Multi-Scale Transformer-CNN Network for Brain Tumor Segmentation and Survival Prediction

Indrajit Mazumdar
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
Kharagpur, West Bengal, India
indrajit.mazumdar97@gmail.com

Jayanta Mukhopadhyay
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
Kharagpur, West Bengal, India
jay@cse.iitkgp.ac.in

Abstract

Accurately segmenting brain tumors and predicting survival is crucial for diagnostic and personalized treatment plans. Recently, convolutional neural networks (CNNs) based on U-Net architecture have been extensively used to segment brain tumors. It is essential to capture both local and global dependencies to segment brain tumors accurately. Therefore, several studies have considered U-Net variants by combining a CNN and a Transformer. However, the skip connections in these Transformer-based U-Net variants do not incorporate the features from multiple scales available at the encoder. To address this issue, we propose a new 3D U-Net variant called Multi-Scale Transformer-CNN Network (MTC-Net) that incorporates a Multi-Scale Transformer Convolution (MTC) block into the skip connections. The MTC block extracts local multi-scale features from the CNN encoder blocks and global features using Swin Transformer blocks to handle the wide variability in tumor size and enhance the feature representation capability. The segmentation mask predicted by the proposed MTC-Net is subsequently used to predict a patient's overall survival time. Most studies on overall survival prediction have used handcrafted radiomic features that lack the ability to fully model complex tumor patterns. In contrast, deep features are specifically adapted to brain tumors. We observed that instead of using only a CNN, combining it with a Transformer produces deep features that provide more accurate predictions. For this purpose, we extract deep features using our Transformer-based network MTC-Net to construct a regression model for predicting overall survival. Comprehensive experimentation on the BraTS 2020 and 2021 benchmark datasets proved the efficacy of the proposed components. MTC-Net outperformed the CNN- and Transformer-based state-of-the-art segmentation networks. Moreover, our approach outperformed the state-of-the-art survival prediction systems.

CCS Concepts

• **Computing methodologies → Neural networks**; **Image segmentation**; **Supervised learning by regression**.

ICVGIP 2024, December 13–15, 2024, Bengaluru, India
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1075-9/24/12
https://doi.org/10.1145/3702250.3702257

Keywords

Brain tumor segmentation, Deep learning, CNN, Transformer, Survival prediction, Machine learning

ACM Reference Format:
Indrajit Mazumdar and Jayanta Mukhopadhyay. 2024. Multi-Scale Transformer-CNN Network for Brain Tumor Segmentation and Survival Prediction. In *Indian Conference on Computer Vision Graphics and Image Processing (ICVGIP 2024), December 13–15, 2024, Bengaluru, India.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3702250.3702257

1 Introduction

The typical kind of brain tumor is glioma, which is sorted as high-grade glioma (HGG) or low-grade glioma (LGG) [8, 22, 31]. Patients with HGG survive for two years or less, whereas those with LGG survive for several years [27, 31]. Gliomas are treated using surgery, radiotherapy, and chemotherapy [8, 13, 22]. They comprise three tumor structures: enhancing tumor (ET), necrotic tumor core (NCR), and edema (ED) [3, 7, 27]. According to clinical applications, these structures are categorized into ET, whole tumor (WT), and tumor core (TC) sub-regions [3, 7, 27]. WT encompasses all tumor structures, whereas TC comprises the NCR and ET structures. Since magnetic resonance imaging (MRI) is noninvasive and uses multiple modalities to deliver complementary information, it is extensively used for detecting and analyzing gliomas [8, 22]. Accurately segmenting gliomas in MRI scans is immensely valuable in diagnostics and planning treatment. Nevertheless, manually segmenting gliomas is laborious and subjective [8, 22]. Hence, it is clinically extremely beneficial to fully automate the process of segmentation [22]. However, this task is challenging owing to considerable variances in gliomas' appearances and shapes. After the tumor is segmented, imaging features are extracted based on the predicted segmentation mask to predict overall survival time, which varies highly across patients. Consequently, overall survival prediction is of enormous clinical importance as it aids in developing personalized treatment strategies. Nevertheless, this task is challenging because it is difficult to identify relevant imaging features. Moreover, accurate glioma segmentation is a prerequisite for extracting accurate imaging features.

Recently, for segmenting gliomas, employing deep learning algorithms like convolutional neural networks (CNNs) has become popular. Encoder-decoder CNNs with skip connections like 2D U-Net [35] and 3D U-Net [12], are widely employed to conduct end-to-end segmentation. The top-performing methods for segmenting gliomas are based on U-Net. For example, No-New-Net [21] is a marginally enhanced U-Net variant that achieved good performance. Myronenko [30] employed an asymmetrical U-Net and regularized the encoder by incorporating a variational autoencoder branch. OM-Net [44] is a U-Net modification incorporating three different segmentation tasks. DeepSCAN [26] is another U-Net variant containing dense blocks [18] and is trained with label-uncertainty loss. Additionally, the nnU-Net [19] is a general-purpose U-Net variant that was subsequently utilized to segment brain tumors by integrating BraTS challenge-specific changes [20]. Moreover, ESA-Net [25] is an efficient U-Net variant comprising depthwise separable convolutions [37] and is trained with a composite loss. The above CNN-based U-Net variants utilize convolution operations that have a good ability to extract local information but have difficulty in modeling global (or long-range) information. It is essential to capture global dependencies to segment tumors accurately.

Currently, Transformer [39] has gained huge prominence for its capacity to extract global information employing the self-attention technique. Vision Transformer (ViT) [14] converts an image into patches and employs a Transformer to capture the relationships between these patches, achieving good performance on image classification. However, ViT has quadratic computational complexity, which makes it inefficient. In contrast, Swin Transformer [24] is an efficient Transformer with linear computational complexity that computes self-attention locally in each window using a shifted window scheme. Taking inspiration from this, Swin-Unet [9] used Swin Transformer blocks to build the decoder and encoder of 2D U-Net. Nonetheless, Transformers are unable to effectively extract local information. It is essential to model both local and global dependencies to conduct accurate semantic segmentation. Consequently, several works have combined CNN and Transformer to exploit their benefits. For instance, TransUNet [10] applies a ViT in the bottleneck of 2D U-Net for conducting multi-organ segmentation. Likewise, TransBTS [41] incorporates a Transformer in the bottleneck of 3D U-Net for glioma segmentation. UNETR [17] modifies U-Net by replacing the CNN encoder with a Transformer encoder. Although the above Transformer-based networks have delivered promising results, it is still a challenge to effectively extract local and global information. Moreover, the skip pathways in these Transformer-based U-Net variants directly combine the encoder and decoder features at the same scale, even though features from multiple scales are available at the encoder. The incorporation of these multi-scale features into the skip connections may take into consideration the size, location, and shape features of tumors at different scales to tackle the variations in tumor size. To cope with the abovementioned problems, we introduce a novel 3D U-Net variant called Multi-Scale Transformer-CNN Network (MTC-Net) to segment brain tumors with improved accuracy. Specifically, MTC-Net introduces a novel Multi-Scale Transformer Convolution (MTC) block into the skip pathways to capture local multi-scale information from the CNN encoder blocks and global features using Swin Transformer blocks. The MTC block effectively uses local multi-scale and global features to handle the wide variability in tumor size and improve the feature representation capability.

In most recent studies, predicting the overall survival of patients with glioma has been made by extracting handcrafted radiomic features and feeding them into a machine learning regression model to predict survival time. Although end-to-end CNN delivers good segmentation performance, using it to perform direct regression results in poor performance owing to overfitting because the dataset comprises few training samples [34]. Besides handcrafted features, clinical information such as resection status and age have been in use to predict survival. For instance, Feng et al. [15] extracted surface area and volume features from all tumor sub-regions and merged those features with age and resection status for training a linear regression method. Wang et al. [40] used age, volume ratio, and surface area features to train a multilayer perceptron (MLP) regressor. Parmar et al. [32] employed shape, volume, and age features for constructing a random forest (RF) model. Pei et al. [34] extracted shape and age features and fed those into an RF regressor. Agravat et al. [1] used statistical, shape, and age features to train an RF regressor. Anand et al. [2] employed texture, shape, and first order features for training an RF regressor. Miron et al. [29] developed a risk-based feature and used it along with age, volume ratio, and shape features to train an Extra-Trees [16] model. Although most studies have used handcrafted radiomic features, these features have limitations of being low-level and generalized, lacking the ability to fully model complex tumor patterns. In contrast, deep features are high-level and specifically adapted to brain tumors. A recent work by Russo et al. [36] extracted deep features from a CNN and fed those into a linear model for predicting survival. However, instead of using only a CNN, combining it with a Transformer may produce deep features that are more accurate and powerful. Therefore, we extract deep features from our Transformer-based network MTC-Net to construct a machine learning regression model for predicting overall survival.

This work introduces a new Transformer-CNN Segmentation Survival (TCSS) model that contains a segmentation component to accurately segment gliomas and a survival prediction component to predict overall survival. The proposed MTC-Net is used as the segmentation component. The survival prediction component uses deep features extracted from our MTC-Net to build a machine learning model for predicting overall survival. Comprehensive experiments were conducted on the BraTS 2020 and 2021 benchmark datasets [3–7, 27] to validate the efficacy of the introduced components. Our MTC-Net outperforms the CNN- and Transformer-based state-of-the-art segmentation approaches. Moreover, TCSS outperforms the state-of-the-art survival prediction approaches. This study has the following primary contributions:

- The proposed TCSS model comprises segmentation and survival prediction components for accurately segmenting gliomas and predicting overall survival.
- The segmentation component of TCSS is the proposed MTC-Net, which introduces an MTC block into the skip pathways. The MTC block extracts local multi-scale information from the CNN encoder blocks and global features using Swin Transformer blocks to tackle the large size variations of tumors and enhance the feature representation capability.
- Our Transformer-based network MTC-Net is used for extracting accurate and powerful deep features to construct a regression model for the survival prediction component of TCSS.

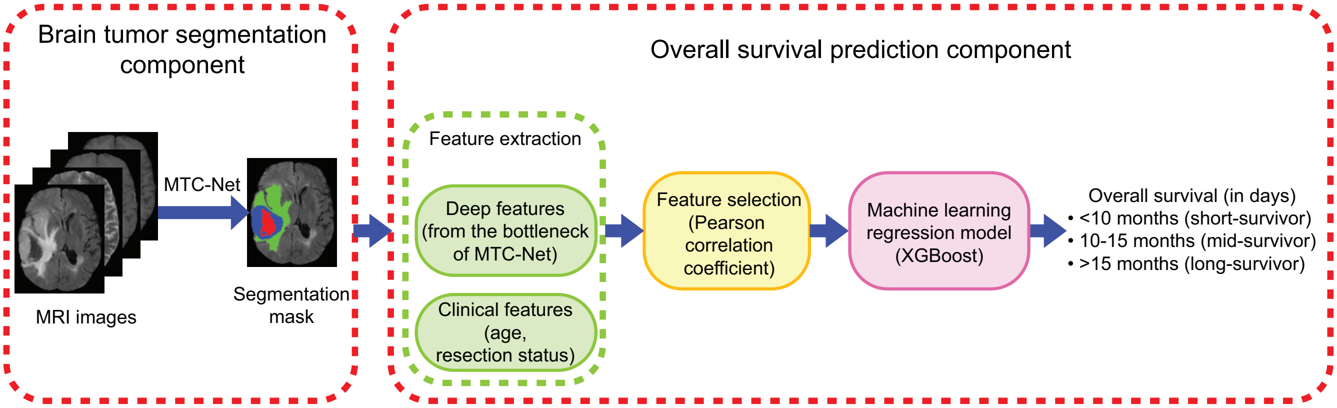
Brain tumor segmentation component
MRI images
MTC-Net
Segmentation mask
Overall survival prediction component
Feature extraction
Deep features (from the bottleneck of MTC-Net)
Clinical features (age, resection status)
Feature selection (Pearson correlation coefficient)
Machine learning regression model (XGBoost)
Overall survival (in days)
• <10 months (short-survivor)
• 10-15 months (mid-survivor)
• >15 months (long-survivor)

Figure 1: Overview of TCSS comprising a brain tumor segmentation component followed by an overall survival prediction component.

2 Methods

2.1 Transformer-CNN Segmentation Survival (TCSS) Model

This section describes the proposed TCSS model to segment gliomas and predict overall survival, depicted in Figure 1. It comprises two components: one for segmentation of gliomas and the other for prediction of overall survival. The brain tumor segmentation component uses our proposed MTC-Net to perform segmentation. In the overall survival prediction component, feature extraction is initially performed by extracting deep features from the MTC-Net encoder. In addition, clinical features were used. Next, the most important features are identified utilizing feature selection. Finally, the identified features are employed to construct a machine learning regression approach for predicting overall survival.

2.2 Multi-Scale Transformer-CNN Network (MTC-Net)

Here, we present the proposed MTC-Net, which is the segmentation component of TCSS. Figure 2 illustrates MTC-Net. It takes an input volume of size $H \times W \times D \times 4$ having a spatial resolution of $H \times W \times D$ with 4 MRI modalities merged along the channel dimension. Each MTC-Net encoder and decoder block comprises two convolutional blocks, each consisting of a $3 \times 3 \times 3$ convolution preceding a group normalization (GN) [43] and ReLU.

A Multi-Scale Transformer Convolution (MTC) block is introduced into the skip connections to effectively capture local multi-scale and global information to address the wide variances in the size of tumors and enhance the feature representation capability. Figure 3 illustrates the MTC block at level j of our MTC-Net. The term level is used to denote a particular vertical depth of the network, with level 1 having the highest spatial resolution; in higher levels, the spatial resolution decreases. First, the MTC block captures multi-scale information. Let j indicate the level in which the MTC block is present. It takes the features generated by the encoder blocks at multiple scales as input. We represent the feature maps produced by the encoder blocks as $\mathbf{E}_i \in \mathbb{R}^{H_i \times W_i \times D_i \times C_i}, i = 1, \ldots, L-1$, where i represents the level of the encoder block and L indicates the total level count ($L = 5$ for our MTC-Net). Here, H_i, W_i, D_i, and C_i indicate the height, width, depth, and channel count of the encoder activation at level i, respectively. Next, the input encoder features from multiple levels are resized to match the encoder feature size at level j using a resize function $\mathbf{R}_{i,j}(\cdot)$ where i and j designate the levels of the encoder and MTC blocks, respectively. When $i > j$, $\mathbf{R}_{i,j}(\mathbf{E}_i)$ employs strided transpose convolution to upsample $\mathbf{E}_i$ by a factor of $2^{(i-j)}$. When $i < j$, $\mathbf{R}_{i,j}(\mathbf{E}_i)$ employs strided convolution to downsample $\mathbf{E}_i$ by a factor of $2^{(j-i)}$. When $i = j$, $\mathbf{R}_{i,j}(\mathbf{E}_i)$ retains $\mathbf{E}_i$. Next, all the resized feature maps are concatenated as follows:

$$\mathbf{X}_j = Concat\left(\mathbf{R}_{1,j}(\mathbf{E}_1), \mathbf{R}_{2,j}(\mathbf{E}_2), \mathbf{R}_{3,j}(\mathbf{E}_3), \mathbf{R}_{4,j}(\mathbf{E}_4)\right), \tag{1}$$

where $\mathbf{X}_j \in \mathbb{R}^{H_j \times W_j \times D_j \times 4C_j}$ is the concatenated feature map containing the extracted multi-scale features. Subsequently, $\mathbf{X}_j$ is fed to a $1 \times 1 \times 1$ convolution succeeded by GN and ReLU to further combine the multi-scale features and adjust the channel count as follows:

$$\mathbf{Y}_j = \delta\left(GN\left(Conv_{1\times1\times1}\left(\mathbf{X}_j\right)\right)\right), \tag{2}$$

where $\mathbf{Y}_j \in \mathbb{R}^{H_j \times W_j \times D_j \times C_{jm}}$ is the feature map containing multi-scale information and δ denotes the ReLU. Here, C_{jm} represents the channel count of the MTC block at level j. We set C_{jm} to 48 at level 1, and at each consecutive level, its value doubles. After capturing multi-scale information, the MTC block uses Swin Transformer blocks [24] to capture global dependencies. For this purpose, the feature map $\mathbf{Y}_j$ is supplied as input to two successive Swin Transformer blocks. A Swin Transformer block consists of an MLP, a multi-head self-attention (MSA) module, and a layer normalization (LN) layer. The window-based multi-head self-attention (W-MSA) and shifted window-based multi-head self-attention (SW-MSA) modules are utilized in the two Transformer blocks in succession. These blocks were extended from 2D to 3D in order to employ them in MTC-Net. We compute the successive blocks as follows:

$$\hat{\mathbf{s}}^{j,1} = \text{W-MSA}\left(\text{LN}\left(\mathbf{Y}_j\right)\right) + \mathbf{Y}_j, \tag{3}$$

$$\mathbf{s}^{j,1} = \text{MLP}\left(\text{LN}\left(\hat{\mathbf{s}}^{j,1}\right)\right) + \hat{\mathbf{s}}^{j,1}, \tag{4}$$

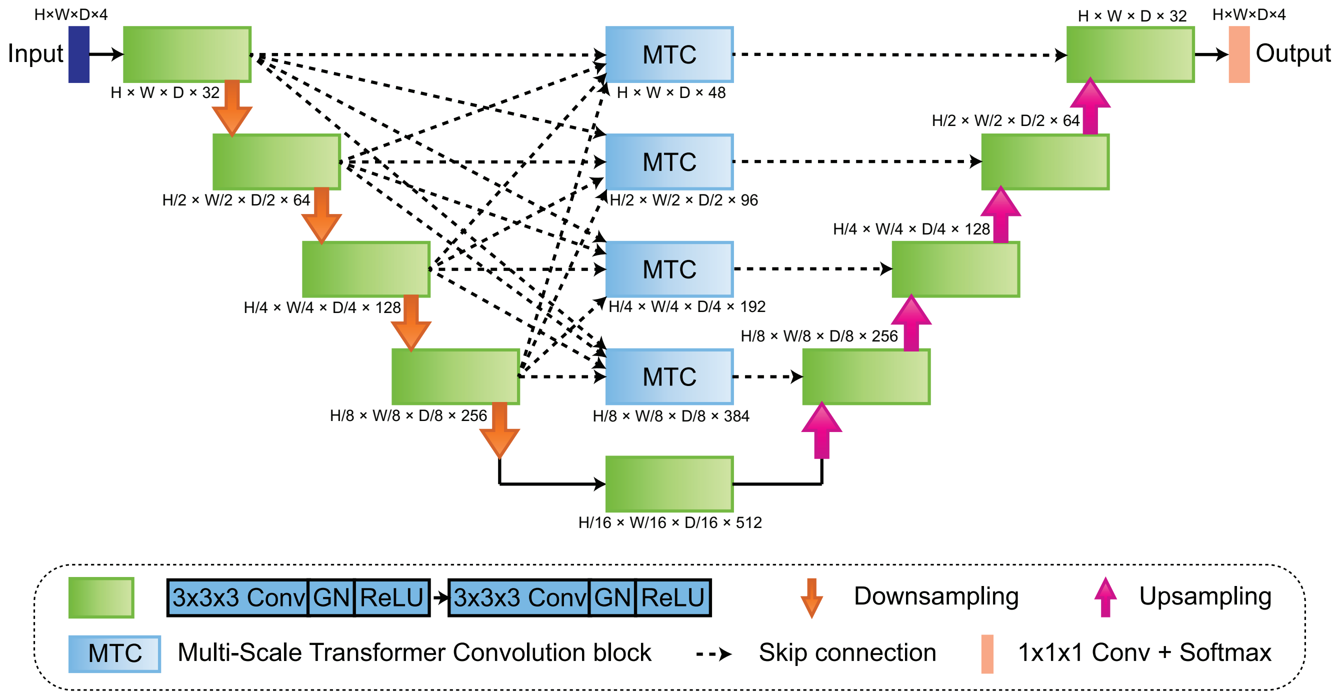

Figure 2: Proposed MTC-Net architecture.

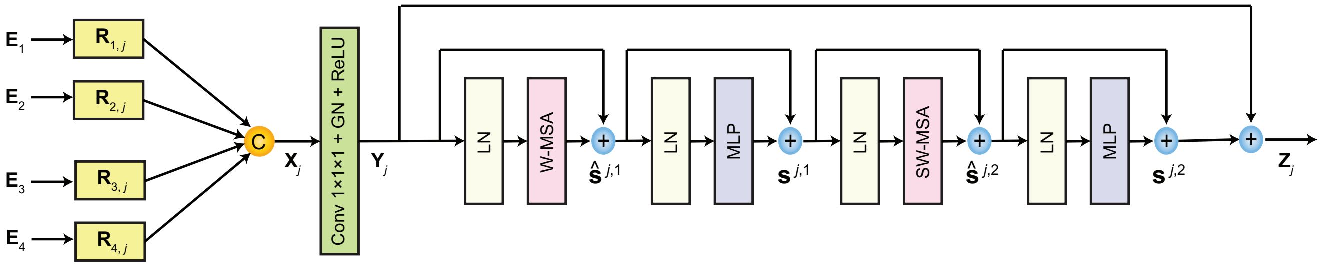

Figure 3: Proposed MTC block at level j of our MTC-Net.

$$\hat{\mathbf{s}}^{j,2} = \text{SW-MSA}\left(\text{LN}\left(\mathbf{s}^{j,1}\right)\right) + \mathbf{s}^{j,1}, \tag{5}$$

$$\mathbf{s}^{j,2} = \text{MLP}\left(\text{LN}\left(\hat{\mathbf{s}}^{j,2}\right)\right) + \hat{\mathbf{s}}^{j,2}, \tag{6}$$

where $\hat{\mathbf{s}}^{j,1}$ and $\mathbf{s}^{j,1}$ indicate the output activations of the W-MSA and the MLP modules for the first block, respectively; further $\hat{\mathbf{s}}^{j,2}$ and $\mathbf{s}^{j,2}$ indicate the output activations of the SW-MSA and the MLP modules for the second block, respectively. Finally, a residual connection is utilized to combine the local multi-scale features with the global features. Specifically, the feature $\mathbf{Y}_j$ is added to the feature $\mathbf{s}^{j,2}$ skipping the intermediate layers using a residual connection as follows:

$$\mathbf{Z}_j = \mathbf{s}^{j,2} + \mathbf{Y}_j, \tag{7}$$

where $\mathbf{Z}_j \in \mathbb{R}^{H_j \times W_j \times D_j \times C_{jm}}$ is the output of the MTC block at level j that is subsequently fed to the decoder block at the same level.

2.3 Overall Survival Prediction Component of TCSS

Here, we introduce the overall survival prediction component of TCSS. Most studies have used handcrafted radiomic features, which have limitations of being low-level and generalized, lacking the ability to fully model complex tumor patterns. In contrast, deep features are high-level and specifically adapted to brain tumors. Therefore, we use deep features to build a machine learning regression model as the survival prediction component of TCSS.

First, deep features are extracted from our Transformer-based network MTC-Net. For this purpose, the tumor region is cropped

Table 1: Ablation analysis of MTC-Net on the BraTS 2021 dataset.

Method	Dice				HD95 (mm)			
	ET	WT	TC	Average	ET	WT	TC	Average
3D U-Net	0.821	0.892	0.851	0.854	12.16	6.16	8.65	8.99
3D U-Net + MTC (only multi-scale)	0.849	0.915	0.883	0.882	9.46	4.53	5.89	6.62
3D U-Net + MTC (only level 1)	0.861	0.920	0.893	0.891	8.35	4.11	5.28	5.91
MTC-Net	**0.880**	**0.933**	**0.907**	**0.906**	**6.89**	**3.51**	**4.46**	**4.95**

from the four MRI modalities according to the predicted segmentation mask. Then, each cropped image is resized to $64 \times 64 \times 64$ and fed as input to MTC-Net. Finally, from the end of the bottleneck of MTC-Net, deep features are extracted and subsequently flattened. Thus, a total of $4 \times 4 \times 4 \times 512 = 32768$ deep features are extracted.

Moreover, we compare our deep features with handcrafted radiomic features. To this end, handcrafted radiomic features are extracted according to the predicted segmentation mask utilizing the PyRadiomics tool [38]. The extracted features comprise 74 texture, 18 first order, and 14 shape features. Detailed feature definitions are available at [45]. These features are extracted for each tumor part. In addition, clinical features such as resection status and age are used.

After feature extraction, each feature is normalized employing Z-score normalization. Owing to the high number of extracted features and few training samples, feature selection is required to prevent overfitting, which is performed using the Pearson correlation coefficient. The selected features are utilized for training a machine learning regression model, which in our case is XGBoost [11] to predict overall survival.

3 Experiments and Results

3.1 Datasets

The BraTS 2020 and 2021 datasets [3–7, 27] were utilized in the experiments. They comprise four MRI images for every patient: contrast-enhanced T1-weighted (T1ce), fluid-attenuated inversion recovery (FLAIR), T2-weighted (T2), and T1-weighted (T1). These MRI modalities were interpolated to 1 mm^3 resolution, skull-stripped, and co-registered. A size $155 \times 240 \times 240$ applies to each MRI scan. Each patient was manually segmented with the class labels comprising ED, NCR, ET, and background. For the segmentation task, the datasets contain both HGG and LGG patients. The BraTS 2020 and 2021 datasets contain 369 and 1251 patients, respectively. The datasets were randomly partitioned into 70% for training, 10% for validation, and 20% for testing, which is consistent with recent work [42]. To analyze the segmentation efficacy, we primarily utilized the BraTS 2021 dataset because it is considerably larger than the BraTS 2020 dataset. However, the BraTS 2021 dataset does not include survival prediction data. Therefore, the BraTS 2020 dataset was used for the task of survival prediction because it includes additional information regarding the age, resection status, and overall survival time (in days) of 236 HGG patients, of which 10 have subtotal resection (STR), 119 have gross total resection (GTR), and 107 do not have resection status available. We utilized 5-fold cross-validation to assess the effectiveness of survival prediction.

3.2 Implementation Details

PyTorch [33] was utilized to implement the segmentation network. Each MRI sequence was normalized during pre-processing to ensure unit variance and zero mean. Randomly extracted 3D patches $128 \times 128 \times 128$ in size were used for training. Random flipping was employed to augment data. Furthermore, soft Dice loss [28], batch size of 1, L2 regularization of 10^{-5}, a learning rate of 10^{-4}, and Adam optimizer [23] were utilized to train the network for 300 epochs. Additionally, for survival prediction, the deep features were extracted using the trained segmentation model.

3.3 Evaluation Metrics

Segmentation performance is evaluated for the TC, WT, and ET tumor regions [3, 7, 27]. Following the BraTS challenges [3, 7], the Dice score and 95th percentile Hausdorff distance (HD95) metrics were utilized for assessing the segmentation performance, which is in accordance with previous studies [17, 20, 21, 25, 26, 30, 41, 44]. Survival prediction performance is evaluated for HGG patients with GTR using classification and regression performance evaluation methods [7]. For the classification evaluation method, the survival time is utilized to categorize the patients into long-survivors (>15 months), mid-survivors (10–15 months), and short-survivors (<10 months) classes, with the performance being evaluated using the accuracy metric [7]. For the regression evaluation method, an error analysis is performed between the predicted and actual survival times, and the performance is evaluated using the mean square error (MSE) metric [7]. Precisely, following the BraTS challenge [7], accuracy and MSE metrics were utilized for assessing the survival prediction performance, which aligns with existing studies [1, 2, 15, 29, 32, 34, 36, 40].

3.4 Efficacy of the Multi-Scale Transformer-CNN Network

To examine the efficacy of MTC-Net, we did an ablation analysis on the BraTS 2021 dataset in Table 1. The top performances are indicated in bold in the tables. The baseline network is 3D U-Net, which was generated by removing the MTC blocks from MTC-Net and using plain skip connections. We subsequently added to the skip pathways of 3D U-Net our MTC blocks incorporating only the multi-scale features without the Swin Transformer blocks (3D

Table 2: Ablation analysis of MTC-Net on the BraTS 2020 dataset.

Method	Dice				HD95 (mm)			
	ET	WT	TC	Average	ET	WT	TC	Average
3D U-Net	0.770	0.875	0.806	0.817	16.07	7.26	8.20	10.51
3D U-Net + MTC (only multi-scale)	0.802	0.899	0.837	0.846	12.06	5.41	5.87	7.78
3D U-Net + MTC (only level 1)	0.814	0.905	0.847	0.855	10.75	4.86	5.29	6.96
MTC-Net	**0.832**	**0.917**	**0.864**	**0.871**	**8.91**	**4.11**	**4.51**	**5.84**

U-Net + MTC (only multi-scale)). This addition improved the average Dice by 2.8% and the average HD95 by 26.3%. This occurs since incorporating the features from multiple scales into the skip pathways utilizes the shape, location, and size features of gliomas at different scales to handle the variations in tumor size. Next, we added the MTC block only to the skip pathway at level 1 of 3D U-Net (3D U-Net + MTC (only level 1)). Finally, we added the MTC blocks to all the skip pathways at multiple levels of 3D U-Net to produce the proposed MTC-Net. We observed that MTC-Net outperformed 3D U-Net + MTC (only multi-scale) in average Dice and HD95 by 2.4% and 25.2%, respectively. It reveals that capturing global features using Swin Transformer blocks and combining them with complementary local multi-scale features improves the feature representation capability, thereby enhancing performance. Moreover, MTC-Net was found to outperform 3D U-Net + MTC (only level 1) in average Dice and HD95 by 1.5% and 16.2%, respectively. This happens because only the level 1 decoder block in 3D U-Net + MTC (only level 1) uses the local multi-scale and global features passed through the skip connections, whereas decoder blocks at all levels in MTC-Net use these features to enhance the segmentation performance further. Moreover, in Table 2, we used the BraTS 2020 dataset to compare the performance of the above networks and found the performances to corroborate the Table 1 findings. It is evident that the MTC blocks effectively utilize local multi-scale and global features to handle the wide variability in tumor size and enhance the feature representation capability, thereby boosting performance.

3.5 Efficacy of the Overall Survival Prediction Component of TCSS

We analyze the efficacy of the overall survival prediction component of TCSS by comparing different feature sets in Table 3. The baseline model is obtained by replacing the deep features in TCSS with handcrafted radiomic features. Next, we replaced our MTC-Net in the segmentation component of TCSS with the 3D U-Net and employed it to extract deep features. We observed that deep features outperformed handcrafted features. To be precise, the deep features from 3D U-Net produced 6.9% higher accuracy and 47.1% lower MSE than handcrafted features. This is because handcrafted features are low-level and generalized, whereas deep features are high-level and specifically adapted to brain tumors. Thus, deep features can better model complex tumor patterns than handcrafted features, leading to enhanced performance. Finally, we used the proposed TCSS that contains MTC-Net in the segmentation component, which is utilized to extract deep features. It is observed that the usage of MTC-Net in place of 3D U-Net for extracting deep features improves the accuracy by 1.6% and MSE by 12.8%. This shows that compared to the CNN-based network 3D U-Net, our Transformer-based network MTC-Net effectively exploits the benefits of both CNN and Transformer to produce more accurate and powerful deep features, resulting in improved performance.

Table 3: Analysis of the overall survival prediction component of TCSS on the BraTS 2020 dataset.

Method	Accuracy	MSE
TCSS (Handcrafted radiomic features)	0.643	83023.70
TCSS (Deep features from 3D U-Net)	0.712	43895.33
TCSS	**0.728**	**38274.54**

3.6 Comparison With Other Methods

Here, the brain tumor segmentation component of TCSS, MTC-Net, is initially compared with state-of-the-art segmentation networks comprising CNN-based techniques (3D U-Net, nnU-Net [20], and ESA-Net [25]) and Transformer-based networks (TransUNet [10], TransBTS [41], UNETR [17], and SwinUNet [9]). The segmentation performances are listed in Table 4. We found MTC-Net to beat all the CNN- and Transformer-based networks, demonstrating its superiority. Precisely, MTC-Net outperformed the best-competing networks in average Dice and HD95 by 0.1% and 7.9%, respectively. The qualitative comparison with various networks is shown in Figure 4. We can see that MTC-Net segmented the tumor parts accurately, with its segmentations being most close to the ground truth.

Furthermore, the TCSS is compared with state-of-the-art overall survival prediction approaches [1, 2, 29, 32, 34, 36]. Table 5 lists the results. Among these approaches, Parmar et al. [32], Pei et al. [34], Agravat et al. [1], Anand et al. [2], and Miron et al. [29] used handcrafted radiomic features, whereas Russo et al. [36] used deep features extracted from a CNN. It is observed that TCSS outperformed all these approaches, obtaining the best performance. Specifically, TCSS outperformed the top-competing approaches in terms of accuracy and MSE by 2.5% and 19.2%, respectively.

Table 4: Comparison with different segmentation techniques on the BraTS 2021 dataset.

Method	Dice				HD95 (mm)			
	ET	WT	TC	Average	ET	WT	TC	Average
3D U-Net	0.821	0.892	0.851	0.854	12.16	6.16	8.65	8.99
nnU-Net [20]	0.879	0.929	**0.907**	0.905	9.05	**3.43**	5.17	5.88
ESA-Net [25]	**0.888**	0.924	0.901	0.904	7.36	3.71	5.07	5.38
TransUNet [10]	0.802	0.885	0.870	0.852	11.45	6.76	9.35	9.18
TransBTS [41]	0.855	0.920	0.899	0.891	9.32	3.91	5.56	6.26
UNETR [17]	0.849	0.917	0.886	0.884	9.80	4.59	6.21	6.86
SwinUNet [9]	0.820	0.894	0.865	0.859	12.20	6.20	8.82	9.07
MTC-Net	0.880	**0.933**	**0.907**	**0.906**	**6.89**	3.51	**4.46**	**4.95**

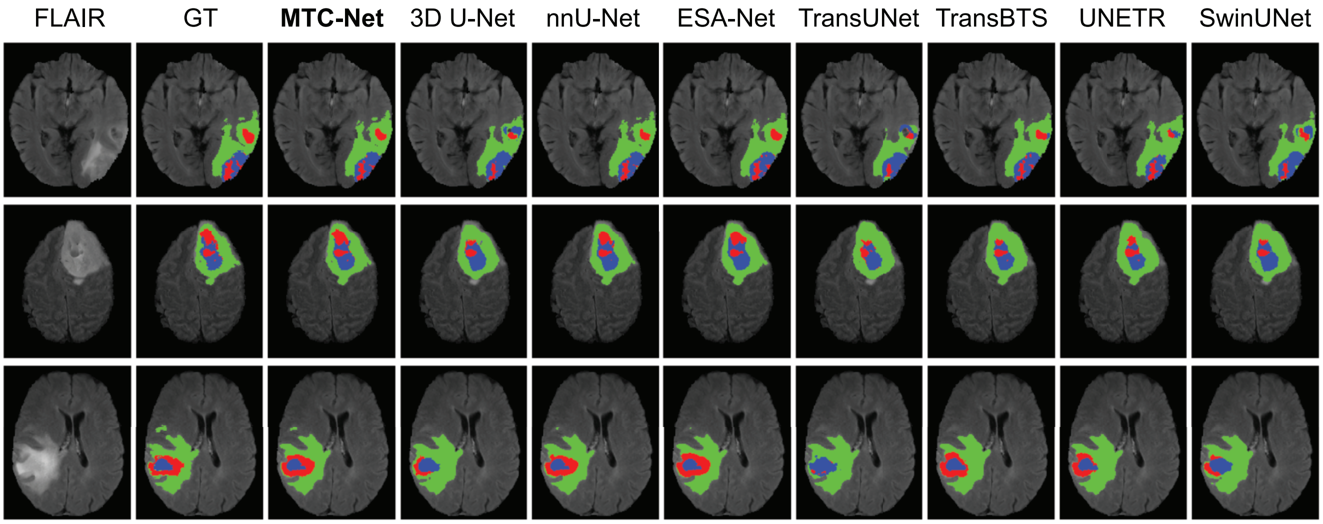

Figure 4: Representative segmentations for various approaches. The first to the third rows show the segmentations for patients "BraTS2021_00112," "BraTS2021_01266," and "BraTS2021_01216". The ED, ET, and NCR are marked in green, blue, and red, respectively.

Table 5: Comparison with various overall survival prediction approaches on the BraTS 2020 dataset.

Method	Accuracy	MSE
Parmar et al. [32]	0.482	100814.50
Pei et al. [34]	0.551	104754.82
Agravat et al. [1]	0.558	85184.79
Anand et al. [2]	0.576	47417.52
Miron et al. [29]	0.703	63608.43
Russo et al. [36]	0.483	166974.04
TCSS	**0.728**	**38274.54**

4 Conclusions

This study introduced a novel TCSS model to accurately segment gliomas and predict overall survival. TCSS comprises two components: one for segmentation and the other for predicting survival. A 3D U-Net variant called MTC-Net was proposed as the segmentation component of TCSS. MTC-Net introduces an MTC block into the skip pathways to capture local multi-scale information from the CNN encoder blocks and global features using Swin Transformer blocks. The MTC block effectively utilizes local multi-scale and global features to handle the wide variances in the size of tumors and enhance the feature representation capability, thereby improving the segmentation performance. For the overall survival prediction component of TCSS, we used our Transformer-based network MTC-Net to extract accurate and powerful deep features to construct a regression model for predicting overall survival. Exhaustive experimentation done on the BraTS 2020 and 2021 datasets proved the efficacy of the introduced components. MTC-Net outperformed the CNN- and Transformer-based state-of-the-art segmentation techniques. Moreover, TCSS outperformed the state-of-the-art survival prediction approaches. In the future, we will consider extending MTC-Net to other segmentation tasks. Finally, the accuracy of our survival prediction method can be further increased by using more clinical information and a large dataset.

References

[1] Rupal R Agravat and Mehul S Raval. 2021. 3D Semantic Segmentation of Brain Tumor for Overall Survival Prediction. In *International MICCAI Brainlesion Workshop (BrainLes)*, Alessandro Crimi and Spyridon Bakas (Eds.). Springer, Cham, 215–227. https://doi.org/10.1007/978-3-030-72087-2_19

[2] Vikas Kumar Anand, Sanjeev Grampurohit, Pranav Aurangabadkar, Avinash Kori, Mahendra Khened, Raghavendra S Bhat, and Ganapathy Krishnamurthi. 2021. Brain Tumor Segmentation and Survival Prediction Using Automatic Hard Mining in 3D CNN Architecture. In *International MICCAI Brainlesion Workshop (BrainLes)*, Alessandro Crimi and Spyridon Bakas (Eds.). Springer, Cham, 310–319. https://doi.org/10.1007/978-3-030-72087-2_27

[3] Ujjwal Baid and et al. 2021. The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification. *arXiv:2107.02314* (2021). https://doi.org/10.48550/arXiv.2107.02314

[4] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin Kirby, John Freymann, Keyvan Farahani, and Christos Davatzikos. 2017. Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-GBM collection. https://doi.org/10.7937/K9/TCIA.2017.KLXWJJ1Q

[5] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin Kirby, John Freymann, Keyvan Farahani, and Christos Davatzikos. 2017. Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-LGG collection. https://doi.org/10.7937/K9/TCIA.2017.GJQ7R0EF

[6] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S. Kirby, John B. Freymann, Keyvan Farahani, and Christos Davatzikos. 2017. Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific Data* 4, 1 (2017), 170117. https://doi.org/10.1038/sdata.2017.117

[7] Spyridon Bakas and et al. 2018. Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge. *arXiv:1811.02629* (2018). https://doi.org/10.48550/arXiv.1811.02629

[8] Stefan Bauer, Roland Wiest, Lutz-P Nolte, and Mauricio Reyes. 2013. A survey of MRI-based medical image analysis for brain tumor studies. *Physics in Medicine and Biology* 58, 13 (2013), R97–R129. https://doi.org/10.1088/0031-9155/58/13/R97

[9] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. 2023. Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation. In *European Conference on Computer Vision Workshops (ECCVW)*, Leonid Karlinsky, Tomer Michaeli, and Ko Nishino (Eds.). Springer, Cham, 205–218. https://doi.org/10.1007/978-3-031-25066-8_9

[10] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. 2021. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv:2102.04306* (2021). https://doi.org/10.48550/arxiv.2102.04306

[11] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 785–794. https://doi.org/10.1145/2939672.2939785

[12] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 2016. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Sebastien Ourselin, Leo Joskowicz, Mert R. Sabuncu, Gozde Unal, and William Wells (Eds.). Springer, Cham, 424–432. https://doi.org/10.1007/978-3-319-46723-8_49

[13] Lisa M. DeAngelis. 2001. Brain Tumors. *New England Journal of Medicine* 344, 2 (2001), 114–123. https://doi.org/10.1056/NEJM200101113440207

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=YicbFdNTTy

[15] Xue Feng, Nicholas J. Tustison, Sohil H. Patel, and Craig H. Meyer. 2019. Brain Tumor Segmentation Using an Ensemble of 3D U-Nets and Overall Survival Prediction Using Radiomic Features. In *International MICCAI Brainlesion Workshop (BrainLes)*, Alessandro Crimi, Spyridon Bakas, Hugo Kuijf, Farahani Keyvan, Mauricio Reyes, and Theo van Walsum (Eds.), Vol. 14. Springer, Cham, 279–288. https://doi.org/10.1007/978-3-030-11726-9_25

[16] Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Machine Learning* 63, 1 (2006), 3–42. https://doi.org/10.1007/s10994-006-6226-1

[17] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. 2022. UNETR: Transformers for 3D Medical Image Segmentation. In *Workshop on Applications of Computer Vision (WACV)*. IEEE, 1748–1758. https://doi.org/10.1109/WACV51458.2022.00181

[18] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2261–2269. https://doi.org/10.1109/CVPR.2017.243

[19] Fabian Isensee, Paul F Jaeger, Simon A A Kohl, Jens Petersen, and Klaus H Maier-Hein. 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods* 18, 2 (2021), 203–211. https://doi.org/10.1038/s41592-020-01008-z

[20] Fabian Isensee, Paul F Jäger, Peter M Full, Philipp Vollmuth, and Klaus H Maier-Hein. 2021. nnU-Net for Brain Tumor Segmentation. In *International MICCAI Brainlesion Workshop (BrainLes)*, Alessandro Crimi and Spyridon Bakas (Eds.). Springer, Cham, 118–132. https://doi.org/10.1007/978-3-030-72087-2_11

[21] Fabian Isensee, Philipp Kickingereder, Wolfgang Wick, Martin Bendszus, and Klaus H. Maier-Hein. 2019. No New-Net. In *International MICCAI Brainlesion Workshop (BrainLes)*, Alessandro Crimi, Spyridon Bakas, Hugo Kuijf, Farahani Keyvan, Mauricio Reyes, and Theo van Walsum (Eds.). Springer, Cham, 234–244. https://doi.org/10.1007/978-3-030-11726-9_21

[22] Ali Işın, Cem Direkoğlu, and Melike Şah. 2016. Review of MRI-based Brain Tumor Image Segmentation Using Deep Learning Methods. *Procedia Computer Science* 102 (2016), 317–324. https://doi.org/10.1016/J.PROCS.2016.09.407

[23] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.1412.6980

[24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *International Conference on Computer Vision (ICCV)*. IEEE, 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986

[25] Indrajit Mazumdar and Jayanta Mukherjee. 2022. Fully automatic MRI brain tumor segmentation using efficient spatial attention convolutional networks with composite loss. *Neurocomputing* 500 (2022), 243–254. https://doi.org/10.1016/j.neucom.2022.05.050

[26] Richard McKinley, Michael Rebsamen, Raphael Meier, and Roland Wiest. 2020. Triplanar Ensemble of 3D-to-2D CNNs with Label-Uncertainty for Brain Tumor Segmentation. In *International MICCAI Brainlesion Workshop (BrainLes)*, Alessandro Crimi and Spyridon Bakas (Eds.). Springer, Cham, 379–387. https://doi.org/10.1007/978-3-030-46640-4_36

[27] Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elizabeth Gerstner, Marc-Andre Weber, Tal Arbel, Brian B. Avants, Nicholas Ayache, Patricia Buendia, D. Louis Collins, Nicolas Cordier, Jason J. Corso, Antonio Criminisi, Tilak Das, Herve Delingette, Cagatay Demiralp, Christopher R. Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan M. Iftekharuddin, Raj Jena, Nigel M. John, Ender Konukoglu, Danial Lashkari, Jose Antonio Mariz, Raphael Meier, Sergio Pereira, Doina Precup, Stephen J. Price, Tammy Riklin Raviv, Syed M. S. Reza, Michael Ryan, Duygu Sarikaya, Lawrence Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos A. Silva, Nuno Sousa, Nagesh K. Subbanna, Gabor Szekely, Thomas J. Taylor, Owen M. Thomas, Nicholas J. Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koen Van Leemput. 2015. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging* 34, 10 (2015), 1993–2024. https://doi.org/10.1109/TMI.2014.2377694

[28] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. 2016. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In *International Conference on 3D Vision (3DV)*. IEEE, 565–571. https://doi.org/10.1109/3DV.2016.79

[29] Radu Miron, Ramona Albert, and Mihaela Breaban. 2021. A Two-Stage Atrous Convolution Neural Network for Brain Tumor Segmentation and Survival Prediction. In *International MICCAI Brainlesion Workshop (BrainLes)*, Alessandro Crimi and Spyridon Bakas (Eds.). Springer, Cham, 290–299. https://doi.org/10.1007/978-3-030-72087-2_25

[30] Andriy Myronenko. 2019. 3D MRI Brain Tumor Segmentation Using Autoencoder Regularization. In *International MICCAI Brainlesion Workshop (BrainLes)*, Alessandro Crimi, Spyridon Bakas, Hugo Kuijf, Farahani Keyvan, Mauricio Reyes, and Theo van Walsum (Eds.). Springer, Cham, 311–320. https://doi.org/10.1007/978-3-030-11726-9_28

[31] Hiroko Ohgaki and Paul Kleihues. 2005. Population-Based Studies on Incidence, Survival Rates, and Genetic Alterations in Astrocytic and Oligodendroglial

Gliomas. *Journal of Neuropathology & Experimental Neurology* 64, 6 (2005), 479–489. https://doi.org/10.1093/jnen/64.6.479

[32] Bhavesh Parmar and Mehul Parikh. 2021. Brain Tumor Segmentation and Survival Prediction Using Patch Based Modified 3D U-Net. In *International MICCAI Brainlesion Workshop (BrainLes)*, Alessandro Crimi and Spyridon Bakas (Eds.). Springer, Cham, 398–409. https://doi.org/10.1007/978-3-030-72087-2_35

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Conference on Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 8026–8037. https://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library

[34] Linmin Pei, A K Murat, and Rivka Colen. 2021. Multimodal Brain Tumor Segmentation and Survival Prediction Using a 3D Self-ensemble ResUNet. In *International MICCAI Brainlesion Workshop (BrainLes)*, Alessandro Crimi and Spyridon Bakas (Eds.). Springer, Cham, 367–375. https://doi.org/10.1007/978-3-030-72084-1_33

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (Eds.). Springer, Cham, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

[36] Carlo Russo, Sidong Liu, and Antonio Di Ieva. 2021. Impact of Spherical Coordinates Transformation Pre-processing in Deep Convolution Neural Networks for Brain Tumor Segmentation and Survival Prediction. In *International MICCAI Brainlesion Workshop (BrainLes)*, Alessandro Crimi and Spyridon Bakas (Eds.). Springer, Cham, 295–306. https://doi.org/10.1007/978-3-030-72084-1_27

[37] Laurent Sifre. 2014. Rigid-Motion Scattering for Image Classification. *Ph.D. thesis, École polytechnique, Palaiseau, France* (2014). https://www.di.ens.fr/data/publications/

[38] Joost J.M. van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G.H. Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo J.W.L. Aerts. 2017. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Research* 77, 21 (2017), e104–e107. https://doi.org/10.1158/0008-5472.CAN-17-0339

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Conference on Neural Information Processing Systems (NeurIPS)*, I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett (Eds.). Curran Associates, Inc., 5998–6008. http://papers.nips.cc/paper/7181-attention-is-all-you-need

[40] Feifan Wang, Runzhou Jiang, Liqin Zheng, Chun Meng, and Bharat Biswal. 2020. 3D U-Net Based Brain Tumor Segmentation and Survival Days Prediction. In *International MICCAI Brainlesion Workshop (BrainLes)*, Alessandro Crimi and Spyridon Bakas (Eds.). Springer, Cham, 131–141. https://doi.org/10.1007/978-3-030-46640-4_13

[41] Wenxuan Wang, Chen Chen, Meng Ding, Hong Yu, Sen Zha, and Jiangyun Li. 2021. TransBTS: Multimodal Brain Tumor Segmentation Using Transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Marleen de Bruijne, Philippe C Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert (Eds.). Springer, Cham, 109–119. https://doi.org/10.1007/978-3-030-87193-2_11

[42] Zirui Wang and Yi Hong. 2023. A2FSeg: Adaptive Multi-modal Fusion Network for Medical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor (Eds.). Springer, Cham, 673–681. https://doi.org/10.1007/978-3-031-43901-8_64

[43] Yuxin Wu and Kaiming He. 2018. Group Normalization. In *European Conference on Computer Vision (ECCV)*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Springer, Cham, 3–19. https://doi.org/10.1007/978-3-030-01261-8_1

[44] Chenhong Zhou, Changxing Ding, Xinchao Wang, Zhentai Lu, and Dacheng Tao. 2020. One-Pass Multi-Task Networks with Cross-Task Guided Attention for Brain Tumor Segmentation. *IEEE Transactions on Image Processing* 29 (2020), 4516–4529. https://doi.org/10.1109/TIP.2020.2973510

[45] Alex Zwanenburg, Stefan Leger, Martin Vallières, and Steffen Löck. 2016. Image biomarker standardisation initiative. *arXiv:1612.07003* (2016). https://doi.org/10.48550/arXiv.1612.07003