CKD-TransBTS: Clinical Knowledge-Driven Hybrid Transformer with Modality-Correlated Cross-Attention for Brain Tumor Segmentation

Jianwei Lin, Jiatai Lin, Cheng Lu, Hao Chen, Huan Lin, Bingchao Zhao, Zhenwei Shi, Bingjiang Qiu, Xipeng Pan, Zeyan Xu, Biao Huang, Changhong Liang, Guoqiang Han, Zaiyi Liu, Chu Han, *Member, IEEE*

***Abstract*— Brain tumor segmentation (BTS) in magnetic resonance image (MRI) is crucial for brain tumor diagnosis, cancer management and research purposes. With the great success of the ten-year BraTS challenges as well as the advances of CNN and Transformer algorithms, a lot of outstanding BTS models have been proposed to tackle the difficulties of BTS in different technical aspects. However, existing studies hardly consider how to fuse the multi-modality images in a reasonable manner. In this paper, we leverage the clinical knowledge of how radiologists diagnose brain tumors from multiple MRI modalities and propose a clinical knowledge-driven brain tumor segmentation model, called CKD-TransBTS. Instead of directly concatenating all the modalities, we re-organize the input modalities by separating them into two groups according to the imaging principle of MRI. A dual-branch hybrid encoder with the proposed modality-correlated cross-attention block (MCCA) is designed to extract the multi-modality image features. The proposed model inherits the strengths from both Transformer and CNN with the local feature representation ability for precise lesion boundaries and long-range feature extraction for 3D volumetric images. To bridge the gap between Transformer and CNN features, we propose a Trans&CNN Feature Calibration block (TCFC) in the decoder. We compare the proposed model with six CNN-based models and six transformer-based models on the BraTS 2021 challenge dataset. Extensive experiments demonstrate that the proposed model achieves state-of-the-art brain tumor segmentation performance compared with all the competitors.**

***Index Terms*— Clinical knowledge-driven, Brain tumor segmentation, Transformer, Multi-modal fusion**

This work was supported by Key-Area Research and Development Program of Guangdong Province (No. 2021B0101420006), Guangdong Provincial Key Laboratory of Artificial Intelligence in Medical Image Analysis and Application (No. 2022B1212010011), the National Key R&D Program of China (No. 2021YFF1201003), the National Science Fund for Distinguished Young Scholars (No.81925023), Regional Innovation and Development Joint Fund of National Natural Science Foundation of China (No.U22A20345), the National Science Foundation for Young Scientists of China (No. 62102103, 62002082, 82202142 and 82102034), the National Natural Science Foundation of China (No. 82272084, 82271941, 82071892, 82272084 and 82071871), High-level Hospital Construction Project (No. DFJHBF202105), China Postdoctoral Science Foundation (No. 2021M690753 and 2022M710843).

Jianwei Lin, Jiatai Lin and Guoqiang Han are with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, 510006, China.

Cheng Lu, Huan Lin, Bingchao Zhao, Zhenwei Shi, Bingjiang Qiu, Xipeng Pan, Zeyan Xu, Biao Huang, Changhong Liang, Zaiyi Liu and Chu Han are with the Department of Radiology, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, 510080, China; Guangdong Provincial Key Laboratory of Artificial Intelligence in Medical Image Analysis and Application, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, 510080, China.

Hao Chen is with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong.

Corresponding author: Zaiyi Liu, Guoqiang Han, Chu Han.

The first three authors contributed equally.

I. Introduction

Glioma is the most common malignant tumor in the central nervous system [1]. Magnetic resonance imaging (MRI) is the routine examination for glioma diagnosis. Conventional MRI, including pre- and post-contrast T1-weighted images (T1 and T1Gd), T2-weighted (T2) and T2-fluid-attenuated inversion recovery (T2FLAIR) images, provides valuable information for clinical diagnosis, therapeutic planning, and follow-up of gliomas [2]. Generally, radiologists integrate the diagnostic information across imaging modalities when assessing glioma, of which the enhancing regions, tumor necrosis, and peritumoral edema receive the most attention. For example, it is well accepted that the higher intensity of enhancement, larger area of necrosis and edema is associated with higher-grade gliomas, with worse prognosis. Therefore, automatically and precisely segmenting the lesion is a vital step for precision medicine in neurology, including treatment planning, quantitative analysis and research purposes.

Due to the strong feature representation capability, convolutional neural networks (CNNs) have been widely used in the medical image segmentation tasks [3] and achieved promising performance, including brain tumor segmentation (BTS) [4]. Recently, vision transformer (ViT) [5] brings the most powerful technique in natural language processing to the fields of computer vision and medical imaging [6]. Thanks to the self-attention mechanism, Transformer can capture long-range information in the 3D volumetric data. Thus, it has been rapidly adapted to brain tumor segmentation in 3D MRI sequences [7], [8]. Based on these two popular techniques, plenty of outstanding approaches have been proposed for brain tumor segmentation to tackle the following challenges, including lesion location and morphological uncertainty [9], low contrast [10] and annotation bias [11]. However, the existing works overlook an important point on how to fuse the

multi-modality images in a reasonable manner. Most of them fuse the modalities in either the input level or the feature level.

In recent AI research, multi-modal fusion has become one of the hottest topics [12]. Many studies try to build the semantic connection between two different modalities, like image and language. However, brain MRI sequences are quite different from the other ones. In brain MRI images, there exists a very strong structural correlation between paired image modalities, which is the clue for brain tumor assessment. To be more specific, T1Gd is obtained based on the T1 with intravenous gadolinium contrast, and the enhancing regions indicate the disruption (or lack) of the blood-brain barrier, which is consistent with viable tumor and tumor-infiltrated brain. T2 and T2-FLAIR are often jointly interpreted. This clinical knowledge could be very useful in brain tumor segmentation.

Inspired by this, we propose a clinical knowledge-driven BTS model, named CKD-TransBTS. Instead of directly concatenating all the modalities, we simply re-organize the input modalities into two groups (T1 & T1Gd) and (T2 & T2FLAIR), according to their imaging principle. In the encoding phase, we design a dual-branch hybrid encoder with our proposed Modality-Correlated Cross-Attention block (MCCA) for multi-modal fusion and feature extraction. The hybrid encoder leverages the strengths of both Transformer and CNN. Transformer captures long-range information from adjacent slices in the 3D volumetric images. CNN introduces inductive bias for more precise lesion boundaries. In the decoding phase, we propose a Trans&CNN Feature Calibration block (TCFC) in order to alleviate the bias of the features extracted from Transformer and CNN.

Extensive experiments are conducted to evaluate the effectiveness of the proposed CKD-TransBTS in the BraTS challenge 2021 dataset [13]. Our model outperforms six CNN-based models (including the 1st in the BraTS21 challenge validation phase) and six transformer-based models and achieved state-of-the-art BTS performance. Several ablation studies are introduced to validate the technical novelties, especially the clinical knowledge-driven formulation. The main contributions of this study are three-fold:

- We propose a clinical knowledge-driven BTS model by considering the structural correlation between different image modalities and re-grouping the input images in a more reasonable manner.
- We propose two technical novelties in CKD-TransBTS. First, a dual-branch hybrid encoder with the novel Modality-Correlated Cross-Attention block (MCCA) is designed for multi-modal fusion and feature extraction. Second, a novel Trans&CNN Feature Calibration block (TCFC) is proposed to bridge the gap and alleviate the bias of the features between Transformer and CNN.
- We have conducted a series of experiments on the BraTS21 dataset. Our proposed model achieves SOTA performance compared with six CNN-based models and six transformer-based models.

II. Related Works

Thanks to the great efforts of the Radiological Society of North America (RSNA), the American Society of Neuroradiology (ASNR), the Medical Image Computing and Computer Assisted Interventions (MICCAI) society and the BraTS challenge organizers [13], [14], more and more standardized and well-labeled MR images have been released to promote the BTS algorithms. Currently, convolutional neural networks have dominated brain tumor segmentation [15]. With the great success of the multi-head self-attention mechanism, transformer-based models are equally important in medical image segmentation [6], especially for 3D volumetric images.

In this section, we first introduce the existing works in CNN-based BTS models and Transformer-based BTS models. Since we design a novel multi-modal fusion model, multi-modal fusion models are also reviewed in the third part.

A. CNN-based BTS Models

Recent CNN-based models have demonstrated promising performance in brain tumor segmentation [4]. Restricted by the computational and memory resources, earlier CNN-based approaches [16], [17] segment the 2D MRI in a slice-by-slice manner. However, 2D approaches neglect the 3D sequential information.

Currently, more and more 3D BTS models are proposed to leverage 3D spatial information. nnU-Net [18] is a general and adaptive baseline model for both 2D and 3D medical image segmentation, which derives a series of nnU-Net-based BTS models [19], [20]. Liu *et al*. [21] propose CANet to capture the sequential information by introducing feature interaction graphs. Combining feature interaction graphs with convolutional space, CANet can capture the discriminative features with contexts. Zhou *et al*. [22] enlarges the receptive field to capture the contextual information around the lesion and proposes lossless feature computation by employing the 3D atrous-convolution layer. They incorporate the multi-scale contexts and lesion information by an atrous-convolution feature pyramid for brain tumor segmentation. OM-Net [23] integrates three correlated tasks into one network to achieve coarse-to-fine segmentation in a lightweight way and handles the different contributions of each channel for categories by using a CGA module.

Since BTS involves 3D volumetric images from four MRI modalities, researchers now try to innovate the BTS models in the following two aspects. 1) How to leverage the 3D sequential information and the locality information. 2) How to fuse the multi-modal images.

B. Transformer-based BTS Models

Transformer [24] is the most popular technology in the Natural Language Processing (NLP) field thanks to the multi-head self-attention mechanism. ViT [25], [26] extends transformer to computer vision by tokenizing the images.

Chen [27] takes advantage of both U-net structure and transformer and proposed a TransUNet for medical image segmentation. TransBTS [28] extracts global features by applying transformer blocks in the bottleneck layer. CoTr [29] captures the long-range dependency between encoder and decoder by introducing a deformable self-attention mechanism of the DeTrans-encoder. UNETR [7] learns contextual and

long-range information by a Transformer-based encoder and fused localized information and global information in the skip connection. VT-UNet [30] simultaneously encodes local and global cues and captures fine detail for boundary refinement by the volumetric transformer encoder-decoder structure. nnFormer [31] interleaves convolution and self-attention operations to give full play to their strengths to eliminate the gap between features of encoder and decoder by using skip attention. Swin-UNETR [32] utilizes a Swin Transformer-based encoder to learn multi-scale contextual representations and model long-range dependencies.

The above studies provide excellent insights on how to associate transformer with CNN in brain tumor segmentation tasks. In this paper, we design a hybrid transformer model for BTS from different perspectives. 1) We formulate the clinical knowledge in the multi-modal fusion. 2) We introduce the inductive bias and the locality inside the transformer modal design, for a more harmonious mix of transformer and CNN. 3) We calibrate the features extracted from transformer and CNN, in order to bridge the gap between them.

C. Multi-modal Fusion Models

Multi-modal data provides richer information than a single modality does, which has attracted more and more attention in both natural data processing and medical data processing fields, such as visual question answering (VQA) [33], RGB-D object recognition [34] and pathogenomics for prognosis analysis [35]. For most of the above tasks, there exists a huge gap between two different modalities, such as {text, image, speech} and {whole slide image, genomic data}. These data share the same semantic features but different structural features. Different from that, MRI modalities in BTS can achieve perfect pixel-level alignment by image registration. Therefore, it is hard to directly apply the aforementioned multi-modal fusion methods to BTS.

In early BTS models, most of them do not consider the multi-modal fusion problem. They just simply concatenate all the modalities before feeding them into the model. Now they try to fuse multiple modalities in a more reasonable manner. Zhang *et al*. [36] introduce a learnable weight to define the contribution of each modality. Wang *et al*. [37] design the model to learn the complementary information effectively by leveraging two same structural densely-connected branches to map two pairs of modalities. Zhou *et al*. [38] introduce four independent branches for four modalities and fuse the features in the latent space. Zhang *et al*. [39] learn the cross-modal feature representation by a GAN-based generation model.

In this study, we introduce clinical knowledge of the imaging principles of different MRI sequences to guide the multi-modal fusion in a reasonable manner. By simply re-grouping the input images with a dual-branch hybrid encoder, the proposed model can learn better cross-modal feature representation.

III. Methodology

In this section, we first demonstrate the insight of the clinical knowledge-driven formulation and the overall architecture of our model in Sec. III-A. Then, we introduce the knowledge-based dual-branch hybrid encoder with the proposed Modality-Correlated Cross-Attention Module (MCCA) block in Sec. III-B. The details of the feature calibration decoder are given in Sec. III-C with Trans&CNN Feature Calibration (TCFC) module. Sec. III-D shows the details of implementation in this study.

A. Formulation and Model Architecture

Before going into the details of the model, let us start by answering the question of '*how radiologists diagnose brain tumor?*', which inspires and motivates the formulation of our proposed model.

1) How Radiologists Diagnose Brain Tumor?: MRI is the routine clinical examination for brain tumor diagnosis. It usually contains four imaging sequences (modalities), including T1-weighted, T1Gd, T2-weighted and T2FLAIR. Generally, radiologists integrate the diagnostic information across imaging modalities when assessing brain tumors. T1-weighted is the pre-contrast sequence which is the basic imaging modality to pre-locate the brain tumor. T1Gd is the post-contrast sequence with the infused gadolinium enhancing the vascular structures and whether the blood-brain barrier is broken down. Therefore, T1-weighted and T1Gd are usually paired to define the tumor core. T2-weighted imaging is used to detect the free water. In brain tumors, T2-weighted and T2FLAIR images are often jointly interpreted. For the T2-weighted hyperintense, non-enhancing regions in glioblastoma, those that contain free water (e.g. tumor necrosis) frequently present with T2FLAIR hypointensity, while those contain bound water (e.g. vasogenic edema) appear as hyperintense T2FLAIR signal [2].

2) Clinical Knowledge-Driven Formulation: Inspired by the way that radiologists assess brain tumors in MRI images, we want the model to learn the spatial and structural correlation between two correlated sequences. Given four image modalities $\{\mathcal{X}_{\mathrm{T1}}, \mathcal{X}_{\mathrm{T1Gd}}, \mathcal{X}_{\mathrm{T2}}, \mathcal{X}_{\mathrm{T2FLAIR}}\}$ and the segmentation model f with the model parameter θ. Most of the existing BTS models simply concatenate all the input modalities and feed them into the segmentation model at once to predict the segmentation result $\mathcal{S}$.

$$\mathcal{S} = f(\theta, \{\mathcal{X}_{\mathrm{T1}}, \mathcal{X}_{\mathrm{T1Gd}}, \mathcal{X}_{\mathrm{T2}}, \mathcal{X}_{\mathrm{T2FLAIR}}\}) \tag{1}$$

In this study, we re-organize the order of the input images by grouping each two correlated imaging modalities $\{\mathcal{X}_{\mathrm{T1}}, \mathcal{X}_{\mathrm{T1Gd}}\}$ and $\{\mathcal{X}_{\mathrm{T2}}, \mathcal{X}_{\mathrm{T2FLAIR}}\}$, according to the clinical knowledge.

$$\mathcal{S} = f_{\mathrm{ours}}(\theta, (\{\mathcal{X}_{\mathrm{T1}}, \mathcal{X}_{\mathrm{T1Gd}}\}, \{\mathcal{X}_{\mathrm{T2}}, \mathcal{X}_{\mathrm{T2FLAIR}}\})) \tag{2}$$

By grouping two correlated image modalities together, our model can learn the inherent correlation between two image modalities, resulting in a better cross-modal feature representation.

3) Model Architecture: Fig. 1 (a) demonstrates the model architecture of the proposed CKD-TransBTS. The same with most of the segmentation models, CKD-TransBTS keeps the U-Net [40] like structure with skip connections. Since MRI images are 3D volumetric data, we use swim transformer [41]

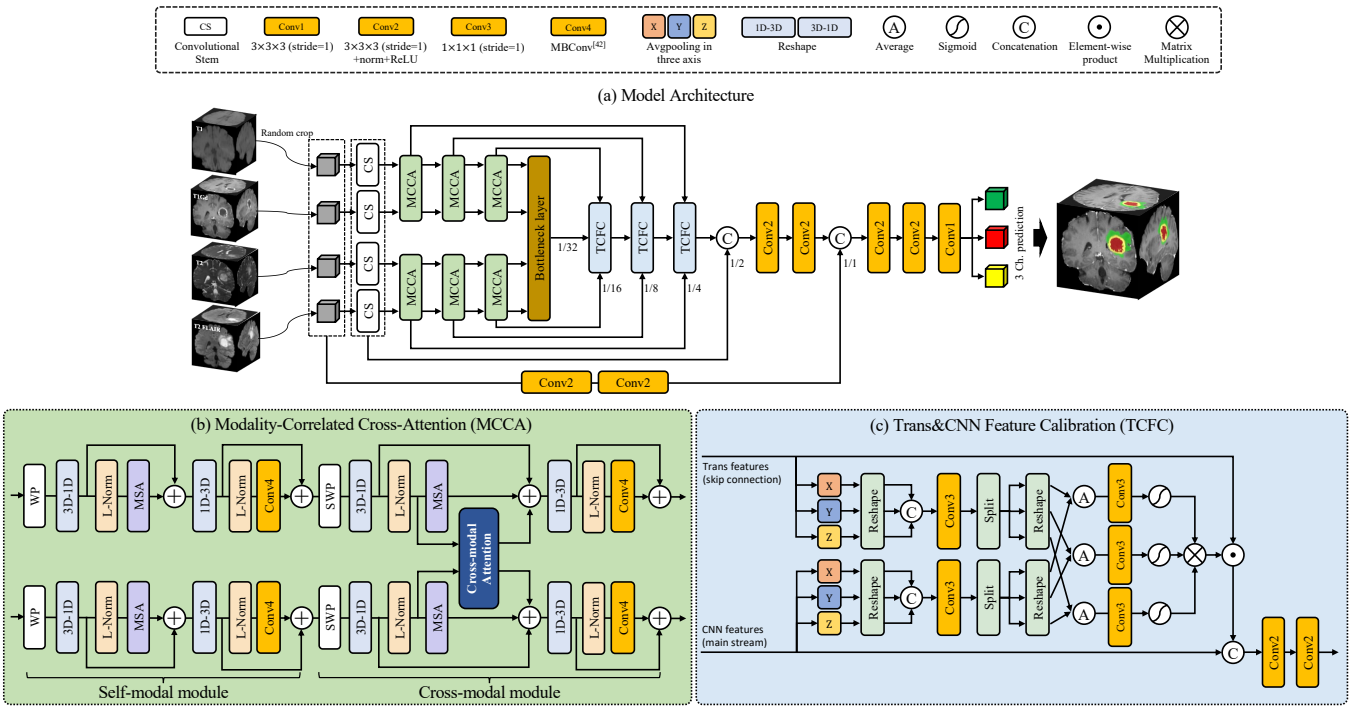

Fig. 1. The architecture of the proposed **CKD-TransBTS**. (a) This model is a U-Net-like structure with a dual-branch hybrid encoder and a feature calibration decoder. According to the clinical knowledge of the MRI in brain tumor diagnosis, we separate the input images into two groups (T1 & T1Gd) and (T2 & T2FLAIR). Convolutional stem is introduced at the beginning. The encoder comprises several MCCA blocks ((b) Modality-Correlated Cross-Attention) which enables cross-modal interactions in a reasonable manner. The decoder consists of several TCFC blocks ((c) Trans&CNN Feature Calibration) to bridge the semantic gap between the features extracted by transformer and CNN. After several convolutional blocks, the model predicts the final brain tumor segmentation results. Note that, in the encoding (decoding) phase, we downsample (upsample) the feature maps by a convolutional (deconvolutional) layer at the end of each stage. In this figure, we omit the downsample and upsample operations for simplification. The resolutions of the feature maps are specified at each stage by the scaling factors.

as the basic architecture of the proposed model to capture the long-range information. In order to bring inductive bias and encourage better local feature representation, we associate transformer with CNN by introducing convolutional layers inside the transformer model.

Since there are two groups of input images, we design a dual-branch hybrid encoder with a convolutional stem and several MCCA blocks. The MCCA block exchanges the information between two correlated image modalities by the cross-modal attention. All the multi-modal features are finally fused in a bottleneck layer. A feature calibration decoder with several TCFC blocks and convolutional layers is designed to obtain the final segmentation results.

B. Dual-Branch Hybrid Encoder

As shown in Fig. 1, a dual-branch hybrid encoder is designed to extract the features from two groups of image modalities. Since two branches are identical (but do not share weights), we only show one branch with the image pair $\{\mathcal{X}_{\mathrm{T1}}, \mathcal{X}_{\mathrm{T1Gd}}\}$ as the example for simplification. The other image pair $\{\mathcal{X}_{\mathrm{T2}}, \mathcal{X}_{\mathrm{T2FLAIR}}\}$ is processed in the same way by the other branch.

1) Convolutional Stem (CS): Downsampling is a common way to reduce the input dimension to save computational and memory resources when processing MRI sequences. However, conventional image downsampling approaches, such as nearest neighbor or bilinear interpolation, will cause information loss.

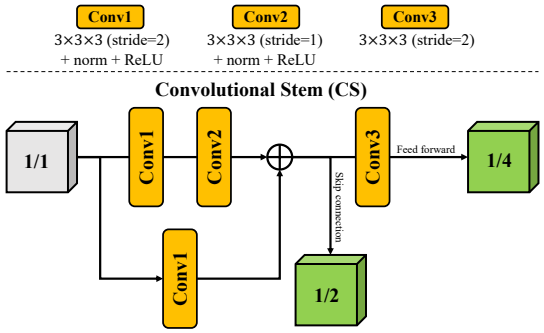

Fig. 2. Convolution stem comprises several convolutional layers with different configurations to downsample the input image in a softer way. In this figure, the gray and green volumes represent the original input image and the intermediate outputs, respectively. $\{1/1, 1/2, 1/4\}$ indicate the scaling factors of the volumes.

And for the BTS task with four volumetric input images, existing approaches tend to downsample the input images by four. In order to reduce the input dimension in a softer way, we introduce the convolutional stem (CS) [42] for each image modality.

As shown in Fig. 2, CS comprises several convolutional blocks with different configurations. Given the input image $\mathcal{X}_{\mathrm{T1}}$, CS outputs two feature volumes $\mathcal{C}^{\frac{1}{2}}_{\mathrm{T1}}$ and $\mathcal{C}^{\frac{1}{4}}_{\mathrm{T1}}$ under two different scales step-by-step, shown as follows.

$$\{\mathcal{C}^{\frac{1}{2}}_{\mathrm{T1}}, \mathcal{C}^{\frac{1}{4}}_{\mathrm{T1}}\} = \mathbf{CS}(\mathcal{X}_{\mathrm{T1}}) \tag{3}$$

where $\mathcal{C}_{\mathrm{T1}}^{\frac{1}{2}}$ is the feature volume for skip connection and $\mathcal{C}_{\mathrm{T1}}^{\frac{1}{4}}$ is the one for feed forward to the MCCA block.

There are two advantages of CS in the BTS task. First, compared with directly downsampling input images by four, CS provides two different scales of feature volumes to help recover the information in the decoding phase. The second advantage originates from Xiao *et al*. [42] that early convolution operation can increase the optimization stability for ViT.

2) Modality-Correlated Cross-Attention (MCCA) Block: We design a modality-correlated cross-attention (MCCA) block to extract the cross-modal features. The MCCA block requires a paired inputs and generates a paired outputs. To consider the locality and the long-range information simultaneously, we combine transformer and CNN in the MCCA block. As shown in Fig. 1 (b), MCCA block consists of two identical branches which are used to extract features from two modalities individually. Each branch is composed of two cascaded modules, the self-modal module and the cross-modal module.

Self-modal module: The self-modal module serves for feature extraction of each single modality, which is a hybrid Transformer-CNN module. We first use the transformer to capture the long-range information. And then we introduce inductive bias and locality by replacing the MLP layer by the convolution layers.

$$\begin{cases} \mathcal{F}_{\mathrm{T1}}^{l} = \mathrm{MSA}(\mathrm{LN}(\mathcal{F}_{\mathrm{T1}}^{l-1})) + \mathcal{F}_{\mathrm{T1}}^{l-1} \\ \mathcal{F}_{\mathrm{T1}}^{l+1} = \mathrm{MBConv}(\mathrm{LN}(\mathcal{F}_{\mathrm{T1}}^{l})) + \mathcal{F}_{\mathrm{T1}}^{l} \end{cases} \tag{4}$$

$$\begin{cases} \mathcal{F}_{\mathrm{T1Gd}}^{l} = \mathrm{MSA}(\mathrm{LN}(\mathcal{F}_{\mathrm{T1Gd}}^{l-1})) + \mathcal{F}_{\mathrm{T1Gd}}^{l-1} \\ \mathcal{F}_{\mathrm{T1Gd}}^{l+1} = \mathrm{MBConv}(\mathrm{LN}(\mathcal{F}_{\mathrm{T1Gd}}^{l})) + \mathcal{F}_{\mathrm{T1Gd}}^{l} \end{cases} \tag{5}$$

For simplicity, we omit the reshape operations in Eq. 4 and Eq. 5. $\mathrm{MSA}(\cdot)$ is the multi-head self-attention (MSA) with window partition. $\mathrm{LN}(\cdot)$ represents the layer normalization and $\mathrm{MBConv}(\cdot)$ is from EfficientNet [43].

Cross-modal module: The cross-modal module follows the swim transformer [41] with shifted window partition and also replaces MLP layer by the MBConv. It exchanges the information between two correlated modalities by a cross-modal attention $\mathrm{CM\text{-}MSA}(\cdot)$, defined as follows.

$$\mathcal{M}_{\mathrm{T1}}, \mathcal{M}_{\mathrm{T1Gd}} = \mathrm{CM\text{-}MSA}(\mathrm{LN}(\mathcal{F}_{\mathrm{T1}}^{l+1}), \mathrm{LN}(\mathcal{F}_{\mathrm{T1Gd}}^{l+1})) \tag{6}$$

$$\mathcal{M}_{\mathrm{T1}} = \mathrm{SoftMax}(\frac{Q_{\mathrm{T1}} K_{\mathrm{T1Gd}}^{T}}{\sqrt{d}} + B) V_{\mathrm{T1Gd}} \tag{7}$$

$$\mathcal{M}_{\mathrm{T1Gd}} = \mathrm{SoftMax}(\frac{Q_{\mathrm{T1Gd}} K_{\mathrm{T1}}^{T}}{\sqrt{d}} + B) V_{\mathrm{T1}} \tag{8}$$

The design of cross-modal module is shown as follows.

$$\begin{cases} \mathcal{F}_{\mathrm{T1}}^{l+2} = \mathrm{MSA}(\mathrm{LN}(\mathcal{F}_{\mathrm{T1}}^{l+1})) + \mathcal{M}_{\mathrm{T1}} + \mathcal{F}_{\mathrm{T1}}^{l+1} \\ \mathcal{F}_{\mathrm{T1}}^{l+3} = \mathrm{MBConv}(\mathrm{LN}(\mathcal{F}_{\mathrm{T1}}^{l+2})) + \mathcal{F}_{\mathrm{T1}}^{l+2} \end{cases} \tag{9}$$

$$\begin{cases} \mathcal{F}_{\mathrm{T1Gd}}^{l+2} = \mathrm{MSA}(\mathrm{LN}(\mathcal{F}_{\mathrm{T1Gd}}^{l+1})) + \mathcal{M}_{\mathrm{T1Gd}} + \mathcal{F}_{\mathrm{T1Gd}}^{l+1} \\ \mathcal{F}_{\mathrm{T1Gd}}^{l+3} = \mathrm{MBConv}(\mathrm{LN}(\mathcal{F}_{\mathrm{T1Gd}}^{l+2})) + \mathcal{F}_{\mathrm{T1Gd}}^{l+2} \end{cases} \tag{10}$$

where $\mathcal{F}_{\mathrm{T1}}^{l+3}$ and $\mathcal{F}_{\mathrm{T1Gd}}^{l+3}$ are the outputs of the MCCA block.

3) Bottleneck Layer: After three MCCA blocks, we concatenate the features of four modalities and feed them into a bottleneck layer which is introduced to bridge the encoder and the decoder. The bottleneck layer shares the same structure with the single branch of the MCCA block without cross-modal attention. The output of the bottleneck layer is defined as $\mathcal{F}_{\mathrm{BNL}}$.

C. Feature Calibration Decoder

In this part, we design a feature calibration decoder to predict the final segmentation results. As shown in Fig. 1, the intermediate features extracted by the encoder are passed to the decoder by skip connections. Since the encoder is a hybrid model which is composed of both transformer and CNN. And the decoder is a pure CNN-based design. There exist a semantic gap between the features of the encoder and the decoder. To bridge the gap, we propose a Trans&CNN Feature Calibration block (TCFC). The feature calibration decoder contains three consecutive TCFC blocks, several convolutional blocks and a segmentation head.

1) Trans&CNN Feature Calibration Block (TCFC): Fig. 1 (c) demonstrates the architecture of the TCFC block whose purpose is to bridge the semantic gap between MCCA and TCFC blocks by providing pixel-wise spatial attention to the features extracted by the MCCA block. Let us denote the feed-forward feature tensor, the transformer feature tensor from skip connections and the output feature tensor as $\mathcal{F}$, $\mathcal{F}_{\mathrm{trans}}$ and $\mathcal{F}'$, TCFC block is formulated by Eq. 11.

$$\begin{aligned} \mathcal{F}' &= \mathrm{TCFC}(\mathcal{F}_{\mathrm{trans}}, \mathcal{F}) \\ \mathcal{F}_{\mathrm{trans}} &= \mathrm{Concate}(\mathcal{F}_{\mathrm{trans}}^{b1}, \mathcal{F}_{\mathrm{trans}}^{b2}) \end{aligned} \tag{11}$$

where $\mathcal{F}_{\mathrm{trans}}^{b1}$ and $\mathcal{F}_{\mathrm{trans}}^{b2}$ denote the tensors from two branches in the dual-branch hybrid encoder. In the first TCFC block, $\mathcal{F} = \mathcal{F}_{\mathrm{BNL}}$.

Since the input and the feature tensors are 3D volume. In order to make full use of the 3D information, average pooling is applied in three directions separately for both $\mathcal{F}$ and $\mathcal{F}_{\mathrm{trans}}$.

$$\begin{cases} \mathcal{F}^{X} = \mathrm{Avgpool}(\mathcal{F}), \mathcal{F}^{X} \in \mathbb{R}^{c \times x \times 1 \times 1} \\ \mathcal{F}^{Y} = \mathrm{Avgpool}(\mathcal{F}), \mathcal{F}^{Y} \in \mathbb{R}^{c \times 1 \times y \times 1} \\ \mathcal{F}^{Z} = \mathrm{Avgpool}(\mathcal{F}), \mathcal{F}^{Z} \in \mathbb{R}^{c \times 1 \times 1 \times z} \end{cases} \tag{12}$$

$$\begin{cases} \mathcal{F}_{\mathrm{trans}}^{X} = \mathrm{Avgpool}(\mathcal{F}_{\mathrm{trans}}), \mathcal{F}_{\mathrm{trans}}^{X} \in \mathbb{R}^{c \times x \times 1 \times 1} \\ \mathcal{F}_{\mathrm{trans}}^{Y} = \mathrm{Avgpool}(\mathcal{F}_{\mathrm{trans}}), \mathcal{F}_{\mathrm{trans}}^{Y} \in \mathbb{R}^{c \times 1 \times y \times 1} \\ \mathcal{F}_{\mathrm{trans}}^{Z} = \mathrm{Avgpool}(\mathcal{F}_{\mathrm{trans}}), \mathcal{F}_{\mathrm{trans}}^{Z} \in \mathbb{R}^{c \times 1 \times 1 \times z} \end{cases} \tag{13}$$

where $x = y = z$ since the input images are cubes in this study.

By reshaping vectors of three directions into the same shape, we concatenate them and compress the channels by a $1 \times 1 \times 1$ convolutional layer. Then we can obtain new feature vectors for three directions $\{\hat{\mathcal{F}}_{\mathrm{trans}}^{X}, \hat{\mathcal{F}}_{\mathrm{trans}}^{Y}, \hat{\mathcal{F}}_{\mathrm{trans}}^{Z}\}$ and $\{\hat{\mathcal{F}}^{X}, \hat{\mathcal{F}}^{Y}, \hat{\mathcal{F}}^{Z}\}$ by splitting them back to the original dimension. Next, we aggregate the features from transformer

TABLE I

QUANTITATIVE COMPARISON WITH SOTA METHODS IN BRATS21 DATASET († MEANS CNN-BASED MODELS). THE TOP-3 RESULTS ARE IN RED, BLUE AND GREEN. T-TEST IS PERFORMED BETWEEN EACH BASELINE MODEL AND OUR MODEL. * MEANS P-VALUE P < 0.05.

Models	(Year) Pub	Dice ↑				HD95 (mm) ↓				Sensitivity ↑		
		ET	TC	WT	Mean	ET	TC	WT	Mean	ET	TC	WT
†VNet [45]	(16) 3DV	0.7820*	0.8051*	0.8402*	0.8091	20.80*	25.08*	15.69*	20.52	0.7951*	0.8080*	0.8520*
†ResUNet [46]	(17) GRSL	0.8143*	0.8472*	0.9027*	0.8547	14.30*	9.33*	10.12*	11.25	0.8266*	0.8466*	0.9024*
†LSTM-CNN [47]	(17) CVPR	0.8347*	0.8628*	0.9058*	0.8678	12.86*	8.77	8.41*	10.01	0.8505*	0.8599*	0.8932*
†UNet++ [48]	(19) TMI	0.7813*	0.8225*	0.8493*	0.8177	23.88*	17.04*	8.64*	16.52	0.8057*	0.8373*	0.8612*
†AttentionU [49]	(19) MIA	0.7897*	0.8373*	0.8566*	0.8278	18.47*	11.84*	15.97*	15.43	0.8266*	0.8510*	0.8980*
†DynUNet [20]	(21) arXiv	0.8581*	0.8971	0.9288	0.8946	13.03*	6.71	6.99	8.91	0.8767*	0.8980	0.9081*
TransBTS [28]	(21) MICCAI	0.8181*	0.8500*	0.8795*	0.8494	16.79*	11.14*	12.78*	13.57	0.8139*	0.8452*	0.8768*
TransUNet [27]	(21) arXiv	0.8182*	0.8772*	0.9191*	0.8715	13.09*	7.34	6.16	8.86	0.8660*	0.9026	0.9130*
VTNet [30]	(21) arXiv	0.8551*	0.8822*	0.9134*	0.8837	9.01*	6.92*	8.22*	8.05	0.8143*	0.8301*	0.8890*
UNETR [7]	(22) WACV	0.8520*	0.8664*	0.9220*	0.8803	12.26*	7.73*	7.78*	9.26	0.8705*	0.8835*	0.9252*
SegTransVAE [50]	(22) ISBI	0.8622*	0.8999	0.9254*	0.8958	10.59	5.88	7.71*	8.06	0.8691*	0.8927*	0.9269*
Swin UNETR [32]	(22) arXiv	0.8681*	0.8998	0.9273*	0.8984	11.09*	6.89	7.33*	8.44	0.8810*	0.9100	0.9347
Ours	(22) -	0.8850	0.9016	0.9333	0.9066	5.93	6.54	6.20	6.22	0.8997	0.9055	0.9334

and CNN in a direction-wise manner.

$$\begin{cases} \bar{\mathcal{F}}^X = \text{Sigmoid}(\text{Conv}(\text{Avg}(\hat{\mathcal{F}}^X_{\text{trans}}, \hat{\mathcal{F}}^X))) \\ \bar{\mathcal{F}}^Y = \text{Sigmoid}(\text{Conv}(\text{Avg}(\hat{\mathcal{F}}^Y_{\text{trans}}, \hat{\mathcal{F}}^Y))) \\ \bar{\mathcal{F}}^Z = \text{Sigmoid}(\text{Conv}(\text{Avg}(\hat{\mathcal{F}}^Z_{\text{trans}}, \hat{\mathcal{F}}^Z))) \end{cases} \tag{14}$$

By a matrix multiplication operation of three vectors, we can obtain a calibrated attention tensor A.

$$\mathcal{A} = \bar{\mathcal{F}}^X \bar{\mathcal{F}}^Y \bar{\mathcal{F}}^Z \tag{15}$$

Then the output of TCFC block can be obtained by concatenating the calibrated transformer features and the feed forward main stream features.

$$\mathcal{F}' = \text{Concate}(\mathcal{A}\mathcal{F}_{\text{trans}}, \mathcal{F}) \tag{16}$$

D. Implementation Details

We implement all the experiments on the PyTorch and MONAI, and train the models on a workstation with an NVIDIA 3090 GPU. We set the learning rate to $1e-4$ and adjust it using the cosine annealing algorithm [44]. The number of epochs for model training is 500 and the Dice loss is used as the objective function. For each modality, the initial sub-volume resolution is $4 \times 4 \times 4$ and the initial embedding size is 32. In the training phase, we first obtain the minimum bounding box of the volume and then partition it randomly into a volume size of 128×128×128. To make the data distribution more complex and alleviate the over-fitting problem, we apply several data augmentation methods, including random zoom, random flip in three directions, Gaussian noise, Gaussian blur and random contrast. All the data augmentation methods were applied to all the four modalities with the same setting. In the test phase, we use the sliding window method with an overlap rate of 0.6. The source code is available at `https://github.com/sword98/CKD-TransBTS`

IV. EXPERIMENT

In this section, we first describe the dataset and the evaluation metrics in Sec. IV-A. Then, we compare our proposed model with existing SOTA approaches in Sec. IV-B. In Sec. IV-C, we conduct ablation studies to evaluate the effectiveness of each technical novelty.

TABLE II

COMPARISON ON BRATS 2021 CHALLENGE LEADERBOARD (VALIDATION SET). THE TOP-3 RESULTS ARE IN RED, BLUE AND GREEN.

(Rank) Models	Dice ↑			HD95 (mm) ↓		
	ET	TC	WT	ET	TC	WT
(1) Luu *et al.* [51]	84.51	87.81	92.75	20.73	7.623	3.47
(4) Siddiquee *et al.* [52]	86.00	88.68	92.65	9.05	5.84	3.60
(6) Kotowski *et al.* [53]	81.98	87.84	92.72	-	-	-
(8) Jia *et al.* [54]	84.80	87.96	92.54	14.18	5.86	3.45
Ours	84.76	88.07	92.33	3.16	4.39	4.23

A. Dataset and Evaluation Metrics

BraTS Challenge 2021: The dataset provides a large amount of annotated brain tumor MRI data, mainly from The Cancer Imaging Archive [13], [55]. Since the validation and test data in BraTS challenge 2021 is private, we split the training set (1251 3D MRI images) provided by the challenge organizers for all the experiments. The training, validation and test set contains 834, 208 and 209 samples respectively. All the MRI images were skull stripped and resampled to $1mm^3$. Each patient's MRI includes four modalities: T1-weighted (T1), post-contrast T1-weighted (T1Gd), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (T2FLAIR), with co-registered to T1 anatomical template. Annotation is divided into three sub-regions: Gd-enhancing tumor (ET), the peritumoral edematous/invaded tissue (ED) and the necrotic tumor core (NCR). Suggested by the challenge organizers, these sub-regions can be clustered into three more segmentation-friendly regions which are used to evaluate the segmentation performance, including enhanced tumor (ET), tumor core (TC) (joining ET and NCR), and whole tumor (WT) (joining ED to TC).

Evaluation metrics: In our experiments, we use Dice score, 95% Hausdorff distance (HD95) and sensitivity to evaluate the segmentation results. Since the healthy area dominates the 3D volumetric image in most of the cases, the specificity values of most baseline models are over 0.99. Due to the limited space of the table, we only show sensitivity in all the quantitative comparisons.

B. Comparisons with SOTA Models

We compare our proposed CKD-TransBTS model with several SOTA models, including six CNN-based models

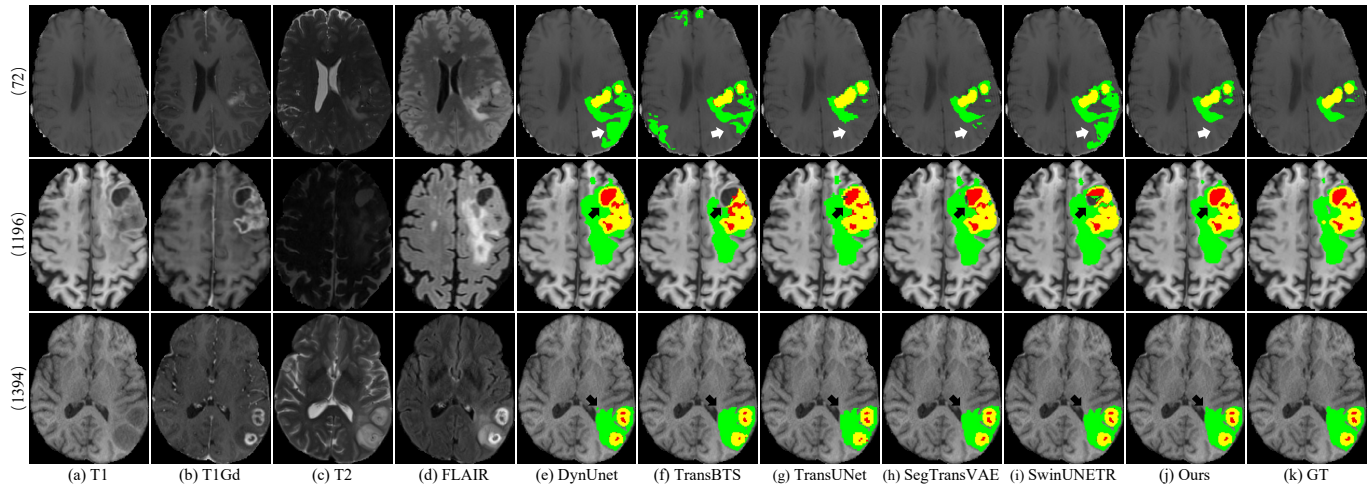

Fig. 3. Visualization of quantitative comparison of SOTA methods on the BraTS21 dataset. From top to bottom are BraTS_72, BraTS_1196 and BraTS_1394 respectively. Green, yellow and red regions indicate ED, ET and NCR. White arrows highlight some false positive results in the results of the baseline models. Black arrows demonstrate the superior regions of our results compared with the baseline models.

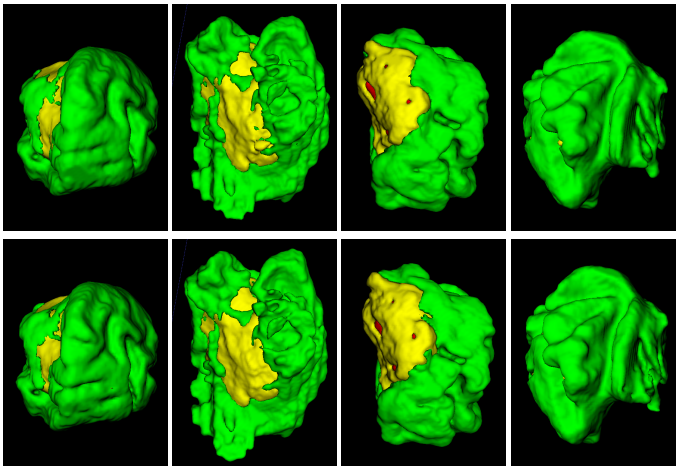

Fig. 4. 3D visualization of our s tumor segmentation results. The first row shows our segmentation results. The second row demonstrates the ground truth. The numbers indicate the indices of the samples.

(VNet [45], ResUNet [46], LSTM-CNN [47], UNet++ [48] AttentionUNet [49] and DynUNet [20]) and six transformer-based models (TransBTS [28], TransUNet [27], UNETR [7], VTNet [30], SegTransVAE [50] and Swin UNETR [32]). For each baseline model in this experiment, we directly run the code if it has been released. For the baseline models without code, we implement them exactly following the details in the corresponding papers. All the baseline models and our proposed model are trained under the same computer hardware with the same dataset split. The best models are selected by the validation set. To be fair, all the quantitative and qualitative results are the direct output of the model without any post-processing procedure.

Table I demonstrates the quantitative results of all the models. The upper part and lower part demonstrate the CNN-based models (marked by †) and the transformer-based models respectively. We color-code the top-3 approaches for every score in red, blue and green. As demonstrated in Table I, our proposed model outperforms all the CNN-based models and transformer-based models and achieves SOTA segmentation performance. Specifically, the Dice scores of the proposed CKD-TransBTS for ET, TC and WT all outperform the competitors. The HD95 metric for ET achieves the best (5.93 mm) which is 3 mm lower than the second one (9.01 mm). The HD95 metrics for TC and WT still maintain the top-2 performance with less than 1 mm performance gap. DynUNet, the top-1 of BraTS challenge 2021 in the validation set, achieves the best performance among the CNN-based models. However, the performance of ET are less effective than our proposed model (Dice: 0.8581 v.s. 0.8850, HD95: 13.03 v.s. 5.93, Sensitivity: 0.8767 v.s. 0.8997). It proves the superiority of the clinical knowledge-driven formulation of our proposed model. Since the enhanced tumor (ET) is defined by comparing the difference of the hyperintense regions between pre- and post- contrast images (T1 and T1Gd). By simply re-grouping the input images, our model can achieve much better segmentation results in ET compared with other baseline models. The same observation can also be found when compared with transformer-based models.

Fig. 3 demonstrates the qualitative results compared with several representative baseline models. As can be seen in the first row in Fig. 3, our model successfully suppresses the false positive results compared with other baseline models (marked by the white arrows). In the second row, our model generates more precise segmentation results locally with fewer noises compared with all the other competitors, pointed by the black arrows. Therefore, it results in a lower HD95 distance. In the third row, our model achieves more complete and precise boundaries. Thanks to the clinical knowledge-driven design, CKD-TransBTS can aware of the correlations between the correlated modalities, to ensure spatial and contextual correctness. Since the clinical knowledge-driven idea originates from how radiologists diagnose brain tumors, this formulation also makes the results closer to the radiologists. The hybrid model of transformer and CNN inherits the strengths from both of them. Capturing long-range information can avoid some false

TABLE III
ABLATION STUDIES ON MULTI-MODAL FUSION (F), FEATURE CALIBRATION (C) AND HYBRID ENCODER (H). T-TEST IS PERFORMED BETWEEN EACH BASELINE MODEL AND OUR MODEL. * MEANS P-VALUE $P < 0.05$.

Models	Ablation			Dice ↑				HD95 ↓				Sensitivity ↑		
	F	C	H	ET	TC	WT	Mean	ET	TC	WT	Mean	ET	TC	WT
(1)				0.8351*	0.8753*	0.9080*	0.8728	15.53*	8.42*	8.37*	10.78	0.8295*	0.8784*	0.8821*
(2)	✓			0.8720*	0.8848*	0.9278*	0.8949	9.20	6.85	7.18*	7.74	0.8623*	0.8852*	0.9172*
(3)		✓		0.8566*	0.8807*	0.9229*	0.8868	13.96*	6.70	7.80	9.49	0.8482*	0.8545*	0.9165*
(4)			✓	0.8548*	0.8889*	0.9213*	0.8885	13.18*	7.41*	7.48*	9.36	0.8623*	0.8987*	0.9201*
(5)	✓		✓	0.8767*	0.8827*	0.9283*	0.8959	7.60	6.96*	8.24*	7.60	0.8853	0.9096	0.9232*
(6)	✓	✓		0.8568*	0.8841*	0.9262*	0.8890	13.39*	9.02	8.09*	10.17	0.8880	**0.9141***	0.9262*
(7)		✓	✓	0.8681*	0.8954	0.9321	0.8985	12.39*	**6.35**	6.35	8.37	0.8803*	0.9088	0.9285*
(8)	✓	✓	✓	**0.8850**	**0.9016**	**0.9333**	**0.9066**	**5.93**	6.54	**6.20**	**6.22**	**0.8997**	0.9055	**0.9334**

positive results far from the lesions. Inductive bias and locality can achieve locally more precise segmentation results. We also demonstrate the 3D volumetric segmentation results compared with ground truth. Fig. 4 shows that our model can generate 3D segmentation results very close to the ground truth.

In the meantime, we also submit our results on the validation set to the BraTS 2021 challenge portal. We compare our proposed model with several top-tier teams in the BraTS 2021 challenge. In Table II, the left-hand side shows the ranking of each team on the test set. The right-hand side shows the quantitative results of the validation set, provided by their papers. We can observe that our proposed model achieves SOTA performance compared with the top-tier teams in BraTS 2021 challenge.

C. Ablation Studies

We also conduct several ablation studies to evaluate the superiority of each technical novelty in our proposed model, including the clinical knowledge-driven multi-modal fusion, the hybrid encoder of transformer and CNN, and the feature calibration decoder. We compare our final model with several models with different configurations. (1) The baseline model without cross-modal fusion, the hybrid encoder and feature calibration. (2)-(4) are the baseline models with only one technical novelty. (5)-(7) are the baseline models with two technical novelties. (8) is our final model. The quantitative results are illustrated in Table III.

1) Effectiveness of Multi-modal Fusion: In this paper, we introduce the clinical knowledge-driven formulation by first re-grouping the modalities before feeding them into the model. And then the proposed MCCA block is used to extract the cross-modal features. To prove the effectiveness of the multi-modal fusion strategy, we remove the cross-modal attention in the MCCA blocks.

Compared with Model (1), Model (2) with the clinical knowledge-driven multi-modal fusion strategy achieves an obvious improvement in both Dice score and HD95 distance. It even outperforms most of the SOTA methods shown in Table I. The mean HD95 distance of Model (2) is in the 2nd place (Ours: 6.22, Model (2): 7.74, VTNet: 8.05), while the mean Dice score is in the 3rd place (Ours: 0.9066, Swin UNETR: 0.8984, Model (2): 0.8949). It proves that the multi-modal fusion strategy effectively improves the segmentation results with more precise boundaries. When combining our multi-modal fusion strategy with the hybrid encoder and feature calibration, the segmentation results are further improved.

2) Effectiveness of Hybrid Encoder: In order to introduce both long-range information and inductive bias, we design a hybrid encoder of transformer and CNN. To demonstrate the effectiveness of the hybrid encoder, we compare our model with the one that replaces the hybrid branches in the MCCA block with the swim transformer blocks.

Quantitative results show that the hybrid encoder (Model (4)) generally improves the segmentation results with better Dice scores compared with the baseline model (Model (1)). However, due to the lack of the multi-modal fusion strategy, HD95 distance of Model (4) is larger than that of Model (2). When combining the multi-modal fusion and hybrid encoder in Model (5), both Dice score (Model (4): 0.8885, Model (5): 0.8959) and HD95 distance (Model (4): 9.36, Model (5): 7.60) are improved. This observation means that the hybrid encoder has a positive impact on the segmentation results while the lack of the multi-modal fusion strategy may lead to imprecise segmentation boundaries.

3) Effectiveness of Feature Calibration: The TCFC block is designed to bridge the gap between the skip connection (transformer) features and the mainstream (CNN) features. Even though the encoder is a hybrid one, the gap between the features of the encoder and the decoder still exists. In this experiment, we disable the feature calibration by replacing the TCFC block with a conventional CNN decoder and directly concatenate the mainstream features and the skip connection features.

With the hybrid encoder only in Model (3), TCFC can also improve the segmentation results compared with Model (1) without it. Actually, combining transformer with CNN in the encoding phase can introduce the inductive bias, which can also bridge the gap between the features extracted from transformer and CNN. When associating hybrid encoder with feature calibration (TCFC) in Model (7), the segmentation performance is further improved in both Dice score (Model (3): 0.8868, Model (4): 0.8885, Model (7): 0.8985) and HD95 distance (Model (3): 9.49, Model (4): 9.36, Model (7): 8.37).

To summarize the ablation studies, the effectiveness of all the technical novelties has been evaluated. Each one of them alone can improve the brain tumor segmentation performance. Multi-modal fusion mainly serves for finding the correlation between the correlated modalities, resulting in more precise boundaries with the lower HD95 distance. Hybrid encoder

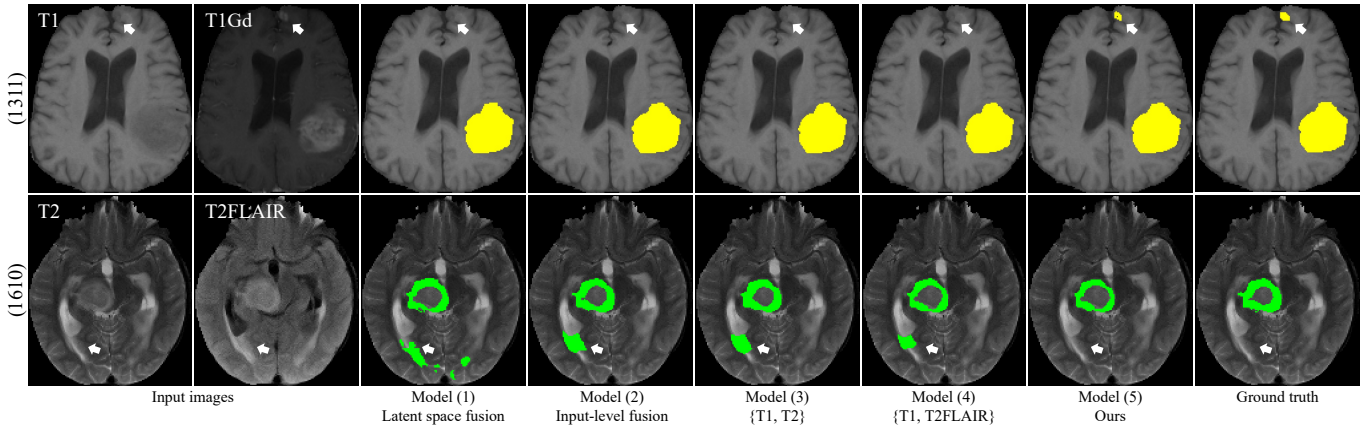

Fig. 5. Visualization of the segmentation results in different categories. The top row shows the segmentation results of the model in the enhance tumor. The bottom row shows the segmentation results of the model in the edema region. We only show one specific categories for better visualization. The leftmost numbers indicate the IDs of the samples.

TABLE IV

COMPARISON WITH DIFFERENT MULTI-MODAL FUSION STRATEGIES. MODALITIES IN THE SAME GROUP ARE MARKED BY THE SAME SYMBOL ⋆ OR ○. T-TEST IS PERFORMED BETWEEN EACH BASELINE MODEL AND OUR MODEL. * MEANS P-VALUE P < 0.05.

Models	T1	T1Gd	T2	T2F	Dice ↑				HD95 ↓				Sensitivity ↑		
					ET	TC	WT	Mean	ET	TC	WT	Mean	ET	TC	WT
(1) Latent	Feature-level fusion				0.8783	0.8989	0.9314	0.9029	10.71*	6.36	6.70	7.92	0.8903*	0.9052	0.9314
(2) Input	Input-level fusion				0.8681*	0.8954	0.9321	0.8985	12.39*	**6.35**	6.35	8.37	0.8803*	0.9088	0.9286*
(3) Ours	○	⋆	○	⋆	0.8753*	0.8933	0.9285*	0.8990	9.30	6.71	6.87	7.63	0.8747*	0.8875*	0.9186*
(4) Ours	○	⋆	⋆	○	0.8786*	0.8865*	0.9292*	0.8981	7.40	6.93	7.03	7.12	0.8841*	**0.9223***	**0.9342**
(5) Ours*	○	○	⋆	⋆	**0.8850**	**0.9016**	**0.9333**	**0.9066**	**5.93**	6.54	**6.20**	**6.22**	**0.8997**	0.9055	0.9334

TABLE V

TIME PERFORMANCE OF DIFFERENT MODELS. THE TIME PERFORMANCE OF THE TRAINING PHASE IS THE TOTAL TIME. THE TIME PERFORMANCE OF THE TEST PHASE IS THE AVERAGE TIME OF EACH CASE. THE PERFORMANCE GAPS OF THE DICE SCORE, THE HD95 DISTANCE AND SENSITIVITY BETWEEN OUR MODEL WITH EACH BASELINE MODEL ARE ALSO SHOWN IN THIS TABLE.

Model	Train (e)	Test	Dice ↑	HD ↓	Sen ↑
ResUNet [46]	3d15h (459)	0.34s	+5.19	-5.03	+5.47
DynUNet [20]	1d22h (220)	1.54s	+1.20	-2.69	+1.90
SegTransVAE [50]	2d13h (351)	1.02s	+1.08	-1.84	+1.70
Swin UNETR [32]	2d10h (248)	2.01s	+0.82	-2.22	+0.47
Ours	2d18h (261)	2.65s	-	-	-

leverage the strengths of both transformer and CNN. Associating the hybrid encoder with feature calibration can bridge the gap between transformer and CNN, resulting in more precise segmentation results. The final model with all the technical novelties achieves the best segmentation results quantitatively and qualitatively.

D. Comparison with Different Fusion Strategies

Furthermore, we also compare our model with different multi-modal fusion strategies in Table IV. (1) In this model, we remove the cross-modal attention in the MCCA block. Each modality has an individual branch to extract the features. The latent features are concatenated before feeding into the bottleneck layer. (2) We directly concatenate four modalities at the input level and with only one single transformer branch in the encoder. (3)-(4) The model architecture remains unchanged with different groups of the input modalities. The modalities in the same group are marked by the symbols ⋆ and ○. (5) Ours marked by * is the final model.

As shown in Table IV, two conventional ways of feature-level fusion and input-level fusion can both achieve reasonable results in Dice score. But they are less effective in HD95 distance. Since HD95 distance focuses more on the correctness of the contours. Model (3) and Model (4) with different groups of modalities can also improve the precision of the contours with lower HD95 distance. It reveals that the proposed MCCA block with cross-modal attention is superior to conventional feature-level fusion and input-level fusion ways. MCCA block is easier to find the correlation between two modalities, even they are less correlated. Inspired by the clinical knowledge from the radiologists, grouping two highly correlated modalities (T1 and T1Gd) in Model (5) achieves the best performance, especially for the enhanced tumor (ET) with DICE of 0.8850 and HD95 of 5.93. Because radiologists define the enhanced tumor by comparing the T1 and T1Gd modalities.

Fig. 5 demonstrates the qualitative results of this experiment. We show two distinct cases to visualize the segmentation results which get benefit from the clinical knowledge-driven design. In this figure, we only show the most relevant category in the results for clearer visualization. In the first row, we show the enhanced tumor (ET) results and the corresponding T1 and T1Gd modalities. As we mentioned above, radiologists assess ET by comparing the T1 and T1Gd. Following this clinical knowledge, our multi-modal fusion strategy successfully help the model find out another small lesion pointed by the arrow. While the models with the other fusion strategies or inappro-

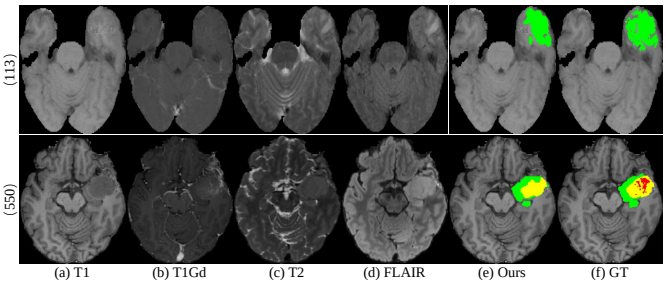

Fig. 6. Examples of imperfect segmentation results.

priate input combinations miss this lesion. The second row show the case of edema, which is assessed by comparing the T2 and T2FLAIR modalities. Our model with a reasonable multi-modal fusion way can suppress the false positive regions pointed by the arrow.

E. Demonstration of Imperfect Results

We also demonstrate some imperfect segmentation results in Fig. 6. The first row shows a hard example. Since radiologists diagnose the brain tumor by considering four MRI modalities. They decide the boundaries of the lesions by the change of the image gradient inside a single modality and the difference among different modalities. For some high-grade brain tumors with less clear lesion boundaries, such as glioblastoma, our model may fail to provide a precise segmentation result compared with experienced radiologists.

The second row demonstrates a limitation of our proposed model with under-segmentation of the necrosis region (red) when the image appearance is in contradiction to the MRI imaging principle. We know that the cells in the necrosis region are mostly broken down, which leads to a lot of free water in this region. And the edema region mostly contains bound water. By the imaging principle of T2 and T2FLAIR, T2 enhances both bound water and free water while T2FLAIR suppresses the free water. The edema region should be high signal in both T2 and T2FLAIR. And the necrosis region should be high signal in T2 and low signal in T2FLAIR. However, in this case, the necrosis region can hardly be observed in T2FLAIR image because the corresponding regions demonstrate high signal. Since our model considers the relationships of {T1, T1Gd} and {T2, T2FLAIR}. It may fail when the difference between paired modalities is not clear or even unobservable.

F. Time Performance

Since transformer-based models typically consume more time than CNN-based models during training and testing. We also conduct a time performance comparison in Table V. We compare our model with four selected representative models of CNN as well as transformer. Table V shows the training and testing time, the number of epochs, the performance gaps of three metrics between our model and each baseline model. As can be seen, ResUNet spent the longest training time and the shortest inference time but with largest performance gap compared with our model. DynUNet improves the performance with shorter training time but with longer inference time. Two transformer models also slightly improve the performance compared with two CNN models. Our model achieves the best segmentation performance with longest inference time and moderate training time.

V. Conclusion

In this paper, we propose a novel clinical knowledge-driven brain tumor segmentation model with four input MRI modalities, named CKD-TransBTS. We deeply analyze the way how radiologists assess and diagnose brain tumor from multiple modalities and introduce the clinical knowledge into our multi-modal fusion strategy. By simply re-grouping the input modalities according to the imaging principles of them, our model can get obvious improvement as shown in the quantitative and qualitative experiments. This multi-modal fusion strategy effectively improves the segmentation results with more precise boundaries and suppresses the false positive. We believe that introducing some relevant prior knowledge or the clinical knowledge into the model design could be an very effective way to benefit the model representation and let model learn the inherent characteristics of the data beyond the labels.

In the technical perspective, we leverage the strengths of both transformer and CNN by proposing a hybrid transformer model. In the encoder phase, we propose a novel Modality-Correlated Cross-Attention (MCCA) block to fuse and extract the multi-modal features. In the decoding phase, a Trans&CNN Feature Calibration (TCFC) block is proposed to bridge the gap and the bias between the features of transformer and CNN. The effectiveness of all the technical novelties have been evaluated by the extensive experiments.

References

[1] D. N. Louis, A. Perry, P. Wesseling, D. J. Brat, I. A. Cree, D. Figarella-Branger, C. Hawkins, H. Ng, S. M. Pfister, G. Reifenberger *et al.*, "The 2021 who classification of tumors of the central nervous system: a summary," *Neuro-oncology*, vol. 23, no. 8, pp. 1231–1251, 2021.

[2] F. John, E. Bosnyák, N. L. Robinette, A. J. Amit-Yousif, G. R. Barger, K. D. Shah, S. K. Michelhaugh, N. V. Klinger, S. Mittal, and C. Juhász, "Multimodal imaging-defined subregions in newly diagnosed glioblastoma: impact on overall survival," *Neuro-oncology*, vol. 21, no. 2, pp. 264–273, 2019.

[3] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: achievements and challenges," *Journal of digital imaging*, vol. 32, no. 4, pp. 582–596, 2019.

[4] Z. Liu, L. Chen, L. Tong, F. Zhou, Z. Jiang, Q. Zhang, C. Shan, Y. Wang, X. Zhang, L. Li *et al.*, "Deep learning based brain tumor segmentation: a survey," *arXiv preprint arXiv:2007.09479*, 2020.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[6] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, "Transformers in medical imaging: A survey," *arXiv preprint arXiv:2201.09873*, 2022.

[7] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 574–584.

[8] A. Hatamizadeh, Z. Xu, D. Yang, W. Li, H. Roth, and D. Xu, "Unetformer: A unified vision transformer model and pre-training framework for 3d medical image segmentation," *arXiv preprint arXiv:2204.00631*, 2022.

[9] P. Wang and A. C. Chung, "Relax and focus on brain tumor segmentation," *Medical Image Analysis*, vol. 75, p. 102259, 2022.

[10] B. Yu, L. Zhou, L. Wang, W. Yang, M. Yang, P. Bourgeat, and J. Fripp, "Learning sample-adaptive intensity lookup table for brain tumor segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 216–226.

[11] Y. Chen and J. Joo, "Understanding and mitigating annotation bias in facial expression recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 980–14 991.

[12] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *arXiv preprint arXiv:2206.06488*, 2022.

[13] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F. C. Kitamura, S. Pati *et al.*, "The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification," *arXiv preprint arXiv:2107.02314*, 2021.

[14] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.

[15] A. Wadhwa, A. Bhardwaj, and V. S. Verma, "A review on brain tumor segmentation of mri images," *Magnetic resonance imaging*, vol. 61, pp. 247–259, 2019.

[16] X. Zhao, Y. Wu, G. Song, Z. Li, Y. Zhang, and Y. Fan, "A deep learning model integrating fcnns and crfs for brain tumor segmentation," *Medical image analysis*, vol. 43, pp. 98–111, 2018.

[17] A. Mehrtash, W. M. Wells, C. M. Tempany, P. Abolmaesumi, and T. Kapur, "Confidence calibration and predictive uncertainty estimation for deep medical image segmentation," *IEEE transactions on medical imaging*, vol. 39, no. 12, pp. 3868–3878, 2020.

[18] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[19] H. M. Luu and S.-H. Park, "Extending nn-unet for brain tumor segmentation," *arXiv preprint arXiv:2112.04653*, 2021.

[20] M. Futrega, A. Milesi, M. Marcinkiewicz, and P. Ribalta, "Optimized u-net for brain tumor segmentation," *arXiv preprint arXiv:2110.03352*, 2021.

[21] Z. Liu, L. Tong, L. Chen, F. Zhou, Z. Jiang, Q. Zhang, Y. Wang, C. Shan, L. Li, and H. Zhou, "Canet: Context aware network for brain glioma segmentation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 7, pp. 1763–1777, 2021.

[22] Z. Zhou, Z. He, and Y. Jia, "Afpnet: A 3d fully convolutional neural network with atrous-convolution feature pyramid for brain tumor segmentation via mri images," *Neurocomputing*, vol. 402, pp. 235–244, 2020.

[23] C. Zhou, C. Ding, X. Wang, Z. Lu, and D. Tao, "One-pass multi-task networks with cross-task guided attention for brain tumor segmentation," *IEEE Transactions on Image Processing*, vol. 29, pp. 4516–4529, 2020.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[25] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Computing Surveys (CSUR)*, 2021.

[26] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu *et al.*, "A survey on vision transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

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

[28] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li, "Transbts: Multimodal brain tumor segmentation using transformer," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 109–119.

[29] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2021, pp. 171–180.

[30] H. Peiris, M. Hayat, Z. Chen, G. Egan, and M. Harandi, "A volumetric transformer for accurate 3d tumor segmentation," *arXiv preprint arXiv:2111.13300*, 2021.

[31] H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, and Y. Yu, "nnformer: Interleaved transformer for volumetric segmentation," *arXiv preprint arXiv:2109.03201*, 2021.

[32] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," *arXiv preprint arXiv:2201.01266*, 2022.

[33] S. He, C. Han, G. Han, and J. Qin, "Exploring duality in visual question-driven top-down saliency," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 7, pp. 2672–2679, 2019.

[34] X. Xu, Y. Li, G. Wu, and J. Luo, "Multi-modal deep feature learning for rgb-d object detection," *Pattern Recognition*, vol. 72, pp. 300–313, 2017.

[35] Z. Ning, Z. Lin, Q. Xiao, D. Du, Q. Feng, W. Chen, and Y. Zhang, "Multi-constraint latent representation learning for prognosis analysis using multi-modal data," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[36] D. Zhang, G. Huang, Q. Zhang, J. Han, J. Han, Y. Wang, and Y. Yu, "Exploring task structure for brain tumor segmentation from multi-modality mr images," *IEEE Transactions on Image Processing*, vol. 29, pp. 9032–9043, 2020.

[37] Y. Wang, Y. Zhang, F. Hou, Y. Liu, J. Tian, C. Zhong, Y. Zhang, and Z. He, "Modality-pairing learning for brain tumor segmentation," in *International MICCAI Brainlesion Workshop*. Springer, 2020, pp. 230–240.

[38] T. Zhou, S. Ruan, Y. Guo, and S. Canu, "A multi-modality fusion network based on attention mechanism for brain tumor segmentation," in *2020 IEEE 17th international symposium on biomedical imaging (ISBI)*. IEEE, 2020, pp. 377–380.

[39] D. Zhang, G. Huang, Q. Zhang, J. Han, J. Han, and Y. Yu, "Cross-modality deep feature learning for brain tumor segmentation," *Pattern Recognition*, vol. 110, p. 107562, 2021.

[40] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[41] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[42] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick, "Early convolutions help transformers see better," *Advances in Neural Information Processing Systems*, vol. 34, pp. 30 392–30 400, 2021.

[43] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 096–10 106.

[44] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[45] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 565–571.

[46] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018.

[47] K.-L. Tseng, Y.-L. Lin, W. Hsu, and C.-Y. Huang, "Joint sequence learning and cross-modality convolution for 3d biomedical segmentation," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 6393–6400.

[48] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE transactions on medical imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.

[49] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical image analysis*, vol. 53, pp. 197–207, 2019.

[50] Q.-D. Pham, H. Nguyen-Truong, N. N. Phuong, K. N. Nguyen, C. D. Nguyen, T. Bui, and S. Q. Truong, "Segtransvae: Hybrid cnn-transformer with regularization for medical image segmentation," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2022, pp. 1–5.

[51] H. M. Luu and S.-H. Park, "Extending nn-unet for brain tumor segmentation," in *International MICCAI Brainlesion Workshop*. Springer, 2022, pp. 173–186.

[52] M. M. R. Siddiquee and A. Myronenko, "Redundancy reduction in semantic segmentation of 3d brain tumor mris," *arXiv preprint arXiv:2111.00742*, 2021.

[53] K. Kotowski, S. Adamski, B. Machura, L. Zarudzki, and J. Nalepa, "Coupling nnu-nets with expert knowledge for accurate brain tumor segmentation from mri," in *International MICCAI Brainlesion Workshop*. Springer, 2022, pp. 197–209.

[54] H. Jia, C. Bai, W. Cai, H. Huang, and Y. Xia, "Hnf-netv2 for brain tumor segmentation using multi-modal mr imaging," *arXiv preprint arXiv:2202.05268*, 2022.

[55] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle *et al.*, "The cancer imaging archive (tcia): maintaining and operating a public information repository," *Journal of digital imaging*, vol. 26, no. 6, pp. 1045–1057, 2013.