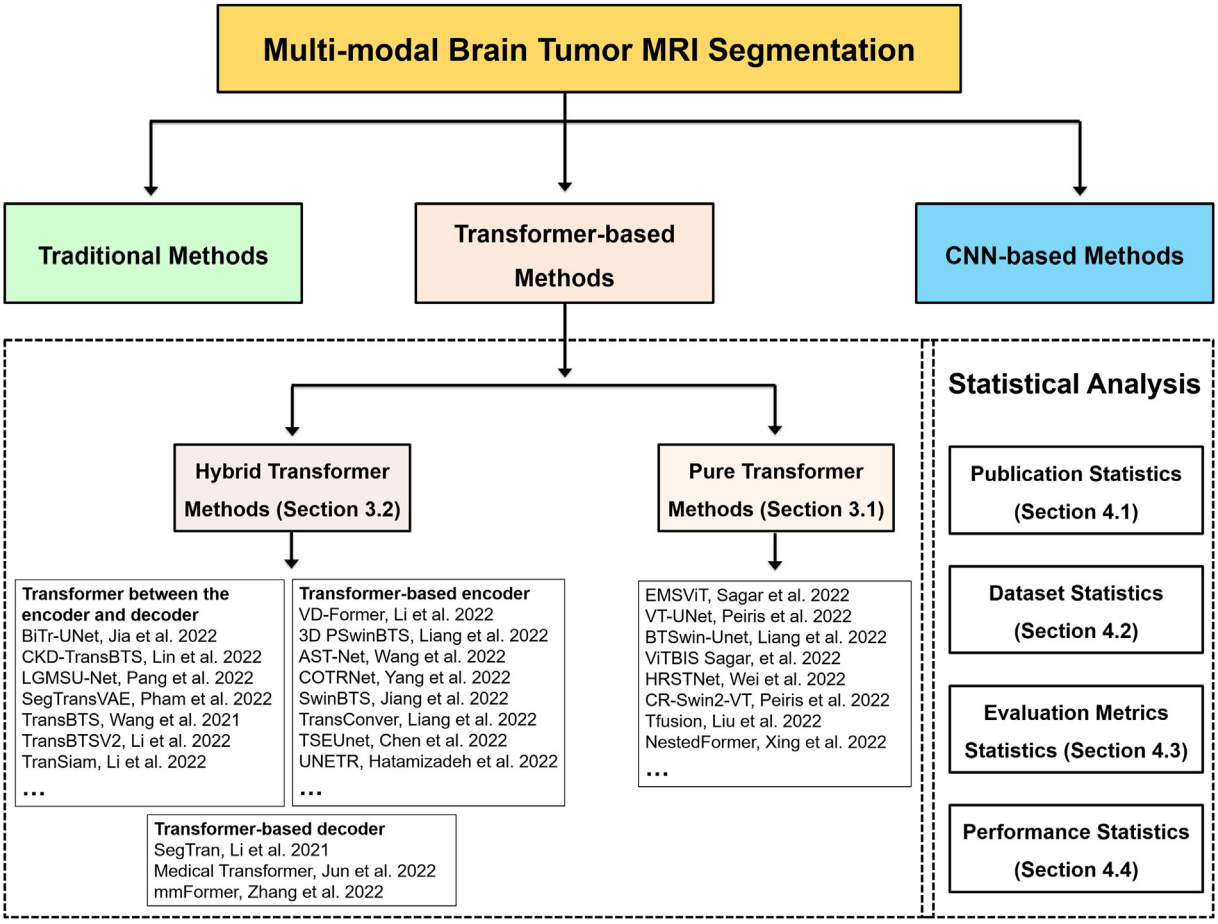

Fig. 3. An overview of the transformer-based methods for multi-modal brain tumor MRI segmentation.

connections. Here, the up- and down-sampling of features are implemented by convolutions and deconvolutions, respectively. Furthermore, BTSwin-UNet introduces a self-supervised learning scheme, which randomly masks some regions on input images and applies an additional reconstruction decoder to recover these regions. Considering that the masked autoencoder is a powerful feature extractor, the above self-supervised learning scheme effectively promotes the encoder to extract high-quality features, which enables BTSwin-UNet to achieve better brain tumor segmentation compared with other methods.

To further improve segmentation performance, some studies present the variants of SwinT. Specifically, Peiris et al.[33] propose a UNet-shaped Volumetric Transformer (VT-UNet). In the encoder, VT-UNet utilizes the classical SwinT blocks. In the decoder, VT-UNet designs VT block that includes two parallel SwinT blocks, where the first and the second SwinT blocks are applied to process the encoder and decoder features respectively, and the final output is obtained by fusing two SwinT blocks outputs and the Fourier feature positional encodings. Then, Peiris et al.[34] develop CR-Swin2-VT, which uses parallel CSwinT and SwinT blocks as the encoder to improve the feature extraction quality, thereby improving brain tumor segmentation results. In addition, CR-Swin2-VT presents a feature fusion manner based on learnable weights to combine the outputs of CSwinT and SwinT blocks at each stage. The above two methods successfully utilize the advantages of SwinT, i.e., they model local feature relations by using memory-efficient window self-attentions, while establishing global feature relations via shifted windows, resulting in pleasant brain tumor segmentation results. Following the style of HRNet, Wei et al.[35] build a High-Resolution Swin Transformer Network (HRSTNet) using standard SwinT blocks. Different from the classical U-shaped encoder-decoder structure, HRSTNet adopts a parallel multi-resolution path to effectively preserve high-resolution information within each network stage. Although such paths increase some computations, they also effectively reduce missing spatial information caused by feature re-sampling operations. Different from the above methods, HRSTNet takes more consideration on high-resolution features and applies multi-resolution feature fusion in each stage. With this design, HRSTNet commendably promotes the interaction between different-scale features, thus achieving excellent segmentation performance in pure transformer methods.

On the other hand, some scholars design high-performance transformers by effectively using multi-modal MRI. Xing et al.[36] propose a Nested Modality-Aware Transformer (NestedFormer), which synthetically considers the feature relations within and between modalities to improve segmentation performance. NestedFormer constructs multiple transformer-based encoders to extract the model-specific features, and presents a Nested Modality-aware Feature Aggregation (NMaFA) scheme to achieve the multi-modal feature fusion for simultaneously modeling the intra-modal and inter-modal feature relations. In specific, NMaFA combines the tri-oriented spatial attention and the cross-modality attention, where the former can model the spatial feature relations within a modality from different views, and the latter can model the global feature relations between all modalities. To enhance skip connections, NestedFormer designs a Modality-Sensitive Gating (MSG), which learns the importance of different modalities in each stage fusion features, and produces more representative features for skip connections by assigning weights to different modalities. Overall, NestedFormer successfully establishes the correlations between MRI modalities, which is a novel and workable scheme compared with transformer structure advances to further improve brain tumor segmentation.

Aiming at the problem of missing modalities that usually occur in multi-modal learning, Liu et al.[37] propose a transformer-based multi--modal feature fusion method called TFusion. Unlike previous works, TFusion does not need to perform zero-padding on missing modalities or generate missing modality features. Concretely, TFusion learns the relations between existing modality features through multiple transformer layers, and then adaptively generates modal-specific pixel-wise weight maps by a modal softmax operation to fuse existing modality features for generating multi-modal fusion features. TFusion further considers the missing modality situations in multi-modal learning, and achieves implicit modal-invariant feature extraction to boost model robustness and effectiveness. In addition, TFusion is plug-and-play, which can be effectively applied in various transformer-based methods and bring satisfactory performance improvements.

In summary, most pure transformer methods aim to improve the local relation modeling ability and reduce the model complexity by proposing novel basic transformer blocks. Their experimental results also show that pure transformer methods perform favorably against state-of-the-art CNNs. As shown in Fig. 4, we exhibit the architectures of some typical pure transformer methods for brain tumor segmentation.

3.2. Hybrid transformer methods

Compared with pure transformer methods, many studies prefer to combine CNN which can well model local feature relations, and transformer which can well model global feature relations for accurate brain tumor segmentation. According to the combination manners of CNN and transformer, hybrid transformer methods are usually divided into three classes: (i) methods with the transformer-based encoder; (ii) methods with the transformer between the encoder and decoder; (iii) methods with the transformer-based decoder.

The first class methods aim to employ the transformer's advantage in modeling long-range dependencies to extract higher-quality encoding features. Therefore, some studies directly adopt transformers as the encoder. UNETR[42] is a representative work, which has a typical encoder-decoder structure. In UNETR, the encoder is constructed based on ViT, and the decoder is built using multiple 3×3 convolutions and 2×2 deconvolutions similar to U-Net. Compared with pure transformer methods, only using the transformer encoder can decrease some computational costs, but it will not affect segmentation performance. Unlike UNETR, Hatamizadeh et al.[43] propose UNetFormer, which adopts 3D SwinT as the encoder, and combines 3D CNN and SwinT as the decoder. In addition, UNetFormer aligns each scale decoder output with the ground truth through a deep supervision mechanism. By comparison, UNetFormer has a light-weight model, where the use of deep supervision improves the quality of decoding features as well as brain tumor segmentation performance.

Considering the SwinT superiorities in terms of performance and computation compared with ViT, many studies construct SwinT-based encoders. Among them, Hatamizadeh et al.[44] directly use 3D SwinT as the encoder and CNN as the decoder to build a U-shaped brain tumor segmentation network. Liang et al.[45] improve SwinT by proposing a parallel shifted window-based transformer block, which first performs MSA in three directions (horizontal, longitudinal and vertical), and then aggregates MSA outputs by the shifted window operation. Moreover, they also introduce a semantic-prior prediction branch parallel to the encoder, which effectively improves the encoder's ability to capture high-level semantic features by drawing closer to the predicted semantic maps and ground truths. In this approach, multi-direction MSA is a key implementation to improve segmentation performance, which borrows the human habit of observing objects, that is, humans always obtain comprehensive object information from multiple views. This scheme is also applicable to other computer vision tasks. Inspired by GoogLeNet,[74] Liang et al.[46] propose TransConver, which designs Transformer-Convolution Inception (TC-Inception) module in the encoder. TC-Inception module consists of the parallel convolution block and SwinT block, and a Cross-Attention Fusion with Global and Local feature (CAFGL) mechanism. In the encoder, the convolution and transformer blocks extract local and global features respectively, and CAFGL mechanism can fuse global-local features by the cross-modal attention. Finally, the cross-attention fusion mechanism is also applied in skip connections to enhance model performance. TransConver makes a structural advance to improve brain tumor segmentation, where TC-Inception module extracts complementary local and global features, and the CAFGL module enables high-quality local-global feature fusion. Lin et al.[47] propose SwinBTS, which adopts the alternating 3D SwinT block and 3D DW convolution block to construct the encoder as well as the decoder. In addition, SwinBTS develops an enhanced transformer module composed of the convolution and MSA between the encoder and decoder to boost the extracted global-local features. Similar to SwinBTS, in COnvolution-and-TRansformer Network (COTRNet), Yang et al.[48] build both the encoder and decoder using alternating ResBlock and transformer block, and introduce a topology-aware loss to learn the topological information, thereby improving segmentation results. Experiments show that the above alternate setting for transformer and convolution also brings explicit performance improvements on brain tumor segmentation,

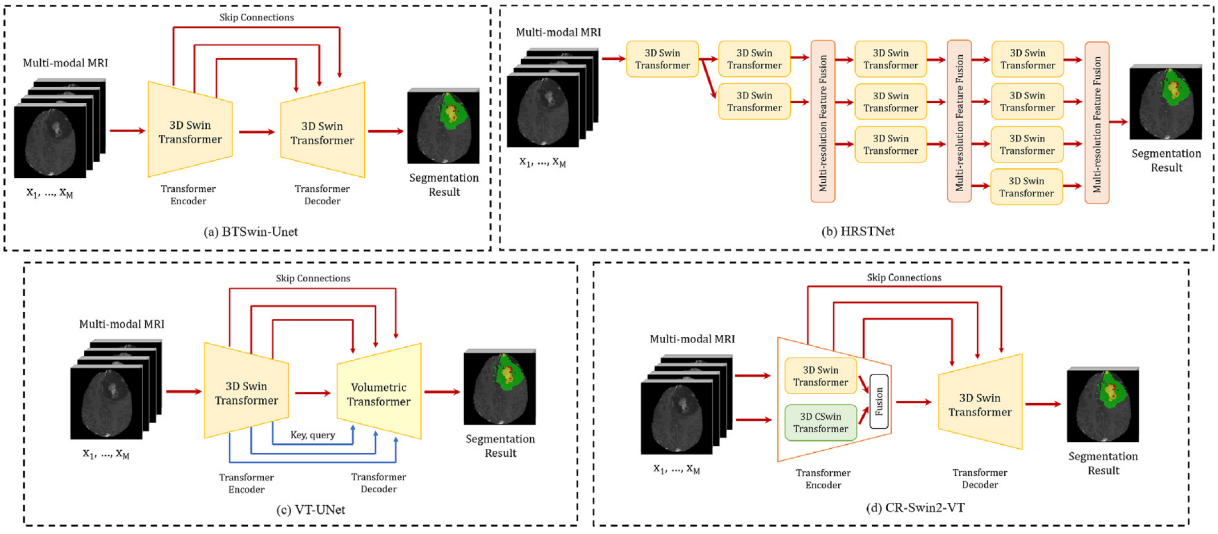

Fig. 4. Typical pure transformer methods for brain tumor segmentation, (a) architecture of BTSwin-Unet (Liang et al.[32]); (b) architecture of HRSTNet (Wei et al.[35]); (c) architecture of VT-UNet (Peiris et al.[33]); (d) architecture of CR-Swin2-VT (Peiris et al.[34]).

we consider that this scheme integrates the extraction of local and global information, thereby improving model robustness.

Taking 3D view information as the starting point, Li et al.[49] propose a brain tumor segmentation model, which consists of a transformer-based pyramid network and multiple cascade RCNNs. Its core component is View-Disentangled Transformer (VD-former), which first rotates a 3D MRI image of size C×H×W×T into the T×W view, T×H view and H×W view images, and then add the original image to three view images to generate the input of VD-former. Note that the view-disentangled design in VD-former further confirms that multi-view MSA is beneficial to improve transformer-based methods. Wang et al.[54] design a light-weight hybrid transformer method named AST-Net. In AST-Net, they comprehensively consider the Y-axis, X-axis and Z-axis information of 3D MRI images, and construct Axial-Spatial Transformer (AST) module in modeling the global feature relations as well as Hierarchical Decoupled Convolution (HDC) in modeling the local feature relations. By comparison, the proposed AST and HDC modules enable AST-Net further reduce the model complexity. In order to reduce computational costs, Liu et al.[50] present a 2D Transition Net to segment 3D brain tumor MRI. This hybrid transformer first adopts a transition head to fuse multi-modal MRI images and compresses the feature dimension to generate input tokens to feed the transformer. Then, Transition Net builds the encoder based on SwinT blocks and the decoder based on convolution blocks. Through the above designs,[50] shows a clear advantage in model complexity compared with existing methods.

After that, some studies propose combining the transformer-based encoder and the CNN-based encoder to jointly extract features. Dhamija et al.[51] directly set up the parallel ViT-style encoder and the U-Net style encoder for feature extraction, and then integrate local and global features through a fully convolutional ensemble decoder to predict segmentation results. Following this idea, Chen et al.[52] propose TSEUNet for brain tumor segmentation. They also build the parallel transformer-based encoder and the CNN-based encoder to model the local and global feature relations simultaneously. In addition, the deep supervision and SE-attention mechanisms are utilized in TSEUNet to improve segmentation performance. Compared with methods using alternate transformers and convolutions in a single encoder, the parallel encoder designs usually show better segmentation results since the task of each encoder is assigned more clearly. Hu et al.[53] present Efficient R-Transformer Network (ERTN) with a dual encoder for brain tumor segmentation. In ERTN, the parallel R-Transformer and U-Net encoders are developed to extract feature-wise and patch-wise information. Benefiting from the proposed rank-attention mechanism to extract the most important top-k queries after ranking for reducing transformer computations, R-Transformer well balances the transformer performance and complexity.

Considering that image boundary information is crucial for brain tumor segmentation, Gai et al.[56] present Residual Mix Transformer Fusion Net (RMTF-Net), which can progressively fuse the features extracted by the ResBlock and the mix transformer block to balance the global and local information in the encoder. Here, RMTF-Net adopts an overlapping patch embedding strategy to alleviate the impact of missing boundary information, and designs a Global Feature Integration (GFI) module for skip connections to enhance decoding features. Zhu et al.[55] propose a multi-branch hybrid transformer, which contains a SwinT-based branch to extract semantic information, a CNN-based branch to extract boundary information, and a graph convolution-based module to fuse semantic and boundary information. In the edge detection branch, they design a Sobel-based edge spatial attention block to boost the boundary information extraction. In Ref. [55], explicitly using image boundary information can apply a constraint on segmentation results, which significantly reduces missegmentation caused by irregular changes of pixel intensities.

Different from the above methods, the second class methods insert the transformer between the encoder and decoder. Since the encoding features have lower spatial resolutions and more semantic information, the second class methods can use less computational costs to model efficient global feature relations. Among recent representative works, TransBTS is proposed by Wang et al.,[57] they first adopt a CNN-based encoder to extract features from multi-modal MRI, and then divide the encoding features into a sequence of non-overlapping patches. After adding learnable position embeddings, multiple standard transformer blocks are used to model the global relations in encoding features. Finally, by reshaping the transformer output into 2D feature maps to feed a CNN-based decoder, the final tumor prediction results are generated. Due to limited transformer layers, TransBTS has similar parameters with pure CNN methods, but exhibits more satisfying segmentation results, which proves the effectiveness of the above-mentioned second class methods. On the basis of TransBTS, Dobko et al.[58] replace ResBlocks in the encoder and decoder with the channel attention-based SE-Residual blocks, and introduce a learnable MLP to produce positional embeddings to improve segmentation performance. Then, Pham et al.[59] propose SegTransVAE, which is a multi-task learning framework consisting of a shared CNN-based encoder, a transformer-based feature enhancement module and two task-specific CNN-based decoders. SegTransVAE can jointly achieve brain tumor segmentation and brain tumor image reconstruction. Lyu et al.[60] propose a transformer-based brain tumor segmentation model with the typical U-Net structure. They insert 12 standard transformer blocks and positional encodings between the CNN-based encoder and decoder to capture the non-local relations in encoding features. Overall, although the above methods have similar implementation principles, they demonstrate that various well-established structural innovations such as attentions and transformer variants easily improve performance. Similarly, Huang et al.[61] build a Generative Adversarial Network (GAN) with the encoder-decoder structure to segment brain tumors. In this GAN, four standard transformer blocks are applied to enhance encoding features. Then, Gao et al.[62] present a deep mutual learning scheme to improve brain tumor segmentation results. They first develop a hybrid CNN-transformer model, where CNN acts as the encoder and decoder, and the transformer is inserted between the encoder and decoder. In addition, a deep supervision scheme is introduced to ensure the effectiveness of different-scale decoder features, and the shallow features are employed to supervise the subsequent features for better retaining edge information. Finally, the logits output by the deepest layer is applied to supervise the previous layer logits. With the above progressive deep supervision scheme, each stage semantic information is well enhanced. As expected, experimental results display that this approach outperforms many existing methods.

Using transformers to enhance skip connections is another way to apply the transformer between the encoder and decoder. Among these methods, Jia et al.[63] propose BiTr-UNet, which adopts the standard transformer blocks to enhance the last and penultimate scale encoder features. After that, Pang et al.[64] propose Axial-Deformable Attention Module (ADAM), which introduces the axial information in dynamic MSA to boost each stage encoder features for skip connections. Specifically, ADAM is composed of continuous Vertical-Deformable Attention (VDA) and Horizontal-Deformable Attention (HDA), where the combination of VDA and HDA comprehensively captures the important non-local feature relations. Unlike inserting transformers into the encoder and decoder, enhancing skip connections via transformers seems only to bring limited performance improvements.

In TransBTSV2,[65] a Flexibly Widened Multi-Head Self-Attention (FW-MHSA) block is proposed between the encoder and decoder. Unlike the classical MSA, FW-MHSA is designed as a shallower but wider architecture, which can achieve competitive performance while reducing computations. Furthermore, they propose a plug-and-play and energy-efficient Deformable Bottleneck Module (DBM), which can enhance skip connections by capturing shape-aware feature representations through 3D deformable convolutions. By comparison, TransBTSV2 not only uses transformers between the encoder and decoder to capture non-local feature relations, but also introduces DBM to enhance skip connections. As result, more accurate segmentation performance is obtained. Li et al.[66] propose TransSiam, which includes a dual-branch

encoder-decoder and a TMM block with the cross-attention and self-attention. In the dual-branch encoder, the CNN and transformer-based ICMT blocks are successively applied to extract features from multi-modal MRI. Then, the modal-specific encoding features are fed into TMM for multi-modal feature fusion.

Different from the above methods, Lin et al.[67] consider the clinical knowledge related to brain tumor diagnosis to construct a brain tumor segmentation model called Clinical Knowledge-Driven Hybrid Transformer (CKD-TransBTS). CKD-TransBTS first divides multi-modal MRI images into two groups according to their imaging principles, and then builds a dual-branch hybrid encoder to extract their features respectively. In the encoder, Modality-Correlated Cross-Attention (MCCA) block contains two parallel paths and each path consists of MSAs and convolutions, where the cross-modal attention connects two paths to achieve the feature interaction between multi-modal features. In the decoder, Trans & CNN Feature Calibration (TCFC) block first pools the input multi-modal features in three dimensions (X, Y, and Z), and then establish the spatial attention by modeling the relations between different-dimension features to enhance the decoder features. In summary, CKD-TransBTS not only considers tumor-related clinical information to extract multi-modal MRI features, but also establishes multi-view interactions across modalities to improve segmentation performance.

In addition to the above two class methods, some studies adopt transformers as the decoder to achieve brain tumor segmentation. Considering that the decoder features contain more high-level semantic information, these methods aim to use the transformer's powerful global

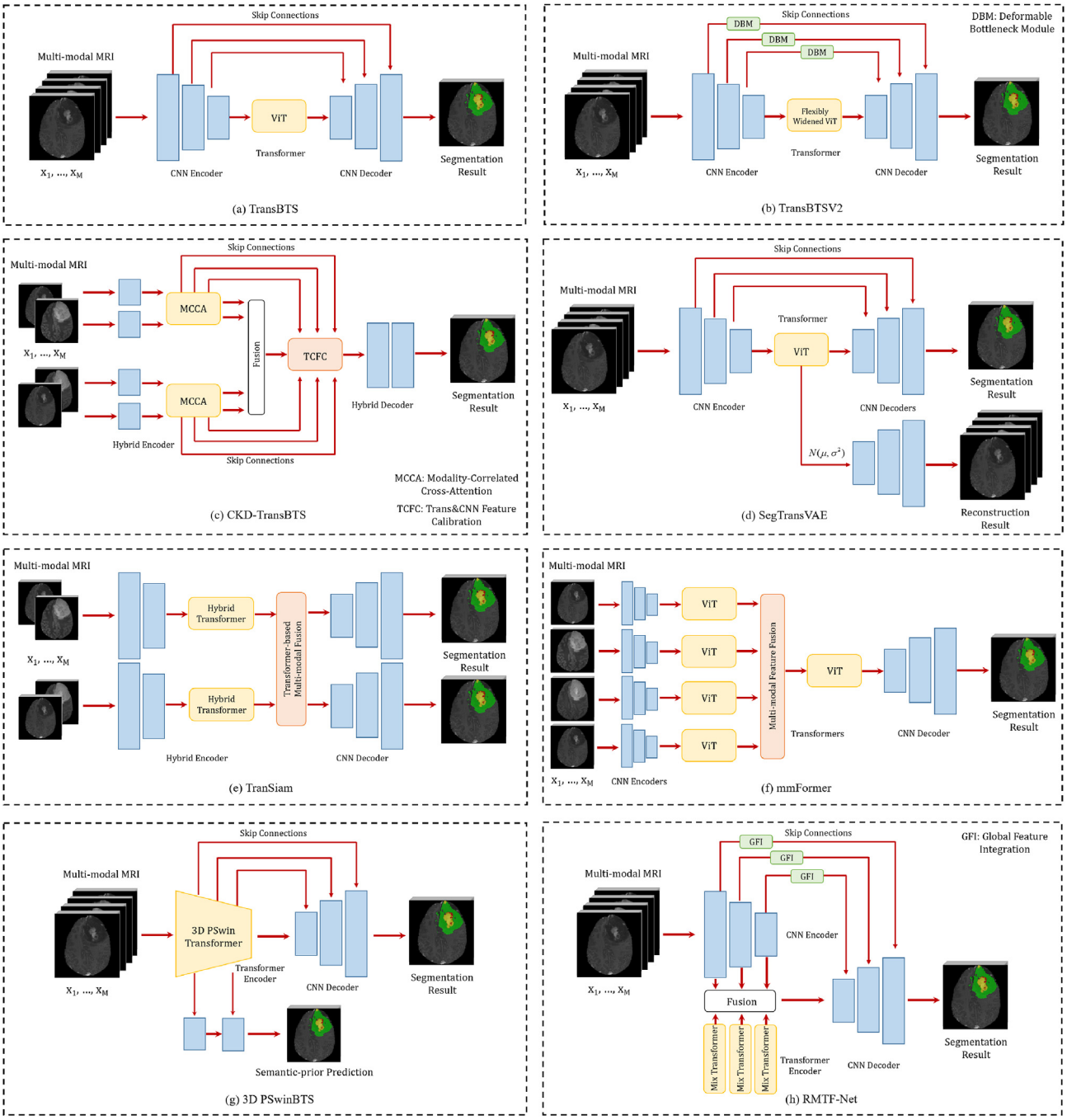

Fig. 5. Typical hybrid transformer methods for brain tumor segmentation, (a) architecture of TransBTS (Wang et al.[57]); (b) architecture of TransBTSV2 (Li et al.[65]); (c) architecture of CKD-TransBTS (Lin et al.[67]); (d) architecture of SegTransVAE (Pham et al.[59]).(e) architecture of TranSiam (Li et al.[66]); (f) architecture of mmFormer (Zhang et al.[70]); (g) architecture of RMTF-Net (Gai et al.[56]); (h) architecture of PSwinBTS (Liang et al.[45]).

modeling capability to explore and learn the relations between tumor regions and backgrounds to improve the segmentation accuracy. Specifically, Li et al.[68] propose Seg-Tran, which first extracts visual features through a CNN-based encoder, and encodes the coordinate information into sinusoidal positional encodings, and then presents a Squeeze-and-Expansion (SE) transformer as the decoder to process spatially flattening feature patches. Different from previous transformer variants, the proposed SE transformer adopts the squeezed attention block to regularize the large-scale attention matrix to reduce computations, and uses the expansion block to combine different-modal features to learn diverse feature representations. Jun et al.[69] propose a transfer learning-based framework named Medical Transformer, which considers the three planes (sagittal, coronal and axial) of multi-modal MRI to comprehensively model high-level representations, and effectively learns multiple relevant tasks including brain disease diagnosis, brain tumor segmentation and brain age prediction. Medical Transformer first rotates MRI inputs at multiple views, and then fed them into three view-specific CNN-based encoders. To capture discriminative information, a transformer-based decoder is utilized to process view-specific encoding features. After fusing different-view features, a prediction network is introduced to achieve the above three tasks. Moreover, both the CNN-based and transformer-based decoders are pre-trained to ensure generalization and effectiveness. Zhang et al.[70] propose a Multi-modal Medical Transformer (mmFormer) to achieve brain tumor segmentation with missing modalities. mmFormer first uses modality-specific CNN-based encoders to extract features from multi-modal MRI separately. Then the modal-specific features are flattened and input into the corresponding transformers. In order to alleviate the influence of missing modalities, a transformer-based modality-correlated encoder is proposed to explore the non-local feature relations between multiple modalities, and realize the multi-modal feature fusion. Finally, the fusion features are reshaped and fed into a CNN-based decoder for incomplete multi-modal brain tumor segmentation. Note that, mmFormer applies the DSC loss-based auxiliary regularizers to improve the effectiveness of both encoder and decoder features. In mmFormer, the potential inter-modality relations on various missing modality situations are comprehensively considered. Therefore, mmFormer shows a robust performance compared with existing brain tumor segmentation methods.

In summary, most current studies are committed to effectively combining CNN and transformer to construct hybrid transformer methods. Compared with pure transformer methods, hybrid transformer methods usually model better local feature relations and have lower model complexity. Experimental results also show that hybrid transformer methods can achieve state-of-the-art segmentation performance. As shown in Fig. 5, we exhibit the architectures of some typical hybrid transformer methods for brain tumor segmentation.

4. Statistical analysis of transformer-based methods

This section conducts a comprehensive statistical analysis for transformer-based brain tumor segmentation methods. Firstly, we show the number of publications in the past three years. Secondly, we report the commonly-used brain tumor segmentation datasets in existing studies. Then, we introduce the well-established evaluation metrics for brain tumor segmentation. Finally, we count and compare the segmentation performance of existing studies.

4.1. The publication statistics

In 2021, Dosovitskiy et al.[26] introduce the transformer that performs well in NLP in computer vision for the first time, called vision transformer. In recent years, many transformer-based methods have been developed for various computer vision tasks such as image classification, segmentation, detection, restoration and generation. In multiple common databases, including IEEE Xplorer, Springer Link, Science Direct, PubMed, Wiley, etc., we use the keywords “brain tumor segmentation”, “multi-modal”, “transformer”, “brain MRI”, etc. for searching and a total of 42 most relevant references were retrieved. These transformer-based brain tumor segmentation methods are published from 2021 to 2023 and we count the number of publications as illustrated in Fig. 6. It can be seen that the earliest transformer-based brain tumor segmentation method appear in 2021, and the publications increase dramatically in 2022. Moreover, the number of publications throughout 2022 shows an increasing trend (14 papers from Jan. to Jun., 21 papers from Jul. to Dec.), and in January 2023, 6 relevant papers have been published. According to the above statistics, transformer-based methods will be further developed for brain tumor segmentation in the next 1–2 years.

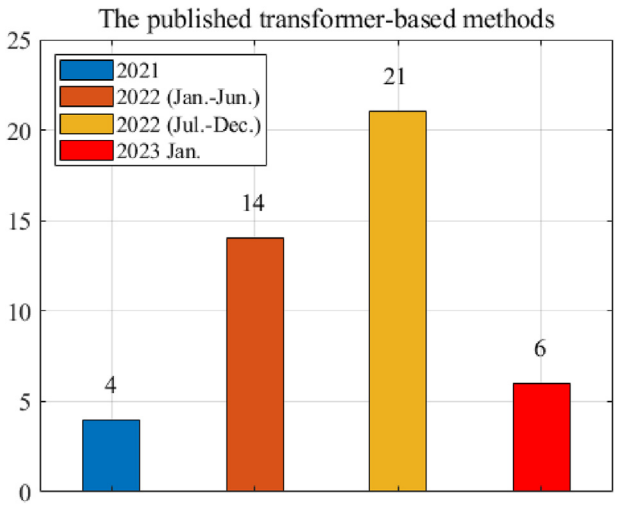

Fig. 6. Statistical results of the publications from 2021 to 2023.

4.2. The commonly-used dataset statistics

Multi-modal MRI datasets are crucial to the training and testing of transformer-based brain tumor segmentation methods. Since 2012, Medical Imaging Computing and Computer-Aided Intervention Association (MICCAI) has launched the multi-modal Brain Tumor Image Segmentation (BraTS) challenge to facilitate the research and comparison of brain tumor segmentation. From 2012 to 2021, MICCAI will publish a new BraTS dataset every year, which contains four modality MRI images (T1, T1ce, T2 and FLAIR) (see Table 5). Benefiting from this challenge, most studies use BraTS datasets to train and test brain tumor segmentation methods. As shown in Fig. 7, we count the multi-modal MRI datasets applied in transformer-based methods over the past three years. It can be observed that more than 90% of studies utilize BraTS datasets. The BraTS 2019, 2020 and 2021 are the most commonly-used datasets, which are closely related to the development times of brain tumor segmentation methods. In addition, few studies[42,49,51,73] adopt the private, MSD, TCIA

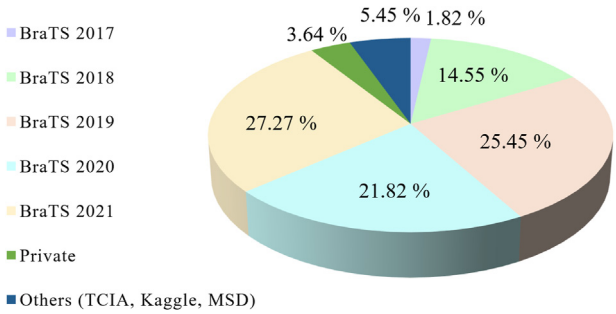

Fig. 7. Statistical results of the commonly-used brain tumor datasets.

and Kaggle datasets to demonstrate the segmentation performance. Although the private datasets enable the comparison of brain tumor segmentation, there is a large gap between the private and public datasets, while the pixel-level annotation of private datasets is difficult and time-consuming. Hence, developing brain tumor segmentation methods on the publicly available BraTS datasets is still the main trend of future studies.

4.3. Evaluation metrics statistics

In brain tumor segmentation, the model's performance can be quantified and compared by various evaluation metrics. In specific, the most widely-used evaluation metrics include Dice Similarity Coefcient (DSC), Hausdorf Distance (HD), Intersection-over-Union (IoU), Accuracy (ACC) and Sensitivity (SE). In the following contents, we will introduce these evaluation metrics in detail. Firstly, DSC can evaluate the coincidence degree between two samples. In fact, for the segmentation result and the ground truth, DSC is the ratio of their twice overlapping region to their total region. The expression of DSC is defined as:

$$\mathrm{DSC} = \frac{2\left|Y_{pre} \cap Y_{GT}\right|}{\left|Y_{pre}\right| + \left|Y_{GT}\right|}, \tag{7}$$

where Y_{pre} represents the segmentation result, and Y_{GT} represents the ground truth.

In the segmentation task, DSC is equivalent to the F1 measure, such that the expression of DSC can be rewritten as:

$$\mathrm{DSC} = \frac{2TP}{2TP + FP + FN}, \tag{8}$$

where TP denotes the true positive, which is the result that the model correctly predicts the tumor region, FP denotes the false positive, which is the result that the model incorrectly predicts the background region, and FN denotes the false negative, which is the result that the model incorrectly predicts the tumor region.

Then, IoU can also measure the overlapping region between the segmentation result and the ground truth. It is the ratio of the intersection of the segmentation result and the ground truth to their corresponding union. The IoU expression can be given as:

$$\mathrm{IoU} = \frac{\left|Y_{pre} \cap Y_{GT}\right|}{\left|Y_{pre} \cup Y_{GT}\right|}. \tag{9}$$

It is worth noting that IoU is equivalent to the Jaccard similarity coefficient, so we can rewrite the IoU expression based on TP, FP and FN:

$$\mathrm{IoU} = \frac{TP}{TP + FP + FN}. \tag{10}$$

Moreover, SE is a common-used evaluation metric, which can evaluate whether the segmentation result is accurate by calculating the proportion of tumor voxels that are correctly classified. Its expression is:

$$SE = \frac{TP}{TP + FN}. \tag{11}$$

Similarly, ACC comprehensively considers the correct classification of both the tumor voxels and the background voxels to evaluate segmentation performance, its expression is defined as:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}. \tag{12}$$

HD is a distance-based evaluation metric for brain tumor segmentation. In HD, we consider the segmentation result and the ground truth as two subsets in the measure space, such that the expression is written as:

$$HD = \max\left\{ \sup_{Y_{pre}} \inf_{Y_{GT}} d(Y_{pre}, Y_{GT}),\ \sup_{Y_{GT}} \inf_{Y_{pre}} d(Y_{pre}, Y_{GT}) \right\}, \tag{13}$$

where *sup* represents the supremum and *inf* represents the infmum.

In summary, for DSC, IoU, SE and ACC, the larger value indicates the segmentation result is closer to the ground truth. When the value is 1, the segmentation result is completely consistent with the ground truth. In contrast, for the distance-based HD, the smaller value represents the better segmentation performance.

4.4. Brain tumor segmentation performance statistics

During 2021–2023, 11 pure transformer methods and 31 hybrid transformer methods are proposed for brain tumor segmentation. As shown in Tables 3 and 4, we summarize the existing transformer-based

Table 3
Statistical results of pure transformer methods.

Author	Year	Segmentation Method	Segmentation Performance			Dataset
			WT	TC	ET	
Liang et al.[32]	2022	3D U-shaped Symmetrical Swin Transformer-based Network (BTSwin-Unet)	DSC = 0.903 HD = 4.977	DSC = 0.817 HD = 6.062	DSC = 0.784 HD = 4.286	BraTS 2018, 2019
Pinaya et al.[39]	2022	Vector Quantised Variational Autoencoder with an Ensemble of Autoregressive Transformer	Avg DSC = 0.537 Avg AUPRC = 0.555			BraTS 2018
Sagar et al.[38]	2022	Efficient Multi-scale Vision Transformer (EMSViT)	DSC = 0.903 HD = 5.621	DSC = 0.822 HD = 7.129	DSC = 0.792 HD = 3.706	BraTS 2019
Peiris et al.[33]	2022	UNet-shaped Volumetric Transformer (VT-UNet)	DSC = 0.919 HD = 3.510	DSC = 0.822 HD = 2.680	DSC = 0.872 HD = 4.100	BraTS 2019
Sagar et al.[30]	2022	Vision Transformer for Biomedical Image Segmentation (ViTBIS)	DSC = 0.903	DSC = 0.822	DSC = 0.792	BraTS 2019
Liu et al.[37]	2022	Transformer based N-to-one Multimodal Fusion (TFusion)	DSC = 0.889	DSC = 0.822	DSC = 0.738	BraTS 2020
Xing et al.[36]	2022	Nested Modality-aware Transformer (NestedFormer)	DSC = 0.920 HD = 4.567	DSC = 0.864 HD = 5.316	DSC = 0.800 HD = 5.269	BraTS 2020
Wei et al.[35]	2022	High-Resolution Swin Transformer Network (HRSTNet)	DSC = 0.921 HD = 4.150	DSC = 0.869 HD = 5.500	DSC = 0.826 HD = 13.910	BraTS 2021
Peiris et al.[34]	2022	Hybrid Window Attention based Transformer (CR-Swin2-VT)	DSC = 0.914 HD = 3.930 SE = 91.23	DSC = 0.854 HD = 11.190 SE = 84.33	DSC = 0.817 HD = 14.810 SE = 82.38	BraTS 2021
Andrade-Miranda et al.[41]	2022	Pure Versus Hybrid Transformer	DSC = 0.896	DSC = 0.867	DSC = 0.856	BraTS 2021
Chen et al.[40]	2022	UNet-shaped Transformer with Convolutional Block Attention Module (CBAM-TransUNet)	DSC = 0.931 HD = 4.910	DSC = 0.915 HD = 4.200	DSC = 0.878 HD = 2.930	BraTS 2021
Karimijafarbigloo et al.[75]	2023	Missing Modality Compensation Transformer (MMCFormer)	DSC = 0.890	DSC = 0.874	DSC = 0.801	BraTS 2018

Table 4
Statistical results of hybrid transformer methods.

Author	Year	Segmentation Method	Segmentation Performance			Dataset
			WT	TC	ET	
Jun et al.[69]	2021	Medical Transformer	DSC = 0.873	DSC = 0.697	DSC = 0.588	BraTS 2019
Li et al.[68]	2021	Alternative Segmentation Framework based on Squeeze-and-expansion Transformer (SegTran)	DSC = 0.895	DSC = 0.817	DSC = 0.740	BraTS 2019
Wang et al.[57]	2021	Multimodal Brain Tumor Segmentation using Transformer (TransBTS)	DSC = 0.901 HD = 9.769	DSC = 0.817 HD = 4.964	DSC = 0.787 HD = 17.947	BraTS 2019, 2020
Liang et al.[46]	2022	U-shaped Segmentation Network based on Convolution and Transformer (TransConver)	DSC = 0.859 HD = 2.587	DSC = 0.838 HD = 1.607	DSC = 0.789 HD = 2.692	BraTS 2018, 2019
Zhang et al.[70]	2022	Multimodal Medical Transformer (mmFormer)	DSC = 0.896 HD = 4.150	DSC = 0.858 HD = 5.500	DSC = 0.776 HD = 13.910	BraTS 2018
Pang et al.[64]	2022	Local Features, Global Features and Multi-scale Features Fused the U-Shaped Network (LGMSU-Net)	Avg DSC = 0.874			BraTS 2018
Chen et al.[52]	2022	3D Neural Network with Fused Transformer and SE-attention (TSEUnet)	DSC = 0.911 SE = 0.914	DSC = 0.873 SE = 0.868	DSC = 0.824 SE = 0.837	BraTS 2018
Wang et al.[44]	2022	Axial-spatial Transformer (AST-Net)	DSC = 0.904 HD = 6.050	DSC = 0.842 HD = 6.120	DSC = 0.778 HD = 30.430	BraTS 2018, 2019, 2020
Liu et al.[50]	2022	3D Multimodal Brain Tumor Image Segmentation Model Transition Net	DSC = 0.913 HD = 20.145	DSC = 0.845 HD = 12.212	DSC = 0.749 HD = 10.092	BraTS 2019
Li et al.[65]	2022	TransBTS V2	DSC = 0.906 HD = 4.272	DSC = 0.845 HD = 5.560	DSC = 0.796 HD = 12.522	BraTS 2019, 2020
Li et al.[66]	2022	Fusing Multimodal Visual Features using Transformer (TranSiam)	Avg DSC = 0.893 Avg HD = 5.650			BraTS 2019, 2020
Gai et al.[56]	2022	Residual Mix Transformer Fusion Net (RMTF-Net)	Avg DSC = 0.935, Avg IoU = 0.882 Avg Sm = 0.948, Avg wFm = 0.941			BraTS 2019, 2020
Jiang et al.[47]	2022	3D Multimodal Brain Tumor Segmentation Using Swin Transformer (SwinBTS)	DSC = 0.918 HD = 3.650	DSC = 0.848 HD = 14.510	DSC = 0.832 HD = 16.030	BraTS 2019, 2020, 2021
Nalawade et al.[71]	2022	Federated Learning Medical Segmentation Transformer	DSC = 0.767 HD = 31.075	DSC = 0.612 HD = 30.089	DSC = 0.628 HD = 29.482	BraTS 2020
Huang et al.[61]	2022	Transformer-based Generative Adversarial Network	DSC = 0.903 HD = 4.909	DSC = 0.815 HD = 7.494	DSC = 0.708 HD = 37.579	BraTS 2020
Liang et al.[45]	2022	Efficient Transformer-based UNet using 3D Parallel Shifted Windows (3D PSwinBTS)	DSC = 0.926 HD = 3.738	DSC = 0.867 HD = 11.084	DSC = 0.826 HD = 17.531	BraTS 2020, 2021
Shi et al.[72]	2022	An Ensemble Approach for Brain Tumor Segmentation	DSC = 0.892 HD = 5.772	DSC = 0.738 HD = 16.874	DSC = 0.819 HD = 16.628	BraTS 2021
Jia et al.[63]	2022	CNN-Transformer Combined Network (BiTr-UNet)	DSC = 0.934 HD = 2.828	DSC = 0.930 HD = 2.236	DSC = 0.889 HD = 1.414	BraTS 2021
Dobko et al.[58]	2022	The Modified TransBTS	DSC = 0.926 HD = 3.374	DSC = 0.869 HD = 11.057	DSC = 0.849 HD = 15.723	BraTS 2021
Pham et al.[59]	2022	Hybrid CNN-Transformer with Regularization (SegTransVAE)	DSC = 0.905 HD = 3.570	DSC = 0.926 HD = 5.840	DSC = 0.855 HD = 2.890	BraTS 2021
Yang et al.[48]	2022	Convolution-and-transformer Network (COTRNet)	DSC = 0.947 HD = 6.164	DSC = 0.952 HD = 9.000	DSC = 0.923 HD = 3.6736	BraTS 2021
Futrega et al.[76]	2022	The Optimized UNETR	Avg DSC = 0.916			BraTS 2021
Hatamizadeh et al.[44]	2022	Swin UNet Transformer (Swin UNETR)	DSC = 0.926 HD = 5.831	DSC = 0.885 HD = 3.770	DSC = 0.858 HD = 6.016	BraTS 2021
Hatamizadeh et al.[43]	2022	Unified Vision Transformer Model (UNetFormer)	DSC = 0.932 HD = 31.075	DSC = 0.921 HD = 30.089	DSC = 0.888 HD = 29.482	BraTS 2021
Dhamija et al.[51]	2022	The Transfused Convolution and Transformer Network (USegTransformer)	Avg DSC = 0.8934, Avg IoU = 0.9746 Avg ACC = 0.9971			TCIA
Wang et al.[73]	2022	TransUNet	Avg DSC = 0.864			Kaggle
Hatamizadeh et al.[42]	2022	UNet Transformers (UNETR)	DSC = 0.789 HD = 8.266	DSC = 0.761 HD = 8.845	DSC = 0.585 HD = 9.354	MSD
Lyu et al.[60]	2022	Transformer-based Brain Tumor Segmentation Network	Avg DSC = 0.887 Avg SE = 0.937			Private
Li et al.[49]	2022	View-disentangled Transformer (VD-Former)	Avg mAP = 0.414 Avg SE = 0.449			Private
Hu et al.[53]	2023	Efficient R-transformer Network (ERTN)	DSC = 0.832 HD = 5.300	DSC = 0.779 HD = 4.600	DSC = 0.726 HD = 5.500	BraTS 2017
Zhu et al.[45]	2023	Deep Semantic and Edge Information Fusion based Brain Tumor Segmentation Network	DSC = 0.910 HD = 4.719	DSC = 0.882 HD = 5.985	DSC = 0.846 HD = 3.051	BraTS 2018, 2019, 2020
Gao et al.[62]	2023	Deep Mutual Learning based Brain Tumor Segmentation Network	DSC = 0.901 HD = 3.282	DSC = 0.840 HD = 4.800	DSC = 0.801 HD = 6.112	BraTS 2019
Gao et al.[62]	2023	Deep Mutual Learning based Brain Tumor Segmentation Network	DSC = 0.901 HD = 3.282	DSC = 0.840 HD = 4.800	DSC = 0.801 HD = 6.112	BraTS 2019
Liu et al.[77]	2023	3D Medical Axial Transformer	DSC = 0.932 HD = 7.130	DSC = 0.919 HD = 3.560	DSC = 0.851 HD = 3.610	BraTS 2018, 2021
Lu et al.[78]	2023	Multi-scale ghost CNN with auxiliary MetaFormer (GMetaNet)	DSC = 0.902 HD = 4.530	DSC = 0.825 HD = 6.400	DSC = 0.785 HD = 3.590	BraTS 2019

methods in terms of authors, publication times, method names, segmentation performance and datasets. From the above tables we can conclude the following contents:

(1) Among existing transformer-based brain tumor segmentation methods, DSC and HD are two most common-used evaluation metrics. For three brain tumor classes (WT, TC and ET) in BraTS

Table 5
The summary of BraTS datasets.

Name	Total Images	Training Set	Validation Set	Test Set	Links
BraTS 2012	50	35	N/A	15	https://www.smir.ch/BRATS/Start2012
BraTS 2013	60	35	N/A	25	https://www.smir.ch/BRATS/Start2013
BraTS 2014	238	200	N/A	38	https://www.smir.ch/BRATS/Start2014
BraTS 2015	253	200	N/A	53	https://www.smir.ch/BRATS/Start2015
BraTS 2016	391	200	N/A	191	https://www.smir.ch/BRATS/Start2016
BraTS 2017	477	285	46	146	https://www.cbica.upenn.edu/BraTS17
BraTS 2018	542	285	66	191	https://www.med.upenn.edu/cbica/sbia/brats2018/tasks.html
BraTS 2019	651	335	125	191	https://www.med.upenn.edu/cbica/brats-2019
BraTS 2020	660	369	125	166	https://www.med.upenn.edu/cbica/brats2020/data.html
BraTS 2021	2000	1251	219	530	https://braintumorsegmentation.org

datasets, most methods show the best segmentation results on WT and the sub-optimal segmentation results on TC.

(2) Pure transformer methods achieve the competitive brain tumor segmentation results, and their segmentation performance is relatively average. Most pure transformer methods obtain the DSC values above 0.9 on the best-segmented WT and above 0.8 on the hardest-segmented ET. Among these methods, better preserving high-resolution information (*e.g.* HRSTNet[35]) seems to effectively improve the segmentation performance of the pure transformer.

(3) Hybrid transformer methods show relatively fluctuating brain tumor segmentation results, and some of them achieve significantly higher performance than pure transformer methods, such as UNetFormer,[43] COTRNet,[48] RMTF-Net,[56] BiTr-UNet[63] and CKD-TransBTS[67] all obtain the DSC values above 0.93 on the best-segmented WT, and above 0.88 on the hardest-segmented ET. Overall, the hybrid methods with the transformer-based decoder have relatively poor segmentation performance (lower than pure transformer methods and other hybrid transformer methods), and the hybrid methods with the transformer between the encoder and decoder well balance the segmentation performance and model complexity. Moreover, the hybrid methods with the transformer-based encoder show state-of-the-art brain tumor segmentation performance since both global and local information is effectively extracted from multi-modal MRI images.

In summary, the above brain tumor segmentation results illustrate that hybrid transformer methods have better performance and research potential. In the next few years, we think that more studies will devote to innovatively combining CNN and transformer for extracting high-quality encoding features. Furthermore, some effective network architectures (such as HRNet) that can retain or recover high-resolution semantics features may further improve brain tumor segmentation performance.

5. Existing challenges and future research directions

Although recent research on brain tumor segmentation delivers remarkable progress and achieves superior performance, they still exist many open challenges that need to be further addressed. Some major challenges include as following:

(1) Limited annotations. Most existing methods depend on a large amount of labeled MRI data and train deep learning models in a fully-supervised manner. However, the collection of extensive labels, i.e., segmentation masks, is expensive and time-consuming, especially in some institutions that are hard to annotate brain tumor cases. In future research, to release the demand for a large number of labeled data, we need to design efficient segmentation methods that can be trained with only limited labels. Specifically, we can employ semi-supervised learning and train brain tumor segmentation models using a small portion of labeled data with a large number of unlabeled data. In this manner, the method exploits the knowledge from massive unlabeled data and produces satisfying segmentation performance without much human labor of manual annotations. Furthermore, we can adopt weak-supervised learning to achieve efficient training for brain tumor segmentation. In particular, instead of pixel-wise segmentation mask, we merely need the box-level annotation, i.e., the bounding box of each tumor region. The segmentation model can be trained by this box-level label and produces comparable performance with pixel-wise labels while saving much annotation labor.

(2) Noisy annotations. Due to the professionality and extensive labor of tumor region annotations, existing brain tumor segmentation datasets contain many noisy annotations, which affects the model training and the performance of brain tumor segmentation. Some researchers propose methods of learning with noisy labels towards image classification including MentorNet, DivideMix, etc.,[79–82] while they are hard to be applied to segmentation task. Since few works focus on noisy annotations towards medical image segmentation, we need to customize specific methods of learning with noisy brain tumor segmentation annotations in the future research.

(3) Incomplete modalities. Most previous works achieve impressive results with complete modalities, while they struggle to perform well in cases with incomplete modalities as inputs. In many clinical practices, obtaining complete modalities is difficult and many institutions have some of the modalities, which suffer from leveraging existing multi-modal brain tumor segmentation methods to perform the satisfying diagnosis. In many clinical practices, medical institutions may contain incomplete MRI modalities due to the limitation of collection devices.[83,84] Most previous works[1–5] on brain tumor segmentation assume that the input MRI data contains complete modalities, while their performance significantly degrades in the cases of incomplete modality input. To tackle this issue, we need to design robust segmentation methods against the cases with incomplete modalities. Although some latest works8[83–86] propose frameworks to handle this problem, most of them a target to only one case of incomplete modalities and they are hard to be applied to various cases. Therefore, in the future research, it is necessary to present a unified framework that is robust for all cases with complete and incomplete modalities.

(4) Class imbalance. Brain tumors occupy only a small portion of the brain, making it difficult to handle the class imbalance in the MRI data. This can lead to biased segmentation results in favor of the larger class (i.e., the healthy tissue) and affect the recognition for brain tumors with small regions. To alleviate this issue, class re-weighting techniques are commonly used to assign higher weights to the minority class and lower weights to the majority class during the training phase, which can help the model focus more on the minority class, i.e., small tumor regions, during training. However, in brain tumor segmentation, many MRI data contain extremely small tumors and the class re-weighting approach may degrade on those tumor regions. Many commonly used methods for the class imbalance issue on classification and

recognition tasks, such as data augmentation, data sampling, model ensemble, etc., are difficult to be applied to brain tumor segmentation. To this end, we need to extend the class re-weighting methods and design an adaptive re-weighting strategy for better boosting the performance of small tumor regions.

(5) Interpretability. Deep learning techniques are considered as the black-box model and lack interpretability, since it is difficult to understand the reasoning behind the predictions. In many real-life applications, especially the clinical practice, it is important to know how the deep models work and the reasons for the decision. One way to tackle this issue is by visualizing the feature maps to highlight the dominant regions towards outputs. Researchers have developed various methods for visualizing the intermediate layers of deep learning models, such as activation maximization, CAM, saliency maps, and t-SNE embeddings. Some latest works adopt feature attribution methods to identify the features that are most relevant for a given prediction made by a deep learning model, including gradient-based attribution, perturbation-based attribution, and activation-based attribution. Future works can be targeted to design specific interpretability methods for brain tumor segmentation with vision Transformer.

(6) Model efficiency and deployment. In clinical practice, the well-trained deep models are deployed in terminal devices with limited resources, which require efficient deep models. To this end, during the training stage, the deep models are expected to be efficient and lightweight. The model compression aims to reduce the size of the model by pruning the weights, quantifying the weights, distilling large models, or using low-rank approximations that can be applied to reduce the memory and computation requirements of the deep model. Moreover, we can design slim network architectures and suitable training regimes to release the demand of a large number of parameters and remain the superior performance. Few works study the model efficiency and deployment for brain tumor segmentation, which is a vital step for the application of the algorithms in the future clinical practice.

6. Conclusion

We comprehensively introduce the advanced transformer-based methods over 2021–2023 for brain tumor segmentation on multi-modal MRI. In this review, extensive comparisons and systematic analysis report that transformer-based methods show excellent performance and potential to replace the traditional CNNs as the fundamental brain tumor segmentation backbones. In clinics, both pure transformer and hybrid transformer methods can well provide quantitative tumor information for doctors, thereby assisting brain tumor diagnosis as well as treatment planning. According to the statistics for the performance of existing transformer-based methods, in future studies, scholars can pay more attention to exploring the effective combinations between CNNs and transformers for modeling better global and local feature relations. Moreover, some well-designed models (*e.g.* HRNet-style transformer) that are different from the classical encoder-decoder structure show promising performance since these models better retain high-resolution semantic information. Hence, the novel transformer structures are expected to further explore. It is worth noting that the missing modalities are common in clinics. To address this problem, scholars should comprehensively consider different missing modalities, and develop more general and robust transformer-based methods.

Authorship Statement

Pengyu Wang: Investigation, Data collation, Writing - Original Draft; Qiushi Yang: Investigation, Writing - Original Draft; Zhibin He: Visualization; Yixuan Yuan: Supervision, Conceptualization, Writing - Review & Editing.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. The corresponding author Yixuan Yuan is the Editorial Board Member of the journal but was not involved in the peer review procedure. This paper was handled by another Editor Board member.

Acknowledgments

This work was supported by National Natural Science Foundation of China (62001410), China; Hong Kong Research Grants Council (RGC) Early Career Scheme Grant 21207420, Hong kong, SAR, China; General Research Fund 11211221, Hong kong, SAR, China.

References

1. Havaei M, Davy A, Warde-Farley D, et al. Brain tumor segmentation with deep neural networks. *Med Image Anal*. 2017;35:18–31.
2. Jia H, Xia Y, Cai W, Huang H. Learning high-resolution and efficient non-local features for brain glioma segmentation in mr images. In: *Proc. Medical Image Computing and Computer-Assisted Intervention*. MICCAI); 2020:480–490.
3. Dai C, Wang S, Mo Y, et al. Suggestive annotation of brain tumour images with gradient-guided sampling. In: *Proc. Medical Image Computing and Computer-Assisted Intervention*. MICCAI); 2020:156–165.
4. Ali S, Li J, Pei Y, Khurram R, Rehman KU, Mahmood T. A comprehensive survey on brain tumor diagnosis using deep learning and emerging hybrid techniques with multi-modal MR image. *Arch Comput Methods Eng*. 2022;29(7):4871–4896.
5. Agravat RR, Raval MS. A survey and analysis on automated glioma brain tumor segmentation and overall patient survival prediction. *Arch Comput Methods Eng*. 2021;28:4117–4152.
6. Ranjbarzadeh R, Caputo A, Tirkolaee EB, Ghoushchi SJ, Bendechache M. Brain tumor segmentation of MRI images: a comprehensive review on the application of artificial intelligence tools. *Comput Biol Med*. 2022;152:106405.
7. Liu Z, Tong L, Chen L, et al. Deep learning based brain tumor segmentation: a survey. *Complex Intell Syst*. 2022:1–26.
8. Jyothi P, Singh AR. Deep learning models and traditional automated techniques for brain tumor segmentation in MRI: a review. *Artif Intell Rev*. 2022:1–47.
9. Soomro TA, Zheng L, Afifi AJ, et al. Image segmentation for MR brain tumor detection using machine learning: a review. *IEEE Rev Biol Eng*. 2022;16:70–90.
10. Zhang W, Wu Y, Yang B, Hu S, Wu L, Dhelim S. Overview of multi-modal brain tumor MRI image segmentation. *Healthcare*. 2021;9:1051.
11. Dhole NV, Dixit VV. Review of brain tumor detection from MRI images with hybrid approaches. *Multimed Tool Appl*. 2022;81(7):10189–10220.
12. Rao CS, Karunakara K. A comprehensive review on brain tumor segmentation and classification of MRI images. *Multimed Tool Appl*. 2021;80(12):17611–17643.
13. Guo X, Yang C, Lam PL, Woo PY, Yuan Y. Domain knowledge based brain tumor segmentation and overall survival prediction. In: *Proc. Medical Image Computing and Computer-Assisted Intervention*. MICCAI); 2020:285–295.
14. Yang Q, Yuan Y. Learning dynamic convolutions for multi-modal 3D MRI brain tumor segmentation. In: *Proc. Medical Image Computing and Computer-Assisted Intervention*. MICCAI); 2021:441–451.
15. Guo X, Schwartz L, Zhao B. Semi-automatic segmentation of multimodal brain tumor using active contours. *Proc MICCAI Brainlesion Workshop, Brainlesion: Glioma, Multiple Sclerosis*. 2013;27:27–30.
16. Hamamci A, Unal G. Multimodal brain tumor segmentation using the tumor-cut method on the BraTS dataset. In: *Proc. Medical Image Computing and Computer-Assisted Intervention*. MICCAI); 2012:19–23.
17. Hamamci A, Kucuk N, Karaman K, Engin K, Unal G. Tumor-cut: segmentation of brain tumors on contrast enhanced MR images for radiosurgery applications. *IEEE Trans Med Imag*. 2011;31(3):790–804.
18. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: *Proc. Medical Image Computing and Computer-Assisted Intervention*. MICCAI); 2015:234–241.
19. Liu Z, Shen L. Medical image analysis based on transformer: a review. *arXiv preprint arXiv*. 2022;2208:06643.
20. He K, Gan C, Li Z, et al. Transformers in medical image analysis: a review. *Intell Med*. 2022;3(1):59–78.
21. Shamshad F, Khan S, Zamir SW, et al. Transformers in medical imaging: a survey. *arXiv preprint arXiv*. 2022;2201:09873.
22. Parvaiz A, Khalid MA, Zafar R, Ameer H, Ali M, Fraz MM. Vision transformers in medical computer vision–a contemplative retrospection. *arXiv preprint arXiv*. 2022;2203:15269.
23. Henry EU, Emebob O, Omonhinmin CA. Vision transformers in medical imaging: a review. *arXiv preprint arXiv*. 2022;2211:10043.
24. Ghosh A, Thakur S. Review of brain tumor MRI image segmentation methods for BraTS challenge dataset. In: *Proc. International Conference on Cloud Computing*. Data Science and Engineering (Confluence); 2022:405–410.
25. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: *Proc. Advances in Neural Information Processing Systems (NIPS)*. 30. 2017.

26. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale. In: *Proc. International Conference on Learning Representations*. ICLR); 2020.
27. Liu Z, Lin Y, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: *Proc. International Conference on Computer Vision*. ICCV); 2021: 10012–10022.
28. Chu X, Tian Z, Zhang B, et al. Conditional positional encodings for vision transformers. *arXiv preprint arXiv*. 2021;2102:10882.
29. Wu K, Peng H, Chen M, Fu J, Chao H. Rethinking and improving relative position encoding for vision transformer. In: *Proc. International Conference on Computer Vision*. ICCV); 2021:10033–10041.
30. Sagar A. ViTBIS: vision transformer for biomedical image segmentation. In: *Proc. MICCAI LL-COVID19 Workshop, Clinical Image-Based Procedures, Distributed and Collaborative Learning, Artificial Intelligence for Combating COVID-19 and Secure and Privacy-Preserving Machine Learning*. CLIP); 2021:34–45.
31. Wu Y, Liao K, Chen J, et al. A U-shaped dilated transformer for 3D medical image segmentation. *Neural Comput Appl*. 2022:1–14.
32. Liang J, Yang C, Zhong J, Ye X. BTSwin-Unet: 3D U-shaped symmetrical Swin transformer-based network for brain tumor segmentation with self-supervised pre-training. *Neural Process Lett*. 2022:1–19.
33. Peiris H, Hayat M, Chen Z, Egan G, Harandi M. A robust volumetric transformer for accurate 3D tumor segmentation. In: *Proc. Medical Image Computing and Computer-Assisted Intervention*. MICCAI); 2022:162–172.
34. Peiris H, Hayat M, Chen Z, Egan G, Harandi M. Hybrid window attention based transformer architecture for brain tumor segmentation. *arXiv preprint arXiv*. 2022; 2209:07704.
35. Wei C, Ren S, Guo K, Hu H, Liang J. High-resolution Swin transformer for automatic medical image segmentation. *arXiv preprint arXiv*. 2022;2207:11553.
36. Xing Z, Yu L, Wan L, Han T, Zhu L. Nestedformer: Nested modality-aware transformer for brain tumor segmentation. In: *Proc. Medical Image Computing and Computer-Assisted Intervention*. MICCAI); 2022:140–150.
37. Liu Z, Wei J, Li R. TFusion: transformer based N-to-One multimodal fusion block. *arXiv preprint arXiv*. 2022;2208:12776.
38. Sagar A. EMSViT: efficient multi scale vision transformer for biomedical image segmentation. In: *Proc. MICCAI Brainlesion Workshop, Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. 2022:39–51.
39. Pinaya WH, Tudosiu P-D, Gray R, et al. Unsupervised brain imaging 3D anomaly detection and segmentation with transformers. *Med Image Anal*. 2022;79:102475.
40. Chen X, Yang L. Brain tumor segmentation based on CBAM-TransUNet. In: *Proc. ACM Workshop on Mobile and Wireless Sensing for Smart Healthcare*. MWSSH); 2022:33–38.
41. Andrade-Miranda G, Jaouen V, Bourbonne V, Lucia F, Visvikis D, Conze P-H. Pure versus hybrid transformers for multi-modal brain tumor segmentation: a comparative study. In: *Proc. International Conference on Image Processing*. ICIP); 2022:1336–1340.
42. Hatamizadeh A, Tang Y, Nath V, et al. UNETR: transformers for 3D medical image segmentation. In: *Proc. Winter Conference on Applications of Computer Vision*. WACV); 2022:574–584.
43. Hatamizadeh A, Xu Z, Yang D, Li W, Roth H, Xu D. UNetFormer: a unified vision transformer model and pre-training framework for 3D medical image segmentation. *arXiv preprint arXiv*. 2022;2204:00631.
44. Hatamizadeh A, Nath V, Tang Y, et al. Swin transformers for semantic segmentation of brain tumors in MRI images. In: *Proc. MICCAI Brainlesion Workshop, Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. 2022:272–284.
45. Liang J, Yang C, Zeng L. 3D PSwinBTS: an efficient transformer-based unet using 3D parallel shifted windows for brain tumor segmentation. *Digit Signal Process*. 2022; 131:103784.
46. Liang J, Yang C, Zeng M, Wang X. TransConver: transformer and convolution parallel network for developing automatic brain tumor segmentation in MRI images. *Quant Imag Med Surg*. 2022;12(4):2397.
47. Jiang Y, Zhang Y, Lin X, et al. A method for 3D multimodal brain tumor segmentation using Swin transformer. *Brain Sci*. 2022;12(6):797.
48. Yang H, Shen Z, Li Z, Liu J, Xiao J. Combining global information with topological prior for brain tumor segmentation. In: *Proc. MICCAI Brainlesion Workshop, Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. 2022:204–215.
49. Li H, Huang J, Li G, et al. View-disentangled transformer for brain lesion detection. In: *Proc. International Symposium on Biomedical Imaging*. ISBI); 2022:1–5.
50. Liu J, Zheng J, Jiao G. Transition Net: 2D backbone to segment 3D brain tumor. *Biomed Signal Process Control*. 2022;75:103622.
51. Dhamija T, Gupta A, Gupta S, Katarya R, Singh G. Semantic segmentation in medical images through transfused convolution and transformer networks. *Appl Intell*. 2022:1–17.
52. Chen Y, Wang J. TSEUnet: a 3D neural network with fused transformer and SE-attention for brain tumor segmentation. In: *Proc. International Symposium on Computer-Based Medical Systems*. CBMS); 2022:131–136.
53. Hu Z, Li L, Sui A, Wu G, Wang Y, Yu J. An efficient R-Transformer network with dual encoders for brain glioma segmentation in MR images. *Biomed Signal Process Control*. 2023;79:104034.
54. Wang P, Liu S, Peng J. AST-Net: Lightweight hybrid transformer for multimodal brain tumor segmentation. In: *Proc. International Conference on Pattern Recognition*. ICPR); 2022:4623–4629.
55. Zhu Z, He X, Qi G, Li Y, Cong B, Liu Y. Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI. *Inf Fusion*. 2023;91: 376–387.
56. Gai D, Zhang J, Xiao Y, et al. Residual mix transformer fusion net for 2D brain tumor segmentation. *Brain Sci*. 2022;12(9):1145.
57. Wang W, Chen C, Ding M, Yu H, Zha S, Li J. TransBTS: multimodal brain tumor segmentation using transformer. In: *Proc. Medical Image Computing and Computer-Assisted Intervention*. MICCAI); 2021:109–119.
58. Dobko M, Kolinko D-I, Viniavskyi O, Yelisieiev Y. Combining CNNs with transformer for multimodal 3D MRI brain tumor segmentation with self-supervised pretraining. *arXiv preprint arXiv*. 2021;2110:07919.
59. Pham Q-D, Nguyen-Truong H, Phuong NN, et al. SegTransVAE: hybrid CNN-transformer with regularization for medical image segmentation. In: *Proc. International Symposium on Biomedical Imaging*. ISBI); 2022:1–5.
60. Lyu Q, Namjoshi SV, McTyre E, et al. A transformer-based deep-learning approach for classifying brain metastases into primary organ sites using clinical whole-brain MRI images. *Patterns*. 2022;3(11):100613.
61. Huang L, Chen L, Zhang B, Chai S. A transformer-based generative adversarial network for brain tumor segmentation. *arXiv preprint arXiv*. 2022;2207:14134.
62. Gao H, Miao Q, Ma D, Liu R. Deep mutual learning for brain tumor segmentation with the fusion network. *Neurocomputing*. 2023;521:213–220.
63. Jia Q, Shu H. BiTr-UNet: a cnn-transformer combined network for MRI brain tumor segmentation. In: *Proc. MICCAI Brainlesion Workshop, Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. 2022:3–14.
64. Pang X, Zhao Z, Wang Y, Li F, Chang F. LGMSU-Net: local features, global features, and multi-scale features fused the U-Shaped network for brain tumor segmentation. *Electronics*. 2022;11(12):1911.
65. Li J, Wang W, Chen C, et al. TransBTSV2: towards better and more efficient volumetric segmentation of medical images. *arXiv preprint arXiv*. 2022;2201:12785, 2022.
66. Li X, Ma S, Tang J, Guo F. TranSiam: fusing multimodal visual features using transformer for medical image segmentation. *arXiv preprint arXiv*. 2022;2204:12185.
67. Lin J, Lin J, Lu C, et al. Clinical knowledge-driven hybrid transformer with modality-correlated cross-attention for brain tumor segmentation. *arXiv preprint arXiv*. 2022; 2207:07370.
68. Li S, Sui X, Luo X, Xu X, Liu Y, Goh R. Medical image segmentation using squeeze-and-expansion transformers. *arXiv preprint arXiv*. 2021;2105:09511.
69. Jun E, Jeong S, Heo D-W, Suk H-I. Medical transformer: universal brain encoder for 3D MRI analysis. *arXiv preprint arXiv*. 2021;2104:13633.
70. Zhang Y, He N, Yang J, et al. mmformer: multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In: *Proc. Medical Image Computing and Computer-Assisted Intervention*. MICCAI); 2022:107–117.
71. Nalawade S, Ganesh C, Wagner B, et al. Federated learning for brain tumor segmentation using MRI and transformers. In: *Proc. Medical Image Computing and Computer-Assisted Intervention*. MICCAI); 2022:444–454.
72. Shi Y, Micklisch C, Mushtaq E, Avestimehr S, Yan Y, Zhang X. An ensemble approach to automatic brain tumor segmentation. In: *Proc. MICCAI Brainlesion Workshop, Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. 2022: 138–148.
73. Wang E, Hu Y, Yang X, Tian X. TransUNet with attention mechanism for brain tumor segmentation on MR images. In: *Proc. International Conference on Artificial Intelligence and Computer Applications*. ICAICA); 2022:573–577.
74. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. In: *Proc. Computer Vision and Pattern Recognition*. CVPR); 2015:1–9.
75. Karimijafarbigloo S, Azad R, Kazerouni A, Ebadollahi S, Merhof D. MMCFormer: missing modality compensation transformer for brain tumor segmentation. In: *Proc. Medical Imaging with Deep Learning (MIDL)*. 2023.
76. Futrega M, Milesi A, Marcinkiewicz M, Ribalta P. Optimized U-Net for brain tumor segmentation. In: *Proc. MICCAI Brainlesion Workshop, Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. 2022:15–29.
77. Liu C, Kiryu H. 3d medical axial transformer: a lightweight transformer model for 3D brain tumor segmentation. In: *Proc. Medical Imaging with Deep Learning (MIDL)*. 2023.
78. Lu Y, Chang Y, Zheng Z, et al. GMetaNet: multi-scale ghost convolutional neural network with auxiliary metaformer decoding path for brain tumor segmentation. *Biomed Signal Process Control*. 2023;83:104694.
79. Jiang L, Zhou Z, Leung T, Li L-J, Fei-Fei L. MentorNet: learning data-driven curriculum for very deep neural networks on corrupted labels. In: *Proc. International Conference on Machine Learning*. ICML); 2018:2304–2313.
80. Zhou T, Wang S, Bilmes J. Robust curriculum learning: from clean label detection to noisy label self-correction. In: *Proc. International Conference on Learning Representations*. ICLR); 2021.
81. Li J, Socher R, Hoi SC. DivideMix: learning with noisy labels as semi-supervised learning. In: *Proc. International Conference on Learning Representations*. ICLR); 2020.
82. Liu S, Niles-Weed J, Razavian N, Fernandez-Granda C. Early-learning regularization prevents memorization of noisy labels. In: *Proc. Advances in Neural Information Processing Systems (NIPS)*. 33. 2020:20331–20342.
83. Havaei M, Guizard N, Chapados N, Bengio Y. HeMIS: hetero-modal image segmentation. In: *Proc. Medical Image Computing and Computer-Assisted Intervention*. MICCAI); 2016:469–477.
84. Zhou T, Canu S, Vera P, Ruan S. Latent correlation representation learning for brain tumor segmentation with missing MRI modalities. *IEEE Trans Image Process*. 2021;30: 4263–4274.
85. Yang Q, Guo X, Chen Z, Woo PY, Yuan Y. D2-net: dual disentanglement network for brain tumor segmentation with missing modalities. *IEEE Trans Med Imag*. 2022; 41(10):2953–2964.
86. Chen C, Dou Q, Jin Y, Chen H, Qin J, Heng P-A. Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion. In: *Proc. Medical Image Computing and Computer-Assisted Intervention*. MICCAI); 2019:447–456.