M^{3}AE: Multimodal Representation Learning for Brain Tumor Segmentation with Missing Modalities

Hong Liu[1,2,*], **Dong Wei**[2,*], **Donghuan Lu**[2], **Jinghan Sun**[2,3], **Liansheng Wang**[1,†], **Yefeng Zheng**[2]

[1]School of informatics, Xiamen University, Xiamen, China
[2]Tencent Jarvis Lab, Tencent Healthcare (Shenzhen) Co., Ltd., Shenzhen, China
[3]School of Medicine, Xiamen University, Xiamen, China
{liuhong,jhsun}@stu.xmu.edu.cn, lswang@xmu.edu.cn, {donwei,caleblu,yefengzheng}@tencent.com,

Abstract

Multimodal magnetic resonance imaging (MRI) provides complementary information for sub-region analysis of brain tumors. Plenty of methods have been proposed for automatic brain tumor segmentation using four common MRI modalities and achieved remarkable performance. In practice, however, it is common to have one or more modalities missing due to image corruption, artifacts, acquisition protocols, allergy to contrast agents, or simply cost. In this work, we propose a novel two-stage framework for brain tumor segmentation with missing modalities. In the first stage, a multimodal masked autoencoder (M^{3}AE) is proposed, where both random modalities (i.e., modality dropout) and random patches of the remaining modalities are masked for a reconstruction task, for self-supervised learning of robust multimodal representations against missing modalities. To this end, we name our framework M^{3}AE. Meanwhile, we employ model inversion to optimize a representative full-modal image at marginal extra cost, which will be used to substitute for the missing modalities and boost performance during inference. Then in the second stage, a memory-efficient self distillation is proposed to distill knowledge between heterogenous missing-modal situations while fine-tuning the model for supervised segmentation. Our M^{3}AE belongs to the 'catch-all' genre where a single model can be applied to all possible subsets of modalities, thus is economic for both training and deployment. Extensive experiments on BraTS 2018 and 2020 datasets demonstrate its superior performance to existing state-of-the-art methods with missing modalities, as well as the efficacy of its components. Our code is available at: https://github.com/ccarliu/m3ae.

Introduction

Segmentation and associated volume quantification of heterogeneous histological sub-regions are of great value to the diagnosis/prognosis, therapy planning, and follow-up of brain tumors (Bakas et al. 2018). Multi-parametric magnetic resonance imaging (MRI) is the current standard of care for clinical imaging diagnosis of brain tumors (Iv et al. 2018). Specifically, four MRI modalities (in this work, we refer to MRI sequences as modalities) are commonly used to provide complementary information and support sub-region analysis: T1-weighted (T1), contrast enhanced T1-weighted (T1c), T2-weighted (T2), and T2 fluid attenuation inversion recovery (FLAIR), where the first two highlight tumor core and the last two highlight peritumoral edema (Figs. 1(a) and (b)). In recent years, deep learning methods have greatly advanced the state of the art of brain tumor segmentation with multimodal MRI (Chen et al. 2020; Chen, Ding, and Liu 2019; Ding et al. 2020; Myronenko 2018; Zhou et al. 2020). However, these methods were optimized for the ideal scenario where the full set of all modalities are present. While in practice, scenarios of missing one or more modalities commonly occur due to image corruption, artifacts, acquisition protocols, allergy to contrast agents, or simply cost.

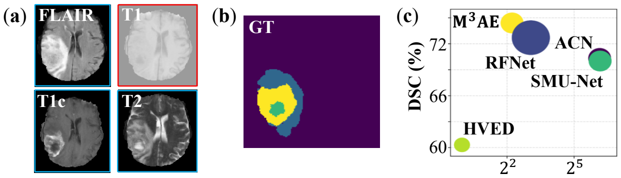

Figure 1: (a) Example images of the four modalities in BraTS 2018, of which one or more may be missing in practice (e.g., T1 here in red box). (b) Corresponding tumor regions: blue: edema; yellow: enhancing tumor; green: necrotic and non-enhancing tumor core. (c) Deployment model size (in millions with $\log_2$ scale) and mean Dice similarity scores (DSCs) across all missing-modal situations on BraTS 2018 test split; circle size indicates GFLOPS. Compared to four up-to-date methods, our M^{3}AE achieves the best performance with a compact and efficient model.

To accommodate the practical scenarios of missing modalities, lots of efforts have been made. A naive approach is to train a 'dedicated' model for each possible subset of modalities. For better performance, the co-training strategy (Blum and Mitchell 1998) was often incorporated to distill knowledge from full-modal to missing-modal networks (Azad, Khosravi, and Merhof 2022; Chen et al. 2021; Hu et al. 2020; Wang et al. 2021b). Despite their decent performance, the dedicated models were time-costly to train and space-costly to deploy, as $2^N - 1$ models were needed for

[*] These authors contributed equally; H. Liu contributed to this work during an internship at Tencent.

[†] Corresponding author.

N modalities. Another approach is to synthesize images of missing modalities for full-modal segmentation (Lee, Moon, and Ye 2020; Yu et al. 2019), where generative adversarial networks (GANs; Goodfellow et al. 2014) were often used. Notwithstanding, the GANs not only were difficult to train for 3D image generation, but also incurred extra overhead for both training and deployment. Currently, the predominant approach is to project the available modalities to a common latent space, where a shared feature representation was learned and then projected to the segmentation space (Havaei et al. 2016; Zhou et al. 2021a,b). This 'catch-all' approach could handle all possible subsets of modalities with a single model, thus was more economic. However, existing catch-all methods often adopted complex designs with multiple encoders (and sometimes multiple decoders, too) and complicated interactions.

In this work, we propose a novel catch-all framework for brain tumor segmentation using MRI with missing modalities, which features innovative integration of multimodal masked autoencoder, model inversion based modal completion, and memory-efficient self distillation in a single straightforward encoder-decoder architecture. Above all, witnessing the recent success of masked autoencoders in learning rich visual representations (He et al. 2022), we propose multimodal masked autoencoder (M^{3}AE), where a random subset of the modalities and random patches of the remaining ones are masked *simultaneously*. The intuition is that, to recover the masked content, the model must effectively utilize the inherent inter-modal correlation both globally and locally, plus the intra-modal local semantics. Accordingly, we name our framework M^{3}AE. Meanwhile, a representative full-modal image is learned via model inversion (Wang et al. 2021a), which is served as the substitute for missing modalities during inference and effective in improving performance. The substitute image is optimized by back propagating the self-supervising M^{3}AE loss, incurring only marginal extra computational cost. To the best of our knowledge, this work is the first attempt to apply model inversion to modality completion of medical images. Lastly, we propose a simple yet efficient self distillation (Ge et al. 2021; Ji et al. 2021) to promote semantic consistency between different modality combinations. To this end, we reduce the memory footprint of co-training dual networks, while still able to effectively distill the semantic information between heterogeneous missing-modal situations. Extensive experiments on two public datasets demonstrate: (1) our framework's robustness to missing modalities and superiority to existing catch-all and dedicated methods (Fig. 1(c)), (2) efficacy of its building components, and (3) competence of its multimodal representation learning for full modalities.

Related Work

Multimodal Brain Tumor Segmentation with Missing Modalities: In this work, we roughly divide existing methods into two categories: dedicated and catch-all. Several methods proposed to train a dedicated model for each targeted missing situation, where the co-training strategy (Blum and Mitchell 1998) was employed to distill knowledge from full-modal to missing-modal networks. Both Hu et al. (2020) and Chen et al. (2021) proposed to distill the knowledge from a multimodal teacher network to monomodal students at the image (i.e., overall semantics) and pixel (i.e., network output) levels. Adversarial co-training network (ACN; Wang et al. 2021b) enhanced the full- to missing-modal distillation by entropy and knowledge adversarial learning for alignment of the latent representations. Style matching U-Net (SMU-Net; Azad, Khosravi, and Merhof 2022) decomposed the common latent space of both full- and missing-modal data into content and style representations, and used a content and style-matching mechanism to distill the informative features from the full-modal network into a missing-modal one. These methods achieved decent performance especially when more than one modality was missing, while at significant computation and memory costs for both training and deployment ($2^N - 1$ models needed for N modalities). In contrast, our framework adopts a catch-all design, i.e., a single model applicable to all missing-modal situations, thus is more economic.

A special group of dedicated methods tackled the problem by synthesizing the missing modalities with fidelity (Lee, Moon, and Ye 2020; Yu et al. 2019), where generative adversarial networks (GANs; Goodfellow et al. 2014) were often used. However, GANs are known to be difficult to train for 3D image generation and may incur extra overhead for both training and deployment. Further, as suggested by Lee, Moon, and Ye (2020), the gadolinium contrast agent was indispensable and the resulting contrast images could not be completely reproduced by the generative models. Instead of synthesizing images of missing modalities perfectly for each subject, we opt to optimize a universal full-modal substitute image at marginal cost, which boosts missing-modal segmentation but does not necessarily look realistic.

The other category of methods attempted to handle all missing-modal situations with a single catch-all model, where modality-specific encoders were commonly employed to embed the modalities into a shared latent space, followed by feature fusion and further processing to yield segmentation (Havaei et al. 2016). On top of the generic paradigm, hetero-modal variational encoder-decoder (HVED; Dorent et al. 2019) incorporated multimodal variational auto-encoders to reconstruct the modalities from the common latent variable, forcing the formulation of a genuinely shared latent representation; Shen and Gao (2019) proposed adversarial training to adapt feature maps of missing modalities to those of full modalities; latent correlation representation learning (Zhou et al. 2021b) modeled inter-modal correlations to estimate the missing modalities' representation in the latent space; Zhou et al. (2021a) explicitly generated a feature-enhanced image to provide necessary feature representations of missing modalities; and region-aware fusion network (RFNet; Ding, Yu, and Yang 2021) relied on a region-aware fusion module to conduct feature fusion from available image modalities according to disparate regions adaptively. All these methods followed complex designs of multiple encoders (and sometimes multiple decoders, too) with complicated interactions. Our framework, while also belonging to the catch-all category, is distinct in that it enables a succinct single-encoder-

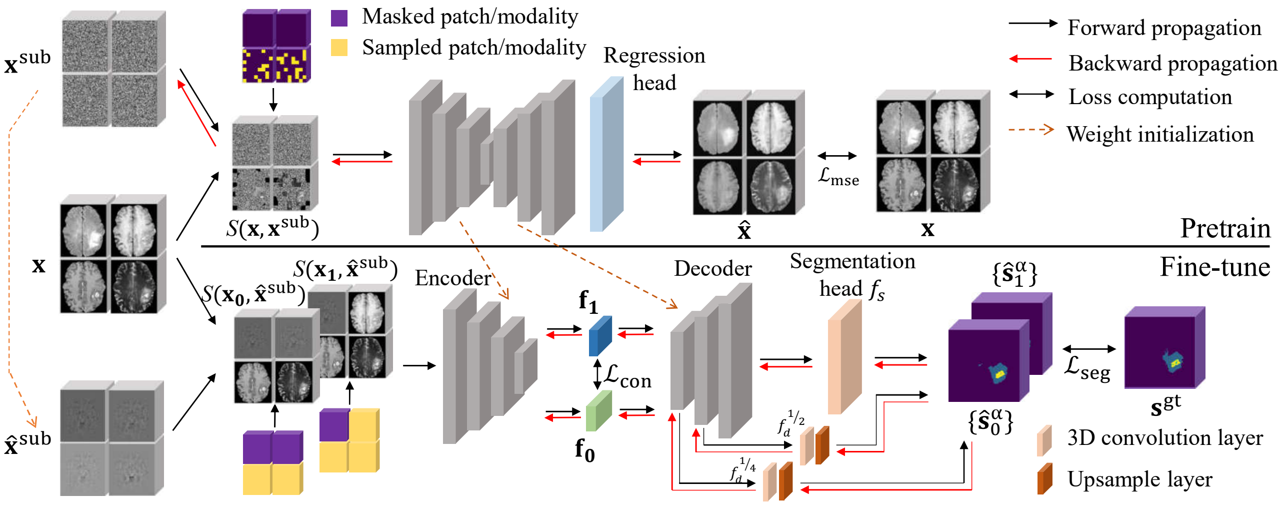
$\mathbf{x}^{\text{sub}}$
🟪 Masked patch/modality
🟨 Sampled patch/modality
Regression head
Forward propagation
Backward propagation
Loss computation
Weight initialization
$S(\mathbf{x}, \mathbf{x}^{\text{sub}})$
$\mathcal{L}_{\text{mse}}$
$\hat{\mathbf{x}}$
$\mathbf{x}$
Pretrain
Fine-tune
$S(\mathbf{x_1}, \hat{\mathbf{x}}^{\text{sub}})$
$S(\mathbf{x_0}, \hat{\mathbf{x}}^{\text{sub}})$
Encoder
Decoder
Segmentation head f_s
$\{\hat{\mathbf{s}}_1^{\alpha}\}$
$\mathbf{f_1}$
$\mathcal{L}_{\text{con}}$
$\mathbf{f_0}$
$f_d^{1/2}$
$f_d^{1/4}$
$\mathcal{L}_{\text{seg}}$
$\mathbf{s}^{\text{gt}}$
$\{\hat{\mathbf{s}}_0^{\alpha}\}$
$\hat{\mathbf{x}}^{\text{sub}}$
3D convolution layer
Upsample layer

Figure 2: Overview of the proposed framework.

single-decoder architecture (essentially a 3D U-Net) to learn rich multimodal representations and deal with heterogeneous missing-modal situations simultaneously.

Self-Supervised Multimodal Representation Learning for Medical Image Analysis: While quite a number of works proposed effective self-supervised representation learning for monomodal medical images (Taleb et al. 2020; Zhang, Wang, and Zheng 2017, etc.), researchers have just began to explore unique pretext tasks based on the 'multimodality' for multimodal medical images. Recently, Taleb et al. (2021) introduced a novel cross-modal jigsaw puzzle (CMJP) task to learn a modality-agnostic feature embedding. Despite its effectiveness, CMJP was proposed for 2D networks and did not consider the practical situations of missing modalities, and how to extend it for 3D networks or missing-modal situations is not straightforward. The regularizing modality reconstruction task in HVED (Dorent et al. 2019) with modality dropout was an effective self-supervising task in preparing for missing modalities. However, it only focused on global inter-modal correlations but ignored the local structural integrity, which is valuable in learning powerful representation for segmentation. In contrast, our M^{3}AE learns rich multimodal representation by modeling both global inter-modal correlations and local intra-modal anatomical integrity, from input with both randomly dropped modalities and randomly masked patches, respectively. To this end, M^{3}AE is inherently robust to missing modalities while suited for fine-scale semantic learning.

Knowledge Distillation: Knowledge distillation (KD; Hinton, Vinyals, and Dean 2015) was originally proposed to compress knowledge from one or more teacher networks (often large complex models or model ensemble) to a student one (often lightweight models). For multimodal segmentation with missing modalities, several works (Hu et al. 2020; Wang et al. 2021b; Chen et al. 2021; Azad, Khosravi, and Merhof 2022) proposed to transfer the 'dark knowledge' of the full-modal network to missing-modal ones via co-training (Blum and Mitchell 1998). Although achieving decent performance, the co-training strategy incurred non-negligible memory cost for training due to the dual-network architecture. In addition, each pair of co-training networks only focused on a fixed correlation between the full modalities and a specific type of missing modalities (e.g., full to T1 alone), failing to exploit the common semantics shared by all different missing-modal situations. Adopting self distillation (Ge et al. 2021; Ji et al. 2021), our framework distills the shared semantics between *heterogeneous* missing-modal situations (including the special case of full-modal) within a *single* network, and achieving better performance for both missing- and full-modal segmentation while consuming less resources for training than previous methods.

Method

The overview of our framework is shown in Fig. 2, including a pretraining and a fine-tuning stage. In the pretraining stage, a novel multimodal masked autoencoder (M^{3}AE) is proposed for self-supervised learning of a robust representation against missing modalities. Meanwhile, a full-modal substitute for missing modalities is learned by model inversion via back propagating the training loss of the M^{3}AE ($\mathcal{L}_{\text{mse}}$). Then in the second stage, a memory-efficient self distillation strategy is proposed to distill the shared semantics between heterogeneous missing-modal situations via a consistency loss ($\mathcal{L}_{\text{con}}$), while fine-tuning the network for brain tumor segmentation using the supervised loss $\mathcal{L}_{\text{seg}}$. The trained segmentation network serves as a 'catch-all' model that can be used for any subset as well as the full set of the modalities. Next, we first describe the newly proposed building components of our framework, including M^{3}AE, model inversion, and self distillation in details, followed by the training and inference procedures integrating them.

Self-Supervised Multimodal Representation Learning via M^{3}AE: Masked autoencoders (MAEs) have been proven successful as scalable self-supervised vision learners (He et al. 2022), where the pretext task is to reconstruct the original signal given its partial observation. Being inspired, we propose a multimodal masked autoencoder (M^{3}AE) for medical images. Consider a multimodal image $\mathbf{x} \in \mathbb{R}^{N\times D\times H\times W}$, where W, H, and D are the width, height, and depth of the image, respectively, and N is number of modalities. In practice, any subset of the N modalities can be missing. Therefore, we sample a random subset of the modalities for masking to mimic the real situation, in addition to randomly masking 3D patches of the remaining modalities as in the original MAE for natural images. Recovering the modalities masked as a whole requires the network to exploit the global inter-modal correlation, whereas recovering the masked patches requires to exploit both intramodal structural integrity and local inter-modal correlation. Thus, our M^{3}AE facilitates self-supervised learning of both anatomical knowledge and inter-modal correlation at the same time. The mean squared error between the reconstructed and original images ($\hat{\mathbf{x}}$ and $\mathbf{x}$ in Fig. 2) is used as the loss function for the M^{3}AE, denoted by $\mathcal{L}_{\text{mse}}$.

A notable difference between the original MAE for natural images and our M^{3}AE is that, masked patches of the former can only be inferred from surrounding context, whereas those of the latter can be additionally inferred from other modalities and thus expected to be easier. Therefore, we empirically set an even higher combined masking rate of 87.5% (compared to 75% used by He et al. (2022)) in our M^{3}AE to make the self-supervising task nontrivial.

Model Inversion based Modality Completion: Most existing approaches to modality completion resorted to GANs to synthesize images of the missing modalities (Lee, Moon, and Ye 2020; Yu et al. 2019), resulting in an extra model and associated training and deployment overheads in addition to the segmentation networks. Via model inversion, we in this work propose space- and time-efficient synthesis of a full-modal substitute image from the M^{3}AE training process at a marginal cost. Model inversion has long been used for explainable deep learning, to synthesize images most representative of certain network predictions, e.g., saliency maps for classification (Simonyan, Vedaldi, and Zisserman 2014). Specifically, we optimize an image $\mathbf{x}^{\text{sub}} \in \mathbb{R}^{N\times D\times H\times W}$ that can lead to smaller reconstruction errors when used to substitute for the masked content (both the whole modalities and intramodal patches) of $\mathbf{x}$:

$$\hat{\mathbf{x}}^{\text{sub}} = \arg\min_{\mathbf{x}^{\text{sub}}} \mathcal{L}_{\text{mse}}(\mathbf{x}, F(S(\mathbf{x}, \mathbf{x}^{\text{sub}}))) + \gamma\mathcal{R}(\mathbf{x}^{\text{sub}}), \quad (1)$$

where $S(\mathbf{x}, \mathbf{x}^{\text{sub}})$ is the operation of replacing the masked content of $\mathbf{x}$ with the location-corresponding content in $\mathbf{x}^{\text{sub}}$, F is the reconstruction function cascading the backbone network f and a regression head, $\mathcal{R}$ is a regularization term, and γ is a weight. Following Nguyen et al. (2016), we use a small amount of L_2 regularization for $\mathcal{R}$ with $\gamma = 0.005$. Here, we make a modification to the original MAE (He et al. 2022) by replacing the masked contents with $\mathbf{x}^{\text{sub}}$ and processing them in the same way as non-masked ones, instead of discarding them. The intuition is that, in order to yield better reconstruction, the optimal substitute must capture most representative modality-specific patterns, which are expected to help with the target task of multimodal segmentation, too. For implementation, $\mathbf{x}^{\text{sub}}$ is updated by back propagation, along with the update to the network parameters. In this way, there is no need to introduce any extra module, and the optimization of $\mathbf{x}^{\text{sub}}$ only incurs marginal cost. An example of optimized $\hat{\mathbf{x}}^{\text{sub}}$ is shown in Fig. 3. Note that only one $\hat{\mathbf{x}}^{\text{sub}}$ is learned for a given training dataset, which can be considered as a form of multimodal representation of the training data and is applicable to all subjects.

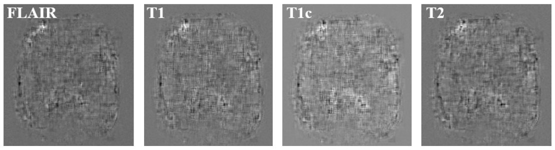

Figure 3: Example full-modal substitute image $\hat{\mathbf{x}}^{\text{sub}}$ optimized via model inversion on the BraTS 2018 dataset.

Fine-Tune with Heterogeneous Missing-Modal Self Distillation for Tumor Segmentation: Implemented via co-training, knowledge distillation from the full-modal to the missing-modal network has proven effective in multimodal segmentation with missing modalities (Hu et al. 2020; Wang et al. 2021b; Chen et al. 2021; Azad, Khosravi, and Merhof 2022), although at the great cost of substantial computational overhead due to the pairing network. Inspired by the self distillation strategy (Ge et al. 2021; Ji et al. 2021), we propose a memory-efficient self distillation strategy to distill knowledge between heterogeneous missing-modal situations within a single network. Specifically, in each batch, we randomly sample two different missing-modal situations (including the special case of full-modal) of one subject via modality dropout as the network input, and encourage consistent semantic features between them with a consistency loss $\mathcal{L}_{\text{con}}$:

$$\mathcal{L}_{\text{con}}(\mathbf{x}_0, \mathbf{x}_1, \hat{\mathbf{x}}^{\text{sub}}) = \mathcal{L}_{\text{mse}}(\mathbf{f}_0, \mathbf{f}_1), \quad (2)$$

where $\mathbf{x}_0$ and $\mathbf{x}_1$ are the two random missing-modal instantiations of $\mathbf{x}$; $\mathbf{f}_0$ and $\mathbf{f}_1 \in \mathbb{R}^{C\times D'\times H'\times W'}$ are the corresponding feature maps extracted from $S(\mathbf{x}_0, \hat{\mathbf{x}}^{\text{sub}})$ and $S(\mathbf{x}_1, \hat{\mathbf{x}}^{\text{sub}})$, respectively; and C, D', H', and W' are the channel number, depth, height, and width of the feature maps, respectively. The mutual knowledge transfer via Eqn. (2) is two-way beneficial: the knowledge transfer from the more to less modalities encourages recovery of the lost information of the missing modalities, and that in the reverse direction (especially from monomodal to multimodal) enhances modality-specific features. In addition, as $\mathbf{x}_0$ and $\mathbf{x}_1$ are obtained by random modality dropout in each epoch, our self distillation transfers knowledge between heterogeneous missing-modal situations, instead of between fixed ones as in paired co-training. Following Hu et al. (2020) and Wang et al. (2021b), we distill in the latent space at the bottleneck of the network (see Fig. 2 bottom).

Training and Inference Procedures: A two-stage (pretraining and fine-tuning) training scheme is employed. In the first stage, the M^{3}AE is trained with both random modality and random patch replacement, with the substitute image

$\mathbf{x}^{\text{sub}}$ optimized at the same time. The optimization problem can be formulated as:

$$\min_{F,\mathbf{x}^{\text{sub}}} \mathcal{L}_{\text{mse}}(\mathbf{x}, F(S(\mathbf{x}, \mathbf{x}^{\text{sub}}))) + \gamma\mathcal{R}(\mathbf{x}^{\text{sub}}). \quad (3)$$

This stage serves as self-supervised pretraining in which the inherent inter-modal correlation and anatomical integrity are learned along with the full-modal substitute image. Then in the second stage, two missing-modal volumes $\mathbf{x}_0$ and $\mathbf{x}_1$ are instantiated for each case by randomly dropping zero to three modalities. Also, the regression head used in the first stage is replaced with a randomly initialized segmentation head f_s. The optimization problem now becomes:

$$\min_{f,f_s,\{f_d\}} \lambda\mathcal{L}_{\text{con}}(\mathbf{x}_0, \mathbf{x}_1, \hat{\mathbf{x}}^{\text{sub}}) + \sum\nolimits_{i=0}^{1} \mathcal{L}_{\text{seg}}(\mathbf{s}^{\text{gt}}, \mathbf{x}_i, \hat{\mathbf{x}}^{\text{sub}}), \quad (4)$$

where λ is a weight, $\mathcal{L}_{\text{seg}}$ is the segmentation loss with deep supervision (Lee et al. 2015):

$$\mathcal{L}_{\text{seg}}(\mathbf{s}^{\text{gt}}, \mathbf{x}_i, \hat{\mathbf{x}}^{\text{sub}}) = \sum\nolimits_{\alpha\in\{1,\frac{1}{2},\frac{1}{4}\}} \mathcal{L}(\mathbf{s}^{\text{gt}}, \hat{\mathbf{s}}_i^{\alpha}), \ i \in \{0, 1\}, \quad (5)$$

where $\mathcal{L}$ is the Dice loss (Milletari, Navab, and Ahmadi 2016) plus cross entropy loss commonly used for medical image segmentation, $\mathbf{s}^{\text{gt}}$ is the segmentation ground truth, and $\hat{\mathbf{s}}_i^{\alpha}$ is the network prediction for $S(\mathbf{x}_i, \hat{\mathbf{x}}^{\text{sub}})$ at a specific scale α with respect to the input size, which is upsampled (if needed) to match the size of $\mathbf{s}^{\text{gt}}$. Specifically, a $1\times1\times1$ convolution (denoted by f_d^{α}) followed by trilinear upsampling is used to yield the intermediate prediction for $\alpha = \frac{1}{2}$ and $\frac{1}{4}$. Therefore, the second stage tunes the network for the target task of multimodal segmentation with missing modalities, with the help of the self distilling consistency loss. As to inference, we simply substitute $\hat{\mathbf{x}}^{\text{sub}}$ for the missing modalities (if any), feed the resulting image to the trained model, and obtain the segmentation by f_s. Pseudo code of the above-described procedures (assuming each minibatch consisting of a single *subject*) is shown in Algorithm S1.

Experiments and Results

Datasets and Evaluation Metrics: We evaluate the proposed framework with two widely used multimodal Brain Tumor Segmentation (BraTS) datasets: BraTS 2018 and 2020 (Bakas et al. 2018). The BraTS datasets comprise multi-contrast MRI exams with four sequences: T1, T1c, T2, and FLAIR. The scans were preprocessed by the organisers, including skull-stripping, re-sampling to a unified resolution (1 mm^3), and co-registration to the same template. Following the challenge, four intra-tumor structures (edema, enhancing tumor, necrotic and non-enhancing tumor core) are grouped into three tumor regions for evaluation: (1) whole tumor, including all tumor tissues, (2) tumor core, composed of the enhancing tumor, necrotic and non-enhancing tumor core, and (3) enhancing tumor. The BraTS 2018 and 2020 datasets include 285 and 369 training cases with ground truth publicly available, respectively, for which we follow the splits of 199:29:57 (training:validation:testing) and 219:50:100 cases in (Ding, Yu, and Yang 2021), respectively. The model is trained on the training splits and tuned (including hyperparameters and other settings) according to the performance on the validation splits, whereas the testing splits are only used for final model evaluation. Following the BraTS challenge, we use the Dice similarity coefficient (DSC) and the 95th percentile of the Hausdorff distance (HD95) for performance quantification.

Implementation: The PyTorch framework (1.7.1; Paszke et al. 2019) is used for experiments. We use the same backbone network as Wang et al. (2021b), which is essentially a 3D U-Net comprising a single encoder and a single decoder employing residual blocks (He et al. 2016) and group normalization (Wu and He 2018) (more details provided in the supplement). As to the regression (for pretraining) and segmentation heads, $1\times1\times1$ convolutions without and with sigmoid are used, respectively. No additional post-processing is conducted. We use two NVIDIA RTX 2080 Ti GPUs for training, with a batch size of two volumes, i.e., two volumes of two subjects for pretraining, and two random missing-modal instantiations of a subject for fine-tuning. The Adam (Kingma and Ba 2014) optimizer is employed with an initial learning rate of 0.0003 and a cosine decay scheduler (Loshchilov and Hutter 2016), for both pretraining (600 epochs) and fine-tuning (300 epochs). To standardize all volumes, we clip the volumes to the [1st, 99th] percentiles of the intensity values followed by min-max scaling, and randomly crop them to a fixed size of $128\times128\times128$ voxels for training. Side length of the random 3D patches is set to 16 voxels following He et al. (2022). $\mathbf{x}^{\text{sub}}$ is initialized to Gaussian noise. Common data augmentation is conducted for training, including: random cropping (from $240\times240\times155$ to $128\times128\times128$ voxels); random intensity shift within [-0.1, 0.1] and scaling within [0.9, 1.1]; and random flipping along the axial, coronal, and sagittal axes with a probability of 0.5. The weight λ and masking ratio are empirically set to 0.1 and 0.875, respectively, based on experimental results on the validation split of BraTS 2018 (see the sensitivity analyses in supplementary Fig. S2).

Comparison with State of the Art (SOTA): Table 1 and Table 2 compare the performance of our framework on the BraTS 2018 and 2020 datasets, respectively, with that of four up-to-date approaches to brain tumor segmentation with missing modalities[1]: U-Net based HVED (HVED; Dorent et al. 2019), ACN (Wang et al. 2021b), SMU-Net (Azad, Khosravi, and Merhof 2022), and RFNet (Ding, Yu, and Yang 2021), with ACN and SMU-Net being dedicated and the other two being catch-all. We reproduce the results of HVED, ACN, and SMU-Net on our data splits by running the authors' codes, and those of RFNet by using the model checkpoint provided by the authors[2].

As we can see, the proposed M^{3}AE yields the strongest performance for all the three evaluated tumor regions and

[1] We also compare to general-purpose multimodal pretraining methods (Geng et al. 2022; Poklukar et al. 2022) in the supplement.

[2] As we use the data splits in (Ding, Yu, and Yang 2021), it is valid to directly use the authors' checkpoints. Meanwhile, since we would like to compare the results without post-processing—to be fair to all compared methods, we reproduce the results instead of directly reporting the numbers with post-processing in (Ding, Yu, and Yang 2021).

Modality				Whole tumor					Tumor core					Enhancing tumor				
FLAIR	T1	T1c	T2	HVED	ACN	SMU-Net	RFNet	M^{3}AE	HVED	ACN	SMU-Net	RFNet	M^{3}AE	HVED	ACN	SMU-Net	RFNet	M^{3}AE
○	○	○	●	77.5*	84.8	84.3	**85.1**	84.8	45.0*	70.4	**70.8**	66.9	69.4	18.7*	42.6	43.4	43.0	**47.6**
○	○	●	○	57.9*	75.7	**77.0***	73.6	75.8	62.7*	81.7	82.5	80.3	**82.9**	55.5*	69.6	69.8	67.7	**73.7**
○	●	○	○	56.9*	**75.1***	74.9	74.8	74.4	40.7*	65.2	63.9	65.2	**66.1**	8.9*	37.1	**38.8**	32.3	37.1
●	○	○	○	77.0*	86.1*	86.4*	85.8*	**88.7**	41.3*	**67.9**	61.6*	62.6*	66.4	17.8*	**38.2**	36.1	35.5	35.6
○	○	●	●	81.7*	84.2*	85.6	85.6	**86.3**	74.4*	78.9*	82.3	82.4	**84.2**	63.3*	67.8	70.6	70.6	**75.3**
○	●	●	○	65.8*	75.7	**77.7***	77.5	77.2	68.9*	79.8	80.4*	81.3	**83.4**	59.5*	68.6	70.5	68.5	**74.7**
●	●	○	○	83.7*	85.4*	84.9*	**89.0**	**89.0**	51.7*	60.6*	61.0*	**72.2**	70.8	16.3*	35.0*	36.1*	38.5	**41.2**
○	●	○	●	80.5*	84.0	85.1	85.4	**86.7**	52.9*	69.7	70.8	71.1	**71.8**	19.3*	42.0	43.3	42.9	**48.7**
●	○	○	●	85.2*	85.8*	86.3*	89.3	**89.9**	51.4*	66.8	66.9	**71.8**	70.9	22.1*	40.1	41.5	**45.4**	**45.4**
●	○	●	○	84.0*	85.5*	86.7*	89.4	**89.7**	71.5*	77.3*	74.8*	81.6*	**84.4**	61.4*	67.2	68.2	72.5	**75.0**
●	●	●	○	85.9*	85.5*	84.4*	**89.9**	88.9	74.1*	78.9*	76.8*	82.3*	**84.1**	61.9*	65.8*	66.2	71.1	**74.0**
●	●	○	●	86.5	84.2*	83.8*	**90.0**	89.9	56.1*	63.9*	60.4*	**74.0**	72.7	22.6*	38.3*	35.5*	**46.0**	44.8
●	○	●	●	87.6*	85.6*	84.4*	**90.4**	90.2	75.1*	79.6*	75.4*	82.6	**84.6**	62.9*	66.1*	67.2*	73.1	**73.8**
○	●	●	●	82.5*	84.9*	83.2*	**86.1**	85.7	75.8*	81.3*	78.8*	82.9	**84.4**	63.6*	67.5	70.4	70.9	**75.4**
●	●	●	●	88.0*	86.2*	85.4*	**90.6***	90.1	76.2*	79.3*	78.3*	82.9	**84.5**	63.0*	67.4	68.1	71.4	**75.5**
Mean				78.7*	83.2*	83.3*	85.5	**85.8**	61.2*	73.4*	72.3*	76.0*	**77.4**	41.1*	54.2*	55.0*	56.6*	**59.9**

Table 1: Performance comparison (DSC % in mean.) with SOTA methods, including HVED (Dorent et al. 2019), ACN (Wang et al. 2021b), SMU-Net (Azad, Khosravi, and Merhof 2022), and RFNet (Ding, Yu, and Yang 2021), on the testing split of BraTS 2018. Present and missing modalities are denoted by ● and ○, respectively. $*$: $p < 0.05$ by Wilcoxon signed rank test for pairwise comparison with our method.

Method	HVED	ACN	SMU-Net	RFNet	M^{3}AE
Whole tumor	80.7*	85.4*	85.3*	86.7	**86.9**
Tumor core	66.5*	77.9	77.7	78.2*	**79.1**
Enhancing tumor	46.7*	59.9*	59.7*	59.7*	**61.7**

Table 2: Performance comparison with SOTA methods (see Table 1 for references) on the testing split of BraTS 2020. The mean performance (DSC % in mean.) of all modal combinations is shown here (due to space limit the detailed performance of each modal combination is given in supplementary Table S3). $*$: $p < 0.05$ by Wilcoxon signed rank test for pairwise comparison with our method.

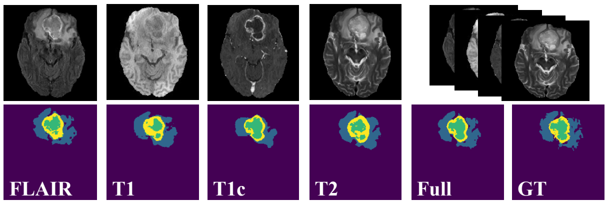

Figure 4: Example segmentation results of the proposed M^{3}AE framework on BraTS 2018 using four individual modalities and all of them. Blue: edema; yellow: enhancing tumor; and green: necrotic and non-enhancing tumor core.

on both datasets, with the highest mean DSCs averaged over all modal combinations. It is worth mentioning that as a catch-all method, our M^{3}AE substantially outperforms the two dedicated methods (ACN and SMU-Net), while using only a single trained model. In contrast, the latter two require 15 models for all the modal combinations. This makes M^{3}AE more efficient to both train and deploy in practice, in addition to being superior in performance. We conjecture that two important facts play key roles here: (1) the co-training strategies employed in the dedicated methods only modelled the one-to-one correlation between the full modalities and each missing-modal situation, whereas our self distillation implicitly models the versatile correlations between all heterogeneous missing-modal situations; and (2) the random modality dropout and patch masking in M^{3}AE are likely to serve as effective data augmentation that helps model training, which is unavailable in paired co-training where the absent modalities are fixed. Meanwhile, our M^{3}AE outperforms RFNet, too, which is the previous best performing method and also a catch-all method. Taking BraTS 2018 for example, compared with RFNet, the mean DSCs of M^{3}AE are slightly higher in whole tumor (85.8% versus 85.5%), apparently higher in tumor core (77.4% versus 76.0%), and substantially higher in enhancing tumor (59.9% versus 56.6%). Besides, using a vanilla encoder-decoder architecture, our M^{3}AE framework is also more memory-economic and computation-efficient to deploy than RFNet, which employed multiple encoders with substantially more parameters and GFLOPS (Fig. 1(c)). In conclusion, the M^{3}AE framework sets a new SOTA for multimodal brain tumor segmentation with missing modalities, while at the same time using an efficient and economic architecture for deployment. Figure 4 shows example segmentation results by M^{3}AE.

Full-Modal Performance: To objectively assess the performance of our framework on full modalities, we also evaluate it on the official validation sets of BraTS 2018 and 2020 online (https://ipp.cbica.upenn.edu/). The models are trained with the same setting as the offline missing-modal evaluation, except that all the cases with public ground truth are used for training without further split (note that the official validation data are kept from training to avoid leakage). The comparison with other methods on BraTS 2018 is shown in

Ablation	Pretrain		Completion			Self	DSC (%) ↑				HD95 (mm) ↓			
config.	Mod.Drop	M^{3}AE	Mean	Zero	Mod.Inv.	distil.	Whole	Core	Enhancing	Mean	Whole	Core	Enhancing	Mean
(a)	×	×	×	×	✓	✓	85.7*	79.9*	53.1*	72.9*	7.4	9.5*	7.9*	8.2*
(b)	✓	×	×	×	✓	✓	**86.8***	80.8*	56.8*	74.8*	7.1	9.1*	5.9	7.4
(c)	×	✓	×	✓	×	✓	86.6	80.8*	55.9*	74.4*	**6.7**	8.5	6.4	7.2
(d)	×	✓	✓	×	×	✓	48.6*	47.9*	39.3*	45.2*	36.0*	36.9*	12.4*	28.4*
(e)	×	✓	×	×	✓	×	86.3	80.8	58.2*	75.1*	7.5*	9.2*	6.3	7.6*
Full	×	✓	×	×	✓	✓	86.5	**81.4**	**59.7**	**75.9**	6.9	**8.4**	**5.8**	**7.0**

Table 3: Ablation studies on effectiveness of our framework's newly proposed components on the validation split of BraTS 2018, by removing or replacing each component from the 'Full' model at a time, including: M^{3}AE self-supervised pretraining, model inversion (Mod.Inv.) based modal completion, and self distillation between heterogeneous missing modalities. The mean performance of all modal combinations is used. The 'Mod.Drop' pretraining refers to removing the patch masking from our M^{3}AE, i.e., using modality dropout (Shen and Gao 2019) alone. *: $p < 0.05$ by Wilcoxon signed rank test for pairwise comparison with the full model.

Method	DSC (%) ↑			HD95 (mm) ↓		
	Whole	Core	Enhancing	Whole	Core	Enhancing
HVED†	88.2*	76.5*	70.9	5.2*	10.5*	6.7
ACN†	89.9*	83.3*	77.9	6.5*	9.4*	4.5
SMU-Net†	90.1	82.3*	79.2	6.8	9.5	6.2*
RFNet†	90.1*	84.6	77.2*	4.8*	**6.9**	5.3*
ModGen†	90.2	82.5*	77.0*	5.6	8.3	4.2
CMJP‡	89.7	84.5	79.7	NA	NA	NA
M^{3}AE	**90.5**	**86.1**	**81.0**	**4.6**	**6.9**	**2.6**
Challenge‡	90.4	86.0	81.5	4.5	8.3	3.8

Table 4: Full-modal performance comparison on the *online* validation set of BraTS 2018, including HVED, ACN, SMU-Net, RFNet (see Table 1 for references), Models Genesis (ModGen; Zhou et al. 2019), and cross-modal jigsaw puzzle (CMJP; Taleb et al. 2021). Single-model performance of the challenge winner (Myronenko 2018) is also included for reference. Best numbers (excluding the challenge entry) are bolded. †: reproduced based on the authors' codes; ‡: provided by the authors; *: $p < 0.05$ by Wilcoxon signed rank test for pairwise comparison with our method; NA: not available. Format: mean, if available.

Table 4 (that on BraTS 2020 is given in supplementary Table S1 due to page limit), including: HVED (Dorent et al. 2019), ACN (Wang et al. 2021b), SMU-Net (Azad, Khosravi, and Merhof 2022), RFNet (Ding, Yu, and Yang 2021), Models Genesis (ModGen, a generic self-supervised representation learning method for medical image analysis; Zhou et al. 2019), and CMJP (Taleb et al. 2021). Performance of the challenge winners (Isensee et al. 2020; Myronenko 2018) is included, too, for reference. Table 4 shows that compared with other non-challenge approaches, our M^{3}AE achieves the best performance in both evaluation metrics for all three tumor regions. It is also mostly comparable and sometimes better than the challenge winner, which involved heavy engineering, e.g., exhaustive parameter tuning. These results indicate that the multimodal representations learned by our framework are not only robust against missing modalities, but also effective with full modalities.

Ablation Study: To validate the efficacy of our framework's novel building components, we conduct ablative experiments where each component is removed or replaced from the complete model. The results are shown in Table 3. In rows (a) and (b), both removing the M^{3}AE pretraining entirely and removing the patch masking (i.e., keeping the modality dropout alone) result in apparent drops in average performance, indicating the indispensable effect of M^{3}AE in learning robust representations of both anatomical and multimodal information against missing modalities. Compared to row (c), substituting our model inversion optimized full-modal image for missing modalities brings obvious improvements in DSCs upon the zero-filling baseline, whereas substituting the mean image of the training set deteriorates the performance sharply (row (d)). These results suggest that our optimized substitute image captures useful modal patterns that can complement the missing modalities for brain tumor segmentation, although looks less realistic than the population mean. Lastly, compared to row (e), the proposed framework is modestly better in both evaluation metrics and for all tumor regions, validating the effectiveness of the two-way hetero-modal knowledge distillation. In addition, the self distillation strategy saves ∼4.7 million parameters compared to co-training with dual networks.

Conclusion

This work presented M^{3}AE, a new framework for brain tumor segmentation using MRI with missing modalities. M^{3}AE featured three novel components: multimodal masked autoencoding for self-supervised learning of robust representations against missing modalities, model inversion based modality completion, and memory-efficient self distillation between heterogeneous missing-modal situations. As a 'catch-all' model, M^{3}AE could accommodate all possible combinations of missing modalities with a single trained model. Extensive experiments on two public benchmark datasets showed that our framework established a new state of the art for brain tumor segmentation with missing modalities and that it was competent in multimodal representation learning. In addition, our ablative experiments validated the efficacy of M^{3}AE's three novel components. In the future, we plan to apply M^{3}AE to completely different modalities (e.g., MRI and CT) and other benchmarks beyond BraTS.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant 2020AAA0109500/ 2020AAA0109501, and in part by the National Key Research and Development Program of China (2019YFE0113900).

References

Azad, R.; Khosravi, N.; and Merhof, D. 2022. SMU-Net: Style matching U-Net for brain tumor segmentation with missing modalities. *arXiv preprint arXiv:2204.02961.*

Bakas, S.; et al. 2018. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv preprint arXiv:1811.02629.*

Blum, A.; and Mitchell, T. 1998. Combining labeled and unlabeled data with co-training. In *Proc. Ann. Conf. on Comput. Learning Theory*, 92–100.

Chen, C.; Dou, Q.; Jin, Y.; Liu, Q.; and Heng, P. A. 2021. Learning with privileged multimodal knowledge for unimodal segmentation. *IEEE Trans. Med. Imag.*, 41(3): 621–632.

Chen, H.; Qin, Z.; Ding, Y.; Tian, L.; and Qin, Z. 2020. Brain tumor segmentation with deep convolutional symmetric neural network. *Neurocomputing*, 392: 305–313.

Chen, S.; Ding, C.; and Liu, M. 2019. Dual-force convolutional neural networks for accurate brain tumor segmentation. *Pattern Recognit.*, 88: 90–100.

Ding, Y.; Gong, L.; Zhang, M.; Li, C.; and Qin, Z. 2020. A multi-path adaptive fusion network for multimodal brain tumor segmentation. *Neurocomputing*, 412: 19–30.

Ding, Y.; Yu, X.; and Yang, Y. 2021. RFNet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 3975–3984.

Dorent, R.; Joutard, S.; Modat, M.; Ourselin, S.; and Vercauteren, T. 2019. Hetero-modal variational encoder-decoder for joint modality completion and segmentation. In *Proc. Int. Conf. MICCAI*, 74–82. Springer.

Ge, Y.; et al. 2021. Self-distillation with batch knowledge ensembling improves ImageNet classification. *arXiv preprint arXiv:2104.13298.*

Geng, X.; Liu, H.; Lee, L.; Schuurams, D.; Levine, S.; and Abbeel, P. 2022. Multimodal Masked Autoencoders Learn Transferable Representations. *arXiv preprint arXiv:2205.14204.*

Goodfellow, I.; et al. 2014. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.*, 27.

Havaei, M.; Guizard, N.; Chapados, N.; and Bengio, Y. 2016. HeMIS: Hetero-modal image segmentation. In *Proc. Int. Conf. MICCAI*, 469–477. Springer.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 16000–16009.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 770–778.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531.*

Hu, M.; et al. 2020. Knowledge distillation from multi-modal to mono-modal segmentation networks. In *Proc. Int. Conf. MICCAI*, 772–781. Springer.

Isensee, F.; Jäger, P. F.; Full, P. M.; Vollmuth, P.; and Maier-Hein, K. H. 2020. nnU-Net for brain tumor segmentation. In *International MICCAI Brainlesion Workshop*, 118–132. Springer.

Iv, M.; Yoon, B. C.; Heit, J. J.; Fischbein, N.; and Wintermark, M. 2018. Current clinical state of advanced magnetic resonance imaging for brain tumor diagnosis and follow up. *Semin. Roentgenol.*, 53(1): 45–61.

Ji, M.; Shin, S.; Hwang, S.; Park, G.; and Moon, I.-C. 2021. Refine myself by teaching myself: Feature refinement via self-knowledge distillation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 10664–10673.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980.*

Lee, C.-Y.; Xie, S.; Gallagher, P.; Zhang, Z.; and Tu, Z. 2015. Deeply-supervised nets. In *Artif. Intell. Stat.*, 562–570. PMLR.

Lee, D.; Moon, W.-J.; and Ye, J. C. 2020. Assessing the importance of magnetic resonance contrasts using collaborative generative adversarial networks. *Nat. Mach. Intell.*, 2(1): 34–42.

Loshchilov, I.; and Hutter, F. 2016. SGDR: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983.*

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proc. Int. Conf. 3D Vis.*, 565–571. IEEE.

Myronenko, A. 2018. 3D MRI brain tumor segmentation using autoencoder regularization. In *Int. MICCAI Brainlesion Workshop*, 311–320. Springer.

Nguyen, A.; Dosovitskiy, A.; Yosinski, J.; Brox, T.; and Clune, J. 2016. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *Adv. Neural Inf. Process. Syst.*, 29.

Paszke, A.; et al. 2019. PyTorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.*, 32.

Poklukar, P.; Vasco, M.; Yin, H.; Melo, F. S.; Paiva, A.; and Kragic, D. 2022. Geometric Multimodal Contrastive Representation Learning. In *Int. Conf. Mach. Learn.*, 17782–17800. PMLR.

Shen, Y.; and Gao, M. 2019. Brain tumor segmentation on MRI with missing modalities. In *Proc. Int. Conf. IPMI*, 417–428. Springer.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at Int. Conf. Learning Representations*.

Taleb, A.; Lippert, C.; Klein, T.; and Nabi, M. 2021. Multimodal self-supervised learning for medical image analysis. In *Proc. Int. Conf. IPMI*, 661–673. Springer.

Taleb, A.; et al. 2020. 3D self-supervised methods for medical imaging. *Adv. Neural Inf. Process. Syst.*, 33: 18158–18172.

Wang, P.; Li, Y.; Singh, K. K.; Lu, J.; and Vasconcelos, N. 2021a. IMAGINE: Image synthesis by image-guided model inversion. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 3681–3690.

Wang, Y.; et al. 2021b. ACN: Adversarial co-training network for brain tumor segmentation with missing modalities. In *Proc. Int. Conf. MICCAI*, 410–420. Springer.

Wu, Y.; and He, K. 2018. Group normalization. In *Proc. Eur. Conf. Comput. Vis.*, 3–19.

Yu, B.; Zhou, L.; Wang, L.; Shi, Y.; Fripp, J.; and Bourgeat, P. 2019. Ea-GANs: edge-aware generative adversarial networks for cross-modality MR image synthesis. *IEEE Trans. Med. Imag.,*, 38(7): 1750–1762.

Zhang, P.; Wang, F.; and Zheng, Y. 2017. Self supervised deep representation learning for fine-grained body part recognition. In *IEEE Int. Symp. Biomed. Imaging*, 578–582.

Zhou, C.; Ding, C.; Wang, X.; Lu, Z.; and Tao, D. 2020. One-pass multi-task networks with cross-task guided attention for brain tumor segmentation. *IEEE Trans. Image Process.*, 29: 4516–4529.

Zhou, T.; Canu, S.; Vera, P.; and Ruan, S. 2021a. Feature-enhanced generation and multi-modality fusion based deep neural network for brain tumor segmentation with missing MR modalities. *Neurocomputing*, 466: 102–112.

Zhou, T.; Canu, S.; Vera, P.; and Ruan, S. 2021b. Latent correlation representation learning for brain tumor segmentation with missing MRI modalities. *IEEE Trans. Image Process.*, 30: 4263–4274.

Zhou, Z.; et al. 2019. Models Genesis: Generic autodidactic models for 3D medical image analysis. In *Proc. Int. Conf. MICCAI*, 384–393. Springer.