Learning Multi-Modal Brain Tumor Segmentation from Privileged Semi-Paired MRI Images with Curriculum Disentanglement Learning

Zecheng Liu[a], Jia Wei[a,*], Rui Li[b] and Jianlong Zhou[c]

[a]South China University of Technology, Guangzhou, China
[b]Golisano College of Computing and Information Sciences, Rochester Institute of Technology, Rochester, NY, USA
[c]Data Science Institute, University of Technology Sydney, Ultimo, NSW 2007, Australia

ARTICLE INFO

Keywords:
Brain tumor segmentation
Privileged semi-paired images
Curriculum disentanglement learning

ABSTRACT

Since the brain is the human body's primary command and control center, brain cancer is one of the most dangerous cancers. Automatic segmentation of brain tumors from multi-modal images is important in diagnosis and treatment. Due to the difficulties in obtaining multi-modal paired images in clinical practice, recent studies segment brain tumors solely relying on unpaired images and discarding the available paired images. Although these models solve the dependence on paired images, they cannot fully exploit the complementary information from different modalities, resulting in low unimodal segmentation accuracy. Hence, this work studies the unimodal segmentation with privileged semi-paired images, i.e., limited paired images are introduced to the training phase. Specifically, we present a novel two-step (intra-modality and inter-modality) curriculum disentanglement learning framework. The modality-specific style codes describe the attenuation of tissue features and image contrast, and modality-invariant content codes contain anatomical and functional information extracted from the input images. Besides, we address the problem of unthorough decoupling by introducing constraints on the style and content spaces. Experiments on the BraTS2020 dataset highlight that our model outperforms the competing models on unimodal segmentation, achieving average dice scores of 82.91%, 72.62%, and 54.80% for WT (the whole tumor), TC (the tumor core), and ET (the enhancing tumor), respectively. Finally, we further evaluate our model's variable multi-modal brain tumor segmentation performance by introducing a fusion block (TFusion). The experimental results reveal that our model achieves the best WT segmentation performance for all 15 possible modality combinations with 87.31% average accuracy. In summary, we propose a curriculum disentanglement learning framework for unimodal segmentation with privileged semi-paired images. Moreover, the benefits of the improved unimodal segmentation extend to variable multi-modal segmentation, demonstrating that improving the unimodal segmentation performance is significant for brain tumor segmentation with missing modalities. Our code is available at https://github.com/scut-cszcl/SpBTS.

1. Introduction

Biomedical technology is crucial to human health and life. Extensive research and application using Deep Learning (DL) in the biomedical domain have significantly improved big medical data analysis, disease diagnosis, and prognostic programs, such as Alzheimer's Disease (AD) [20], Coronavirus (Covid-19) [4, 5], and various tumors [15]. The brain, the most complex human organ, is the primary command and control center. Brain tumor incidence is an important contributor to global mortality. According to the National Brain Tumor Foundation (NBTF) report, in the USA, 29,000 people were diagnosed with primary intracranial tumors, of which 13,000 died [31]. In addition, one in four childhood cancer deaths is caused by brain tumors.

Automatic and accurate segmentation of brain tumors is essential for diagnosis, treatment planning, and follow-up evaluations. [19, 27]. Such a segmentation process requires precisely detecting a tumor's location and extent. However, the tumorous shape, size, and location uncertainty pose a unique challenge, especially in infiltrative tumors like gliomas [10, 32]. A common solution is integrating information acquired from multi-modal paired MRI since different MRI pulse sequences (modalities) provide complementary information on brain tumors from multiple perspectives [37, 40]. As illustrated in Figure 1, T1ce (contrast-enhanced T1-weighted) highlights tumors without peritumoral, but the image contrast of the whole peritumoral edema is enhanced in T2 (native T2-weighted), and Flair (T2 Fluid Attenuated Inversion Recovery) [33, 43]. Although these methods demonstrate a promising performance, they require paired data in training and testing (Figure 2(a)). This hinders their applicability in clinical practice when only unpaired or missing modality images are available.

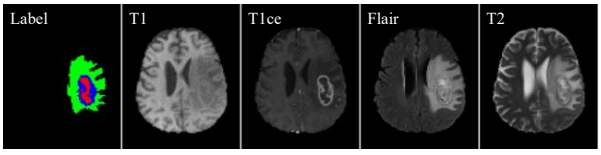

Figure 1: Different brain tumor information of the same subject can be detected from different sequences of brain MRI. In the Label image, the green area indicates the whole peritumoral edema, red and blue area indicate the necrotic tumor core and the enhancing tumor, respectively.

*Corresponding author
scutzcliu@gmail.com (Z. Liu); csjwei@scut.edu.cn (J. Wei); rxlics@rit.edu (R. Li); jianlong.zhou@uts.edu.au (J. Zhou)

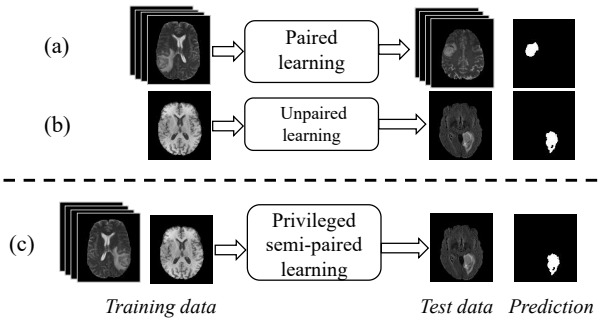

Figure 2: Illustration of (a) Paired learning, (b) Unpaired learning, and (c) Privileged semi-paired learning frameworks.

Spurred by the abovementioned problems, a multi-modal unpaired learning method is proposed for medical image segmentation [34, 39] (Figure 2(b)). For instance, Yuan *et al*. [39] propose a two-stream translation and segmentation unified attentional generative adversarial network. The model is trained with unpaired data and performs predictions on unpaired images by capturing and calibrating complementary information from translation to improve segmentation. However, the image translation quality is poor without supervising the paired images, especially for brain tumor areas. This is because the method cannot effectively exploit the complementary information from different modalities, such as varying shapes of brain tumors. Therefore, such a strategy leads to unsatisfactory unimodal segmentation performance.

In this work, we propose a privileged semi-paired learning framework, with Figure 2(c) revealing that limited paired images are introduced in training data. Unlike unpaired learning methods, we exploit the complementary information from paired images to improve unimodal segmentation performance. Specifically, we extract modality-specific style codes and modality-invariant content codes from the input images with a multi-task disentanglement model. For a complete decoupling, we propose a two-step curriculum disentanglement learning strategy that adds constraints on the content and style spaces. Finally, we extend our model's application for variable multi-modal brain tumor segmentation through a designed fusion block.

The contributions of this paper are as follows:

- We propose a privileged semi-paired learning framework for brain tumor segmentation. Introducing limited paired images enhances our model's ability to capture and exploit complementary information between the modalities.
- We propose a two-step (intra-modality and inter-modality) curriculum disentanglement learning strategy to effectively separate the input images' style and content.
- We qualitatively and quantitatively evaluate our method on brain tumor segmentation tasks on the BraTS2020 [2] and BraTS2018 [3] datasets. The results demonstrate our method's superiority over current state-of-the-art unpaired medical image segmentation methods.
- We further demonstrate the superior performance of our model on variable multi-modal brain tumor segmentation, demonstrating that unimodal segmentation performance is significant for brain tumor segmentation with missing modalities.

2. Related work

The methods for multi-modal brain tumor segmentation can be broadly separated into two categories: segmentation through paired learning and segmentation through unpaired learning. Table 1 shows a comparison overview between these works, including a summary of strengths and weaknesses.

2.1. Segmentation through paired learning

Multiple imaging modalities have been widely used in medical image segmentation due to its ability to provide complementary information to reduce information uncertainty. During the past few years, most researches focused on the multi-modal fusion strategies, such as input-level fusion and layer-level fusion. These methods either concatenate multi-modality images as multi-channel inputs [18, 36, 41] or fuse the features from different networks trained by different modalities [12, 9]. The improvement in the accuracy of brain tumor segmentation relies on the exploitation of complementary information. However, these methods rely on paired data in both training and test, and it hinders their applicability in clinical practice, where only unpaired or missing modality images are available.

Recently, to mitigate performance degradation when inferencing, medical image segmentation with missing modality has been extensively studied [1]. The most popular approach is to fuse the available modalities in a latent space to learn a shared feature representation for segmentation. A variable number of input modalities are mapped to a unified representation by computing the first and second moments [14], mean function [23], or fusion block [7, 42]. Moreover, Shen *et al*. [30] utilize synthesizd images as multi-channel inputs to obtain shared representation for segmentation with a multi-modal image completion and segmentation disentanglement network called ReMIC. Furthermore, Chen *et al*. [8] propose a privileged knowledge learning framework with the "Teacher Student" architecture. Privileged information is transferred from a multi-modal teacher network to a unimodal student network for unpair images. However, the method also requires a large amount of paired images for multi-modal teacher network training. Our method, instead, utilizes privileged semi-paired images, where only limited paired images are available for training.

2.2. Segmentation through unpaired learning

For medical image segmentation, in order to utilize all available data for training even when the images are

Table 1
A comparative overview between different networks.

Strategy	Networks	Strength	Weakness
Segmentation through paired learning	"Cascaded Anisotropic Convolutional Neural Network" [36] OM-Net [41] HyperDenseNet [12] "Modality-Pairing Network" [37] "Cross-modality deep feature learning for brain tumor segmentation" [40] nnU-Net [17]	Fully exploit the complementary information provided by different modalities.	Relye on paired images in both training and test, which aggravates the problem of data scarcity and leads to limited application scenarios.
Segmentation with missing modality	HeMIS [14] "Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion" [7] "Latent correlation representation learning for brain tumor segmentation with missing mri modalities" [42] ReMIC [30]	Wide applicability. Having the ability to deal with any combinatorial subset of available modalities with a unified model.	Utilize paired images in training. Uniform training on all missing scenarios indiscriminately makes it hard to learn the most difficult unimodal segmentation.
Segmentation through unpaired learning	"Multi-modal learning from unpaired images" [34] UAGAN [39]	Reduce data usage requirements, which alleviates data scarcity. Improve unimodal segmentation performance. It is critical for expanding to missing modality scenarios.	Use only unpaired data for training. Correlations between different modalities cannot be learned directly.

unpaired, an X-shaped multiple encoder-decoder network is proposed [34]. The model extracts modality-independent features to improve segmentation accuracy by sharing the last layers of the encoders. Information from one modality is captured in the shared network to improve the performance of segmentation task on another modality. Furthermore, Yuan *et al*. [39] propose a two-stream translation and segmentation network called UAGAN. The network captures inferred complementary information from modality translation task to improve segmentation performance. The above methods do not require any paired images, and utilize easily accessible unpaired images for training, instead. However, these methods cannot integrate complementary information without paired images. On the contrary, our method can effectively leverage complenentary information of limited number of paired images by encoding them into a modality-invariant content space through content consistency constraint and supervised translation for brain tumor segmentation.

3. Methodology

3.1. Proposed model

The suggested model considers a multi-task disentanglement framework that effectively extracts modality-invariant content codes for brain tumor segmentation by fully exploiting multi-modal complementary information from privileged semi-paired images. Content and style codes are decoupled for the unpaired images through image reconstruction and modality translation task learning. Considering limited paired images containing complementary information, the model's ability to learn multi-modal correlations is enhanced by converting the modality translation task from unsupervised to supervised and applying content consistency constraints.

As shown in Figure 3, we use paired images as an example to illustrate our framework. Given images x_a, x_b from the same subject and different modalities. We adopt one-hot vectors to represent their modality label and expands them to the same image size, denoted as m_a and m_b. Given the depth-wise concatenation (x_a, m_a) and (x_b, m_b), our goal is to trian a single generator G that can simultaneously accomplish the following tasks: (1) Reconstructing the input images x_a and x_b as $x_{a\to a}$ and $x_{b\to b}$, respectively. (2) Translating x_a of modality m_a to the corresponding output image $x_{a\to b}$ of modality m_b, and x_b of modality m_b to $x_{b\to a}$ of modality m_a. (3) Generating brain tumor segmentation masks x_a^{seg} and x_b^{seg} of the input images x_a and x_b, respectively. We denote it as $G((x_a, m_a), (x_b, m_b))\to(x_{a\to a}, x_{a\to b}, x_a^{seg}, x_{b\to b}, x_{b\to a}, x_b^{seg})$. The architecture of our model is composed of two modules described below.

We design the generator G with four shared networks (E_s,E_c,D_s,D_t) based on feature disentanglement. Given the input concatenation (x_a, m_a), the network E_s generates its style code s_{x_a} which is a vector with dimension n_s, and the network E_c generates its content code c_{x_a} which is a feature map, denoted as $E_s((x_a, m_a))\to s_{x_a}$ and $E_c((x_a, m_a))\to c_{x_a}$. Similarly, s_{x_b} and c_{x_b} are also obtained from the input (x_b, m_b), denoted as $E_s((x_b, m_b))\to s_{x_b}$ and $E_c((x_b, m_b))\to c_{x_b}$. Then, we perform image reconstruction, translation, and segmentation based on these disentangled

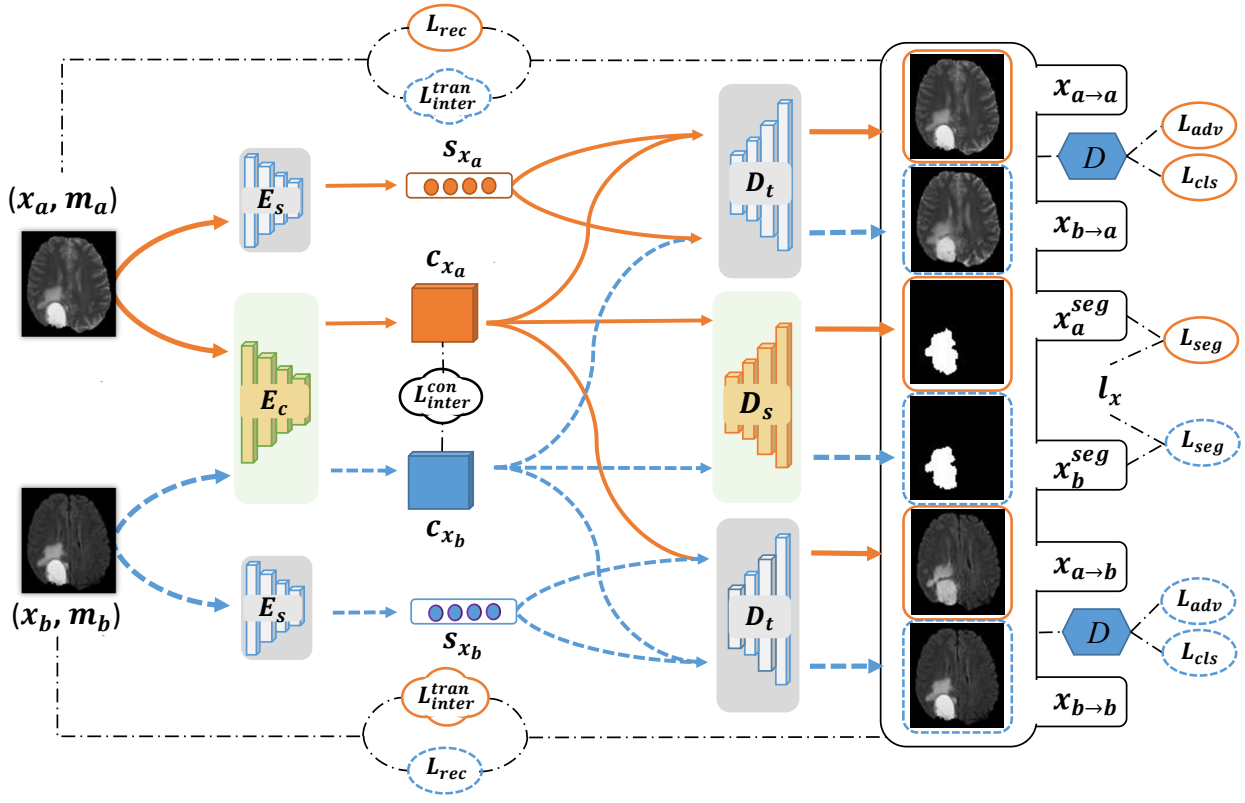

Figure 3: Illustration of the proposed framework. Paired images for the inter-modality learning scheme are depicted in this example. All the networks are unified, including the content and style encoders (E_c, E_s), the translation and segmentation decoders (D_t, D_s), and the discriminator D. The data stream of images x_a and x_b are drawn as solid orange arrows and dotted blue arrows, respectively. Losses are computed by the corresponding generated images and segmentation maps (orange solid box for x_a, blue dotted box for x_b). Note that image modality translation loss L_{inter}^{tran} and content consistency loss L_{inter}^{con} are only applied to paired images in the inter-modality learning scheme.

representations. For image reconstruction, given the content code and the style code obtained from the same input image, the decoder D_t generates the corresponding reconstruction image, denoted as $D_t(c_{x_a}, s_{x_a})\to x_{a\to a}$ and $D_t(c_{x_b}, s_{x_b})\to x_{b\to b}$. For image translation, given the content code and the style code obtained from different input images, the decoder D_t translates the source image of one modality (corresponding to the content code) to the target image of the other modality (corresponding to the style code). We denoted it as $D_t(c_{x_a}, s_{x_b})\to x_{a\to b}$ and $D_t(c_{x_b}, s_{x_a})\to x_{b\to a}$. For image segmentation, given the content code, the decoder D_s generates a binary mask to identify and highlight the tumor area of the corresponding input image, denoted as $D_s(c_{x_a})\to x_a^{seg}$ and $D_s(c_{x_b})\to x_b^{seg}$.

The probability distributions produced by the discriminator D distinguish whether the generated images from G are real or fake, and determine which modality they are from.

3.2. Curriculum Disentanglement Learning

We propose a novel two-step curriculum disentanglement learning method to leverage privileged semi-paired images for brain tumor segmentation, as shown in Figure 2(c), when limited paired images are only available in training. Compared to previous feature disentanglement learning models (DRIT [24], MUNIT [16]), the proposed model focuses on effective separation of style and content. As shown in Figure 4(c), the previous models only use unpaired inter-modality learning scheme for training. Unpaired images x_a (from subject x of modality a) and y_b (from subject y of modality b) are mapped into the same content space but different style spaces. However, there are no specific constraints for these spaces, so the disentanglement mapping tends to incur large variations, which results in the problem of ambiguous separation between style and content. Ouyang et al. [26] investigate the problem. They require that the content representations from the same patient with different modalities should be as similar as possible. However, under the constraint, the style and other task-unrelated components (e.g., noise and artifacts) tend to corrupt the content representations, which fails to reduce the ambiguity of the content and style. On the contrary, we contend that different modalities of a given patient essentially reflect the inherent anatomy of the patient, which is consistent even though its appearance may be diverse across different modalities. Therefore, to solve this problem, we propose a curriculum disentanglement learning strategy with two steps:

(1) In the first step, as shown in Figure 4(a), we generate two style-consistent images x_a^1 and x_a^2 from the same original

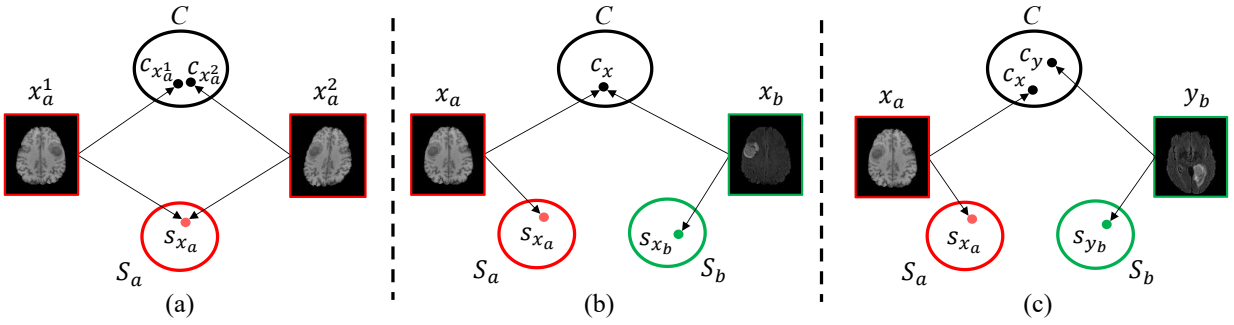

Figure 4: Different feature disentanglement learning schemes. (a) Style-consistent learning scheme. Style-consistent images obtained by horizontal flip (x_a^1) and elastic deformation (x_a^2) are given in this example. (b) Paired inter-modality learning scheme. (c) Unpaired inter-modality learning scheme.

image x_a with different image processing methods (such as horizontal flip and elastic deformation). We then define a style consistency loss to map the images to the same point s_{x_a} in the style space S_a.

(2) The second step consists of reconstruction, unsupervised/supervised translation, and segmentation based on unpaired and paired inter-modality learning schemes. The unpaired inter-modality learning scheme, as shown in Figure 4(c), maps unpaired images x_a and y_b obtained from different subjects to different points in the content space C. The paired inter-modality learning scheme, as shown in Figure 4(b), maps paired images x_a and x_b obtained from the same subject x to the same point c_x in the content space C with a content consistency loss.

Through the two steps, our proposed method can separate modality-specific style codes and modality-invariant content codes from the input images. In particular, the modality-specific style codes describe attenuation of tissue features and image contrast, and the modality-invariant content codes contain consistent inherent anatomical and functional information. The effective disentanglement of the two codes is critical for brain tumor segmentation.

Algorithm 1 The curriculum disentanglement learning

Training input: intra-modality augmented style-consistent images $(x_a^1, x_a^2)_1, \ldots, (x_a^1, x_a^2)_i$, paired inter-modality images $(x_a, x_b)_1, \ldots, (x_a, x_b)_j$, and unpaired inter-modality images $(x_a, y_b)_1, \ldots, (x_a, y_b)_k$

Training output: generator G

while not converged **do** // In the first step
    // Style-consistent pattern
    Updata G and D using Eq. (9) and Eq. (6)
end while
while not converged **do** // In the second step
    // Paired inter-modality pattern
    Updata G and D using Eq. (12) and Eq. (6)
    // Unpaired inter-modality pattern
    Updata G and D using Eq. (7) and Eq. (6)
end while

Test input: unpaired images $x_1, \ldots, x_n$

Test output: segmentation results $x_1^{seg}, \ldots, x_n^{seg}$

Calculate $\forall r$, $x_r^{seg} \leftarrow G((x_r, m_r))$
// m_r is the modality label of x_r

3.3. Constructing the objective function

Our loss function consists of three parts: (1) common losses for both curriculum disentanglement learning steps; (2) losses for intra-modality disentanglement step; (3) losses for inter-modality disentanglement step. For simplicity, we only describe the losses for image x_a, since the loss function for x_b is the same. Alg. 1 summarizes the overall procedure of the curriculum disentanglement learning.

3.3.1. Common losses

Losses for both curriculum disentanglement learning steps include an adversarial loss, a modality classification loss, a reconstruction loss, and a segmentation loss.

Adversarial loss: To minimize the difference between the distributions of generated images and real images, we define the adversarial loss as:

$$L_{adv} = \mathbb{E}_{x_a}[\log D_{src}(x_a)] + \frac{1}{2}\mathbb{E}_{x_{a\to a}}[\log(1 - D_{src}(x_{a\to a}))] + \frac{1}{2}\mathbb{E}_{x_{a\to b}}[\log(1 - D_{src}(x_{a\to b}))] \tag{1}$$

where D_{src} denotes probability distributions, given by discriminator D, of real or fake images [11]. The discriminator D maximizes this objective to distinguish between real and fake images, while the generator G tries to generate more realistic images to fool the discriminator.

Modality classification loss: To allocate the generated image to correct modality, the modality classification loss is imposed to G and D. It contains two terms: modality classification loss of real images which is used to optimize D, denoted as L_{cls}^r, and the loss of fake images used to

optimize G, denoted as L_{cls}^{f}.

$$L_{cls}^{r} = \mathbb{E}_{x_a}[-\log D_{cls}(m_a|x_a)] \tag{2}$$

$$\begin{aligned} L_{cls}^{f} = &\frac{1}{2}\mathbb{E}_{x_{a\to a}}[-\log(D_{cls}(m_a|x_{a\to a}))] \\ &+\frac{1}{2}\mathbb{E}_{x_{a\to b}}[-\log(D_{cls}(m_b|x_{a\to b}))] \end{aligned} \tag{3}$$

where D_{cls} represents the probability distributions over modality labels and input images [11].

Reconstruction loss: To prevent the omission of detailed information, we employ the reconstruction loss to constrain the recovered images:

$$L_{rec} = \mathbb{E}_{x_a, x_{a\to a}}[\|x_{a\to a} - x_a\|_1] \tag{4}$$

Segmentation loss: The segmentation loss is a dice loss:

$$L_{seg} = -\frac{2\sum_{i=1}^{N} l_x^i x_{seg}^i}{\sum_{i=1}^{N}(l_x^i l_x^i + x_{seg}^i x_{seg}^i) + \epsilon} \tag{5}$$

Here, l_x^i, x_{seg}^i denote ground truth and prediction of voxel i, respectively. The $\epsilon = 1e^{-7}$ is a constant for numerical stability.

Objective functions: By combining the above losses together, our common objective functions are as follows:

$$L_D = -L_{adv} + L_{cls}^{r} \tag{6}$$

$$L_G = L_{adv} + L_{cls}^{f} + \lambda_{rec} L_{rec} + \lambda_{seg} L_{seg} \tag{7}$$

where λ_{rec} and λ_{seg} are hyperparameters to control the relative importance of reconstruction loss and segmentation loss. In addition, we use L2 norm regularization to constrain the style codes to encourage a smooth space and minimise the encoded information [22].

3.3.2. Curriculum losses for the first step

In the first intra-modality disentanglement step, we train the model with style-consistent learning scheme shown in Figure 4(a). The augmented images are created by the following six image processing methods: (1) horizontal flip, (2) vertical flip, (3) rotate random angle (0°∼360°), (4) zoom in to random ratios (0.8∼1.2), (5) elastic deformation [29], and (6) shift a random distance (0px∼20px) in all directions. For each original image x_a in training data (both paired and unpaired image), we randomly use two methods to obtain two style-consistent images denoted as x_a^1 and x_a^2. Note that original image can belong to any modality. Let a indexes a modality, and $s_{x_a^1}$ and $s_{x_a^2}$ denote the style codes of x_a^1 and x_a^2, respectively. We define a style consistency loss to constrain $s_{x_a^1}$ and $s_{x_a^2}$ to be similar:

$$L_{intra}^{sty} = \mathbb{E}_{(s_{x_a^1}, s_{x_a^2})}[\|s_{x_a^1} - s_{x_a^2}\|_1] \tag{8}$$

In the intra-modality step, the objective function to optimize D is as in Eq. (6), while the objective function to optimize G is defined as:

$$L_{G_{intra}} = L_G + \lambda_{sty} L_{intra}^{sty} \tag{9}$$

Here, L_G is as in Eq. (7) and the λ_{sty} is the hyperparameter to control the contribution of L_{intra}^{sty}.

3.3.3. Curriculum losses for the second step

In the second inter-modality disentanglement step, the training data include both paired and unpaired images from different modalities.

The objective function for D in Eq. (6) and the objective function for G in Eq. (7) generate different content codes and different style codes for unpaired image (x_a, y_b).

For paired images (x_a, x_b), they have the same content codes and different style codes, so we construct a content consistency loss to constrain their content codes:

$$L_{inter}^{con} = \mathbb{E}_{(c_{x_a}, c_{x_b})}[\|c_{x_a} - c_{x_b}\|_1] \tag{10}$$

In addition, the image $x_{a\to b}$ generated from the translation task $D_t(c_{x_a}, s_{x_b})$ is expected to be consistent with the image $x_{b\to b}$ generated from the reconstruction task $D_t(c_{x_b}, s_{x_b})$, since that x_a and x_b are paired. Meanwhile, $x_{b\to b}$ is the reconstrcuted image of x_b, $x_{a\to b}$ is expected to be consistent with x_b. Thus, we introduce a translation loss to further constrain c_{x_a} and c_{x_b} as:

$$L_{inter}^{tran} = \mathbb{E}_{x_{a\to b}, x_b}[\|x_{a\to b} - x_b\|_1] \tag{11}$$

Therefore, the objective function to optimize D is the same as in Eq. (6), while the objective function to optimize G is defined as:

$$L_{G_{inter}} = L_G + \lambda_{con} L_{inter}^{con} + \lambda_{tran} L_{inter}^{tran} \tag{12}$$

where, L_G is as in Eq. (7) and the λ_{con} and λ_{tran} are hyperparameters to control the contributions of L_{inter}^{con} and L_{inter}^{tran}, respectively.

4. Experiments and results

In this section, we first introduce the experimental settings, including datasets, baseline methods, evaluaion metrics, and implementation details. Then, we present and discuss quantitative and qualitative results of our method, including brain tumor segmentation, image translation, ablation study, influence of paired subjects, and disentanglement evaluation.

4.1. Experimental settings

4.1.1. Datasets

To validate the proposed model, we conduct experiments on two widely used benchmark datasets of BraTS2020 [2] and BraTS2018 [3] that consist of 369 and 285 subjects, respectively. Each subject consists of one segmentation mask

Table 2
Performance evaluation for the segmentation task of WT, TC and ET on BraTS2020. A better method has higher Dice (Best highlighted in bold).

Metric		Dice(%)↑				
Modality		T1ce	T1	T2	Flair	Aver
WT	nnU-Net [17]	78.36±2.23	74.52±0.25	84.18±1.39	88.22±0.29	81.31±0.21
	UAGAN [39]	75.56±1.36	75.05±2.40	82.62±0.60	84.53±0.23	79.44±0.06
	ReMIC [30]	72.18±1.05	74.63±0.18	76.96±0.04	75.37±3.59	74.78±0.67
	Ours	**79.58±1.13**	**77.80±2.16**	**85.66±0.35**	**88.58±0.08**	**82.91±0.36**
TC	nnU-Net [17]	84.78±1.88	53.35±1.33	66.59±2.70	66.63±0.97	67.84±0.78
	UAGAN [39]	80.76±0.65	58.53±0.67	67.99±0.65	70.13±0.90	69.35±0.71
	ReMIC [30]	80.68±3.17	53.86±4.72	62.19±3.16	55.14±0.86	62.96±2.98
	Ours	**85.37±2.05**	**62.18±0.71**	**71.68±1.44**	**71.27±2.80**	**72.62±0.37**
ET	nnU-Net [17]	82.09±1.07	28.17±0.30	45.08±2.17	40.03±5.30	48.84±2.06
	UAGAN [39]	75.30±3.01	33.64±2.74	47.19±3.56	46.24±5.53	50.64±3.63
	ReMIC [30]	74.98±0.33	32.16±2.26	39.22±3.44	34.68±0.16	45.29±1.34
	Ours	**82.33±1.04**	**37.47±0.80**	**51.61±3.41**	**47.78±6.36**	**54.80±2.50**

Table 3
The dataset distribution.

Dataset	Training (paired+unpaired)	Validation	Test	All
BraTS2020	240 (40+200)	60	69	369
BraTS2018	180 (32+148)	50	55	285

and four modality scans: T1, T1ce, T2, Flair. The segmentation mask contains four labels, namely NCR (label 1: the necrotic tumor core), ED (label 2: the peritumoral edematous/invaded tissue), NET (label 3: the non-enhancing tumor core), and ET (label 4: the enhancing tumor). To better represent the clinical application tasks, different structures have been grouped into three mutually inclusive tumor regions: **ET**: the enhancing tumor, **TC** (Union of labels 1, 3 and 4): the tumor core, and **WT** (Union of all labels): the whole tumor. In BraTS2020, we utilize 240 subjects as semi-paired training data, 60 subjects as unpaired validation data, and 69 subjects as unpaired test data. For semi-paired training data, we use 40 of 240 subjects as paired data, while the rest as unpaired data. In BraTS2018, we utilize 180 subjects as semi-paired training data, 50 subjects as unpaired validation data and 55 subjects as unpaired test data. For semi-paired training data, we use 32 of 180 subjects as paired data, while the rest as unpaired data. The specific number of subjects in the training set, validation set, and test set are shown in Table 3. The subjects are evenly divided between four modalities. For all images, we resize them to 128×128 uniformly.

4.1.2. Baseline methods

Segmentation results are evaluated by comparing with the following methods: (1) nnU-Net [17], which achieves the best performance in the BraTS2020 competition. Since only limited paired images are available in the training data and all the test data are unpaired, we implement it as unpaired learning (Figure 2(b)). The model is trained and tested on four mixed modalities where images are unpaired. (2) UAGAN [39], a recently proposed unpaired brain tumor segmentation model, is a two-stream trainslation and segmentation network. Inferred complementary information are captured in the modality translation task to improve segmentation performance. Since the model is unpaired, we take semi-paired training data as unpaired data for training. (3) ReMIC [30], a recently proposed image completion and segmentation model for random missing modalities, first achieve image completion, and then concatenate the synthesized modalities as multi-channel inputs to obtain shared representation for segmentation. Table 5 presents parameters of different models. Our model is based on disentanglement framework, so the number of parameters (173.97M) is larger than others. Please note that, in the segmentation test, we only need to load part of our network (the content encoder

Table 4
Performance evaluation of WT segmentation on BraTS2018.

Metric	Dice(%)↑				
Modality	T1ce	T1	T2	Flair	Aver
nnU-Net [17]	84.10	76.12	**85.91**	86.16	83.18
UAGAN [39]	78.14	75.61	80.86	80.89	78.95
ReMIC [30]	77.97	72.69	76.75	72.23	74.89
Ours	**86.56**	**82.65**	84.93	**87.86**	**85.51**

Table 5
The parameters of different model for tumor segmentation.

nnU-Net	UAGAN	ReMIC	Ours	Ours-test
18.67M	44.73M	89.43M	173.97M	31.03M

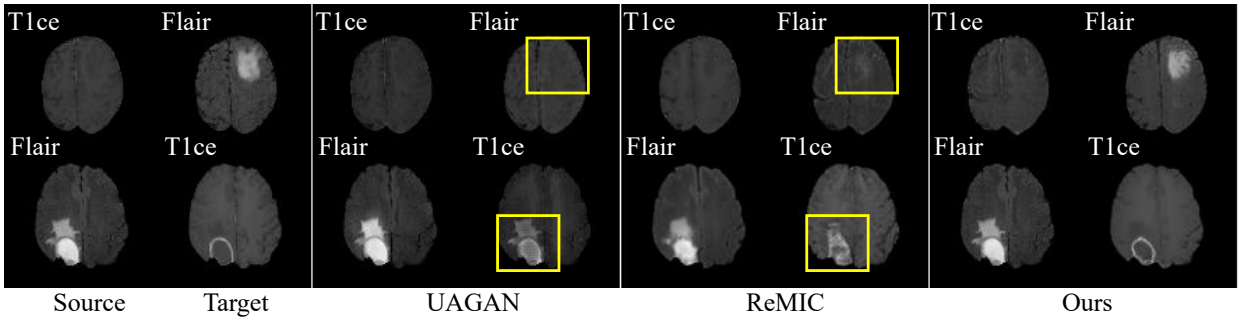

Figure 5: Image translation results (from source to target) between T1ce and Flair. Yellow boxes highlight the failed translation of brain tumors. In each image, reconstructed images are in the left column, and translated images are in the right column.

Table 6
Quantitative evaluations on translated images.

Metric	SSIM↑				
Modality	T1ce	T1	T2	Flair	Aver
UAGAN [39]	0.5153	0.4193	0.3850	0.4328	0.4382
ReMIC [30]	0.7205	**0.7748**	0.7579	0.7061	0.7398
Ours	**0.7741**	0.7701	**0.7905**	**0.7671**	**0.7754**

E_c and segmentation decoder D_s), and its parameter amount is 31.03M.

4.1.3. Evaluation metrics

We evaluate segmentation performance with dice score (Dice). We compute the metric on each modality, and report average values. In the translation tasks, we use structural similarity (SSIM) as an evaluation metric.

4.1.4. Implementation details

The content encoder E_c and the segmentation decoder D_s in the segmentation generator is similar to the U-Net [29]. The style encoder E_s and image generation decoder D_t are adapted from [24]. In our experiments, we set $\lambda_{rec} = 50$, $\lambda_{tran} = 100$, $\lambda_{con} = 10$, $\lambda_{sty} = 10$, and $\lambda_{seg} = 100$. The batch size and training epoch are 8 and 50 respectively. We train the model with 20 epochs in the first step and 30 epochs in the second step, which leads to convergence in practice. The dimensionality of the style code is set to $n_s = 8$. All models are optimized with Adam [21], and the initial learning rates are $1e^{-4}$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The learning rate is fixed in the first 40 epochs, and then linearly declines to $1e^{-6}$. All images are normalized to $[-1, 1]$ prior to the training and testing. Our implementation is on an NVIDIA RTX 3090 (24G) with PyTorch 1.8.1.

4.2. Results and analyses

4.2.1. Brain tumor segmentation

We first evaluate the brain tumor segmentation performance of our model on BraTS2020 segmentation tasks (WT: the while tumor, TC: the tumor core, and ET: the enhancing tumor). Quantitative results are shown in Table 2. Our model achieves the best overall performance and outperforms the others in all cases. Compared with the best performer of the state-of-the-art methods, our method improves the average dice score from 81.31% to 82.91%, 69.35% to 72.62% and 50.64% to 54.80% on WT, TC and ET segmentation tasks, respectively. In addition, we further evaluate our model on BraTS2018 WT segmentation task. The dice score results are shown in Table 4, our method achieves superior performance in most cases, and improves the average dice score from 83.18% to 85.51% compared with nnU-Net. The segmentation results indicate that our method can effectively exploit complementary information by leveraging privileged semi-paired images through the curriculum disentanglement learning model.

4.2.2. Image translation

Since the Flair is the most informative modality for the segmentation of WT, and T1ce is for the segmentation of TC and ET [14, 42], we discuss the results of image translation between these two modalities in Figure 5. Note that the translation performances are only evaluated against UAGAN and ReMIC, since there is no translation for nnU-Net. For the translation between modalities T1ce and Flair that convey different biological information (Figure 1, for example) of brain tumors, our model is superior to others, particularly for the tumor areas translation. Furthermore, as shown in Table 6, our model outperforms other methods on translation task in terms of SSIM, which suggests that our model can produce more realistic images, and more effectively exploit accurate complementary information to improve segmentation.

4.2.3. Ablation study

In this section, we assess the contribution of different components in WT segmentation on BraTS2020. We denote the **C**ontent **c**onsistency loss (Eq. (10)), the inter-modality **T**ranslation loss (Eq. (11)), **F**eature disentanglement framework and **C**urriculum **d**isentanglement learning as Cc, T, F and Cd, respectively. As shown in Table 7, we describe the ablation experiments as follows: (1) **Ours w/o Cc** denotes that our model is trained without the content consistency loss. (2) **Ours w/o T** denotes that our model is trained without the inter-modality translation loss. (3) **Ours w/o F** denotes that the style encoder E_s and image generation decoder D_t are deactivated. In this experment, our model

Table 7
Performance evaluation of the WT segmentation for ablation study on components. w/o means without.

Metric	Dice(%)↑				
Modality	T1ce	T1	T2	Flair	Aver
Ours w/o Cc, T, Cd, F [1]	76.09	72.53	80.04	84.78	78.36
Ours w/o Cc, T, Cd	75.78	74.35	84.04	87.32	80.38
Ours w/o Cc, T	77.88	76.68	85.06	87.97	81.90
Ours w/o Cd	77.25	74.69	84.46	87.29	80.93
Ours w/o T	78.50	77.66	85.16	87.95	82.32
Ours w/o Cc	78.76	77.81	85.21	88.15	82.47
Ours end-to-end	78.27	78.44	84.62	87.85	82.30
Ours	79.58	77.80	85.66	88.58	82.91

[1] Cc, T, F and Cd denote the **C**ontent **c**onsistency loss (Eq. (10)), the inter-modality **T**ranslation loss (Eq. (11)), **F**eature disentanglement framework and **C**urriculum **d**isentanglement learning, respectively.

is trained without feature disentanglement. Since Cc, T and Cd are based on feature disentanglement framework, these components cannot be applied in Ours w/o F, and we denote it as **Ours w/o Cc, T, Cd, F** in Table 7. (4) **Ours w/o Cd** denotes that we train our model only at inter-modality step for all 50 epochs. Table 7 shows that the best results are achieved with all components. The performance is significantly improved from 78.36% (**Ours w/o Cc, T, Cd, F**) to 82.91%. Compared to the Dice scores of 78.36% (**Ours w/o Cc, T, Cd, F**) and 80.38% (**Ours w/o Cc, T, Cd**), the performance improvement is due to using **F** reduces the disturbance of modality-specific information. By utilizing **Cd**, the segmentation accuracy increased from 80.38% (**Ours w/o Cc, T, Cd**) to 81.90% (**Ours w/o Cc, T**), which shows the importance of thorough decoupling. The performance degradation, from 82.91% (**Ours**) to 80.93% (**Ours w/o Cd**), can also reflect this. Compared to the Dice scores of 82.91% (**Ours**), 82.32% (**Ours w/o T**), 82.47% (**Ours w/o Cc**), and 81.90% (**Ours w/o Cc, T**), introducing **T** and **Cc**, which is work only for paired images, benifits the model to fully exploit information between modalities. In addition, the model can also be trained with both style-consistent learning and inter-modality learning for 50 epochs in an end-to-end manner (**Ours end-to-end**). Compared with the two-step training scheme, the average Dice score dropped from 82.91% to 82.30%.

4.2.4. Influence of paired subjects

We conduct a ratio test to investigate the effect of the paired subjects. We keep the number of training subjects fixed, and assign different numbers of subjects as paired data. As shown in Figure 6, introducing paired data in training does improve the performance of our model. Note that our model is still better than the state-of-the-art methods when the Number of Paired Subjects (NPS) equals to 0, and can get satisfactory results when NPS is relatively small (40, for example), which is a good news for clinical practice.

4.2.5. Disentanglement evaluation

We qualitatively examine the effect of each dimension of style code s with latent space arithmetics [6] on 10 subjects. We set the style code size to $n_s = 8$ as suggested

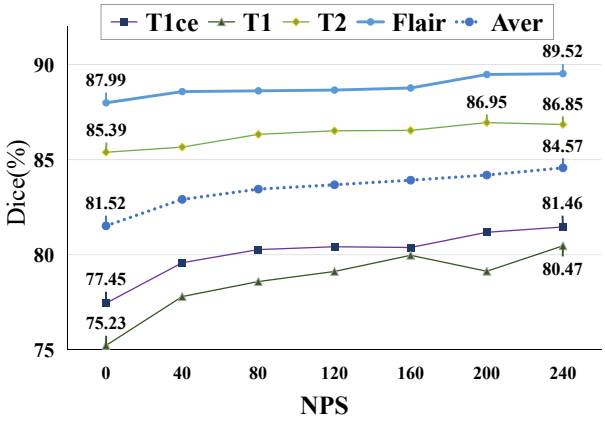

Figure 6: The ratio test for the WT segmentation task on BraTS2020. Only the header and tail values as well as the best values are displayed. NPS: the Number of Paired Subjects.

by related work [6, 24]. We conduct statistical analysis on style codes obtained from all the test images, and the max, min and avearge values are 0.189, -0.678 and -0.014, respectively. Note that, we use L2 norm regularization to constrain the style codes. Therefore, interpolating in the range [-0.7, 0.2] covers the possible space. We discover that image style are controlled by the 3rd dimension. As shown in Figure 7, images of each column are generated by interpolating the values of the 3rd dimension with the rest fixed. In addition, we change the value of the 3rd dimension with others fixed, and compare the synthetic images with the corresponding four real images. The SSIM values for T1ce, T1, T2 and Flair get the maximum of 0.7203, 0.7017, 0.6668 and 0.6963 when the value of the 3rd dimension is set to 0.1, -0.3, 0.0 and -0.1, respectively. The anatomy of the brain is clearer in T1ce and T1, while the lesioned tissue is more prominent in T2 and Flair. The former SSIM value is more affected by structural similarity, while the latter is more affected by image brightness and contrast. We think that the reason for generating more similar T1ce and T1 images is that the model can accurately extract modality-independent brain anatomical information. However, it is hard to generate images with the same brightness and contrast as real images.

Table 8
Variable multi-modal brain tumor segmentation results of WT task on BraTS2020 [2]. The table shows the Dice score for different MRI modalities being either absent (∘) or present (•), in order of T1ce, T1, T2, Flair. A better method has higher Dice (Best highlighted in bold)

Modalities				Dice(%)↑						
$T1ce$	$T1$	$T2$	$Flair$	U_hemis [14]	Rmbts [7]	Lmcr [42]	ReMIC [30]	Ours _TF	Ours _TF_UB	nnU-Net _Oracle[17]
•	∘	∘	∘	72.23	74.04	52.67	72.18	**79.58**	81.46	–
∘	•	∘	∘	71.03	74.22	56.70	74.63	**77.80**	80.47	–
∘	∘	•	∘	82.84	78.16	79.59	76.96	**85.66**	86.85	–
∘	∘	∘	•	85.07	86.24	79.26	75.37	**88.58**	89.52	–
•	•	∘	∘	76.80	78.14	69.10	71.87	**82.94**	83.73	–
•	∘	•	∘	85.33	85.24	81.99	75.80	**88.47**	88.98	–
•	∘	∘	•	87.53	88.59	83.98	71.75	**90.36**	90.76	–
∘	•	•	∘	84.84	84.05	77.37	68.48	**87.69**	88.23	–
∘	•	∘	•	86.42	87.78	84.03	70.16	**89.75**	90.32	–
∘	∘	•	•	86.64	87.94	83.74	73.32	**89.16**	90.32	–
•	•	•	∘	85.92	85.82	80.40	73.83	**88.39**	89.14	–
•	•	∘	•	88.10	88.84	86.79	73.56	**90.28**	90.78	–
•	∘	•	•	88.81	89.51	86.40	74.51	**90.46**	90.94	–
∘	•	•	•	88.40	87.30	85.90	74.53	**89.94**	90.39	–
•	•	•	•	88.92	89.20	87.29	76.41	**90.65**	90.99	92.86
Average				83.92	84.34	78.35	73.56	**87.31**	88.19	–

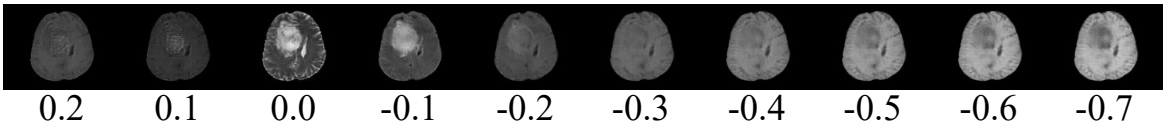

Figure 7: Evaluation of the effect of the style codes. Images of each column are generated by interpolating the values of 3rd dimension with the rest fixed.

4.2.6. Variable multi-modal brain tumor segmentation

We further evaluate the variable multi-modal brain tumor segmentation performance of our model by introudcing a Transformer [35] based fusion block, called TFusion (details can be found in B). The content codes extracted from available modalities are fused into a common content code for prediction. As shown in Table 8, U_hemis [14] (3D U-Net version), Rmbts [7] and Lmcr [42] are the networks for variable multi-modal brain tumor segmentation. Ours_TF denotes the proposed model integrated with TFusion block. The results demonstrate that our proposed model performs well on variable multi-modal brain tumor segmentation by integrating the fusion block and achieves significant improvement when some modalities are missing during testing. In addition, we evaluate the performance of Ours_TF_UB and nnU-Net_Oracle, which are trained without missing modality as upper bound and oracle respectively. Compared with Ours_TF_UB, Ours_TF achieves competitive results when the paired images are limited (40 paired images for training in the experiments). Compared with nnU-Net_Oracle, Ours_TF can handle any situations with missing modalities while nnU-net_Oracle, which is an ad-hoc method, fails to do that.

5. Discussion

Most automatic brain tumor segmentation methods use paired multi-modal images because images of different modalities provide complementary brain tumor information for more accurate segmentation. However, high-quality multi-modal public datasets, such as BraTS [2], force current research to ignore the scarcity of paired images, which is a practical problem in the real-world clinical environment. To address this problem, the unimodal methods propose solely relying on unpaired images for segmentation. However, these methods ignore the complementary information the available paired images provide. Therefore, this work studies brain tumor segmentation from privileged semi-paired images, where limited paired images are introduced during training. Specifically, we focus on improving the accuracy of unimodal segmentation by fully exploring the complementary information between multiple modalities with limited paired images. Table 2 compares our method with other unimodal methods and reveals that the developed scheme achieves a higher unimodal segmentation accuracy than state-of-the-art methods across different segmentation tasks and modalities.

NnU-Net [17] is the champion of the BraTS2020 challenge, demonstrating its ability to adequately capture complementary information when the full set of modalities is available. However, in the unimodal segmentation tasks, its segmentation accuracy decreases due to the absence of paired images. UAGAN [39] is the state-of-the-art unimodal method, which captures the modal-invariant information with only unpaired images by introducing the translation task. However, its segmentation performance is limited because the available paired images are not exploited. ReMIC [30] is a classic image completion and segmentation model for missing modalities. This method predicts missing modalities and segments the brain tumors by exploiting the completed modalities. However, the quality of the recovered images directly affects the performance, especially if a single modality is available. Therefore, Table 2 highlights that ReMIC has the worst performance in unimodal segmentation.

The modality-missing scenarios involve 15 image combinations of four modalities that may be provided in actual applications. Considering these situations, we propose a fusion block, TFusion, which fuses the missing multimodal features. By integrating this block, we obtain the fused content codes for segmentation under different missing scenarios. Table 8 presents our method's segmentation performance on 15 possible cases, achieving higher accuracy than the missing modality methods. U_hemis [14] achieves 83.92% average accuracy, extracts the features of each available modality, and fuses them by computing the first and second moments for segmentation. All available modalities contribute equally, and their latent correlations are neglected. The Dice score of Rmbts [7] is 84.34%, which employs a gated feature fusion block. The features extracted from the available modalities are fused automatically to exploit the correlation between multiple modalities. However, they simulate the features of missing modalities with zero values, inevitably introducing a computation bias and degrading the performance. Furthermore, the segmentation performance of these methods, when trained directly for missing modalities, is unsatisfactory in unimodal cases. In particular, Lmcr [42] attains only 52.67% and 56.70% accuracy with only T1ce and T1, respectively. We argue that this may be because it focuses on fusing information from multimodal images while neglecting to extract more beneficial information from a single modality.

In modality-missing scenarios, unimodal segmentation performance is significant, and therefore our approach starts with improving the performance of unimodal segmentation. By integrating the designed fusion block, we improve the segmentation performance in different multi-modal segmentation cases while retaining the superiority of unimodal segmentation. Therefore, we increase the average Dice score of the 15 possible combinations from 84.34% to 87.31%. This result shows the importance of improving the performance of unimodal segmentation.

6. Conclusion

This paper proposes a novel framework that leverages privileged semi-paired images for multi-modal brain tumor segmentation. Specifically, we develop a two-step curriculum disentanglement learning model that can be trained with semi-paired images and make predictions with unpaired images as inputs. The two steps extract the style and content of the input images separately. Furthermore, with limited paired images, we incorporate the supervised translation and content consistency loss to enhance the exploitation of the encoded complementary information. The quantitative and qualitative evaluations show the superiority of our proposed model compared to the state-of-the-art methods.

Our work presents some limitations that inspire future research directions. Specifically, the provided medical data is 3D. Compared to 2D, it can include richer semantic information, such as hierarchical information. Therefore, extending the model to the 3D data domain can further promote the need for automatic brain tumor segmentation. The fusion block (TFusion) is also based on Swin Transformer [25], which uses a self-attention mechanism. Hence, the high GPU memory requirements pose a limitation. Therefore, designing a more lightweight and efficient fusion block is a future research direction. For example, using PoolFormer [38] to replace the Swin Transformer can be appealing. It reduces the computational complexity by replacing the self-attention module with an embarrassingly simple spatial pooling operator. Its effectiveness is verified in vision tasks. Finally, our work is based on supervised learning, while obtaining high-quality annotated data requires professionals and is time-consuming. Given that the scarcity of annotation data is a common problem in medical image processing, we will further study brain tumor segmentation in a semi-supervised and privileged semi-paired learning setting to alleviate the scarcity of labeled data.

7. Conflict of interest statement

The authors declared that they have no conflicts of interest to this work.

8. Acknowledgment

This work is supported in part by the Guangzhou Science and Technology Planning Project (202201010092), the Guangdong Provincial Natural Science Foundation (2020A1515010717), NSF-1850492 and NSF-2045804.

A. Dataset parameters

In BraTS2020 [2] and BraTS2018 dataset [3], all modalities (T1ce, T1, T2 and Flair) from the same subject are co-registered to a common anatomical template SRI [28], resampled to isotropic $1mm^3$ and skull-stripped following manual revision. Each modality volume contains 155 slices with the size of 240×240.

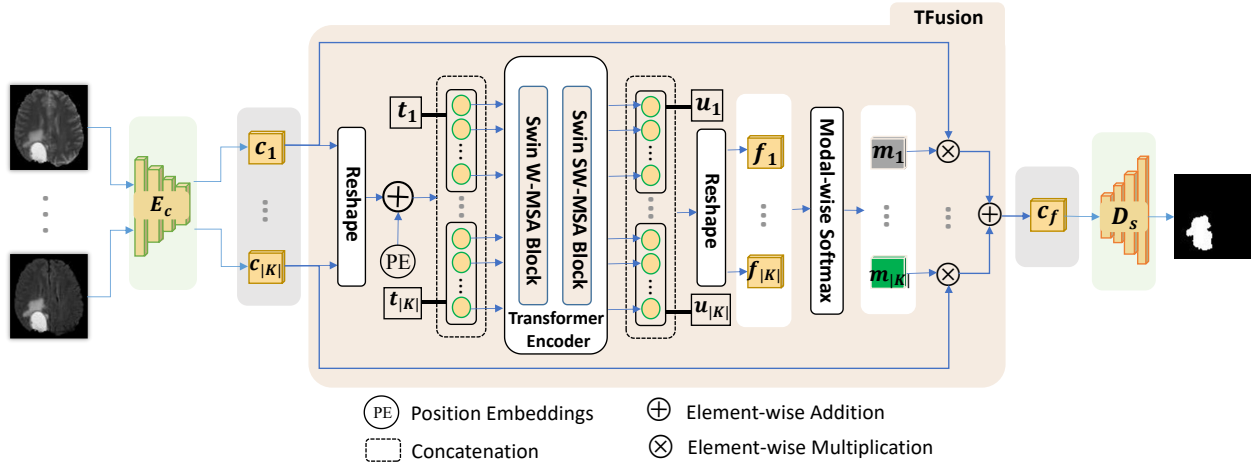

Figure 8: The illustration of the proposed TFusion. E_c and D_s are the content encoder and segmentation decoder of our model, respectively.

B. Details of TFusion block

As shown in Figure 8, we propose a TFusion block for variable multi-modal brain tumor segmentation, which is a transformer based N-to-One fusion block at the voxel level.

Let $K \subseteq \{1, 2, \dots, S\}$ denotes the available modality set of K, where S is the number of all possible modalities. The content codes $(c_1, \dots, c_{|K|}$, $|K|$ denotes the number of available modalities) extracted from available modalities are fused into a common content code c_f for prediction. In the TFusion block, inspired by ViT [13], the input content codes are reshaped as a sequence of token embeddings $(t_1, \dots, t_{|K|})$ by flattening their spatial dimensions into one dimension and

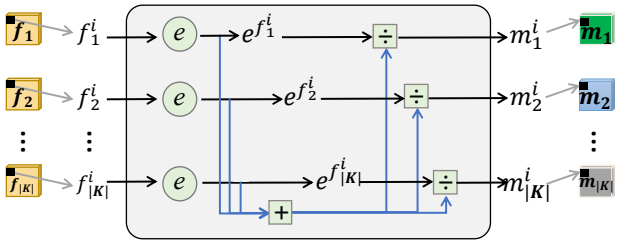

Figure 9: The illustration of modal-wise and voxel-level softmax function.

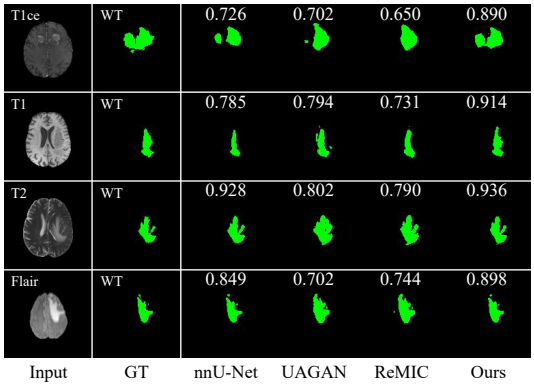

Figure 10: Visualization of WT segmentation results and the corresponding Dice scores. Rows: different input modalities. Columns: all the methods.

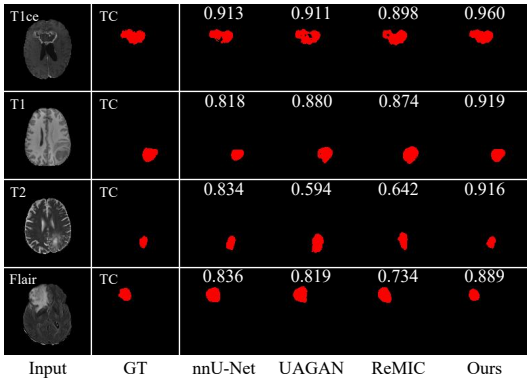

Figure 11: Visualization of TC segmentation results and the corresponding Dice scores. Rows: different input modalities. Columns: all the methods.

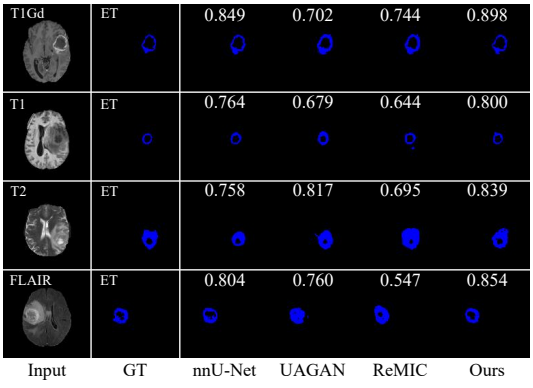

Figure 12: Visualization of ET segmentation results and the corresponding Dice scores. Rows: different input modalities. Columns: all the methods.

Table 9
Performance evaluation for the segmentation task of WT, TC and ET on BraTS2020. A better method has lower ASSD (Best highlighted in bold).

Metric		ASSD(mm)↓				
Modality		T1ce	T1	T2	Flair	Aver
WT	nnU-Net [17]	2.90±0.52	3.18±0.21	2.10±0.27	1.50±0.06	2.42±0.01
	UAGAN [39]	3.83±0.08	3.84±0.58	2.77±0.09	2.40±0.15	3.21±0.07
	ReMIC [30]	3.87±0.18	3.77±0.09	3.48±0.09	3.81±0.03	3.74±0.08
	Ours	**2.65±0.04**	**3.00±0.49**	**1.96±0.11**	**1.42±0.06**	**2.26±0.08**
TC	nnU-Net [17]	**1.98±0.63**	**4.72±0.36**	3.70±0.64	3.33±0.02	3.43±0.08
	UAGAN [39]	2.66±0.20	5.13±0.34	3.53±0.13	3.44±0.04	3.69±0.06
	ReMIC [30]	2.67±0.36	5.71±0.91	4.77±0.40	5.04±0.17	4.54±0.46
	Ours	2.03±0.66	4.88±0.16	**3.15±0.33**	**3.25±0.64**	**3.33±0.11**
ET	nnU-Net [17]	1.12±0.48	4.33±0.30	3.07±0.59	**3.15±0.57**	2.92±0.49
	UAGAN [39]	1.64±0.29	4.52±0.78	3.15±0.28	3.61±0.88	3.22±0.55
	ReMIC [30]	2.01±0.27	4.99±0.45	4.25±0.37	4.40±0.16	3.92±0.13
	Ours	**1.02±0.28**	**3.71±0.53**	**2.98±0.41**	3.73±1.51	**2.86±0.68**

Table 10
Performance evaluation of WT segmentation on BraTS2020 with different values of n_s.

	Dice(%)↑				
n_s	T1ce	T1	T2	Flair	Aver
1	77.99±2.70	78.30±1.00	85.83±0.71	88.88±0.64	82.75±1.26
2	78.00±4.02	77.47±0.93	85.44±0.22	88.56±0.37	82.37±1.39
4	79.07±4.33	78.34±1.42	85.08±0.35	88.34±0.74	82.70±1.70
8	79.58±1.13	77.80±2.16	85.66±0.35	88.58±0.08	82.91±0.36
16	78.43±4.52	77.73±1.06	85.79±0.57	88.42±0.11	82.59±1.56

combining with the sinusoidal position embeddings [35]. Then, the token embeddings are fed into the transformer encoder, which consists of blocks W-MSA and SW-MSA [25], to learn latent multi-modal correlations $(u_1, \dots, u_{|K|})$. By reshaping the correlations, we get the transformed feature maps $(f_1, \dots, f_{|K|})$, which have the same size as the input content codes. As shown in Figure 9, we denote the i-th voxel of f_k and m_k as f_k^i and m_k^i $(k \in K)$, respectively. e is the natural logarithm. Through a modal-wise and voxel-level softmax function, we obtain weight maps $(m_1, \dots, m_{|K|})$ for fusion. By element-wise multiplying input content codes with the corresponding weight maps and summing all of them, we can obtain a fused content code c_f for prediction.

Since the sum of $m_1^i, \dots, m_{|K|}^i$ is 1, the value range of fused content code c_f remains stable to improve the robustness of the model for variable input modalities. Moreover, the relative sizes of $f_1^i, \dots, f_{|K|}^i$ are retained in the corresponding weights, which contain the latent multi-modal correlation learned from transformered encoder. **In particular**, when only one modality is available, all the values of the weight map is 1, which means $c_f = c_k (k \in K, |K| = 1)$. In this case, the input content code remains unchanged, which maintains the performance of the model for brain tumor segmentation with single modalitiy.

It is worth noting that, TFusion block is a flexible data-dependent fusion strategy. It does not need to simulate missing modalities (e.g. zero-padding and synthetic modality).

C. Other metrics on segmentation results

We further use average symmetric surface distance (ASSD) for evaluation. The ASSD metric is introduced to evaluate the average symmetric surface distance.

$$ASSD = \frac{\sum_{a \in A} min_{b \in B} d(a,b) + \sum_{b \in B} min_{a \in A} d(a,b)}{N_A + N_B} \tag{13}$$

Here, A and B denote the boundary voxel set of prediction and ground truth volumes, and $d(a,b)$ represents the Euclidean distance between voxel a and b. N_A and N_B denote the number of voxels in A and B. A better method has lower ASSD. The results of WT, TC and ET segmentation on BraTS2020 are shown in Table 9. We obtain the best average results (WT: 2.26mm, TC: 3.33mm, ET: 2.86mm), i.e., lowest ASSD values, in all three segmentation tasks, while ReMIC performs worst (WT: 3.74mm, TC: 4.54mm, ET: 3.92mm). It is consistent with the Dice score. Although nnU-Net outperforms our model in some cases, our results

are also competitive. For example, in the TC segmentation with T1ce, the ASSD value of nnU-Net is 1.98mm, while the value of our method (2.03mm) is slightly higher. Visualization of segmentation results are shown in Figure 10, Figure 11 and Figure 12, it illustrates that our model can identify more accurate details of irregular shape brain tumors to achieve high dice scores.

D. Extra experiments for n_s

We experimented with different values of n_s on WT segmentation of BraTS2020. As shown in Table 10, best average dice are achieved when the dimensionality of the style code is set to $n_s = 8$.

References

[1] Azad, R., Khosravi, N., Dehghanmanshadi, M., Cohen-Adad, J., Merhof, D., 2022. Medical image segmentation on mri images with missing modalities: A review. arXiv preprint arXiv:2203.06217 .

[2] Bakas, S., Menze, B., Davatzikos, C., Kalpathy-Cramer, J., Farahani, K., Bilello, M., Mohan, S., Freymann, J.B., Kirby, J.S., Ahluwalia, M., Statsevych, V., Huang, R., Fathallah-Shaykh, H., Wiest, R., Jakab, A., Colen, R.R., Kotrotsou, A., Marcus, D., Milchenko, M., Nazeri, A., Weber, M.A., Mahajan, A., Baid, U., 2020. MICCAI Brain Tumor Segmentation (BraTS) 2020 Benchmark: "Prediction of Survival and Pseudoprogression". URL: https://doi.org/10.5281/zenodo.3718904, doi:10.5281/zenodo.3718904.

[3] Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., et al., 2018. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. CoRR abs/1811.02629. URL: http://arxiv.org/abs/1811.02629, arXiv:1811.02629.

[4] Bhosale, Y.H., Patnaik, K.S., 2022. Application of deep learning techniques in diagnosis of covid-19 (coronavirus): A systematic review. Neural Processing Letters , 1–53.

[5] Bhosale, Y.H., Patnaik, K.S., 2023. Puldi-covid: Chronic obstructive pulmonary (lung) diseases with covid-19 classification using ensemble deep convolutional neural network from chest x-ray images to minimize severity and mortality rates. Biomedical Signal Processing and Control 81, 104445.

[6] Chartsias, A., Papanastasiou, G., Wang, C., Semple, S., Newby, D.E., Dharmakumar, R., Tsaftaris, S.A., 2020. Disentangle, align and fuse for multimodal and semi-supervised image segmentation. IEEE transactions on medical imaging 40, 781–792.

[7] Chen, C., Dou, Q., Jin, Y., Chen, H., Qin, J., Heng, P.A., 2019. Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 447–456.

[8] Chen, C., Dou, Q., Jin, Y., Liu, Q., Heng, P.A., 2021. Learning with privileged multimodal knowledge for unimodal segmentation. IEEE Transactions on Medical Imaging , 1–1doi:10.1109/TMI.2021.3119385.

[9] Chen, L., Wu, Y., DSouza, A.M., Abidin, A.Z., Wismüller, A., Xu, C., 2018. Mri tumor segmentation with densely connected 3d cnn, in: Medical Imaging 2018: Image Processing, International Society for Optics and Photonics. p. 105741F.

[10] Cheng, J., Liu, J., Kuang, H., Wang, J., 2022. A fully automated multimodal mri-based multi-task learning for glioma segmentation and idh genotyping. IEEE Transactions on Medical Imaging .

[11] Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J., 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[12] Dolz, J., Gopinath, K., Yuan, J., Lombaert, H., Desrosiers, C., Ayed, I.B., 2018. Hyperdense-net: a hyper-densely connected cnn for multi-modal image segmentation. IEEE transactions on medical imaging 38, 1116–1126.

[13] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929.

[14] Havaei, M., Guizard, N., Chapados, N., Bengio, Y., 2016. Hemis: Hetero-modal image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 469–477.

[15] Hu, Z., Tang, J., Wang, Z., Zhang, K., Zhang, L., Sun, Q., 2018. Deep learning for image-based cancer detection and diagnosis- a survey. Pattern Recognition 83, 134–149.

[16] Huang, X., Liu, M.Y., Belongie, S., Kautz, J., 2018. Multimodal unsupervised image-to-image translation, in: Proceedings of the European conference on computer vision (ECCV), pp. 172–189.

[17] Isensee, F., Jäger, P.F., Full, P.M., Vollmuth, P., Maier-Hein, K.H., 2021. nnu-net for brain tumor segmentation, in: Crimi, A., Bakas, S. (Eds.), Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Springer International Publishing, Cham. pp. 118–132.

[18] Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., Maier-Hein, K.H., 2017. Brain tumor segmentation and radiomics survival prediction: Contribution to the brats 2017 challenge, in: International MICCAI Brainlesion Workshop, Springer. pp. 287–297.

[19] Jiang, H., Diao, Z., Yao, Y.D., 2021. Deep learning techniques for tumor segmentation: a review. The Journal of Supercomputing , 1–45.

[20] Khojaste-Sarakhsi, M., Haghighi, S.S., Ghomi, S.F., Marchiori, E., 2022. Deep learning for alzheimer's disease diagnosis: A survey. Artificial Intelligence in Medicine , 102332.

[21] Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

[22] Lai, M., 2015. Deep learning for medical image segmentation. arXiv preprint arXiv:1505.02000 .

[23] Lau, K., Adler, J., Sjölund, J., 2019. A unified representation network for segmentation with missing modalities. arXiv:1908.06683.

[24] Lee, H.Y., Tseng, H.Y., Huang, J.B., Singh, M., Yang, M.H., 2018. Diverse image-to-image translation via disentangled representations, in: Proceedings of the European Conference on Computer Vision (ECCV).

[25] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022.

[26] Ouyang, J., Adeli, E., Pohl, K.M., Zhao, Q., Zaharchuk, G., 2021. Representation disentanglement for multi-modal brain mri analysis, in: International Conference on Information Processing in Medical Imaging, Springer. pp. 321–333.

[27] Rafi, A., Khan, Z., Aslam, F., Jawed, S., Shafique, A., Ali, H., 2022. A review: Recent automatic algorithms for the segmentation of brain tumor mri. AI and IoT for Sustainable Development in Emerging Countries , 505–522.

[28] Rohlfing, T., Zahr, N.M., Sullivan, E.V., Pfefferbaum, A., 2010. The sri24 multichannel atlas of normal adult human brain structure. Human brain mapping 31, 798–819.

[29] Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, Cham. pp. 234–241.

[30] Shen, L., Zhu, W., Wang, X., Xing, L., Pauly, J.M., Turkbey, B., Harmon, S.A., Sanford, T.H., Mehralivand, S., Choyke, P.L., Wood, B.J., Xu, D., 2021. Multi-domain image completion for random missing input data. IEEE Transactions on Medical Imaging 40, 1113–1122. doi:10.1109/TMI.2020.3046444.

[31] Singh, V., Gourisaria, M.K., GM, H., Rautaray, S.S., Pandey, M., Sahni, M., Leon-Castro, E., Espinoza-Audelo, L.F., 2022. Diagnosis

of intracranial tumors via the selective cnn data modeling technique. Applied Sciences 12, 2900.

[32] Tripathi, P.C., Bag, S., 2022. A computer-aided grading of glioma tumor using deep residual networks fusion. Computer Methods and Programs in Biomedicine 215, 106597.

[33] Upadhyay, N., Waldman, A., 2011. Conventional mri evaluation of gliomas. The British journal of radiology 84, S107–S111.

[34] Valindria, V.V., Pawlowski, N., Rajchl, M., Lavdas, I., Aboagye, E.O., Rockall, A.G., Rueckert, D., Glocker, B., 2018. Multi-modal learning from unpaired images: Application to multi-organ segmentation in ct and mri, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 547–556. doi:10.1109/WACV.2018.00066.

[35] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I., 2017. Attention is all you need, in: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[36] Wang, G., Li, W., Ourselin, S., Vercauteren, T., 2017. Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks, in: International MICCAI brainlesion workshop, Springer. pp. 178–190.

[37] Wang, Y., Zhang, Y., Hou, F., Liu, Y., Tian, J., Zhong, C., Zhang, Y., He, Z., 2020. Modality-pairing learning for brain tumor segmentation. arXiv:2010.09277.

[38] Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., Yan, S., 2022. Metaformer is actually what you need for vision, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10819–10829.

[39] Yuan, W., Wei, J., Wang, J., Ma, Q., Tasdizen, T., 2019. Unified attentional generative adversarial network for brain tumor segmentation from multimodal unpaired images, in: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, Springer International Publishing, Cham. pp. 229–237.

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

[41] Zhou, C., Ding, C., Lu, Z., Wang, X., Tao, D., 2018. One-pass multi-task convolutional neural networks for efficient brain tumor segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 637–645.

[42] Zhou, T., Canu, S., Vera, P., Ruan, S., 2021. Latent correlation representation learning for brain tumor segmentation with missing mri modalities. IEEE Transactions on Image Processing 30, 4263–4274.

[43] Zhou, T., Ruan, S., Canu, S., 2019. A review: Deep learning for medical image segmentation using multi-modality fusion. Array 3, 100004.