D^2-Net: Dual Disentanglement Network for Brain Tumor Segmentation With Missing Modalities

Qiushi Yang, Xiaoqing Guo, *Graduate Student Member, IEEE*, Zhen Chen, *Graduate Student Member, IEEE*, Peter Y. M. Woo, and Yixuan Yuan, *Member, IEEE*

***Abstract*— Multi-modal Magnetic Resonance Imaging (MRI) can provide complementary information for automatic brain tumor segmentation, which is crucial for diagnosis and prognosis. While missing modality data is common in clinical practice and it can result in the collapse of most previous methods relying on complete modality data. Current state-of-the-art approaches cope with the situations of missing modalities by fusing multi-modal images and features to learn shared representations of tumor regions, which often ignore explicitly capturing the correlations among modalities and tumor regions. Inspired by the fact that modality information plays distinct roles to segment different tumor regions, we aim to explicitly exploit the correlations among various modality-specific information and tumor-specific knowledge for segmentation. To this end, we propose a Dual Disentanglement Network (D^2-Net) for brain tumor segmentation with missing modalities, which consists of a *modality disentanglement stage* (MD-Stage) and a *tumor-region disentanglement stage* (TD-Stage). In the MD-Stage, a spatial-frequency joint modality contrastive learning scheme is designed to directly decouple the modality-specific information from MRI data. To decompose tumor-specific representations and extract discriminative holistic features, we propose an affinity-guided dense tumor-region knowledge distillation mechanism in the TD-Stage through aligning the features of a disentangled binary teacher network with a holistic student network. By explicitly discovering relations among modalities and tumor regions, our model can learn sufficient information for segmentation even if some modalities are missing. Extensive experiments on the public BraTS-2018 database demonstrate the superiority of our framework over state-of-the-art methods in missing modalities situations. Codes are available at https://github.com/CityU-AIM-Group/D2Net.**

***Index Terms*— Contrastive learning, knowledge distillation, modality disentanglement, missing modalities.**

I. Introduction

BRAIN tumor segmentation is of great importance for the diagnosis and prognosis of gliomas [1]–[3]. Magnetic Resonance Imaging (MRI) with various acquisition parameters is the primary imaging device to diagnose tumors since it can visualize soft-issue lesions with superior resolution. In recent years, numerous automatic brain tumor segmentation algorithms [1], [2], [4]–[7] have been proposed to reduce the labor-intensive process of manual segmentation with encouraging results. In daily neurosurgical practice, only a limited number of MRI modalities are performed per scan to sufficiently derive the diagnosis, and comprehensive scanning is often not performed due to the restricted accessibility of MRI in resource-limited public health institutions. However, most prior methods [1], [2], [4]–[6] for brain tumor segmentation require a complete set of MRI data and can be difficult to apply in the real-world setting where missing modalities are often encountered. Therefore, there is an urgent demand to develop a robust segmentation model to address this issue.

Recently, extensive arts [8]–[19] attempt to handle the situations of missing modalities in medical image segmentation and they can generally be broadly grouped into three lines. *Multi-modal fusion* based algorithms [8], and [9] learn features only with all available sets of modalities, neglecting massive useful information from the missing modality data. *Modality generation* based methods [12] [10], [11], [20] adopt generative models to synthesize missing modalities, and then utilize the complete modalities for model optimization. But most of these modality generation based methods are limited in synthesizing desirable data with clear clinical correlations. More recently, *shared knowledge extraction* based approaches [13]–[18], [21]–[24], aiming to learn a latent feature space that is robust against the number of available modalities, have gained popularity. Instead of synthesizing MRI images directly, this strategy extracts shared representations towards tumor segmentation by learning correlations among inherent multi-modal features [13], [15]–[17], [22], [23] or statistical information [14], [21]. The extracted shared features can represent modality-robust information for brain tumor segmentation and perform well even if some modalities are missing.

Despite the remarkable performance improvements of *shared knowledge extraction* based schemes [14], [15], [17], [18], [21]–[23], they capture the correlations of multi-modal features without explicit modality-aware supervision, and they are hard to obtain robust knowledge towards situations of missing modalities. Additionally, in the real-world process of clinical diagnosis, clinicians jointly utilize several MRI images to recognize specific tumor regions, demonstrating the impor-

Manuscript received 24 March 2022; accepted 10 May 2022. Date of publication 16 May 2022; date of current version 30 September 2022. This work was supported by the Innovation and Technology Commission Innovation and Technology Fund ITS/100/20 under Grant CityU 9440276. *(Corresponding author: Yixuan Yuan.)*

Qiushi Yang, Xiaoqing Guo, Zhen Chen, and Yixuan Yuan are with the Department of Electrical Engineering, City University of Hong Kong, Hong Kong, SAR, China (e-mail: qsyang2-c@my.cityu.edu.hk; xqguo.ee@my.cityu.edu.hk; zchen.ee@my.cityu.edu.hk; yxyuan.ee@cityu.edu.hk).

Peter Y. M. Woo is with the Department of Neurosurgery, Kwong Wah Hospital, Hong Kong, SAR, China (e-mail: wym307@ha.org.hk).

Digital Object Identifier 10.1109/TMI.2022.3175478

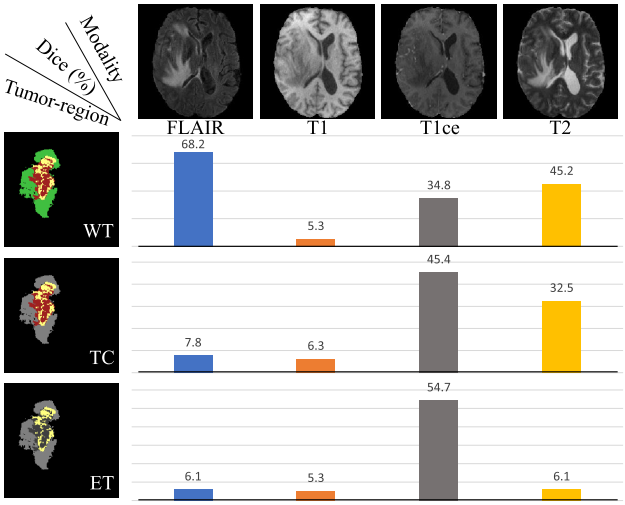

Fig. 1. Comparison of segmentation performance of three tumor-region categories with each modality. Green, yellow and red of segmentation maps refer to peritumoral edema, enhancing tumor, non-enhancing tumor and necrosis, respectively.

tance of multi-modalities for tumor-region segmentation [13], [15]. In Fig. 1, we empirically illustrate the performance comparison of three tumor-region categories with different modalities using a U^2-Net [25]. Specifically, the whole tumor (WT) region is highly distinguishable with FLAIR and T2 modalities, and the enhancing tumor (ET) region exhibits clear structures in T1ce. Hence, it is necessary to disentangle the modality-specific information and content-preserved features from MRI images, and further decouple the features into multiple tumor-specific features. The former disentanglement enables the model to learn relations among modalities, and the later captures task-oriented correlations between modalities and tumor regions. In this way, the model can achieve robust brain tumor segmentation with missing modalities.

Previous methods [13], [16] disentangle modality-specific and modality-independent features via simple auto-encoder networks with reconstruction constraint or adversarial training to deal with the missing modality issues, however, their performances are inadequate to meet the clinical demands. A rational reason is that these commonly used reconstruction constraints may derive degenerate solutions and fail to decompose features well, while adversarial training based models [10]–[12], [20] lack explicit guidance of modality-specific information. Recently, contrastive learning based approaches [26]–[29] have shown state-of-the-art performances in many image recognition tasks. But their applications in medical image segmentation tasks, especially brain tumor segmentation with missing modalities, have yet to be further explored. Moreover, existing disentanglement based methods [13], [16] only consider spatial domain, while frequency domain as another view preserves low-level statistics [30]–[32] and modality-specific information [32]. We therefore propose a novel spatial-frequency jointly modality contrastive (SFMC) learning scheme in the modality disentanglement stage (MD-Stage) to explicitly capture modality-specific information in a self-supervised learning strategy. The proposed SFMC learning scheme can provide explicit supervision towards modality-specific information with joint spatial and frequency domains as a multi-view learning strategy. Furthermore, it can enable the model to obtain modality-invariant content-preserved features that are robust against situations of missing modalities.

In order to decompose tumor-specific features and obtain holistic representations for segmentation, an intuitive solution is to design a novel tumor-region knowledge distillation mechanism with the separated tumor masks as explicit disentanglement supervisions. The knowledge distillation includes a holistic tumor-region branch as the student for all categories of tumor segmentation and a set of disentangled binary branches as the teacher for each tumor category. However, previous distillation based models [33]–[36] usually distill features of student and teacher at manual selected layers, ignoring the distinct contributions for segmentation of multi-level features as well as some potential effective distillation links among features and logits. To address this issue, we propose an affinity-guided dense tumor-region knowledge distillation (ADT-KD) mechanism in the tumor-region disentanglement stage (TD-Stage) to learn holistic and tumor-specific features in a mutual learning way. It exploits the dense and effective links to distill knowledge by learning the relative affinities between features, then controls the distillation intensities of all layer pairs in both the feature- and logit-levels to improve the ability of information communication and segmentation performance.

In this paper, we propose a dual disentanglement framework (D^2-Net) consisting of the MD-Stage and TD-Stage to handle the issue of missing modalities in brain tumor segmentation. The MD-Stage helps the model explicitly exploit the correlations among modality-specific information by decomposing them from MRI images, and the TD-Stage produces decoupled tumor-specific knowledge unrelated to MRI modalities, which can address the missing modality issue. In summary, this work includes three major contributions:

- In contrast to prior arts that implicitly learn modality relations among multi-modal MRI images in the feature space, we propose a novel SFMC learning scheme in the MD-Stage to explicitly decouple the modality-specific information from MRI images and exploit the correlations among modality representations in a content-independent modality space.
- We represent the first effort to decouple features into multiple individual tumor-specific features in the TD-Stage. To provide tumor-specific supervision and enable accurate tumor-region disentanglement, an ADT-KD mechanism is designed to align the features of a disentangled binary teacher with a holistic student networks.
- We conduct extensive experiments on a challenging brain tumor segmentation dataset BraTS-2018 and demonstrate a clear advantage of the proposed framework over state-of-the-art methods in missing modalities situations.

II. Related Work

A. Learning With Missing Modalities

In recent years, various methods for brain tumor segmentation with missing modalities have been investigated [8]–[19],

[21], [22]. Generally, they can be grouped into three lines. The first line is based on *multi-modal fusion* approaches [8], [9], which simply employ feature fusion strategies to aggregate features of all available sets of modalities to obtain fused representations for segmentation. Shen *et al*. [8] introduce a channel-independent encoder and a feature fusion module to learn fused features in an input robust way. Karin *et al*. [9] adopt a fusion layer at the end of the network to combine features from all available modalities. Although they are easy to implement, they ignore information from missing modalities and the potential correlations among all modalities. The second category, *modality generation* based methods [10]–[12], [20], aims to synthesize missing modalities and then leverages complete modalities as inputs to perform segmentation tasks. MGM-GAN [20] utilizes a multi-scale gate emergence-based generative adversarial network (GAN) to synthesize the missing modalities of MRI from available modalities. Hi-Net [11] learns a mapping from existing modalities to missing modalities by exploiting the correlations among multiple modalities. These methods are computationally cumbersome requiring extra networks for synthesis. Since the optimization objective of generation and segmentation tasks are not coordinated, the synthesized data may affect segmentation performance.

Shared knowledge extraction based strategies [13]–[18], [21], [22] as another line target to learn a shared feature representation, i.e., modality-invariant features, by removing modality-specific information. HeMIS [14] computes the first and second moments of feature statistics of each modality, and integrates them for segmentation. Chen *et al*. [13] leverage a disentanglement framework to learn the content and appearance codes of MRI data by a simple auto-encoder network and then fuse shared content codes to segment tumor regions. Some recent arts [22], [23] propose "dedicated" knowledge distillation schemes to learn mono-modal knowledge from a multi-modal model, while they need to train one specialized model for each case with missing modalities, which is considered cost-ineffective for multiple cases. The latest algorithm [15] presents a correlation module to capture the latent multi-source relations and then transforms all available features to latent representations for fusion with an attention module for segmentation.

The *shared knowledge extraction* based methods are simple yet able to model correlations among multi-modal features. However, most of them suffer from the unstable training problem and unsatisfactory feature disentanglement. In this work, we propose a contrastive learning based scheme in both the spatial and frequency domains to explicitly and stably decompose modality-specific information, and introduce a novel affinity-guided dense knowledge distillation mechanism to explicitly learn correlations among tumor-regions and modalities simultaneously.

B. Disentanglement Representation Learning

Disentanglement representation learning [12], [37]–[41] has been commonly studied in unsupervised learning, and it enables deep learning models to be easily interpretable. Recent disentanglement strategies are generally divided into two directions.

The first direction is to model information in latent space utilizing GAN models. GAN-based models reply on the hypothesis that content and modality information could be encoded in deep generative architectures. MixNMatch [12] decomposes object pose, shape, texture and background information of images by leveraging adversarial image and feature distributions aligned to learn the latent attribution encoders. InfoGAN [37] learns decoupled representations by maximizing the mutual information between latent variables and generated images. These approaches are efficient, but are difficult to converge and sensitive to hyper-parameters. An alternative direction is to learn statistical independent latent variables in the embedded space based on an encoder-decoder framework. Cheng *et al*. [38] adopt a Variational Auto-Encoder (VAE) to decompose the shape, appearance and background of images in a hierarchical manner. The β-VAE [41] adjusts the prior matching term of KL divergence with a coefficient β with an adaptive module to learn disentangled information.

Nevertheless, these methods [12], [37]–[41] lack explicit supervision with respect to modality information, and their performance may be unsatisfactory. Moreover, they merely employs unsupervised learning schemes to learn latent disentanglement representation without any task-oriented supervision thus struggling to capture desired information. In contrast, the proposed contrastive loss leverages the inherent characteristic of multi-modal inputs that all modalities are aligned, which can guide the model to learn modality-specific information in a self-supervised manner.

III. Method

In this work, we proposed a dual disentanglement network (D^2-Net) for brain tumor segmentation with missing modalities, as illustrated in Fig. 2. It consists of a modality disentanglement stage (MD-Stage) for modality-specific information decoupling and a tumor-region disentanglement stage (TD-Stage) for tumor-specific knowledge separation.

In the MD-Stage, D^2-Net takes the MRI image $\boldsymbol{x}$ with multi-modalities as inputs to obtain fused tumor feature $\boldsymbol{z}$ via the tumor encoder E_T. Meanwhile, each modal image is separately input the corresponding modality encoders E_M with a global average pooling (GAP) layer to produce the modality-specific codes $\boldsymbol{c}$ under the constraint of subtly designed spatial-frequency jointly modality contrastive (SFMC) learning scheme. The disentangled modality-specific codes $\boldsymbol{c}$ are then transferred to desired tumor-specific codes $\hat{\boldsymbol{c}}$ via a tumor modality projection network E_{TM} to discover the correlations among modalities and tumor-region features in the latent space. In the TD-Stage, the modality-content reconstruction network E_R transfers the tumor feature $\boldsymbol{z}$ and tumor-specific codes $\hat{\boldsymbol{c}}$ into tumor-specific features $\boldsymbol{F}$. The tumor-region features $\boldsymbol{F}$ are decoupled via a disentangled binary decoder D^T and then boost the holistic features through a novel affinity-guided dense tumor-region knowledge distillation (ADT-KD) mechanism with a holistic multi-class tumor-region decoder D^S and the disentangled binary decoder D^T. Finally, the brain tumor segmentation is obtained by holistic tumor-region student D^S. The overall

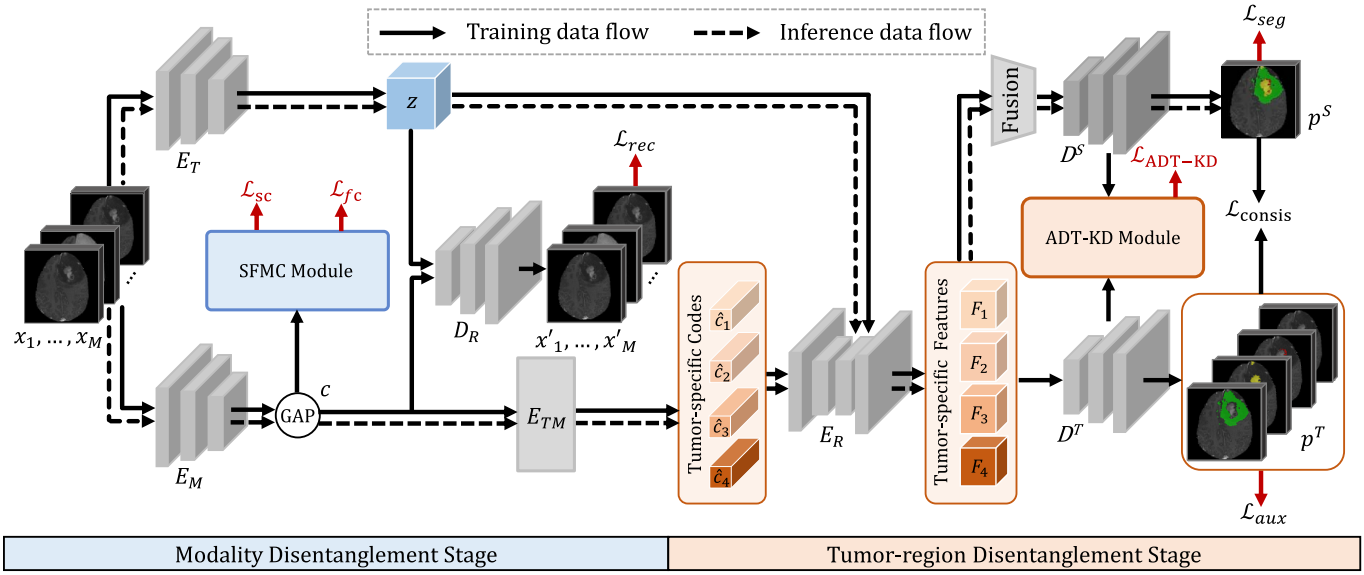

Fig. 2. Overview of the proposed dual disentanglement network (D^2-Net). **Left**: Modality disentanglement stage with the spatial-frequency jointly modality contrastive (SFMC) module. **Right**: Tumor-region disentanglement stage with the affinity-guided dense tumor-region knowledge distillation (ADT-KD) module.

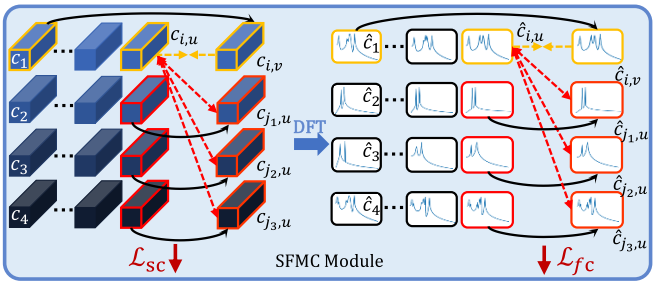

Fig. 3. The illustration of spatial-frequency jointly modality contrastive (SFMC) module. In the spatial domain, it calculates the spatial contrastive loss (left) and in the frequency domain, it conducts the frequency contrastive loss (right).

network is optimized by joint loss functions including the proposed SFMC loss ($\mathcal{L}_{\text{SFMC}}$), ADT-KD loss ($\mathcal{L}_{\text{ADT-KD}}$), one reconstruction loss ($\mathcal{L}_{\text{rec}}$), one consistency loss ($\mathcal{L}_{\text{consis}}$), one binary cross-entropy loss ($\mathcal{L}_{\text{aux}}$) and one standard segmentation loss ($\mathcal{L}_{\text{seg}}$).

A. Modality Disentanglement Stage (MD-Stage)

In the MD-Stage, we propose SFMC learning scheme to decompose modality-specific codes from MRI images in the spatial and frequency domain jointly, as illustrated in Fig. 3. To the best of our knowledge, this represents the first effort to leverage spatial and frequency contrastive learning to decouple the modality-specific information from MRI images for brain tumor segmentation.

Concretely, given the multi-modal 3D MRI images $\boldsymbol{x} = \{\boldsymbol{x}_m\}_{m=1}^{M}, \boldsymbol{x} \in \mathbb{R}^{M\times G\times H\times W}$, where M is the number of modalities, G is the number of slices, H and W are the height and width of inputs, we employ the tumor encoder E_T to obtain the coarse tumor feature $\boldsymbol{z} = E_T(\boldsymbol{x})$. Simultaneously, a set of modality encoders $E_M = \{E_M^1, \ldots, E_M^M\}$ are adopted to extract the modality-specific codes for every modality MRI image $\boldsymbol{c} = \{\boldsymbol{c}_{m,g}\}_{m,d=1}^{M,G} = \{E_M^m(\boldsymbol{x}_m)\}_{m=1}^{M}, \boldsymbol{c} \in \mathbb{R}^{M\times G\times N}$, where N is the dimension of modality-specific codes, for each slice of the MRI images.

1) SFMC: For *spatial modality contrastive learning* in the MD-Stage, our target is to guarantee the obtained modality-specific codes from the same modality are similar and codes from different modalities are distinct. Specifically, given an anchor modality-specific code $\boldsymbol{c}_{i,u}$, with i, u indicating the index of modalities and slices, respectively, we regard codes $\boldsymbol{c}_{i,v,v\neq u}$ belonging to the same modality but different slices as positive ones, and codes $\boldsymbol{c}_{j,u,j\neq i}$ belonging to other modalities whereas the same slice as negative ones. To restrain the anchor codes and positive/negative codes via measuring their similarities, the spatial modality contrastive learning loss is formulated as follows:

$$\mathcal{L}_{\text{sc}}^{i,u,v} = -\log \frac{h_\theta\left(\boldsymbol{c}_{i,u}, \boldsymbol{c}_{i,v}\right)}{h_\theta\left(\boldsymbol{c}_{i,u}, \boldsymbol{c}_{i,v}\right) + \sum_{j=1}^{M} \mathbb{1}_{[j\neq i]} h_\theta\left(\boldsymbol{c}_{i,u}, \boldsymbol{c}_{j,u}\right)}, \tag{1}$$

where $\mathbb{1}_{[j\neq i]} \in \{0, 1\}$ is an indicator function evaluating to 1 if $j \neq i$, and $h_\theta(\cdot,\cdot)$ denotes the affinity metric function and we adopt exponential cosine similarity τ as: $h_\theta(\boldsymbol{p}, \boldsymbol{q}) = \exp\left(\frac{\boldsymbol{p}\cdot\boldsymbol{q}}{\|\boldsymbol{p}\|\cdot\|\boldsymbol{q}\|}\cdot\frac{1}{\tau}\right)$, where τ is the temperature factor facilitating the model to learn from hard negatives. Optimized by this contrastive loss, the model can directly decouple the modality-specific codes from MRI images in the spatial domain.

Considering that the amplitude information in the frequency domain retains low-level statistics related to modality information [30], [42], we further employ the *frequency modality contrastive learning* scheme in the frequency domain to enhance the disentanglement performance. To be precise, we firstly convert modality-specific codes $\boldsymbol{c}$ into frequency domain to get frequency spectrum via Discrete Fourier Transform (DFT):

$$\mathcal{F}(\boldsymbol{c})(k) = \sum_{n=0}^{N'-1} \boldsymbol{c}(n) e^{-j2\pi\left(\frac{n}{N'}k\right)}, \tag{2}$$

where $\mathcal{F}(\boldsymbol{c})(k)$ represents the k-th harmonics of the frequency spectrum $\mathcal{F}(\boldsymbol{c})$. Then, we obtain the discrete

Fourier amplitude spectrum:

$$\boldsymbol{c}^f(k) = \left[R^2(\boldsymbol{c})(k) + I^2(\boldsymbol{c})(k)\right]^{1/2}, \tag{3}$$

where $R(\boldsymbol{c})$ and $I(\boldsymbol{c})$ denote the real and imaginary part of the frequency amplitude spectrum, respectively. In the multi-modal cases, $\boldsymbol{c}^f = \{\boldsymbol{c}^f_m\}_{m=1}^M, \boldsymbol{c}^f \in \mathbb{R}^{M\times G\times N'}$, where N' denotes the dimension of each frequency-wise representation and $\boldsymbol{c}^f$ represents the intensity of each frequency component.

Then, we conduct the modality contrastive learning scheme in the frequency domain to pull the intra-modal modality spectrum and push the inter-modal modality spectrum. Given a modality spectrum $\boldsymbol{c}^f_{i,u}$ as the anchor, modality spectrum belonging to the same modality but different slices are regarded as positive ones $\boldsymbol{c}^f_{i,v,v\neq u}$, and modality spectrum belonging to other modalities whereas the same slice are regarded as negative ones $\boldsymbol{c}^f_{j,u,j\neq i}$, the formulation of the frequency contrastive loss is:

$$\mathcal{L}^{i,u,v}_{\text{fc}} = -\log \frac{h_\theta\left(\boldsymbol{c}^f_{i,u}, \boldsymbol{c}^f_{i,v}\right)}{h_\theta\left(\boldsymbol{c}^f_{i,u}, \boldsymbol{c}^f_{i,v}\right) + \sum_{j=1}^M \mathbb{1}_{[j\neq i]} h_\theta\left(\boldsymbol{c}^f_{i,u}, \boldsymbol{c}^f_{j,u}\right)}, \tag{4}$$

where $\boldsymbol{c}^f_{i,u}, \boldsymbol{c}^f_{i,v}$ and $\boldsymbol{c}^f_{j,u}$ are anchor, positive and negative frequency spectrum, respectively.

For the brain tumor segmentation with M modalities and G slices each modality as MRI inputs, the SFMC loss $\mathcal{L}_{\text{SFMC}}$ is the sum of spatial and frequency contrastive loss and can be formulated as:

$$\mathcal{L}_{\text{SFMC}} = \frac{1}{MG^2}\sum_{i=1}^{M}\sum_{u=1}^{G}\sum_{v=1}^{G}\mathbb{1}_{[v\neq u]}(\mathcal{L}^{i,u,v}_{\text{sc}} + \mathcal{L}^{i,u,v}_{\text{fc}}). \tag{5}$$

To ensure the tumor feature $\boldsymbol{z}$ retain intact content information of MRI images, a set of reconstruction decoders $D_R = \{D_{R_1}, \ldots, D_{R_M}\}$ are adopted to incorporate modality-specific codes $\boldsymbol{c}$ with tumor feature $\boldsymbol{z}$ to reconstruct MRI data $\boldsymbol{x}' = D_R(\boldsymbol{c}, \boldsymbol{z})$, optimized by the reconstruction loss $\mathcal{L}_{\text{rec}}(\boldsymbol{x}, \boldsymbol{x}') = \left\|\boldsymbol{x} - \boldsymbol{x}'\right\|_2$ with pixel-wise mean square error as metric between MRI images $\boldsymbol{x}$ and the reconstructed images $\boldsymbol{x}'$. Hence, the loss function in the MD-Stage can be formulated as

$$\mathcal{L}_{\text{modal-disen}} = \mathcal{L}_{\text{SFMC}} + \mathcal{L}_{\text{rec}}, \tag{6}$$

which enables the model to disentangle the modality-specific information and obtain robust content-preserved features stably. In such a way, D^2-Net can extract shared features for segmentation even if some modalities are missing. Then, it can lead to the projection from original MRI modalities to tumor-region aware modalities.

2) Tumor Modality Transfer: Prior studies [43], [44] suggest that neural networks can encode the original modality-specific codes $\boldsymbol{c}$ and tumor-specific codes $\hat{\boldsymbol{c}}$ into a linear latent modality space. Therefore, we propose a tumor modality projection network E_{TM} to improve the model robustness and learn segmentation task-oriented tumor-specific modality information. E_{TM} transfers the original modality-specific codes $\boldsymbol{c}$ into the tumor-specific codes $\hat{\boldsymbol{c}} = \{\hat{\boldsymbol{c}}_{i,j}\}_{i,j=1}^{C,G}, \hat{\boldsymbol{c}} \in \mathbb{R}^{C\times G\times N}$ by conducting a linear transformation, where C is the number of segmentation categories: $\hat{\boldsymbol{c}} = E_{TM}(\boldsymbol{c})$. E_{TM} processes all modality codes together by fusing them across multiple modalities and it is guided by tumor-specific features $\boldsymbol{F}$, which can preserve tumor-specific knowledge by an auxiliary segmentation loss. Moreover, the tumor feature $\boldsymbol{z}$ contains aggregated features of all types of tumor regions, thus, the diverse categories of tumor information is involved in tumor-specific codes $\hat{\boldsymbol{c}}$. In such a way, the model can obtain the fixed number of tumor-specific codes, more suitable for tumor segmentation than modality-specific codes and unrelated to the number of available input modalities. Therefore, the tumor modality transfer can identify the correlations among modalities and facilitate the model to capture modality-specific information thus improving the model robustness even if some modalities are missing. Then, with the obtained tumor-specific codes $\hat{\boldsymbol{c}}$, the model can use the tumor feature $\boldsymbol{z}$ to generate desirable tumor-specific features for better segmentation.

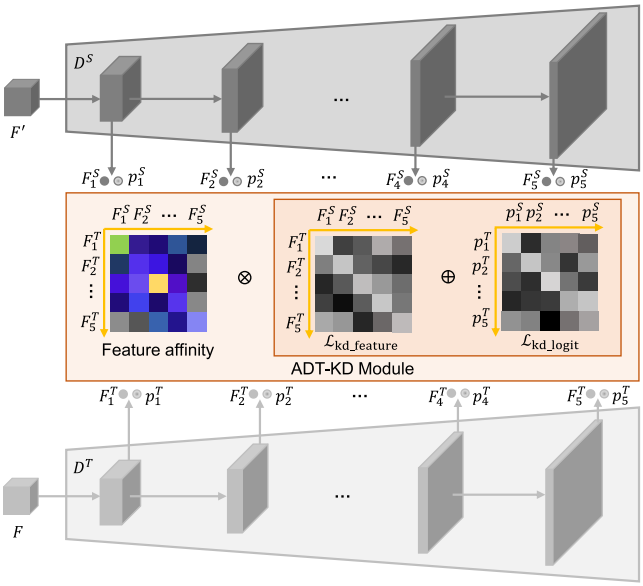

Fig. 4. The structure of the affinity-guided dense tumor-region knowledge distillation (ADT-KD) module. It performs the knowledge distillation in all possible links among pairs of layers in the feature-level and logit-level.

B. Tumor-Region Disentanglement Stage (TD-Stage)

Considering that the features of different tumor-regions have complementary and mutual knowledge, we aim to decouple tumor-specific features and then leverage them to obtain holistic features for segmentation. To this end, as in Fig. 4, in the TD-Stage, we first design a disentanglement branch to generate decoupled tumor-region features, and then propose the ADT-KD in the feature-level and logit-level to obtain holistic tumor features. With a holistic multi-class tumor-region student D^S and a disentangled binary teacher D^T, the ADT-KD can decouple the distinct tumor-specific features and balance the holistic and disentangled information for accurate segmentation. Since the decoupled tumor-specific features are unrelated to the number of available modalities, thus the TD-Stage can address the missing modality issue.

1) Decoding for Tumor-Region Segmentation: To obtain decoupled tumor-region representations, we first construct the features with learned tumor-specific codes and then provide them the explicit supervision with the binary mask of each tumor category. To be specific, we jointly utilize the coarse tumor feature $\boldsymbol{z}$ and the tumor-specific codes $\hat{\boldsymbol{c}}$ learned in the MD-Stage to generate tumor-specific features $\boldsymbol{F} = \{\boldsymbol{F}_c\}_{c=1}^{C} = E_R(\hat{\boldsymbol{c}}, \boldsymbol{z})$, where $\boldsymbol{F} \in \mathbb{R}^{C\times G'\times H'\times W'}$, and E_R is the modality-content reconstruction network with AdaIN [45] layers to incorporate tumor-specific modality information within content representations. G', H', W' are the size of feature maps. Different from the original features $\boldsymbol{z}$ of MRI images, the generated tumor-specific features $\boldsymbol{F}$ preserve the tumor segmentation task-oriented modalities, which are more discriminative and robust. In terms of each tumor category, tumor-specific features $\{\boldsymbol{F}_c\}_{c=1}^{C}$ are then decoded by the decoupled binary teacher D^T to obtain C binary segmentation predictions $\boldsymbol{p}_c^T = D^T(\boldsymbol{F}_c)$, where $\boldsymbol{p}_c^T \in \mathbb{R}^{1\times H\times W}$. Each $\boldsymbol{p}_c^T$ corresponding to one tumor background region, which can be optimized with the binary segmentation mask $\{\boldsymbol{y}_i\}_{i=1}^{C}$, where $\boldsymbol{y}_i = \mathbb{1}_{[i=c]}$ for the ground truth category c of a sample. The binary segmentation mask, as an auxiliary supervision, can guide the model to segment each tumor category by a binary cross-entropy loss:

$$\mathcal{L}_{aux} = \Sigma_{c=1}^{C}\left[\boldsymbol{y}_c \log \boldsymbol{p}_c^T + (1-\boldsymbol{y}_c) log(1-\boldsymbol{p}_c^T))\right]. \quad (7)$$

In parallel, the tumor-specific features $\boldsymbol{F}$ are integrated by a fusion network $Fusion$ to get fused holistic features $\boldsymbol{F}' = Fusion(\boldsymbol{F})$, and then they are fed into the holistic tumor-region student D^S to get segmentation result $\boldsymbol{p}^S = D^S(\boldsymbol{F}')$, where $\boldsymbol{p}^S \in \mathbb{R}^{C\times H\times W}$. In this manner, we can obtain the decoupled tumor-specific features $\{\boldsymbol{F}_c\}_{c=1}^{C}$, holistic features $\boldsymbol{F}'$, and two segmentation predictions $\boldsymbol{p}^S$ and $\boldsymbol{p}^T = \text{concat}(\boldsymbol{p}_1^T, \ldots, \boldsymbol{p}_C^T)$, where concat denotes the concatenation.

2) ADT-KD: To utilize the category-wise and generic knowledge simultaneously as complementary two-view supervisions, we devise the ADT-KD to improve the discrimination of disentangled tumor-specific features and incorporated holistic features for segmentation in a mutual learning manner. Let $\{\boldsymbol{F}_s^S\}_{s=1}^{N_s}$ and $\{\boldsymbol{F}_t^T\}_{t=1}^{N_t}$ be two feature banks, where the feature maps $\boldsymbol{F}_s^S$ and $\boldsymbol{F}_t^T$ are from the holistic tumor-region student D^S and the disentangled binary teacher D^T, respectively. The N_s and N_t denote the numbers of the elements in the feature banks of D^S and D^T.

Specifically, we first resize the $\boldsymbol{F}_s^S$ to match the size of $\boldsymbol{F}_t^T$, then calculate the affinities between them as the scalar $a_{s,t}$:

$$a_{s,t} = \text{GAP}\left(\frac{\boldsymbol{F}_s^S \cdot \boldsymbol{F}_t^T}{\|\boldsymbol{F}_s^S\|_2 \|\boldsymbol{F}_t^T\|_2}\right), \quad (8)$$

where $(\cdot)$ represents the Hadamard product and GAP$(\cdot)$ refers to global average pooling. Then, we can calculate the feature-level ADT-KD losses among all feature maps $\boldsymbol{F}_t^T$ and $\boldsymbol{F}_s^S$ with the affinities $a_{s,t}$ as re-weighting terms. In order to capture the distilled knowledge oriented to segmentation task, we further conduct the logit-level ADT-KD loss via the auxiliary classifier at the side of each feature map. Each auxiliary classifier consists of several convolution layers and outputs a probability score map, (i.e., logit), as $\{\boldsymbol{p}_t^T\}_{t=1}^{N_t}$ or $\{\boldsymbol{p}_s^S\}_{s=1}^{N_s}$, $\boldsymbol{p}_t^T, \boldsymbol{p}_s^S \in \mathbb{R}^{C\times H\times W}$. In this way, the ADT-KD loss including the feature-level and logit-level is formulated as:

$$\mathcal{L}_{\text{ADT-KD}} = \Sigma_{s=1}^{N_s}\Sigma_{t=1}^{N_t} a_{s,t}\left(\left\|\boldsymbol{F}_t^T - \boldsymbol{F}_s^S\right\|_2 + \text{KL}(\boldsymbol{p}_t^T \| \boldsymbol{p}_s^S)\right), \quad (9)$$

where the first term represents the feature-level knowledge distillation loss $\mathcal{L}_{\text{kd\_feature}}$, and the second term indicates the logit-level distillation loss $\mathcal{L}_{\text{kd\_logit}}$. KL$(\cdot\|\cdot)$ is the Kullback-Leibler divergence between two predictions. Different from existing distillation works [33]–[35] that directly align feature maps of the teacher with student networks, our ADT-KD transfers the feature-level and logit-level knowledge between the teacher and the student guided by the identified affinities for all possible combinations of pairs of layers ($N_s \times N_t$ pairs). By adopting the affinity among mediate feature maps, the teacher and student networks can communicate the knowledge between the pair of features according to the feature relation, which can smooth the training process and suppress the knowledge transfer between two features with large discrepancy. In this manner, the student can learn tumor-region knowledge from all layers of the teacher, and the teacher can also learn holistic and task-oriented information from the student.

The loss function in the TD-Stage is summarized as follows:

$$\mathcal{L}_{\text{tumor-disen}} = \mathcal{L}_{\text{ADT-KD}} + \mathcal{L}_{\text{aux}} + \mathcal{L}_{\text{consis}}, \quad (10)$$

where $\mathcal{L}_{\text{consis}} = \left\|\boldsymbol{p}^T - \boldsymbol{p}^S\right\|_2$ is the prediction consistency constraint to further reduce the distance between two final outputs of student and teacher. With the tumor-region disentanglement loss, we can decouple the tumor-specific features and enhance the holistic features for accurate segmentation effectively.

C. Optimization

In D^2-Net, it consists of a MD-Stage and a TD-Stag in order to address the issue of missing modalities for brain tumor segmentation. Therefore, the overall objective function of our D^2-Net framework is:

$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{seg}} + \lambda_1 \mathcal{L}_{\text{modal-disen}} + \lambda_2 \mathcal{L}_{\text{tumor-disen}}, \quad (11)$$

where λ_1 and λ_2 are trade-off factors to balance the contribution of each term. $\mathcal{L}_{seg}$ consists of standard dice loss and multi-class cross-entropy loss for accurate segmentation task.

IV. Experiments

A. Datasets and Pre-Processing

We evaluate our D^2-Net on the 2018 Brain Tumor Segmentation Challenge (BraTS) dataset, which consists of 285 annotated MRI sequences in the training set. Each case contains four aligned modalities including FLAIR, T1, T1ce and T2. There are four categories including the necrotic and non-enhancing tumor core (NET), enhancing tumor (ET), edema (ED) and background (BG), and we evaluate the segmentation performance on WT (the sum of all regions except background), TC (the sum of necrotic, non-enhancing and enhancing tumor) and ET (the enhancing

TABLE I

ROBUSTNESS COMPARISON OF OUR MODEL AGAINST CURRENT STATE-OF-THE-ART METHODS ON THE BRATS-2018 DATASET. THE DICE SCORES (%) ARE PERFORMED AS THE PERFORMANCES ON THREE TUMOR REGIONS (WT, TC AND ET) IN EACH CASE WITH MISSING MODALITIES (M). THE BEST AND SECOND BEST SCORES ARE **HIGHLIGHTED** AND UNDERLINED. (✓) DENOTES THE PRESENT MODALITY AND (–) DENOTES THE MISSING MODALITY

M	Flair	–	–	–	✓	–	–	✓	–	✓	✓	✓	✓	✓	–	✓	Avg	P-value (10^{-2})
	T1	–	–	✓	–	–	✓	✓	✓	–	–	✓	✓	–	✓	✓		
	T1ce	–	✓	–	–	✓	✓	–	–	–	✓	✓	–	✓	✓	✓		
	T2	✓	–	–	–	✓	–	–	✓	✓	–	–	✓	✓	✓	✓		
WT	HeMIS [14]	35.2	2.2	0.4	48.6	47.3	16.4	68.7	48.5	70.2	66.9	72.0	72.4	7.3	40.8	75.4	49.2	0.01
	FDGF [13]	63.4	40.7	10.8	63.2	61.9	35.5	82.4	65.3	85.7	82.2	87.1	79.6	84.2	69.3	85.3	66.4	0.79
	U-HVED [21]	74.2	**50.9**	35.2	76.9	80.6	60.1	85.1	75.0	83.0	87.1	86.2	87.5	87.2	79.6	88.3	75.8	3.71
	KD-Net [22]	31.1	38.9	39.6	29.8	48.7	52.2	49.0	49.1	40.4	48.5	59.1	54.4	54.6	58.5	62.7	47.8	0.02
	CRNet [15]	30.7	30.2	6.8	70.1	52.2	35.7	79.2	40.2	78.5	80.2	83.3	86.0	88.1	62.2	88.5	60.8	0.001
	ACNet [23]	28.9	41.3	**46.4**	38.2	49.8	58.6	58.3	55.3	44.7	50.9	63.4	62.5	56.3	64.8	67.6	52.5	0.07
	D^2-Net	**76.3**	42.8	15.5	**84.2**	**84.1**	**62.1**	**87.3**	**80.1**	**87.9**	**87.5**	**87.7**	**88.4**	**88.8**	**80.9**	**88.8**	**76.2**	-
TC	HeMIS [14]	17.6	4.2	0.1	14.4	40.3	31.8	10.0	17.6	16.5	38.7	50.1	16.6	55.3	47.2	52.0	27.5	0.03
	FDGF [13]	34.2	35.7	10.4	26.6	56.2	53.7	23.9	36.3	30.1	56.2	61.5	33.3	66.1	68.5	78.5	44.7	0.13
	U-HVED [21]	52.4	63.7	30.2	44.2	75.8	72.3	53.8	59.2	61.1	75.3	70.0	58.2	76.6	77.2	78.0	63.2	4.64
	KD-Net [22]	30.1	29.6	27.9	25.6	44.3	39.5	38.6	43.3	39.6	41.9	48.6	51.2	48.5	51.5	56.2	41.1	0.05
	CRNet [15]	14.4	60.4	7.7	44.9	57.2	61.9	54.4	20.7	32.0	67.4	73.1	60.5	78.7	63.3	75.1	51.5	0.60
	ACNet [23]	30.7	35.8	**32.1**	34.1	51.2	47.7	45.8	47.3	47.2	45.1	52.3	56.4	56.6	59.0	61.7	46.9	0.36
	D^2-Net	**56.7**	**65.1**	16.8	**47.3**	**80.3**	**78.2**	**61.6**	**63.2**	**62.6**	**80.8**	**80.9**	**63.7**	**80.7**	**79.0**	**80.1**	**66.5**	-
ET	HeMIS [14]	0.0	9.7	0.4	4.7	47.9	38.6	0.4	0.1	0.4	50.1	58.7	1.4	62.3	56.4	58.8	26.0	0.21
	FDGF [13]	15.2	30.8	6.2	13.2	55.9	49.0	8.8	13.2	16.0	61.5	64.8	11.6	65.4	61.3	60.2	35.5	0.08
	U-HVED [21]	22.4	61.8	14.0	13.1	63.6	60.0	10.4	15.7	12.6	63.2	60.7	16.9	62.7	66.4	67.2	40.7	0.19
	KD-Net [22]	**35.0**	27.3	25.3	19.4	47.2	35.7	33.1	**46.2**	42.5	35.2	41.9	49.7	52.0	52.6	55.8	39.9	0.54
	CRNet [15]	9.6	55.4	5.3	22.4	55.7	53.6	12.6	7.2	12.2	60.2	63.5	14.0	65.5	58.3	66.2	37.5	4.13
	ACNet [23]	34.2	28.9	**26.7**	**26.5**	47.4	38.8	**36.6**	45.6	**46.6**	36.5	42.9	**52.5**	53.1	53.8	56.6	41.8	0.65
	D^2-Net	16.0	**66.3**	8.1	8.1	**68.7**	**70.7**	9.5	16.5	17.4	**64.8**	**65.7**	19.4	**66.4**	**68.3**	**68.4**	**42.3**	-

tumor) regions. We split the training dataset into three folds with 190 sequences as training data and 95 sequences as validation data, then conduct the experiments in three-fold cross-validation manner. Following previous works [2], [4], [13], [15], we perform data pre-processing by normalizing the intensity of each MRI sequence to zero mean and unit variance. Moreover, we use standard data augmentation techniques including random flipping, random rotation and random intensity change. In the training phase, the MRI data of each modality is randomly cropped from $240 \times 240 \times 155$ to $128 \times 128 \times 128$ due to limited memory.

B. Implementation Details

1) Network Structures: In D^2-Net, E_T, D_R, D^S and D^T adopt the encoder/decoder structures of our baseline model U^2-Net [25]. The adaptive instance normalization [45] is embedded in D_R and E_R to incorporate the modality information with tumor-specific representations. The modality encoder E_M consists of four E_M^i and it is used to generate the modality-specific codes $\boldsymbol{c}$ with the dimension of $\mathbb{R}^{4\times1\times1\times8}$. The tumor modality projection network E_{TM} composes of three $1 \times 1 \times 1$ convolution layers, and E_T as the shared tumor encoder across all modalities produces tumor feature $\boldsymbol{z}$ with the shape of $\mathbb{R}^{4\times4\times4\times4}$. E_R consists of four residual $3\times3\times3$ convolution layers attaching adaptive instance normalization [45]. In the fusion network $Fusion$, two convolution layers and a squeeze-to-excite layer [46] with the channel-wise attention are adopted to fuse the disentangled tumor-specific features. In our network, the batch normalization and Leaky ReLU activation are used followed by all convolution layers in E_M,E_T,D^S,D^T and $Fusion$ network. In each feature bank of D^S and D^T, there are four feature maps with the sizes of $\{\mathbb{R}^{8\times8\times8\times64}, \mathbb{R}^{16\times16\times16\times32}, \mathbb{R}^{32\times32\times32\times16}, \mathbb{R}^{64\times64\times64\times8}\}$, where the first three dimensions refer to the shape of each feature map and the last dimension denotes the channel. To balance the contributions of the MD-Stage and TD-Stage, we set the trade-off terms λ_1 and λ_2 both as 1 for the overall loss function.

2) Training Settings: Our framework is implemented with PyTorch and trained on V100 GPUs with 400 epochs. The model is optimized by Adam with an initial learning rate of 3×10^{-4}, weight decay of 5×10^{-5} and a batch size of 2. In the training phase, we randomly set some of modalities as null matrices to mimic the situations of missing modalities as in [13], [15], [17]. In the test phase, for the cases of missing modalities, we set the missing modalities as null matrices.

C. Comparison With the State-of-the-Art Methods on Missing Modalities

We conduct extensive experiments to compare our D^2-Net framework with state-of-the-art methods [13]–[15], [21]–[23] on different cases with missing MRI modalities on the BraTS-2018 dataset. For fair comparison, all models adopt networks with the comparable number of parameters and are trained under the same settings in an end-to-end fashion with randomly discarding input modalities. As illustrated in Table I, our proposed method outperforms existing approaches under most cases with missing modalities, and achieves the best performance of 76.2%, 66.5%, 42.3% in average for dice score on WT, TC, ET, respectively. From the right column of Table I, the P-value scores between D^2-Net and other comparisons are smaller than 0.05, indicating that the proposed method obtains significant improvements over state-of-the-art methods. The model U-HVED [21] with the second best results utilizes a VAE-based framework to learn the statistical information in the latent space and achieves relatively decent performances. Notably, D^2-Net surpasses U-HVED [21] with an average dice score increase of 0.4%, 3.3%, 1.6% on WT, TC and ET regions, and our D^2-Net performs better in nearly all cases with missing one or two modalities, which demonstrates

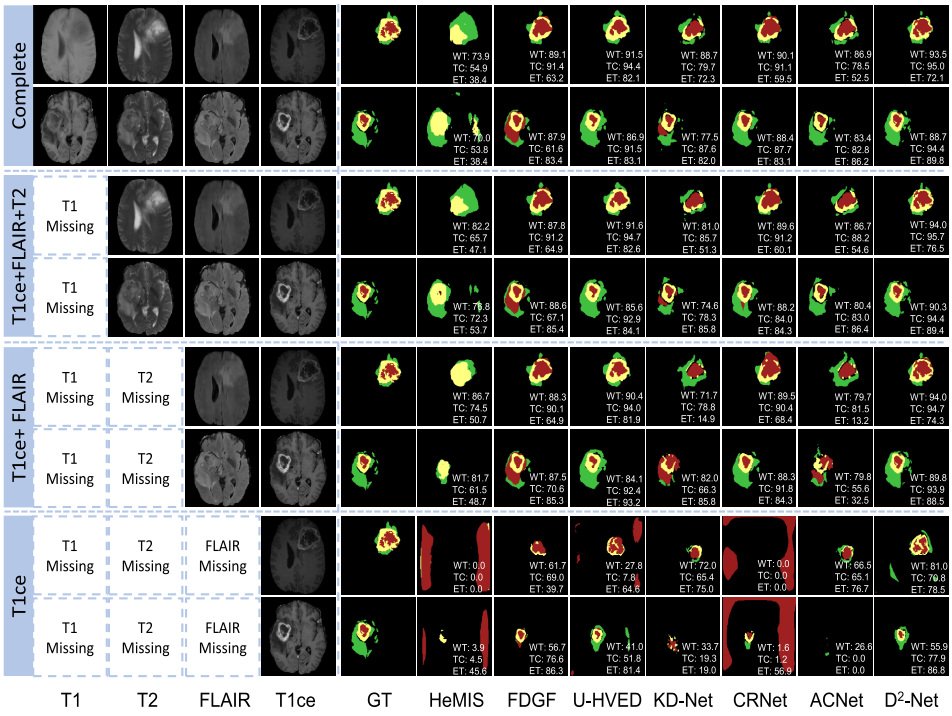

Fig. 5. Comparison of segmentation results on four cases of missing modalities: complete modalities; FLAIR, T1ce, T2; FLAIR, T1ce; T1ce. From the left to right are four MRI modalities: T1, T2, FLAIR and T1ce; The fifth column presents the Ground Truth of two patients, the sixth to eleventh columns perform the results of state-of-the-art approaches, and the right column shows our segmentation results; Red: necrotic and non-enhancing tumor core; Yellow: enhancing tumor; Green: edema.

the effectiveness and robustness of the proposed model. It is because two novel disentanglement components of D^2-Net allow the explicitly capture of latent correlations among modalities and tumor-regions. Note that D^2-Net is superior than other comparisons in the case of complete modalities, with 88.8% on glioma segmentation. These results prove that decoupling modality-specific information and tumor-specific knowledge can improve model robustness against situations with missing modalities. Furthermore, the qualitative results of D^2-Net and other comparison approaches [13]–[15], [21]–[23] in four situations with missing modalities (complete modalities, FLAIR+T1ce+T2, FLAIR+T1ce, T1ce) are shown in Fig. 5. The proposed D^2-Net can produce more accurate segmentation results in most cases.

We further analyze different effects of D^2-Net against multiple situations of missing modalities. From Table I, in all cases including available T1ce, D^2-Net exhibits the best results on TC and ET regions. Particularly, in the cases of missing T2, T1 or Flair (13^{th}, 15^{th}, 16^{th} columns), D^2-Net outperforms the second best one with obvious advantages of 10.9%, 2.0%, 1.8% on TC region, and 0.9%, 0.9%, 1.9% on ET regions. Compared with the results on the case with complete modalities, the performances of D^2-Net in the cases of missing T2 or T1 only have a slight decrease on WT (1.1%, 0%) and ET (2.7%, 2.0%) regions. This validates that D^2-Net is robust against missing T2, T1 or Flair modalities. For all six cases consisting of two available modalities ($7^{th} - 12^{th}$ columns), we achieve the best dice scores with an average increment of 2.6% and 4.8% against the second best scores on WT and TC regions, respectively, which indicates that the proposed D^2-Net can exploit accurate correlations of different modal features when even two modalities are missing. In addition, D^2-Net also achieves the best scores in the cases with only one available modality (3^{rd}, 4^{th}, 6^{th} columns). In particular, D^2-Net performs the best results on WT and TC regions in cases where only T2 or Flair modalities are available (3^{rd}, 6^{th} columns). These experiments results prove D^2-Net is powerful and robust for brain tumor segmentation even when certain modalities are missing.

D. Evaluation of Modality Disentanglement

1) Ablation Study: In order to investigate the effects of our SFMC learning scheme, we present the quantitative results in the $4^{th} - 6^{th}$ rows of Table II. In comparison to the baseline in the 3^{rd} row, the model with spatial modality contrastive learning loss $\mathcal{L}_{\text{sc}}$ gains a remarkable improvements in the average dice score of 9.3%. Then, the model with alternative frequency modality contrastive learning loss $\mathcal{L}_{\text{fc}}$ obtains an increase of 8.6% over the baseline. Moreover, by combining the spatial and frequency modality contrastive losses as SFMC, the model obtains dice scores of 72.9%, 65.5%, 42.2%, with as additional improvement of 14.9%, 23.2% and 4.1% on WT, TC and ET against the baseline. These results indicate the spatial and frequency modality contrastive learning have complementary effects and they can improve the segmentation performances via decoupling modality-specific information.

Furthermore, to verify the effects of tumor modality transfer, we perform the ablations on this transfer process by retaining or discarding tumor modality projection network E_{TM}. From Table IV, the results degrade when we remove the tumor modality transfer or directly use original modality-specific

TABLE II

EVALUATION OF EACH PROPOSED COMPONENT ON THE BRATS-2018 DATASET. THE AVERAGE AND STANDARD DEVIATION OF DICE SCORES (%) ARE PERFORMED ACROSS FIFTEEN CASES OF MISSING MODALITIES. $\mathcal{L}_{SC}$ AND $\mathcal{L}_{FC}$ DENOTE THE SPATIAL AND FREQUENCY MODALITY CONTRASTIVE LOSSES IN THE MD-STAGE. $\mathcal{L}_{KD\_FEATURE}$ AND $\mathcal{L}_{KD\_LOGIT}$ REPRESENT THE FEATURE-LEVEL AND LOGIT-LEVEL LOSSES OF ADT-KD IN THE TD-STAGE

Method				Dice Score			
$\mathcal{L}_{sc}$	$\mathcal{L}_{fc}$	$\mathcal{L}_{kd\_feature}$	$\mathcal{L}_{kd\_logit}$	WT	TC	ET	Avg
-	-	-	-	58.0±20.3	42.3±23.4	38.1±25.5	46.1±22.7
✓	-	-	-	68.5±23.4	58.8±20.9	38.8±28.9	55.4±20.8
-	✓	-	-	68.3±22.9	57.8±20.3	38.1±28.1	54.7±19.5
✓	✓	-	-	72.9±22.9	65.5±20.9	42.2±31.9	60.2±20.6
-	-	✓	-	69.7±22.5	58.1±25.0	41.7±31.5	56.5±22.2
-	-	-	✓	71.7±21.2	59.6±22.0	40.5±30.5	57.3±20.6
-	-	✓	✓	72.2±25.8	61.9±22.9	41.7±31.6	58.6±21.7
✓	✓	✓	✓	**76.2±20.2**	**66.5±16.9**	**42.3±27.1**	**61.7±17.5**

TABLE III

THE IMPACTS OF THE NUMBERS OF POSITIVE AND NEGATIVE SAMPLES

Samples	Number	WT (%)	TC (%)	ET (%)	Avg (%)
Positive	1	76.7	66.7	42.6	62.0
	2	77.4	66.2	43.0	62.2
	4	76.9	66.4	43.6	62.3
	8	77.1	66.2	43.2	62.2
Negative	24	75.9	66.4	42.7	61.7
	48	76.7	66.7	42.6	62.0
	96	76.2	67.2	42.9	62.1
	192	76.4	67.0	42.5	62.0

TABLE IV

THE ABLATION STUDY ON TUMOR MODALITY TRANSFER

Method	WT (%)	TC (%)	ET (%)	Avg. (%)
D^2-Net w/o E_{TM}	75.8	65.9	41.8	61.2
D^2-Net	**76.7**	**66.7**	**42.6**	**62.0**

codes $\boldsymbol{c}$. It suggests tumor-specific codes $\hat{\boldsymbol{c}}$ contain more suitable information compared with original codes $\boldsymbol{c}$ for segmentation, indicating the tumor-specific codes $\hat{\boldsymbol{c}}$ can encode the tumor-specific information.

In addition, we conduct sensitivity analysis on the number of positive and negative samples to evaluate their impacts for the performance of D^2-Net. As shown in Table III, more positive samples can only bring few improvements, and more negative samples rarely improve the performance. It is mainly because that the selected negative samples drawn from nearby slices with highly similar tumor-related content can be regarded as hard negative samples. Similarly, the selected positive samples chosen from far away slices contain the same modality while highly discrepant tumor-related content, therefore, they can be regarded as hard positive samples. According to previous works [47]–[49], utilizing hard positive and negative samples are efficient for contrastive learning. Therefore, in the process of SFMC, we pick up 1 slice that belongs to the same modality and has the interval with 48 slices as the single positive sample, and the slices belonging to other three modalities while locating in the 8 nearby slices on two sides as negative ones, i.e., 48 (48$= 3 \times 8 \times 2$) negative samples.

2) Visualization of Modality-Specific Codes Distribution: In the MD-Stage, D^2-Net produces decoupled modality-specific codes $\boldsymbol{c}$ by SFMC with spatial and frequency contrastive learning. To clearly show the disentanglement effect, we visualize the distributions of them in 2D space via t-SNE [50] in Fig. 6. Fig. 6 (a) shows the codes obtained without SFMC as the baseline, Fig. 6 (b) and (c) present the modality-specific codes decoupled with spatial and frequency contrastive learning, respectively, and Fig. 6 (d) suggests the distribution

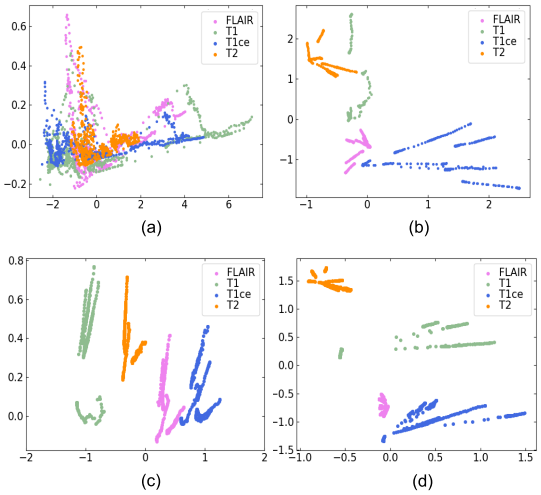

Fig. 6. 2D visualization of modality-specific codes distributions: (a) the baseline, (b) with spatial contrastive learning scheme, (c) with frequency contrastive learning scheme, (d) with both.

obtained with both. It is clear that both spatial and frequency contrastive learning enlarges the distances of inter-modality specific codes, verifying that proposed SFMC enables the model to decouple modality-specific information from MRI.

3) Comparison With Modality Disentanglement Methods: To prove the superiority of the proposed SFMC learning scheme in the MD-Stage, we compare it with other methods [13], [24] of modality disentanglement. To obtain the segmentation results of the MD-Stage, we remove the ADT-KD mechanism, i.e., $\mathcal{L}_{kd\_feature}$ and $\mathcal{L}_{kd\_logit}$, and retain two decoders D^S and D^T in the TD-Stage. As in Table V, the SFMC learning scheme achieves average dice scores of 72.9%, 65.5% and 42.2% among all cases with missing modalities on WT, TC and ET regions, which is better than previous approach [24] with a significant improvement of 7.1%, 16.9%, 3.7% on WT, TC and ET regions. It is mainly because our SFMC provides explicit modality-specific supervision that leads to such accurate and robust segmentation performance.

E. Evaluation of Tumor-Region Disentanglement

1) Ablation Study: We assess the contribution of the proposed ADT-KD mechanism from the quantitative perspective as in the $7^{th}-9^{th}$ rows of Table II. We empirically analyze the role of the TD-Stage by utilizing either feature-level or logit-level ADT-KD loss. As in Table II, the model trained only

TABLE V
EVALUATION OF THE MD-STAGE (MD) AND TD-STAGE (TD) ON THE BRATS-2018 DATASET. THE AVERAGE AND STANDARD DEVIATION OF DICE SCORES (%) ARE PERFORMED ACROSS ALL MISSING CASES

Stage	Method	Dice Score				P-value
		WT	TC	ET	Avg	(10^{-2})
MD	AE [13]	58.0±24.2	42.3±23.6	38.1±28.8	46.1±22.2	0.08
	GAN [24]	65.8±25.0	48.6±19.8	38.5±33.5	51.0±24.7	0.36
	Ours	**72.9±22.9**	**65.5±20.9**	**42.2±31.9**	**60.2±20.6**	-
TD	Multi-KD [23]	71.6±24.1	54.6±26.1	41.4±31.4	55.9±22.6	1.25
	Dense-KD [51]	71.2±17.7	58.4±21.0	**42.5±32.2**	57.3±20.6	4.23
	Ours	**72.2±25.8**	**61.9±22.9**	41.7±31.6	**58.6±21.7**	-

TABLE VI
THE IMPACTS OF DECODER LAYER NUMBERS WITH ADT-KD

Layer number	WT (%)	TC (%)	ET (%)	Avg (%)
1	73.1	65.8	42.2	60.4
2	75.4	66.0	42.3	61.2
3	76.2	66.5	42.5	61.7
4	76.7	66.7	42.6	62.0
5	76.9	66.8	42.6	62.1
6	77.0	66.7	42.7	62.1

with feature-level ADT-KD outperforms the baseline with a dramatic advance of 10.4% dice score, which indicates the effectiveness of distillation for segmentation. We also train the model with only logit-level ADT-KD and obtain the 57.3% average dice score. In addition, incorporating two level distillation losses can obtain the average dice scores of 76.2%, 66.5%, 42.3% on WT, TC and ET regions, respectively, which is better than employing either one for training the model. Based on the model with two modality constrastive components (6^{th} row), the two-level ADT-KD can further improve the performance with 3.3% and 1.0% on WT and TC regions, and 1.5% average increase on those three regions (10^{th} row). These results demonstrate the consistent effectiveness of two-level ADT-KD mechanism, which plays a necessary role for tumor-region segmentation in the TD-Stage.

To investigate the influence of different layer numbers in ADT-KD, we conduct the ablation study on the number of decoder layers in ADT-KD process. As shown in Table VI, utilizing two pairs of layers in ADT-KD process brings 0.8% improvement than the method using only one pair of layers. As the number of decoder layers increases from 1 to 4, the performance shows remarkable improvements, while the results tend to be stable when the number of layers are over 4. As using more layers will bring extra computational cost, we employ 4 layers of the decoder in ADT-KD to balance the performance and the computational cost.

2) Visualization of Tumor-Specific Features Distribution: In the TD-Stage, we extract decoupled tumor-specific features by ADT-KD and binary supervision to explicitly learn the correlations among different tumor-region features. In order to demonstrate the effects of TD-Stage, we visualize the 2D tumor-specific features as shown in Fig. 7. Compared with the baseline Fig. 7 (a), the distribution of tumor-specific features trained with TD-Stages exhibits larger inter-category distance as in Fig. 7 (b), and proves the ability of tumor-region disentanglement of the TD-Stage.

3) Comparison With Distillation Methods: To further verify the superiority of proposed ADT-KD mechanism, we compare it with two state-of-the-art knowledge distillation methods [33], [51] that only consider the feature-level distillation.

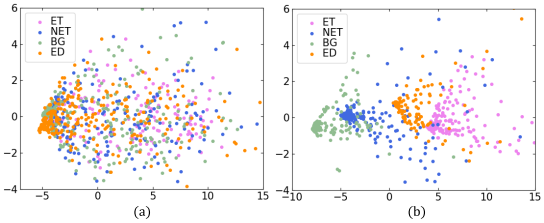

Fig. 7. 2D visualization of tumor-specific features distributions: (a) without TD-Stage, (b) with TD-Stage.

Table V shows that ADT-KD surpasses the Multi-KD [33] and Dense-KD [51] on three tumor regions, with 2.7% and 1.3% improvements on average dice score, which proves that the ADT-KD can exchange useful tumor-specific knowledge in both feature-level and logit-level.

V. DISCUSSION

A. Additional Metrics and Model Complexity

To further verify the effectiveness of the proposed method, we test D^2-Net under four additional evaluation metrics including average symmetric surface distance (ASSD), 95% hausdorff distance (HD95), the model complexity and efficiency. As shown in Table VII, compared with state-of-the-art methods, D^2-Net achieves the best average scores on ASSD and HD95 across all cases with missing modalities. Specifically, we achieve 9.9, 10.3 and 18.4 HD95 scores, outperforming the best second one with 1.3, 1.1 and 1.8, respectively. D^2-Net also obtains the best score of ASSD on WT, TC and ET regions with 8.2, 10.3 and 13.7, respectively. These results further prove the superiority of the proposed method. In addition, the right two columns in Table VII imply results of the model complexity and efficiency. D^2-Net only requires 5.1 million parameters, less then other methods, and the required FLOPs with 4.2×10^{10} is comparable with others, demonstrating the efficiency and effectiveness of D^2-Net.

B. Impact of Each Proposed Loss Function

To study the impact of each proposed loss function in D^2-Net, we conduct sensitivity studies on trade-off factors of loss functions. Note that for the analysis of each proposed loss, we set the factors of all other losses as 1. Fig. 8 shows the average dice scores across three tumor regions with different weights of spatial contrastive loss $\mathcal{L}_{\text{sc}}$, frequency contrastive loss $\mathcal{L}_{\text{fc}}$, reconstruction loss $\mathcal{L}_{\text{rec}}$, feature-level distillation loss $\mathcal{L}_{\text{kd\_feature}}$ and logit-level distillation loss $\mathcal{L}_{\text{kd\_logit}}$. With the weights increasing from 0 to 1.0, all losses perform positive trend on dice score, and the four losses in SFMC and ADT-KD perform stably when factors are equal to 2.0. our chosen weights, i.e., equaled setting as 1.0, achieve the best results in most cases, demonstrating the effectiveness of proposed losses.

C. Analysis of Frequency Contrastive Learning in Different Levels

In D^2-Net, we adopt frequency contrastive learning scheme in the level of the output of modality encoders E_M, i.e., the

TABLE VII

EVALUATION OF OUR METHOD ON THE BRATS-2018 DATASET. THE AVERAGE AND STANDARD DEVIATION OF 95% HAUSDORFF DISTANCE (HD95) AND AVERAGE SYMMETRIC SURFACE DISTANCE (ASSD) (mm) ARE PERFORMED AMONG FIFTEEN CASES OF MISSING MODALITIES. THE NUMBER OF MODEL PARAMETERS (10^6) AND FLOPS (10^{10}) ARE SHOWN IN RIGHT TWO COLUMNS)

Method	HD95			ASSD			Param.	FLOPs
	WT	TC	ET	WT	TC	ET		
HeMIS [14]	21.6 ± 10.4	30.5 ± 17.5	32.3 ± 16.9	12.5 ± 6.2	16.7 ± 6.5	17.2 ± 8.4	**5.1**	**2.8**
FDGF [13]	12.6 ± 7.5	19.5 ± 8.4	26.5 ± 14.1	10.5 ± 5.3	13.3 ± 5.6	16.0 ± 7.5	5.3	4.0
U-HVED [20]	11.2 ± 6.0	11.4 ± 7.2	22.0 ± 12.6	9.1 ± 4.7	11.1 ± 5.1	14.6 ± 7.7	5.9	3.9
KD-Net [21]	21.4 ± 2.6	18.9 ± 6.5	22.7 ± 10.5	12.8 ± 5.4	13.4 ± 5.3	15.0 ± 6.9	7.7	3.0
CRNet [15]	14.2 ± 8.0	16.0 ± 7.8	24.0 ± 13.3	11.2 ± 5.6	12.4 ± 5.0	15.4 ± 7.8	5.5	4.5
ACNet [22]	19.8 ± 3.0	17.4 ± 6.0	20.2 ± 9.3	12.2 ± 5.1	12.9 ± 4.7	14.5 ± 6.5	7.7	3.2
D^2-Net	**9.9 ± 6.9**	**10.3 ± 6.8**	**18.4 ± 11.1**	**8.2 ± 4.9**	**10.3 ± 5.3**	**13.7 ± 7.1**	**5.1**	4.2

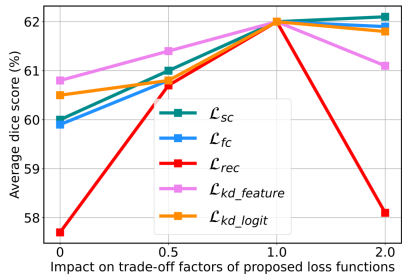

Fig. 8. The impacts on trade-off factors of proposed loss functions.

4^{th}-layer to provide a complementary constraint for spatial contrastive learning. To analyze the effects of the scheme in different levels, we adopt the scheme in each level: input-level, $1^{st} - 3^{rd}$-layer feature maps and 4^{th}-layer codes of the modality encoders, and record the corresponding performance in Table VIII. The frequency contrastive learning adopted in input and $1^{st} - 3^{rd}$ layers of the modality encoders shows obvious decreased performance, and our choice delivers the best results. It demonstrates that the modality-specific codes extracted by modality encoders represent sufficient and compact modality-specific information.

D. Model Compression Effect With ADT-KD

The proposed ADT-KD not only facilitates the model to learn discriminative tumor-specific features, but helps the lightweight tumor-region student D^S to learn compact knowledge from D^T. To investigate the model compression ability of ADT-KD, we compress D^S by reducing the number of feature channels in each layer as the ratio of 1/2 and 1/4 to train D^2-Net, respectively. From Table IX, the small student network with 1/4 compression ratio can also achieve decent performance of 57.1% by distilling knowledge from the teacher network, verifying the model compression ability of ADT-KD.

E. Advantage and Limitation

D^2-Net can explicitly capture correlations across different modalities and tumor regions simultaneously, which improves the model robustness against cases with missing modalities. Although D^2-Net shows superior performance, it may produce unsatisfying results on some special cases. According to Table I, in the cases of missing T1ce modality, D^2-Net may land in a predicament and deliver relative low performance, especially on the ET region. The reason is that D^2-Net mainly depends on T1ce modality, while the modality information

TABLE VIII

THE RESULTS OF D^2-NET USING FREQUENCY CONTRASTIVE LEARNING IN VARIOUS LEVELS

Level	WT (%)	TC (%)	ET (%)	Avg. (%)
MRI Input	73.3	65.5	42.0	60.3
1^{st}-layer feature maps	74.0	65.9	42.1	60.7
2^{nd}-layer feature maps	74.2	66.1	42.4	60.9
3^{rd}-layer feature maps	75.2	66.6	42.5	61.4
4^{th}-layer codes (Ours)	**76.7**	**66.7**	**42.6**	**62.0**

TABLE IX

THE EFFECTS OF DECODER COMPRESSION WITH ADT-KD

Reduction ratio	WT (%)	TC (%)	ET (%)	Avg (%)
1 ×	**76.7**	**66.7**	**42.6**	**62.0**
1/2 ×	73.4	63.6	41.2	59.4
1/4 ×	70.8	60.1	40.5	57.1

carried by T1ce are relatively hard to capture. To alleviate this problem, in future work, we will introduce an extra network pre-trained by complete modalities to transfer modality-aware information to D^2-Net via knowledge distillation to further improve the performance with missing modalities.

VI. CONCLUSION

In this work, we propose a dual disentanglement network (D^2-Net) with the novel MD-Stage and TD-Stage to explicitly capture the correlations among modalities and tumor regions for brain tumor segmentation with missing modalities. The modality disentanglement provides the first effort to employ a SFMC learning scheme to decouple modality-specific information from MRI images in both spatial and frequency domain, which enables the model to directly learn the correlations among multiple modalities. Additionally, to discover the relations among tumor-region features and obtain holistic features for better segmentation, the tumor-region disentanglement leverages a novel ADT-KD mechanism to decompose various tumor-specific knowledge. Extensive experiments on BraTS-2018 dataset show the capability and robustness of our model compared with state-of-the-art methods on brain tumor segmentation with missing modalities.

REFERENCES

[1] M. Havaei *et al.*, "Brain tumor segmentation with deep neural networks," *Med. Image Anal.*, vol. 35, pp. 18–31, Jan. 2017.
[2] H. Jia, Y. Xia, W. Cai, and H. Huang, "Learning high-resolution and efficient non-local features for brain glioma segmentation in MR images," in *Proc. MICCAI*, 2020, pp. 480–490.
[3] C. Dai *et al.*, "Suggestive annotation of brain tumour images with gradient-guided sampling," in *Proc. MICCAI*, 2020, pp. 156–165.
[4] B. Yu *et al.*, "Learning sample-adaptive intensity lookup table for brain tumor segmentation," in *Proc. MICCAI*, 2020, pp. 216–226.
[5] J. Zhang, Y. Xie, Y. Wang, and Y. Xia, "Inter-slice context residual learning for 3D medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 40, no. 2, pp. 661–672, Feb. 2021.

[6] X. Guo, C. Yang, P. L. Lam, P. Y. Woo, and Y. Yuan, "Domain knowledge based brain tumor segmentation and overall survival prediction," in *Proc. MICCAI Workshops*, 2019, pp. 285–295.

[7] Q. Yang and Y. Yuan, "Learning dynamic convolutions for multi-modal 3D MRI brain tumor segmentation," in *Proc. MICCAI Workshops*, 2020, pp. 441–451.

[8] Y. Shen and M. Gao, "Brain tumor segmentation on MRI with missing modalities," in *Proc. IPMI*, 2019, pp. 417–428.

[9] K. van Garderen, M. Smits, and S. Klein, "Multi-modal segmentation with missing MR sequences using pre-trained fusion networks," in *Proc. MICCAI Workshops*, 2019, pp. 165–172.

[10] Y. Wang *et al.*, "3D auto-context-based locality adaptive multi-modality GANs for PET synthesis," *IEEE Trans. Med. Imag.*, vol. 38, no. 6, pp. 1328–1339, Jun. 2018.

[11] T. Zhou, H. Fu, G. Chen, J. Shen, and L. Shao, "Hi-Net: Hybrid-fusion network for multi-modal MR image synthesis," *IEEE Trans. Med. Imag.*, vol. 39, no. 9, pp. 2772–2781, Sep. 2020.

[12] Y. Li, K. K. Singh, U. Ojha, and Y. J. Lee, "MixNMatch: Multifactor disentanglement and encoding for conditional image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8039–8048.

[13] 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 *Proc. MICCAI*, 2019, pp. 447–456.

[14] M. Havaei, N. Guizard, N. Chapados, and Y. Bengio, "HeMIS: Hetero-modal image segmentation," in *Proc. MICCAI*, 2016, pp. 469–477.

[15] T. Zhou, S. Canu, P. Vera, and S. Ruan, "Latent correlation representation learning for brain tumor segmentation with missing MRI modalities," *IEEE Trans. Image Process.*, vol. 30, pp. 4263–4274, 2021.

[16] J. Ouyang, E. Adeli, K. M. Pohl, Q. Zhao, and G. Zaharchuk, "Representation disentanglement for multi-modal brain MRI analysis," in *Proc. IPMI*, 2021, pp. 321–333.

[17] Q. Wang, L. Zhan, P. Thompson, and J. Zhou, "Multimodal learning with incomplete modalities by knowledge distillation," in *Proc. 26th Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 1828–1838.

[18] T. Xia, A. Chartsias, and S. A. Tsaftaris, "Pseudo-healthy synthesis with pathology disentanglement and adversarial learning," *Med. Image Anal.*, vol. 64, Aug. 2020, Art. no. 101719.

[19] Q. Suo, W. Zhong, F. Ma, Y. Yuan, J. Gao, and A. Zhang, "Metric learning on healthcare data with incomplete modalities," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 3534–3540.

[20] B. Zhan, D. Li, X. Wu, J. Zhou, and Y. Wang, "Multi-modal MRI image synthesis via GAN with multi-scale gate mergence," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 1, pp. 17–26, Jan. 2022.

[21] R. Dorent, S. Joutard, M. Modat, S. Ourselin, and T. Vercauteren, "Hetero-modal variational encoder-decoder for joint modality completion and segmentation," in *Proc. MICCAI*, 2019, pp. 74–82.

[22] M. Hu *et al.*, "Knowledge distillation from multi-modal to mono-modal segmentation networks," in *Proc. MICCAI*, 2020, pp. 772–781.

[23] Y. Wang *et al.*, "ACN: Adversarial co-training network for brain tumor segmentation with missing modalities," 2021, *arXiv:2106.14591*.

[24] J. Ouyang, E. Adeli, K. M. Pohl, Q. Zhao, and G. Zaharchuk, "Representation disentanglement for multi-modal brain MR analysis," 2021, *arXiv:2102.11456*.

[25] C. Huang, H. Han, Q. Yao, S. Zhu, and S. K. Zhou, "3D U^2-Net: A 3D universal U-Net for multi-domain medical image segmentation," in *Proc. MICCAI*, 2019, pp. 291–299.

[26] P. Khosla *et al.*, "Supervised contrastive learning," 2020, *arXiv:2004.11362*.

[27] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9729–9738.

[28] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. ICML*, 2020, pp. 1597–1607.

[29] Z. Xie, Y. Lin, Z. Zhang, Y. Cao, S. Lin, and H. Hu, "Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16684–16693.

[30] Q. Xu, R. Zhang, Y. Zhang, Y. Wang, and Q. Tian, "A Fourier-based framework for domain generalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14383–14392.

[31] L. N. Piotrowski and F. W. Campbell, "A demonstration of the visual importance and flexibility of spatial-frequency amplitude and phase," *Perception*, vol. 11, no. 3, pp. 337–346, Jun. 1982.

[32] Y. Yang and S. Soatto, "FDA: Fourier domain adaptation for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4085–4095.

[33] Y. Liu, C. Shu, J. Wang, and C. Shen, "Structured knowledge distillation for dense prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jun. 12, 2020, doi: 10.1109/TPAMI.2020.3001940.

[34] F. R. Valverde, J. V. Hurtado, and A. Valada, "There is more than meets the eye: Self-supervised multi-object detection and tracking with sound by distilling multimodal knowledge," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11612–11621.

[35] Q. Dou, Q. Liu, P. A. Heng, and B. Glocker, "Unpaired multi-modal segmentation via knowledge distillation," *IEEE Trans. Med. Imag.*, vol. 39, no. 7, pp. 2415–2425, Jul. 2020.

[36] J. Jiang *et al.*, "Deep cross-modality (MR-CT) educed distillation learning for cone beam CT lung tumor segmentation," *Med. Phys.*, vol. 48, no. 7, pp. 3702–3713, Jul. 2021.

[37] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. NeurIPS*, 2016, pp. 2180–2188.

[38] Y.-C. Cheng, H.-Y. Lee, M. Sun, and M.-H. Yang, "Controllable image synthesis via SegVAE," in *Proc. ECCV*, 2020, pp. 159–174.

[39] M. Kim, Y. Wang, P. Sahu, and V. Pavlovic, "Bayes-factor-VAE: Hierarchical Bayesian deep auto-encoder models for factor disentanglement," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2979–2987.

[40] S. Bhagat, S. Uppal, Z. Yin, and N. Lim, "Disentangling multiple features in video sequences using Gaussian processes in variational autoencoders," in *Proc. ECCV*, 2020, pp. 102–117.

[41] I. Higgins *et al.*, "beta-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. ICLR*, 2017. [Online]. Available: https://openreview.net/forum?id=Sy2fzU9gl

[42] J. Huang, D. Guan, A. Xiao, and S. Lu, "FSDR: Frequency space domain randomization for domain generalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6891–6902.

[43] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE Trans. Electron. Comput.*, vol. EC-14, no. 3, pp. 326–334, Jun. 1965.

[44] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, "Learnability and the Vapnik-Chervonenkis dimension," *J. ACM*, vol. 36, no. 4, pp. 929–965, Oct. 1989.

[45] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1501–1510.

[46] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[47] S. Ma, Z. Zeng, D. McDuff, and Y. Song, "Active contrastive learning of audio-visual video representations," 2020, *arXiv:2009.09805*.

[48] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman, "With a little help from my friends: Nearest-neighbor contrastive learning of visual representations," 2021, *arXiv:2104.14548*.

[49] Y. Kalantidis, M. B. Sariyildiz, N. Pion, P. Weinzaepfel, and D. Larlus, "Hard negative mixing for contrastive learning," 2020, *arXiv:2010.01028*.

[50] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 1–27, 2008.

[51] M. Ji, B. Heo, and S. Park, "Show, attend and distill: Knowledge distillation via attention-based feature matching," in *Proc. AAAI*, vol. 35, no. 9, 2021, pp. 7945–7952.