ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240522

Anatomical Consistency Distillation and Inconsistency Synthesis for Brain Tumor Segmentation with Missing Modalities

Zheyu Zhang[a], Xinzhao Liu[a], Zheng Chen[a], Yueyi Zhang[a], Huanjing Yue[c], Yunwei Ou[d] and Xiaoyan Sun[a,b,*]

[a]MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition,
University of Science and Technology of China, Hefei, China
[b]Anhui Province Key Laboratory of Biomedical Imaging and Intelligent Processing,
Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei, China
[c]Tianjin University, Tianjin, China
[d]Beijing Tiantan Hospital, Capital Medical University, Beijing, China

Abstract. Multi-modal Magnetic Resonance Imaging (MRI) is imperative for accurate brain tumor segmentation, offering indispensable complementary information. Nonetheless, the absence of modalities poses significant challenges in achieving precise segmentation. Recognizing the shared anatomical structures between mono-modal and multi-modal representations, it is noteworthy that mono-modal images typically exhibit limited features in specific regions and tissues. In response to this, we present Anatomical Consistency Distillation and Inconsistency Synthesis (ACDIS), a novel framework designed to transfer anatomical structures from multi-modal to mono-modal representations and synthesize modality-specific features. ACDIS consists of two main components: Anatomical Consistency Distillation (ACD) and Modality Feature Synthesis Block (MFSB). ACD incorporates the Anatomical Feature Enhancement Block (AFEB), meticulously mining anatomical information. Simultaneously, Anatomical Consistency ConsTraints (ACCT) are employed to facilitate the consistent knowledge transfer, i.e., the richness of information and the similarity in anatomical structure, ensuring precise alignment of structural features across mono-modality and multi-modality. Complementarily, MFSB produces modality-specific features to rectify anatomical inconsistencies, thereby compensating for missing information in the segmented features. Through validation on the BraTS2018 and BraTS2020 datasets, ACDIS substantiates its efficacy in the segmentation of brain tumors with missing MRI modalities.

1 Introduction

Brain tumors have a profound impact on overall health, underscoring the critical necessity for precise segmentation crucial in both diagnosis and the monitoring of treatment outcomes. The advanced segmentation of brain tumors often relies on the integration of various Magnetic Resonance Imaging (MRI) sequences, encompassing FLuid Attenuation Inversion Recovery (FLAIR), contrast-enhanced T1-weighted (T1ce), T1-weighted (T1), and T2-weighted (T2) modalities. These sequences collectively provide complementary information essential for achieving precise diagnostic outcomes. However, the clinical reality presents challenges, as the simultaneous availability of all these modalities cannot be guaranteed. This uncertainty gives rise to difficulties in achieving accurate tumor segmentation. For example, the unavailability of the T1c modality may be attributed to patient allergies to contrast agents, while other modalities may be inaccessible due to disparities in MRI parameters. Consequently, addressing the segmentation of brain tumors in the absence of certain modalities has emerged as a critical imperative in clinical settings [2].

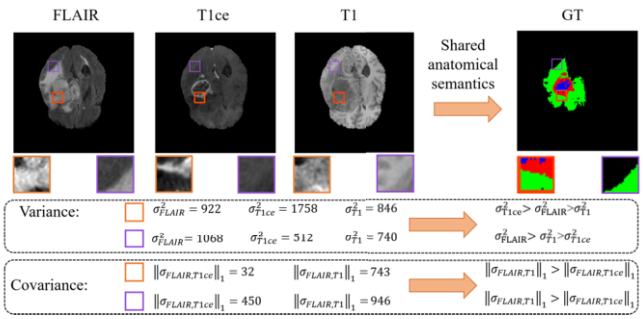

Figure 1. Different modalities share the same anatomical semantics, but they vary in terms of pixel intensity visualization. Typically, 1. The modality with richer information and clearer anatomical structure exhibits greater variance. 2. Greater anatomical similarity between modalities results in higher absolute covariance values.

Current methods for handling missing modalities primarily focus on two strategies: modality synthesis [19, 41, 34, 31, 17] and common latent space modeling [7, 42, 36, 24, 30]. Modality synthesis methods utilize generative adversarial networks or diffusion models to generate missing modalities. Nevertheless, these generated modalities lack comprehensive modality-specific biological information [19], limiting their effectiveness in achieving accurate segmentation. On the other hand, common latent space modeling aims to project each modality into a shared latent space, where modality features are fused to obtain the segmentation results. This strategy incorporates knowledge distillation to guide the extraction and aggregation of mono-modal and multi-modal features [29, 24, 30]. De-

* Corresponding Author. Email: sunxiaoyan@ustc.edu.cn

spite these advancements, current methods tend to uniformly transfer knowledge between mono-modality and multi-modality, neglecting the distinct characteristics that are unique to each modality. Crucially, while these modalities share consistent anatomical semantics like tumor or tissue types [33, 1], they exhibit significant inconsistencies in pixel intensity visualization.

To address these issues, we propose the Anatomical Consistency Distillation and Inconsistency Synthesis (ACDIS) through elaborately combined knowledge distillation and feature synthesis for brain tumor segmentation with missing modalities. Our method comprises two core components: Anatomical Consistency Distillation (ACD) and Modality Feature Synthesis Block (MFSB). ACD focuses on transferring consistent anatomic insights from multi-modality to mono-modality, enhancing the representation of anatomical structures. It utilizes Anatomical Consistency ConsTraints (ACCT), emphasizing the transmission of knowledge based on variance and covariance within a local window, which highlights information richness and anatomical structural similarity. These constraints ensure that the transfer of knowledge is not only rich in detail but also precisely aligned with the structural similarities inherent in anatomical features. The Anatomical Feature Enhancement Block (AFEB) complements this by deeply extracting mono-modal anatomical information, cooperating with ACCT to refine and enrich the mono-modal representations.

On the other hand, MFSB synthesizes modality-specific features that address the inconsistencies often seen in anatomical information due to varying pixel intensities. It generates a modality-specific style for transferring existing mono-modal features to simulate the missing modalities, compensating for the discrepancies in anatomical structure visualization.

Our main contributions can be summarized as follows:

- We investigate the anatomical consistency and inconsistency between the mono-modality and the multi-modality. Based on them, we propose ACDIS that leverages the consistent anatomical insights and synthesizes the inconsistent anatomical features.
- ACDIS incorporates ACD with a focus on variance and covariance in local windows, conveying rich information and emphasizing anatomical structural similarity, thereby enhancing mono-modal representation. MFSB is employed to learn modality-specific styles, addressing the gap in missing information.
- ACDIS consistently outperforms state-of-the-art methods for brain tumor segmentation with missing modalities on BraTS2018 and BraTS2020.

2 Related Work

2.1 Brain Tumor Segmentation

Convolutional Neural Networks (CNNs) and Transformers are both potent architectures for feature extraction in brain tumor segmentation tasks. Within the domain of CNNs, several methods build upon the foundational U-Net model [26], focusing on enhancements such as 2D/3D operations [6], cascaded networks [16, 40], skip connections [43] and automated adaptations [15]. Transformer architectures, known for their ability to model long-range dependencies, offer significant advantages for capturing the global context necessary for accurate segmentation. Chen *et al.* [5] incorporated Transformer layers on top of CNN features to utilize global contextual information. Inspired by the hierarchical Swin Transformer [22], Cao *et al.* [3] developed a U-shaped architecture using Swin Transformer blocks, which facilitates robust feature extraction by leveraging the advantages of both CNNs and Transformers. This integration of CNN and Transformer is followed by numerous methods [35, 28, 20, 12, 11]. However, the phenomenon of missing modalities prevents them from extracting complete complementary information, posing an obstacle for accurate brain tumor segmentation.

2.2 Multi-modal Learning for Missing Modalities

Recently, many multi-modal learning methods adopt modality synthesis and common latent space modeling to overcome the missing modality issue. Modality synthesis [19, 31] methods that generate missing modalities and train a segmentation model with complete modalities. HiNet and MouseGAN++ employ multiple specific generative adversarial networks to separately generate missing modalities [41, 34]. M3AE reconstructs multi-modal substitutes within multi-modal mask image modeling strategy [21]. Kim *et al.* propose 3D latent diffusion models combined with conditioning using the target modality allows generating high-quality target modality in 3D [17, 25]. These synthesis methods often require training their networks on various combinations of missing modalities, which can be computationally intensive. Moreover, the segmentation performance is closely tied to the quality of the generated modalities, which may often contain unexpected noise.

Common latent space modeling methods [7, 18, 27, 38] that project all available modalities into a shared latent space, and then fuse all modalities to produce a segmentation map. RFNet, LCRL, and mmFormer employ mono-modal segmentation or reconstruction tasks to improve the mono-modal representation in the latent space [7, 42, 36]. ShapSpec further enhances modality features through feature domain classification and feature distribution alignment [27]. Some methods, such as [4, 32], disentangle modalities into several attributes and process them separately. Konwer *et al.* adopt meta-learning and adversarial learning strategies to enhance modality-agnostic representations [18]. TMFormer [38] employs token merging strategy to obtain compact representation in mono-modal and multi-modal token sequences. While these methods primarily focus on improving representations in scenarios involving missing modalities, they overlook the intrinsic biological information mined in multi-modal modalities, which can boost the mono-modal representations by modeling data distribution relationships between mono-modality and multi-modality.

2.3 Knowledge Distillation

Cross-modal distillation has been a long-standing approach for transferring specific knowledge between different modalities [10]. KD-Net [14] and ACN [29] transfer feature distributions from the multi-modal network to the mono-modal network, focusing on latent features or soft segmentation masks, respectively. MMANet [30] focuses on distilling information from hard samples that lie near the decision boundary. Additionally, GSS [24] uses an ensemble of mono-modal soft segmentation masks as the teacher's output for distillation. However, these methods obscure the distinction between what should be consistent and what should be inconsistent across modalities. It is crucial to maintain consistency in anatomical structures across mono-modal and multi-modal data while accepting inconsistency in pixel intensities. Introducing conflicting information that contradicts specific imaging technologies adversely affects the student model's performance.

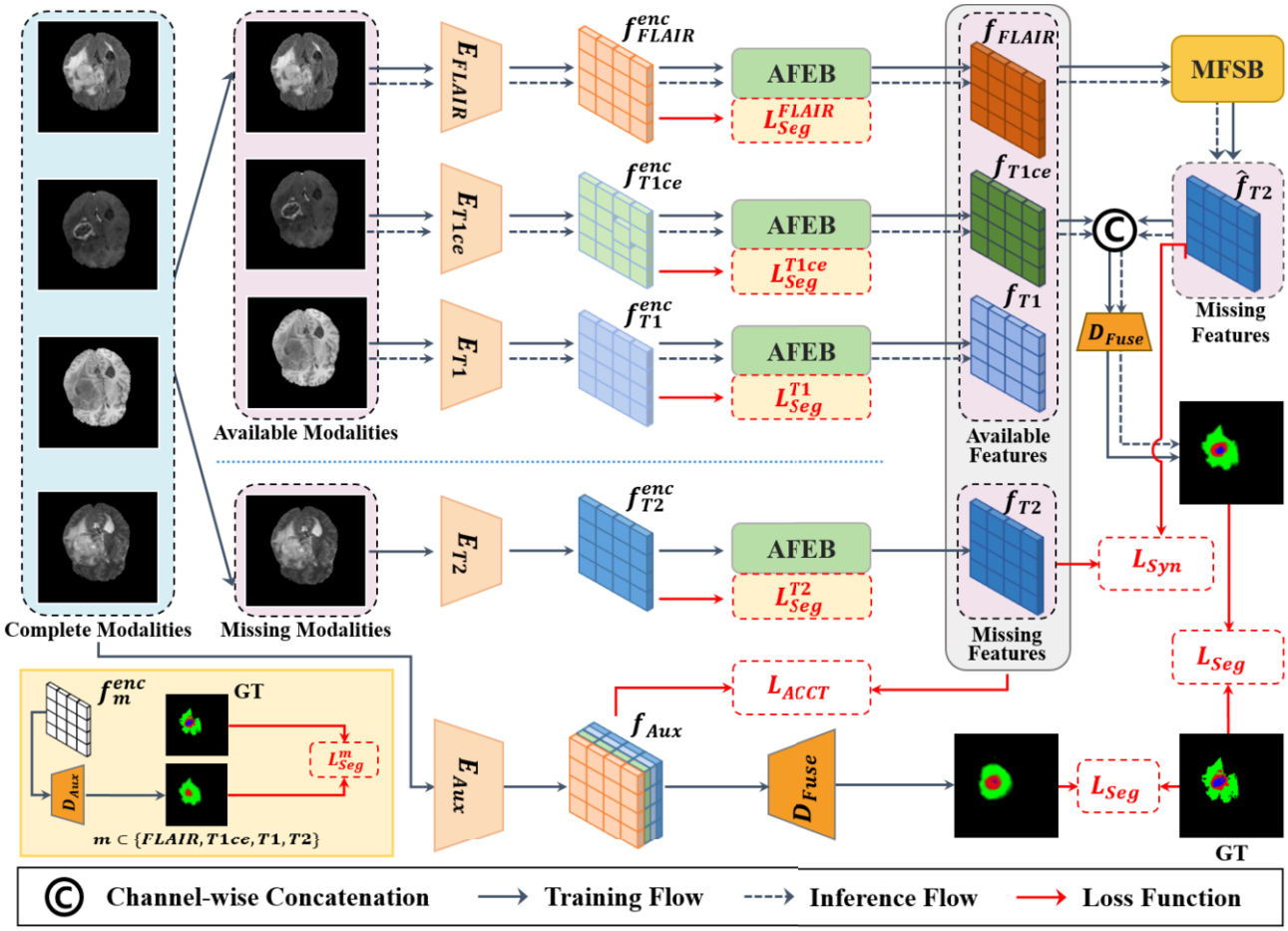

Figure 2. Overall architecture of our ACDIS. During the training phase, ACDIS comprises four mono-encoders for extracting mono-modal features, one auxiliary encoder dedicated to consistency distillation in conjunction with the proposed AFEB and ACCT, one auxiliary decoder for obtaining individual mono-modal segmentation (denoted in the yellow box), one MFSB designed to synthesize features for missing modalities, and one fusion decoder responsible for generating the final segmentation result. The auxiliary encoder and decoder are discarded during the inference phase.

3 Method

Our framework is illustrated in Figure 2. Each modality is sent into individual encoder $\{E_m\}_{m\in\{FLAIR,T1ce,T1,T2\}}$ for extracting hierarchical features, which are then fed into the auxiliary decoder D_{Aux} to yield the mono-modal segmentation. To enhance the representation of mono-modal anatomical structures, the mono-modal hierarchical features are passed through AFEB to elaborately model spatial structural dependencies in collaboration with ACCT. ACCT introduces an auxiliary encoder E_{Aux} that provides multi-modal features with comprehensive anatomical information. ACCT and AFEB are essential components of our ACT for ensuring anatomical consistency. After being processed by ACD, the individual features contain richer anatomical structure information. The MFSB utilizes available mono-modal features to synthesize the missing ones, thus addressing anatomical inconsistencies. The completed multi-modal features are then fed into the decoder D_{Fuse} to obtain the final segmentation.

3.1 Anatomical Consistency Distillation (ACD)

The ACD consists of ACCT and AFEB, where AFEB refines the mono-modal representation cooperating with ACCT.

3.1.1 Anatomical Consistency ConsTraints (ACCT)

While different modalities may exhibit distinct pixel intensities attributable to their respective imaging technologies, they share consistent anatomical semantics in terms of relative positions, especially when registered to standard atlases. Multi-modal data inherently encapsulates a more comprehensive anatomical information compared to mono-modal data. Consequently, our approach focuses on distilling anatomical structural information from the multi-modal data into the mono-modal data.

As shown in Figure 1, windows with higher variances σ^2 typically contain richer information related to the segmentation mask, such as more pronounced contrast at the segmentation boundaries. This observation leads us to utilize variance as a key feature in our distillation process. This is illustrated as

$$L_{Var} = 1 - \frac{2\sigma_{f_m}\sigma_{f_{Aux}} + \epsilon}{\sigma_{f_m}^2 + \sigma_{f_{Aux}}^2 + \epsilon}, \tag{1}$$

where f_m and f_{Aux} represent the mono-modal features and multi-modal features, respectively. The multi-modal features f_{Aux} are fixed to serve as the label for the mono-modal features f_m. The term $\epsilon = 1e\text{-}6$ is included to prevent division by zero.

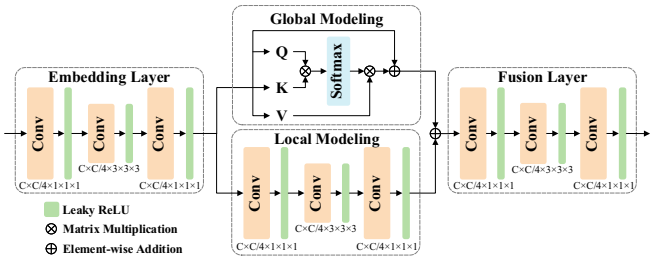

Figure 3. Anatomical Feature Enhancement Block (AFEB).

The constraint on the variance pushes the mono-modal representation to learn information richness of the multi-modal representation. Secondly, we employ the covariance to push the mono-modal representation learn the anatomical structure of the multi-modal representation, which quantifies the variation trends among pixel values. When two local windows exhibit similar gradients in corresponding positions, the covariance tends to increase. The covariance of two windows can be described as

$$\sigma_{f_m f_{Aux}} = \frac{1}{n-1}\sum_{i=1}^{n}(f_{m_i} - \bar{f}_m)(f_{Aux_i} - \bar{f}_{Aux}). \quad (2)$$

Since FLAIR and T1ce MRI modalities often exhibit opposite pixel intensities due to their underlying imaging principles and the use of contrast agents, this may cause the feature pixel intensity demonstrating the opposite variational trend. Therefore, we use $||\cdot||_1$ in the constraint to only focus on the gradient magnitude, which is presented as

$$L_{Covar} = 1 - \frac{||\sigma_{win_1 win_2}||_1 + \epsilon}{\sigma_{win_1}\sigma_{win_2} + \epsilon}. \quad (3)$$

Note that we use Sigmoid function to normalize the mono-modal features f_m and the multi-modal features f_{Aux} before sending into Equations 1 and 3. The loss in ACCT can be presented as

$$L_{ACCT} = L_{Var} + L_{Covar}. \quad (4)$$

3.1.2 Anatomical Feature Enhancement Block (AFEB)

In order to effectively assimilate anatomical structural knowledge from multi-modal data, we devise an Anatomical Feature Enhancement Block (AFEB) to facilitate the extraction of spatial anatomical information within the mono-modal domain and establish data relationships between mono-modal and multi-modal representations. As depicted in Figure 3, we firstly employ embedding layers to project the mono-modal features into a multi-modal space, facilitating the modeling of relationships between mono-modal and multi-modal representations. Subsequently, we employ both an attention layer and a convolutional layer to systematically explore spatial dependencies in both global and local aspects. Finally, we fuse the global and the local features, obtaining the enhanced mono-modal feature f_m, where $m \in \{FLAIR, T1ce, T1, T2\}$.

The enhanced mono-modal feature f_m is further constrained with proposed ACCT loss L_{ACCT}, which distills the multi-modality features to f_m. In this way, the ACD framework enhances the mono-modal representation by leveraging the underlying anatomical structure consistency.

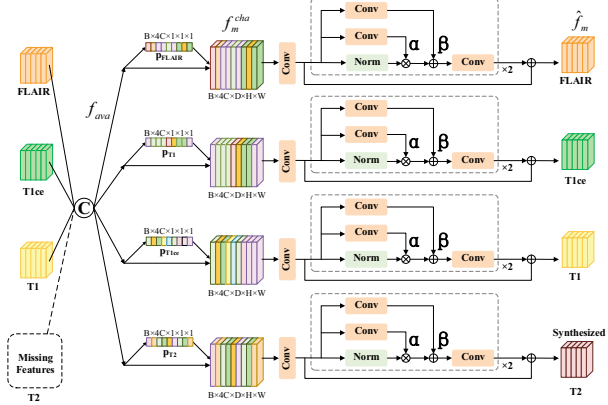

Figure 4. Modality Feature Synthesis Block (MFSB).

3.2 Modality Feature Synthesis Block (MFSB)

Building upon the foundation laid by ACD, we introduce the Multi-Modal Feature Synthesis Block (MFSB) to address the anatomical inconsistencies, specifically focusing on the feature intensities of missing modalities. As illustrated in Figure. 4, all available mono-modal features are concatenated in channel dimension, denoted as $f_{ava} = \text{Concat}(f_{FLAIR}, f_{T1ce}, f_{T1}, f_{T2}) \in \mathcal{R}^{B\times 4C\times D\times H\times W}$. In cases where the corresponding modality f_m is absent, we compensate it by assigning zero values and proceed to synthesize it in the subsequent steps.

For each modality synthesis $\hat{f}_m$, we learn the prior weight p_m, which serves to model inter-modal significance along the channel dimension, thereby amplifying the most relevant mono-modal features. A dimension-reducing convolution operation $W_2 \in \mathcal{R}^{C\times 4C}$ is employed to squeeze the amplified features. The processes can be expressed as

$$p_m = \delta(W_1 \text{GAP}(f_{ava})), \quad (5)$$

$$f_m^{cha} = W_2 p_m f_{ava}, \quad (6)$$

where δ is the sigmoid function, $W_1 \in \mathcal{R}^{4C\times 4C}$, GAP$(\cdot)$ represents the global average pooling function, and $f_m^{cha} \in \mathcal{R}^{B\times C\times D\times H\times W}$. Subsequently, we learn the modal-specific style on the spatial dimension. The process is expressed as

$$\alpha_m = \delta(W_3 f_m^{cha}), \quad (7)$$

$$\beta_m = \delta(W_4 f_m^{cha}), \quad (8)$$

$$\hat{f}_m = (1+\alpha_m) f_m^{cha} + \beta_m, \quad (9)$$

where $\alpha_m \in \mathcal{R}^{B\times C\times D\times H\times W}$, $\beta_m \in \mathcal{R}^{B\times C\times D\times H\times W}$ represent learned affine transformations that encapsulate the modality-specific styles, and $W_3 \in \mathcal{R}^{C\times C}$ and $W_4 \in \mathcal{R}^{C\times C}$ denote corresponding convolutional operations, respectively. In this context, α_m serves to adjust the magnitude of synthesized features for the missing modalities, while β_m facilitates the shifting of these features to a suitable space. In the training stage, we can obtain the complete modality-specific features with complete modalities input. We use the Mean Square Error function to supervise the synthesis of $\hat{f}_m$, prioritizing pixel-wise intensity, aligning with the objective of inconsistency synthesis. The loss is presented as

$$L_{Syn} = ||\hat{f}_m - f_m||_2. \quad (10)$$

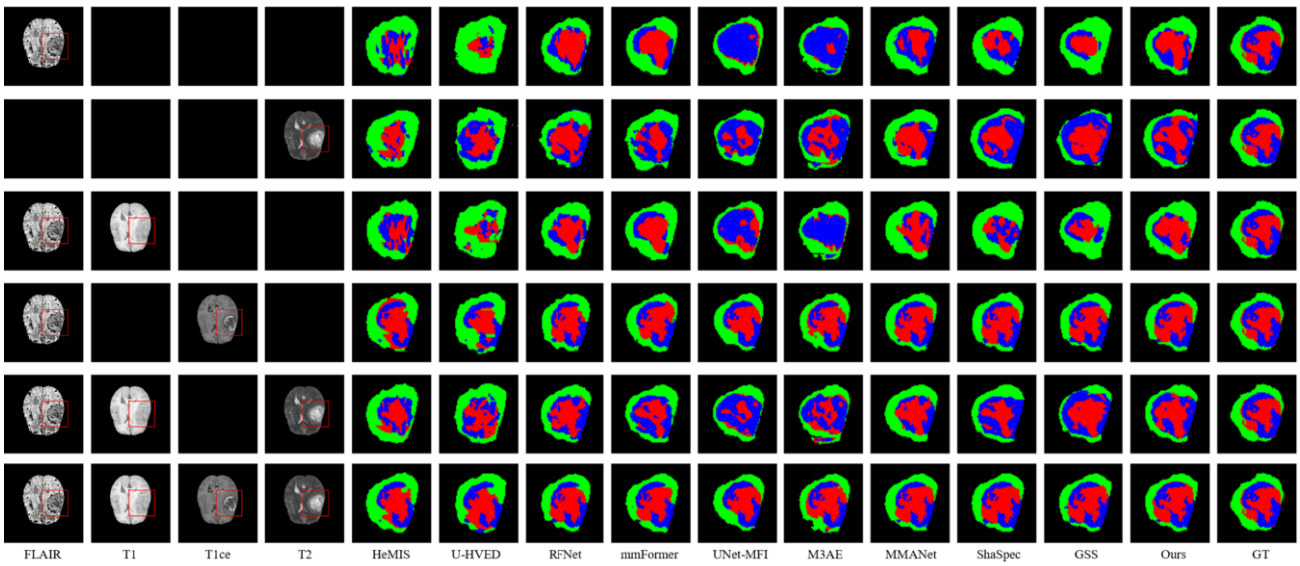

Figure 5. Segmentation results of different methods with various available modalities on BraTS2020.

We do not employ the adversarial loss [9], since the primary objective of the discriminator is to enhance synthesis diversity, while our goal is to generate the missing features with high fidelity. In this manner, we synthesize the feature intensities of missing modalities that compensate the anatomical inconsistent information.

Since we obtain the complete multi-modal features $f_{comp} \in \mathcal{R}^{B\times 4C\times D\times H\times W}$ using our MFSB, we forward f_{comp} to the decoder D_{Fuse} to derive the final segmentation results. The implementation of this decoder is similar to the 3D U-Net architecture.

3.3 The Overall Loss

In line with previous works such as [7, 36], we employ the weighted cross-entropy loss denoted as L_{WCE} and the Dice loss denoted as L_{Dice} to align our predictions with the corresponding ground-truth segmentation maps. These losses are formulated as

$$L_{Seg} = \sum_{i=1}^{N}(L_{WCE}(y_i, \hat{y}_i) + L_{Dice}(y_i, \hat{y}_i)), \tag{11}$$

where $\hat{y}_i$ and y_i are the predicted segmentation and the corresponding ground-truth, respectively. The parameter $N = 6$ signifies that we are predicting six segmentation maps. These include four segmentation maps derived from a shared auxiliary decoder $DAux$, each taking one of the four modalities as input, one segmentation map obtained from a fusion decoder D_{Fuse} that utilizes multi-modal features from the multi-modal encoder E_{Aux}, and one segmentation map generated by the shared fusion decoder D_{Fuse} using compensated features from our MFSB.

Consequently, by combining Equations 4, 10, and 11, the overall loss is expressed as

$$L_{Overall} = L_{ACCT} + L_{Seg} + L_{Syn}. \tag{12}$$

4 Experiments

4.1 Training Strategy

Our ACDIS comprises four mono-modal encoders E_m and one fusion decoder D_{Fuse}, accompanied by an additional auxiliary encoder E_{Aux} for knowledge distillation and an auxiliary decoder D_{Aux} for mono-modal segmentation. The architecture of these encoders and decoders closely resembles that of the 3D U-Net.

We conduct our experiments using the PyTorch framework version 1.13.0. The training process is carried out on two NVIDIA A800-80GB GPUs, spanning 500 epochs, with a batch size of two volumes, consuming in a total 98 GPU hours. We utilize the Adam optimizer with an initial learning rate set to 1×10^{-4} and a weight decay of 1×10^{-4}. During the training phase, the input images are randomly cropped to dimensions of $80\times80\times80$. Data augmentation techniques include random rotations, flips, and intensity adjustments.

It is worth noting that our initial training of ACDIS omitted the anatomical constraints, specifically L_{ACCT}, as well as the synthesis loss L_{Syn}. This is due to the possibility that the auxiliary encoder E_{Aux} may not provide multi-modal features with comprehensive anatomical structures, and the MFSB may encounter difficulties in constructing missing modality features solely based on the limited anatomical information contained within mono-modal features. The anatomical consistency distillation process runs concurrently with our training from the first to the final epoch, while the feature synthesis process starts from the 21st epoch onwards.

4.2 Datasets and Evaluation Metric

We preform our experiments on two datasets from the Multi-modal Brain Tumor Segmentation Challenge [23], i.e., BraTS2018 and BraTS2020, which align with that of previous studies [21, 27]. BraTS2018 and BraTS2020 include 285 and 369 cases with ground truth publicly available, respectively. For BraTS2020, consistent with [7], we randomly split it into $219 : 50 : 100$ for training, validation, and testing, respectively. For BraTS2018, we split it into $199 : 29 : 57$, and incorporate a three-fold validation.

Each case of the datasets has four different modalities, i.e., FLAIR, T1ce, T1, and T2 modalities. These modalities are characterized by a volume size of $240 \times 240 \times 155$, and they capture various properties of brain tumor subregions: GD-enhancing tumor (ET), peritumoral edema (ED), and the necrotic and non-enhancing tumor core (NCR/NET). These subregions of brain tumors are grouped into three

M																	
	FLAIR	●	○	○	○	●	●	●	○	○	○	●	●	●	○	●	AVG
	T1ce	○	●	○	○	●	○	○	●	●	○	●	●	○	●	●	
	T1	○	○	●	○	○	●	○	●	○	●	●	○	●	●	●	
	T2	○	○	○	●	○	○	●	○	●	●	○	●	●	●	●	
WT	HeMiS	71.60	67.71	68.96	68.19	69.17	68.67	69.83	69.01	69.78	69.40	70.21	71.28	70.73	71.58	72.06	69.88
	U-HVED	69.85	46.82	46.77	54.03	61.45	58.25	64.50	62.91	65.76	64.29	66.99	69.70	68.38	70.35	71.41	62.76
	RFNet	86.42	**77.34**	76.46	86.21	89.55	89.30	89.35	81.00	87.45	87.95	90.39	90.20	90.42	88.59	90.77	86.76
	UNet-MFI	82.27	73.18	72.10	82.45	83.64	84.34	84.85	77.30	83.44	83.52	85.45	85.70	85.85	84.20	84.93	82.21
	mmFormer	82.40	74.25	74.37	83.07	84.54	84.61	85.82	77.98	84.05	84.00	85.34	86.11	86.22	84.64	86.38	82.92
	MMANet	83.21	64.46	64.35	82.88	86.74	87.64	87.47	70.98	85.64	85.33	87.92	89.37	87.97	86.38	89.25	82.64
	ShaSpec	84.10	70.36	70.23	83.95	88.65	89.07	88.70	76.02	86.21	87.15	89.47	89.43	90.07	87.62	90.11	84.74
	M3AE	86.53	73.85	**76.71**	86.09	89.48	89.38	89.25	78.11	87.37	87.20	89.99	90.18	90.42	88.61	90.56	86.25
	GSS*	86.36	72.45	73.33	86.76	89.51	90.65	90.91	78.38	88.87	**90.05**	91.14	**93.25**	90.67	89.43	**93.28**	87.00
	Ours	**88.74**	76.59	76.36	**87.41**	**91.14**	**91.71**	**91.49**	**81.66**	**89.83**	89.61	**92.16**	92.40	**92.14**	**90.62**	92.82	**88.31**
TC	HeMiS	53.43	51.41	51.56	51.11	51.70	51.08	51.85	51.88	52.35	51.51	52.95	53.76	52.97	54.38	55.03	52.46
	U-HVED	34.62	35.51	27.30	37.67	42.15	38.26	43.41	44.93	47.53	44.97	49.13	51.30	49.40	52.72	54.17	43.53
	RFNet	65.04	82.37	64.31	68.47	84.69	71.45	72.62	83.15	84.06	72.11	84.71	84.70	74.28	84.11	84.74	77.39
	UNet-MFI	63.94	77.63	59.38	68.05	79.92	68.23	70.72	77.61	80.09	70.21	80.03	80.94	71.40	80.75	81.28	74.01
	mmFormer	66.19	77.96	61.17	69.18	80.36	69.58	71.55	79.93	80.79	70.90	80.18	81.31	72.02	81.12	81.22	74.90
	MMANet	65.79	74.05	57.47	70.21	85.21	72.73	72.65	80.99	85.24	71.88	85.87	85.68	74.48	**86.39**	86.63	76.99
	ShaSpec	66.06	77.34	59.57	68.29	85.02	72.87	72.95	80.84	84.46	72.28	85.83	85.57	74.74	85.47	86.50	77.19
	M3AE	68.04	81.39	**66.00**	70.27	82.01	73.82	74.95	82.39	83.01	72.54	82.44	83.06	75.09	84.06	84.40	77.56
	GSS*	67.33	78.43	61.67	70.10	84.88	73.45	74.17	82.20	86.12	**74.18**	86.50	**88.39**	74.33	86.28	**88.67**	78.45
	Ours	**71.05**	**82.61**	64.08	**72.03**	**85.84**	**75.31**	**75.00**	**84.34**	**86.34**	73.57	**86.95**	86.62	**76.43**	86.35	86.92	**79.56**
ET	HeMiS	**43.77**	42.41	**41.59**	41.45	41.83	40.29	41.19	42.08	42.39	41.00	43.67	44.16	42.95	45.27	46.33	42.69
	U-HVED	12.88	24.94	7.27	24.26	30.02	21.95	29.40	33.64	36.18	32.12	39.39	40.91	38.09	43.18	45.33	30.64
	RFNet	40.47	**74.27**	37.51	43.59	**76.45**	43.81	46.99	75.22	73.94	46.37	77.01	76.38	48.95	76.38	76.64	60.93
	UNet-MFI	39.70	69.42	29.38	**46.00**	70.13	40.06	48.69	69.25	72.32	45.71	71.28	70.88	46.55	72.00	71.41	57.52
	mmFormer	40.47	68.91	33.97	45.61	69.81	43.63	48.09	71.10	70.72	45.92	70.08	71.60	48.38	70.65	71.36	58.02
	MMANet	36.40	66.07	28.22	41.91	70.43	41.50	44.23	68.87	71.86	43.58	71.94	71.98	44.60	71.81	72.43	56.39
	ShaSpec	38.92	66.93	31.75	42.58	70.85	43.22	45.62	69.92	71.04	44.21	72.08	71.09	46.22	71.81	72.02	57.22
	M3AE	40.49	72.43	39.93	45.97	74.66	43.20	47.30	75.42	76.81	46.63	75.94	**77.08**	48.19	77.40	78.00	61.30
	GSS*	42.03	69.46	35.30	45.60	74.20	47.78	49.33	74.41	74.78	**48.94**	76.12	76.32	50.02	76.98	75.95	61.15
	Ours	41.50	73.42	36.88	45.99	76.35	**47.87**	**50.09**	**77.24**	**76.83**	48.84	**79.55**	77.07	**52.75**	**78.36**	**79.05**	**62.79**

Table 1. Performance comparison (DSC%) with SOTA methods, including HeMiS, U-HVED, RFNet, UNet-MFI, mmFormer, MMANet, ShaSpec, M3AE, and GSS on BraTS2020. Available and missing modalities are represented by ● and ○, respectively. The '*' demonstrates that its codes and results are reproduced by ourselves.

Consistency Distillation					Dice		
w/o AFEB	w/ AFEB				WT	TC	ET
	L_{MSE}	L_{KL}	L_{Var}	L_{Covar}			
✓					83.34	75.36	56.22
	✓				84.85	76.68	57.77
		✓			74.35	62.86	47.29
			✓		85.97	77.23	59.71
				✓	86.04	77.98	60.23
			✓	✓	**86.61**	**78.56**	**61.09**

Table 2. Ablation study on consistency distillation.

Inconsistency Synthesis					Dice		
w/o MFSB	w/ MFSB				WT	TC	ET
	L_{Adver}	L_{Var}	L_{Covar}	L_{Syn}			
✓					86.61	78.56	61.09
	✓				86.89	78.73	61.41
		✓	✓		87.16	78.67	61.55
				✓	**88.31**	**79.56**	**62.79**
	✓			✓	88.00	79.21	62.34
		✓	✓	✓	88.11	79.42	62.50

Table 3. Ablation study on inconsistency synthesis.

nested subregions: the whole tumor (WT), the tumor core (TC), and the enhancing tumor (ET).

Dice coefficient is adopted to evaluate our ACDIS, aligning with [7]. The metric is defined as

$$\text{Dice} = \frac{2 \cdot ||\bar{y}_{\bar{k}} \cap y_{\bar{k}}||_1}{||\bar{y}_{\bar{k}}||_1 + ||y_{\bar{k}}||_1}, \tag{13}$$

where $\bar{k}$ represents different tumor classes, i.e., WT, TC, and ET.

4.3 Comparisons with the State-of-the-art

To evaluate the effectiveness of our method, we compare our ACDIS with nine state-of-the-art methods on different cases with missing modalities.The involved methods contain HeMiS [13], U-HVED [8], RFNet [7], UNet-MFI [39], mmFormer [36], MMANet [30], ShaSpec [27], M3AE [21], and GSS [24]. For a fair comparison, all methods are trained under their recommended hyper-parameters within the same dataset split.

As illustrated in Tab. 1, our method achieves preferable results for most combinations of missing modalities. We achieve improvements of 1.3%, 1.1%, and 1.5% over the second-ranked method on the average DSC for WT, TC, and ET. We also provide a visualization comparison in Figure 5, illustrating that our method yields more accurate segmentation results in different combinations of modalities. *More results can be found in the appendix of [37].*

4.4 Ablation Study

We evaluate the proposed components on BraTS2020, employing the average DSC to measure the performance. To maintain the invariance of parameter counts, when removing the AFEB and MFSB blocks, we replace them with $3 \times 3 \times 3$ convolutions in their respective positions, ensuring a fair comparison.

Effect of consistency distillation. We conduct experiments within our framework, excluding MFSB to evaluate the effectiveness of consistency distillation. Without AFEB, the model achieves a Dice score of 83.34% for WT, 75.36% for TC, and 56.22% for ET, serving as a baseline for comparison. Then, we cooperate it with multiple distillation losses, including the mean squared error loss

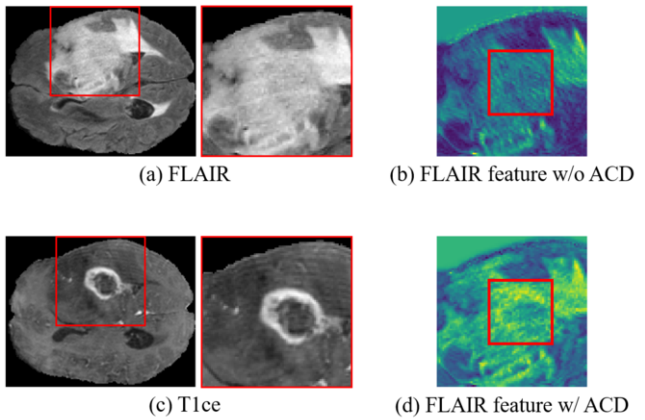

Figure 6. Feature Visualization: (a) depicts the input image of the FLAIR modality. (b) illustrates the average feature derived from our model without our ACD. (c) showcases the input image of the T1ce modality. Finally, (d) denotes the average feature obtained from our model with ACD, highlighting that the FLAIR features have effectively assimilated characteristics of the T1ce modality through the use of ACD.

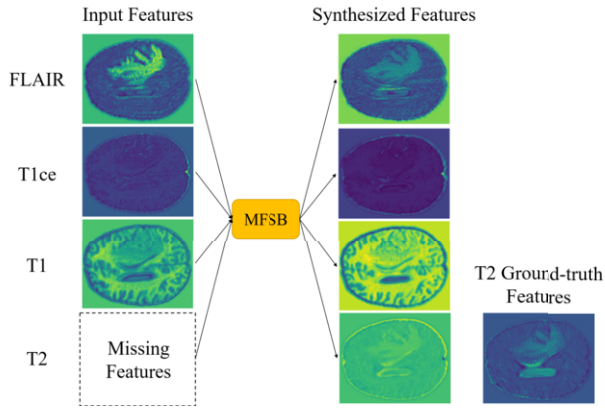

Figure 7. Feature synthesis process of our MFSB. With the available modalities, containing FLAIR, T1ce, and T1, our MFSB effectively synthesizes the features of the missing modality, i.e., T2 modality. T2 ground-truth features are obtained when complete modalities are provided.

(L_{MSE}), Kullback-Leibler divergence loss (L_{KL}), our proposed variance loss (L_{Var}), and our covariance loss (L_{Covar}), to constrain the knowledge transferred from multi-modal representations to individual mono-modal representations. As illustrated in Table 2, our ACCT-based approach consistently yielded the most favourable segmentation results. This can be attributed to the specific focus of each loss function. Both L_{MSE} and L_{KL} primarily target adjustments in pixel-wise intensity and intensity distributions, respectively. These adjustments standardize intensity values across different mono-modal features, which can inadvertently suppress modality-specific information that is critical for accurate segmentation. In contrast, our approach enhances segmentation by enriching the local information richness through the variance loss (L_{Var}) and promoting anatomical structural similarity through the covariance loss (L_{Covar}). These tailored losses improve the consistency between mono-modal and multi-modal features without diminishing the mono-modal inconsistent characteristics.

We provide a visualization in Figure 6 to illustrate the effectiveness of our ACD. By comparing Figure 6 (b) and Figure 6 (d), it is evident that the FLAIR features, with the application of our ACD, are able to capture anatomical structures that are less distinct in the FLAIR modality alone but are prominent in the T1ce modality. This capability demonstrates that, even in the absence of the T1ce modality, our network can successfully extract and highlight these subtle yet consistent anatomical structures from the FLAIR modality alone.

Effect of inconsistency synthesis. We further investigate the effectiveness of inconsistency synthesis within our framework, which incorporates ACD. The comparative results are presented in Table 3. In the absence of the Modality Feature Synthesis Block (MFSB), which is designed for synthesizing features of absent modalities, our model achieves Dice scores of 86.61% for WT, 78.56% for TC, and 61.09% for ET. When integrating MFSB alongside other losses, including adversarial loss (L_{Adver}), variance loss (L_{Var}), covariance loss (L_{Covar}), and mean squared error synthesis loss (L_{Syn}), our model demonstrates significantly improved performance. Notably, the highest performance is observed when employing L_{Syn}, which specifically encourages the MFSB to generate modality-specific features with distinct intensity characteristics that are critical for inconsistency synthesis. However, the combination of L_{Adver} with L_{Syn} results in a performance decline, primarily due to a decrease in data fidelity.

To illustrate our MFSB's feature synthesis process (Figure 7), we demonstrate its capability to synthesize missing T2 modality features. FLAIR, T1ce, and T1 modalities undergo encoding via mono-modal encoders and our ACD, resulting in 'Input Features' obtained by averaging features across channels. Our MFSB learns modality-specific affine transformations and applies non-linear activation to synthesize missing features ('Synthesized Features'). Remarkably, the synthesized T2 features resemble ground-truth counterparts, distinguished by distinct colors from the other mono-modal features indicating captured anatomical inconsistencies, i.e., modality-specific pixel values.

5 Conclusion

In this study, we introduce the Anatomical Consistency Distillation and Inconsistency Synthesis (ACDIS) framework, a novel approach for brain tumor segmentation in cases of missing modalities. ACDIS is composed of two primary components: Anatomical Consistency Distillation (ACD) and the Modality Feature Synthesis Block (MFSB). In ACD, the Anatomical Feature Enhancement Block (AFEB) effectively models the relationship between mono-modal and multi-modal representations, cooperating with Anatomical Consistency ConsTraints (ACCT), which transfer rich multi-modal knowledge and anatomical structural similarities to boost mono-modal representation. Concurrently, MFSB addresses the challenge of inconsistent anatomical information in missing modalities by generating comprehensive modality-specific features from available mono-modal representations. By leveraging both anatomical consistency and inconsistency, ACDIS demonstrates superior segmentation performance for the task of brain tumor segmentation with missing modalities, as validated on BraTS2020 and BraTS2018.

Acknowledgement

This work was in part supported by the National Natural Science Foundation of China under grants 62032006 and 62021001.

References

[1] K. Akeret, M. Weller, and N. Krayenbühl. The anatomy of neuroepithelial tumours. *Brain*, page awad138, 2023.

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

[3] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *Computer Vision–ECCV 2022 Workshops*, pages 205–218. Springer, 2023.

[4] C. Chen, Q. Dou, Y. Jin, H. Chen, J. Qin, and P.-A. Heng. Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI*, pages 447–456. Springer, 2019.

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

[6] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI*, pages 424–432. Springer, 2016.

[7] Y. Ding, X. Yu, and Y. Yang. Rfnet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3975–3984, 2021.

[8] R. Dorent, S. Joutard, M. Modat, S. Ourselin, and T. Vercauteren. Hetero-modal variational encoder-decoder for joint modality completion and segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI*, pages 74–82. Springer, 2019.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[10] S. Gupta, J. Hoffman, and J. Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2827–2836, 2016.

[11] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event*, pages 272–284. Springer, 2022.

[12] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022.

[13] M. Havaei, N. Guizard, N. Chapados, and Y. Bengio. Hemis: Hetero-modal image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI*, pages 469–477. Springer, 2016.

[14] M. Hu, M. Maillard, Y. Zhang, T. Ciceri, G. La Barbera, I. Bloch, and P. Gori. Knowledge distillation from multi-modal to mono-modal segmentation networks. In *Medical Image Computing and Computer Assisted Intervention–MICCAI*, pages 772–781. Springer, 2020.

[15] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

[16] Z. Jiang, C. Ding, M. Liu, and D. Tao. Two-stage cascaded u-net: 1st place solution to brats challenge 2019 segmentation task. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 231–241. Springer, 2020.

[17] J. Kim and H. Park. Adaptive latent diffusion model for 3d medical image to image translation: Multi-modal magnetic resonance imaging study. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7604–7613, 2024.

[18] A. Konwer, X. Hu, J. Bae, X. Xu, C. Chen, and P. Prasanna. Enhancing modality-agnostic representations via meta-learning for brain tumor segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21415–21425, 2023.

[19] D. Lee, W.-J. Moon, and J. C. Ye. Assessing the importance of magnetic resonance contrasts using collaborative generative adversarial networks. *Nature Machine Intelligence*, 2(1):34–42, 2020.

[20] J. Li, W. Wang, C. Chen, T. Zhang, S. Zha, H. Yu, and J. Wang. Transbtsv2: Wider instead of deeper transformer for medical image segmentation. *arXiv preprint arXiv:2201.12785*, 2022.

[21] H. Liu, D. Wei, D. Lu, J. Sun, L. Wang, and Y. Zheng. M3ae: Multimodal representation learning for brain tumor segmentation with missing modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1657–1665, 2023.

[22] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[23] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

[24] Y. Qiu, D. Chen, H. Yao, Y. Xu, and Z. Wang. Scratch each other's back: Incomplete multi-modal brain tumor segmentation via category aware group self-support learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21317–21326, 2023.

[25] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[26] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI*, pages 234–241. Springer, 2015.

[27] H. Wang, Y. Chen, C. Ma, J. Avery, L. Hull, and G. Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15878–15887, 2023.

[28] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li. Transbts: Multimodal brain tumor segmentation using transformer. In *Medical Image Computing and Computer Assisted Intervention–MICCAI*, pages 109–119. Springer, 2021.

[29] Y. Wang, Y. Zhang, Y. Liu, Z. Lin, J. Tian, C. Zhong, Z. Shi, J. Fan, and Z. He. Acn: adversarial co-training network for brain tumor segmentation with missing modalities. In *Medical Image Computing and Computer Assisted Intervention–MICCAI*, pages 410–420. Springer, 2021.

[30] S. Wei, C. Luo, and Y. Luo. Mmanet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20039–20049, 2023.

[31] H. Yang, J. Sun, and Z. Xu. Learning unified hyper-network for multi-modal mr image synthesis and tumor segmentation with missing modalities. *IEEE Transactions on Medical Imaging*, 2023.

[32] Q. Yang, X. Guo, Z. Chen, P. Y. Woo, and Y. Yuan. D 2-net: Dual disentanglement network for brain tumor segmentation with missing modalities. *IEEE Transactions on Medical Imaging*, 41(10):2953–2964, 2022.

[33] Z. Yang and S. Farsiu. Directional connectivity-based segmentation of medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11525–11535, 2023.

[34] Z. Yu, X. Han, S. Zhang, J. Feng, T. Peng, and X.-Y. Zhang. Mousegan++: Unsupervised disentanglement and contrastive representation for multiple mri modalities synthesis and structural segmentation of mouse brain. *IEEE Transactions on Medical Imaging*, 2022.

[35] Y. Zhang, H. Liu, and Q. Hu. Transfuse: Fusing transformers and cnns for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI*, pages 14–24. Springer, 2021.

[36] Y. Zhang, N. He, J. Yang, Y. Li, D. Wei, Y. Huang, Y. Zhang, Z. He, and Y. Zheng. mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In *MICCAI*, pages 107–117. Springer, 2022.

[37] Z. Zhang, X. Liu, Z. Chen, Y. Zhang, H. Yue, Y. Ou, and X. Sun. Anatomical consistency distillation and inconsistency synthesis for brain tumor segmentation with missing modalities. *arXiv preprint arXiv:2408.13733*, 2024.

[38] Z. Zhang, G. Yang, Y. Zhang, H. Yue, A. Liu, Y. Ou, J. Gong, and X. Sun. Tmformer: Token merging transformer for brain tumor segmentation with missing modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7414–7422, 2024.

[39] Z. Zhao, H. Yang, and J. Sun. Modality-adaptive feature interaction for brain tumor segmentation with missing modalities. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 183–192. Springer, 2022.

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

[41] T. Zhou, H. Fu, G. Chen, J. Shen, and L. Shao. Hi-net: hybrid-fusion network for multi-modal mr image synthesis. *IEEE Transactions on Medical Imaging*, 39(9):2772–2781, 2020.

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

[43] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, 39(6):1856–1867, 2019.