

Unsupervised Tumor-Aware Distillation for Multi-Modal Brain Image Translation

Chuan Huang¹, Jia Wei^{1*}, Rui Li²

¹School of Computer Science and Engineering, South China University of Technology, Guangzhou, China
cschuan@mail.scut.edu.cn csjwei@scut.edu.cn

²Golisano College of Computing and Information Sciences, Rochester Institute of Technology, Rochester, NY, USA
rxlics@

Abstract—Multi-modal brain images from MRI scans are widely used in clinical diagnosis to provide complementary information from different modalities. However, obtaining fully paired multi-modal images in practice is challenging due to various factors, such as time, cost, and artifacts, resulting in modality-missing brain images. To address this problem, unsupervised multi-modal brain image translation has been extensively studied. Existing methods suffer from the problem of brain tumor deformation during translation, as they fail to focus on the tumor areas when translating the whole images. In this paper, we propose an unsupervised tumor-aware distillation teacher-student network called UTAD-Net, which is capable of perceiving and translating tumor areas precisely. Specifically, our model consists of two parts: a teacher network and a student network. The teacher network learns an end-to-end mapping from source to target modality using unpaired images and corresponding tumor masks first. Then, the translation knowledge is distilled into the student network, enabling it to generate more realistic tumor areas and whole images without masks. Experiments show that our model achieves competitive performance on both quantitative and qualitative evaluations of image quality compared with state-of-the-art methods. Furthermore, we demonstrate the effectiveness of the generated images on downstream segmentation tasks. Our code is available at <https://github.com/scut-HC/UTAD-Net>.

Index Terms—Brain image translation, Multi-modal, Knowledge distillation, Tumor-aware, Unsupervised learning

I. INTRODUCTION

Multi-modal brain images from MRI (Magnetic Resonance Imaging) scans are widely used in various clinical scenarios[1], [2]. These images are further divided into several modalities(sequences), such as T1-weighted (T1), T1-with-contrast-enhanced (T1ce), T2-weighted (T2), T2-fluid-attenuated inversion recovery (Flair), etc. Each modality exhibits distinct contrasts, providing complementary lesion information from different perspectives. As shown in Figure 1, Flair and T2 images depict the peritumoral edema areas clearly, while T1 images highlight the white and gray matter tissues[3], making them suitable for presenting anatomical structures. T1ce images delineate the structures and the edges of the tumors[4], which is convenient to observe the morphology of different types of tumors. Fully paired multi-modal images assist doctors in achieving more precise diagnoses[5].

The benefits of using multi-modal images to assist medical analysis have been widely recognized[6]. However, physi-

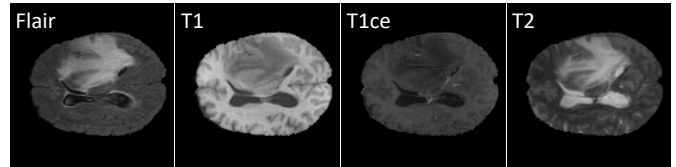


Fig. 1. Examples of brain tumor MRI images from different modalities.

cians usually obtain some of the modalities in practice due to practical considerations such as time, cost, and artifacts. As a result, many images are modality-missing, which can negatively impact the accuracy of diagnoses. Consequently, there has been significant research attention on generating images of missing modalities[7], [8].

Existing methods for multi-modal image translation have shown promising results in natural images. However, when applied to medical images, particularly brain tumor images, the results are often unsatisfactory[3]. Compared with two-dimensional natural images, three-dimensional medical images have more structural information[9], leading to blur or deformation in image translation. Moreover, due to the privacy of patients, a large number of medical images collected by different institutions are private, which increases the difficulty of model training. Therefore, translating brain tumor images remains a challenging task. Some methods propose using segmentation or object detection networks to assist translation[10], [11], [12], while others propose loss functions to preserve tumor information[13], [5], [14]. However, they either require paired images for training or still rely on additional networks or labels during inference.

To solve the problem of brain tumor deformation during translation, we propose an Unsupervised Tumor-Aware Distillation Teacher-student Network called UTAD-Net. Our model comprises a teacher network and a student network. The teacher network consists of two branches: the global branch translates the input whole images into the target modality, while the local branch translates the tumor images obtained through pixel-level multiplication of the whole images and their corresponding tumor masks. Each branch is composed of a pair of encoders and decoders. Unlike other distillation models, the student network adopts an identical network structure as the teacher network but reduces the input information.

* Jia Wei is the corresponding author.

What is inputted into the local branch is not the tumor images, but the whole images. The encoder of the local branch in the student network is trained to perceive the tumor areas and learn features similar to those encoded by the encoder of the teacher network for the tumor images. We also propose a local consistency loss to preserve the anatomical structure of the tumors. All training samples consist of unpaired single-modality images. In this way, the student network can focus on the tumor areas and generate more realistic images without the tumor masks. During the inference phase, we directly employ the student network to translate the input source modality images into the final target modality images.

In summary, the main contributions of this paper are as follows:

- we propose UTAD-Net to accurately perceive the tumor areas and generate more realistic images with clearer textures and richer structural details using unpaired brain tumor images, assisted by additional tumor masks and the local consistency loss.
- To the best of our knowledge, we are the first to propose reducing the input information of the student network in the distillation model. This approach allows the teacher network to capture the tumor areas with the guidance of the tumor masks, while the student network acquires knowledge from the teacher network through distillation and gains the ability to generate images without masks.
- We present qualitative and quantitative evaluations of image quality for multi-modal brain images translation on the BRATS2020 dataset[15]. Our model outperforms state-of-the-art methods. We also demonstrate that the images generated by our model are helpful for improving the segmentation performance through downstream tasks.

II. RELATED WORKS

Cross-modality image translation has been intensively studied in recent years. For instance, Pix2pix[16] presents a cGAN-based[17] method to convert images between source and target modalities. However, it requires paired data for training, which is hard to realize. Therefore, unsupervised image translation using unpaired data has piqued researcher interest. CycleGAN[18] and DiscoGAN[19] propose a cycle consistency loss, which attempts to preserve the crucial information of the images. By constraining the reconstructed image and the source image, the model is available to translate images between the two given modalities with unpaired data. UNIT[20] assumes that the two modalities share the same latent space, and proposes to combine VAE and GAN to form a more robust generative model. The encoder maps the images of different domains to the same distribution to obtain the latent code, and then the decoder maps the latent code back to the image domain. However, these models fail to generate diverse styles in the images. The resulting target modal image for a given input image is singular. In order to solve this problem, MUNIT[21] and DRIT[22] disentangle the latent code into the content code which is shared by different modalities and the style code which is unique for different modalities and

restricted to normal distribution. During inference, the style code can be obtained through sampling, allowing the model to generate images with diverse styles.

However, the above models can only translate images between two modalities. If we want to translate images between n modalities, we need to train the model for $n(n-1)/2$ times. In order to perform in a unified model to translate multi-modality images, StarGAN[23] proposed a single generator to learn the mapping between any two given modalities. The source images and mask vectors are inputted to the generator which then outputs the generated target images. The discriminator needs to not only distinguish whether the images are real or false, but also classified the domain they belong to. Since every mask vector is corresponding to a given condition, the generated images are simplex without style diversity. DRTI++[24] adds domain codes for translation so that any target modality images can be generated by a unified generator. ResViT[25] introduces a transformer-based model that exhibits broad adaptability across diverse configurations of source-target modalities.

Although the above model can achieve multi-modality translation, it can not focus on local targets but only on the whole image. CSCG[13] propose that for unsupervised learning, cycle consistency loss will easily lead to local deformation of the image if there are no other constraints. InstaGAN[26] proposed to add segmentation labels of local instances as additional input information so that the network will pay more attention to the shape of the local instances in the training process and reduces the deformation. DUNIT[12] and INIT[10] respectively propose to use object detection and segmentation to assist translation. Ea-GANs[27] proposes to integrate edge maps that contain critical textural information to boost synthesis quality. TC-MGAN[5] introduces a multi-modality tumor consistency loss to preserve the critical tumor information in the target-generated images but it can only translate the images from the T2 modality to other MR modalities. TarGAN[11] focuses on the target area by using a shape controller. While these models can translate images more effectively, they also require more supervised information.

Some of the above methods can only translate images between two given modalities, and some require paired and labeled data for training, which is not completely consistent with the practical application scenarios that most data are unpaired. We propose UTAD-Net to learn an end-to-end mapping from an arbitrary source modality to the given target modality, which can focus on the local tumor areas and translate better by using unpaired images.

III. METHODS

In this section, we first describe our framework and the pipeline of our approach, then we define the training objective functions.

A. Framework and pipeline

For the teacher network, given an image I_{source} from the source modality s and corresponding mask M , we aim to train

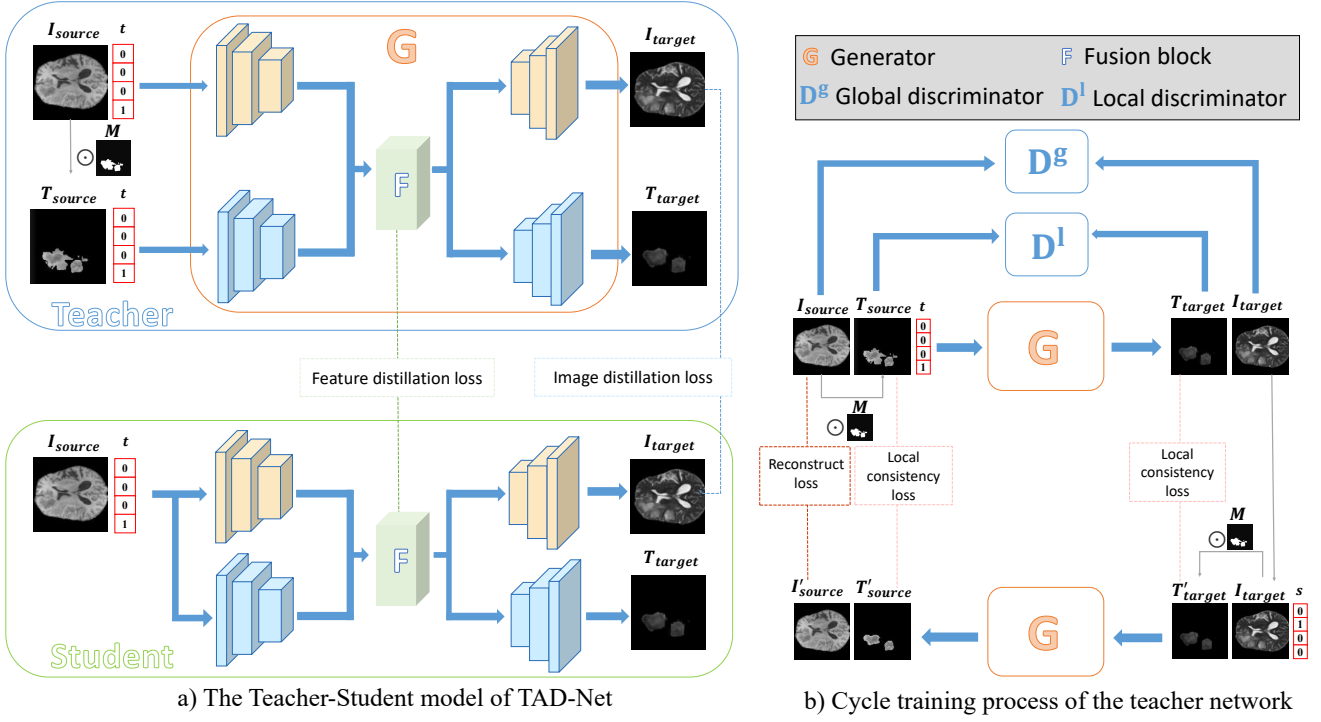


Fig. 2. Overview of the proposed UTAD-Net. a) The generator G comprises two encoder-decoder pairs and a fusion block F to generate the whole image and tumor image. The student network learns the knowledge of the teacher network through distillation learning at both the feature and image levels. b) The generator G tries to reconstruct the whole image and tumor image. The global discriminator D^g determines whether the whole image is real or fake and classifies its modality, while the local discriminator D^l models the tumor image.

a model that can translate both the whole image and the tumor area to the target modality t . To achieve this, we first multiply the two on pixel-level to get the source tumor image T_{source} . Then the source whole image I_{source} is fed to the global branch to generate the target whole image I_{target} , and the source tumor image T_{source} is fed to the local branch to generate the target tumor image T_{target} , respectively. The mapping is denoted as: $(I_{target}, T_{target}) = G(I_{source}, T_{source}, t)$.

The student network adopts the identical network structure as the teacher network, but with a key difference: it does not require the tumor image T_{source} as input. Unlike the teacher network, the local branch of the student network takes the whole image as input instead of the tumor image, allowing the model to generate the target image without a tumor mask, which is denoted as: $(I_{target}, T_{target}) = G(I_{source}, t)$. The teacher-student model is shown in Figure 2.

Generator G . The generator is comprised of two encoder-decoder pairs, one for the global branch and the other for the local branch. We use a fusion block F proposed by Hi-Net[28] to fuse features from the encoders of the two branches. The decoders of the two branches receive the fused feature and respectively generate the target global image I_{target} and the target tumor image T_{target} . Then the generator translates the target whole image I_{target} and its corresponding tumor image (only used in the teacher network) T'_{target} to the reconstructed whole image I'_{source} and tumor image T'_{source} . In this way, a cycle training process is accomplished.

Discriminator D^g and D^l . Similar to StarGAN[23], the discriminator tries to not only distinguish whether the image is real or fake but also judge the modality to which it belong. Our model utilizes two discriminators: D^g is responsible for the whole image in the global branch and D^l is responsible for the tumor image in the local branch respectively.

B. Training objective functions

Adversarial loss. Adversarial loss is designed to make the images generated by the generator more realistic to confuse the discriminator. To train the model more stably, we use the adversarial loss WGAN-GP[29] as follows:

$$\mathcal{L}_{adv}^g = \mathbb{E}_{I_{source}} [D_{src}^g(I_{source})] - \mathbb{E}_{I_{target}} [D_{src}^g(I_{target})] - \lambda_{gp} \mathbb{E}_{\hat{I}} [(\|\nabla D_{src}^g(\hat{I})\|_2 - 1)^2], \quad (1)$$

$$\mathcal{L}_{adv}^l = \mathbb{E}_{T_{source}} [D_{src}^l(T_{source})] - \mathbb{E}_{T_{target}} [D_{src}^l(T_{target})] - \lambda_{gp} \mathbb{E}_{\hat{T}} [(\|\nabla D_{src}^l(\hat{T})\|_2 - 1)^2], \quad (2)$$

where $\hat{I}$ and $\hat{T}$ are uniformly sampled along a straight line connecting the corresponding real images and generated images. λ_{gp} is set as 10.0 in our model. D_{src}^g and D_{src}^l represent the probability distributions of real or fake for the images generated by the global and local branches respectively.

Modality classification loss. Given an image and its target modality vector, we hope the generator can generate images that are as close to the target modality as possible. Similar to StarGAN, the discriminators aim to judge the modality

they belong to. For real images, we define the modality classification loss which is used to optimize the discriminators as follows:

$$\mathcal{L}_{r_cls}^g = \mathbb{E}_{I_{source}, s}[-\log D_{cls}^g(s|I_{source})], \quad (3)$$

$$\mathcal{L}_{r_cls}^l = \mathbb{E}_{T_{source}, s}[-\log D_{cls}^l(s|T_{source})], \quad (4)$$

where the term $D_{cls}^g(s|I_{source})$ and $D_{cls}^l(s|T_{source})$ represent the probability distribution over the modality vector for the whole images and tumor images. Similarly, we define the modality classification loss for fake images which is used to optimize the generator as follows:

$$\mathcal{L}_{f_cls}^g = \mathbb{E}_{I_{target}, t}[-\log D_{cls}^g(t|I_{target})], \quad (5)$$

$$\mathcal{L}_{f_cls}^l = \mathbb{E}_{T_{target}, t}[-\log D_{cls}^l(t|T_{target})], \quad (6)$$

$$\mathcal{L}_{f_cls} = \mathcal{L}_{f_cls}^g + \mathcal{L}_{f_cls}^l. \quad (7)$$

Local consistency loss. We aim to constrain the similarity of the generated tumor images T_{target} and the tumor areas of the generated whole images T'_{target} to alleviate the problem of distortion in brain tumor image translation. The reconstructed tumor images T'_{source} and the source tumor images T_{source} are constrained in the same way. We thus propose a local consistency loss as an extra constraint to improve the translation effect, which is defined as follows:

$$\begin{aligned} \mathcal{L}_{local} = & \mathbb{E}(\|T_{target} - T'_{target}\|_1) \\ & + \mathbb{E}(\|T_{source} - T'_{source}\|_1). \end{aligned} \quad (8)$$

Reconstruct loss. A challenge is to ensure that the generated images I_{target} just simply change the image style information and still contain all the content information of the source images I_{source} . To solve this problem, we feed I_{target} into the translation network for cycle translation to obtain the reconstructed images I'_{source} . To minimize the difference between I'_{source} and I_{source} , we add a reconstruct loss which is defined as follows:

$$\mathcal{L}_{rec} = \mathbb{E}(\|I_{source} - I'_{source}\|_1). \quad (9)$$

Distillation loss. To allow the student network to perceive the tumor areas like the teacher network without masks, we impose constraints on the intermediate layer features and the final generated target images of both networks, so that the student network can better learn the knowledge from the teacher network. The distillation loss is defined as follows:

$$\begin{aligned} \mathcal{L}_{dis} = & \mathbb{E}(\|F^{teacher} - F^{student}\|_1) \\ & + \mathbb{E}(\|I_{target}^{teacher} - I_{target}^{student}\|_1), \end{aligned} \quad (10)$$

where $F^{teacher}$ and $F^{student}$ mean the fused feature of the teacher and student networks.

Total loss. Combining all the losses mentioned above, we finally defined the objective function as follows:

$$\mathcal{L}_D^{g/l} = -\mathcal{L}_{adv}^{g/l} + \mathcal{L}_{r_cls}^{g/l}, \quad (11)$$

$$\mathcal{L}_G^{teacher} = \mathcal{L}_{adv}^g + \mathcal{L}_{adv}^l + \mathcal{L}_{f_cls} + \lambda_1(\mathcal{L}_{rec} + \mathcal{L}_{local}) \quad (12)$$

$$\mathcal{L}_G^{student} = \mathcal{L}_G^{teacher} + \lambda_2 \mathcal{L}_{dis}, \quad (13)$$

where λ_1 and λ_2 are hyper-parameters to balance losses. We set λ_1 and λ_2 to be 10.0 in our experiments.

IV. EXPERIMENTS

A. Settings

1) *Datasets:* We conduct all our experiments on the BRATS 2020 dataset (<https://www.med.upenn.edu/cbica/brats2020/data.html>)[30], which provides brain tumor images of four modalities: Flair, T1, T1ce and T2. All images have been manually segmented by experienced raters. Annotations include the GD-enhancing tumor (ET), peritumoral edema (ED), as well as the necrotic and non-enhancing tumor core (NCR/NET)[31], [15]. The tumors are classified into three distinct regions according to clinical references: Whole Tumor (WT = ET + ED + NCR/NET), Tumor Core (ET + NCR/NET), and Enhancing Tumor (ET). We calculated masks corresponding to these three areas based on the annotations to guide the brain image translation. 180 samples are randomly selected for training, 30 for validation and 159 for testing in the translation task. For each sample, we select images from only one randomly chosen modality. We trained three sets of models using the aforementioned three types of tumor masks, enabling translation between any two of the mentioned four modalities. To further verify the effectiveness of the model, we feed the fully multi-modality images generated by different methods into nnU-Net[32] to compare their segmentation effectiveness as a downstream segmentation task. 120 samples are selected for training and 69 for testing. It is worth noting that none of the samples used for the downstream segmentation task are included in the training set for the translation task, avoiding the resulting metric being biased positively towards our model. All the images are resized to 128×128 pixels.

2) *Evaluation metrics:* For the translation task, we use structural similarity index measure (SSIM), peak-signal-noise ratio (PSNR)[34] and learned perceptual image patch similarity (LPIPS)[35] to measure the similarity between the generated images and ground truth. For the downstream segmentation task, we use Dice similarity coefficient (DSC)[36], average symmetric surface distance (ASSD), 95th percentile of Hausdorff distance (HD95) to measure the integrity of the predicted pseudo masks generated by nnU-Net[32] which is an acknowledged state-of-the-art medical image segmentation model.

3) *Baselines:* We compare our translation results with StarGAN [23], DRIT++[24], Targan[11], ReMIC[33] and ResViT[25]. StarGAN proposes using a unified model to translate images to arbitrary modalities. DRIT++ disentangles an image into the content code and the attribute code during the training and generates images using the content code extracted from the input images and the attribute code sampled from the standard normal distribution. TarGAN utilizes an extra shape controller to alleviate the problem of image deformation in the target area. ReMIC learns shared content and domain-specific style encoding across multiple domains. ResViT proposes a unified synthesis model with a transformer-based generator for medical image synthesis. Note that we implement unsupervised versions for ReMIC and ResViT to enable a fair comparison.

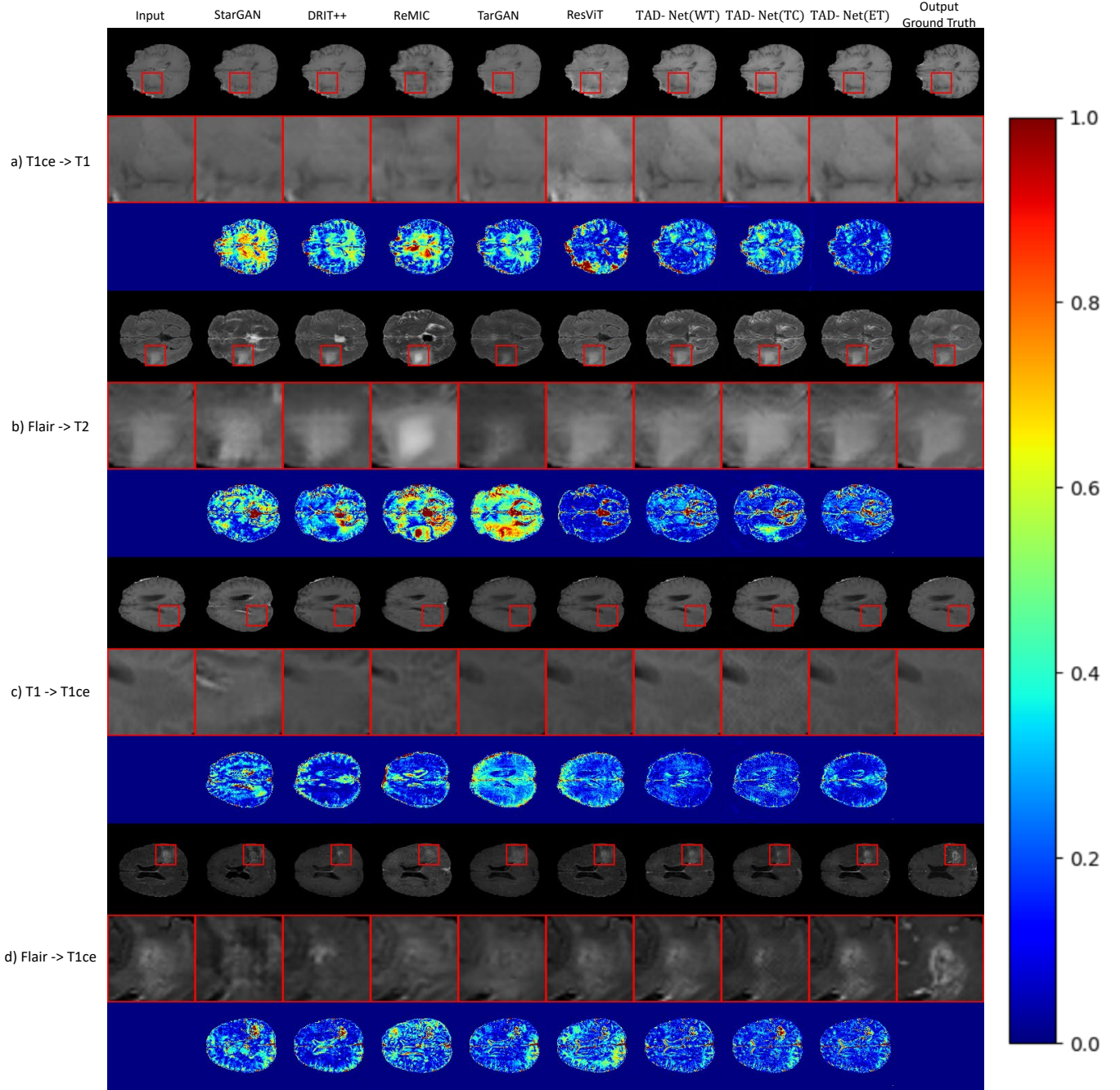


Fig. 3. Qualitative evaluations of our model and the other baselines on the BRATS2020 dataset. The four samples are denoted as: a) T1ce→T1. b) Flair→T2. c) T1→T1ce. d) Flair→T1ce. For each sample, the first row represents the translation of the whole images, the second row depicts the magnified image of the tumor areas, and the third row illustrates the error map between the generated images and the ground truth (GT) images. Red boxes demonstrate that our method generates more realistic images with clearer textures and richer structural details.

4) Implementation details: We implement PatchGAN[16] as the backbone for both the global discriminator and local discriminator, and U-net[37] as the backbone for the generator. We train the teacher network for 100 epochs with a learning rate of 10^{-4} for both the generator and the discriminators during the first 50 epochs and then linearly decay the learning rate to 10^{-6} at the final epoch. Then we train the student network in the same way. Adam[38] optimizer is used with

momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We also adopt data augmentation and normalization for the training samples. All the experiments are conducted on PyTorch 1.8.1 with NVIDIA RTX 3090(24G).

B. Translation results

1) Qualitative evaluation: Figure 3 shows the qualitative results of our model and the other baselines. Our method

TABLE I

QUANTITATIVE EVALUATION OF THE GENERATED IMAGES. WE REPORT THE MEAN VALUE FOR TRANSLATION BETWEEN ANY TWO MODALITIES HERE. THE SYMBOL $\uparrow$ DENOTES HIGHER IS BETTER WHILE THE SYMBOL $\downarrow$ DENOTES LOWER IS BETTER. UTAD-NET(WT) DENOTES OUR MODEL GUIDED BY THE WHOLE TUMOR MASK, AND THE SAME APPLIES TO THE OTHERS.

Method		SSIM $\uparrow$	PSNR $\uparrow$	local SSIM $\uparrow$	local PSNR $\uparrow$	LPIPS $\downarrow$
Baseline	StarGAN[23]	0.8477 $\pm$ 0.0005	28.02 $\pm$ 1.30	0.5886 $\pm$ 0.0008	18.48 $\pm$ 1.40	0.0930 $\pm$ 0.0228
	DRIT++[24]	0.8615 $\pm$ 0.0008	30.24 $\pm$ 1.20	0.6095 $\pm$ 0.0009	19.16 $\pm$ 1.26	0.0802 $\pm$ 0.0276
	TarGAN[11]	0.8588 $\pm$ 0.0006	29.66 $\pm$ 1.24	0.6037 $\pm$ 0.0010	18.78 $\pm$ 1.31	0.0860 $\pm$ 0.0235
	ReMIC[33]	0.8459 $\pm$ 0.0007	29.31 $\pm$ 1.41	0.6010 $\pm$ 0.0013	18.69 $\pm$ 1.14	0.0874 $\pm$ 0.0262
	ResViT[25]	0.8735 $\pm$ 0.0005	30.45 $\pm$ 1.37	0.6068 $\pm$ 0.0009	19.33 $\pm$ 1.20	0.0702 $\pm$ 0.0236
Teacher	UTAD-Net(WT)	0.8757 $\pm$ 0.0004	31.57 $\pm$ 1.14	0.6195 $\pm$ 0.0008	20.72$\pm$1.33	0.0655 $\pm$ 0.0214
	UTAD-Net(TC)	0.8767$\pm$0.0005	31.89$\pm$1.29	0.6199$\pm$0.0007	20.44 $\pm$ 1.22	0.0647$\pm$0.0249
	UTAD-Net(ET)	0.8751 $\pm$ 0.0006	31.61 $\pm$ 1.42	0.6166 $\pm$ 0.0008	20.39 $\pm$ 1.16	0.0690 $\pm$ 0.0243
Student	UTAD-Net(WT)	0.8742 $\pm$ 0.0006	31.48$\pm$1.26	0.6147 $\pm$ 0.0009	20.21$\pm$1.23	0.0664 $\pm$ 0.0230
	UTAD-Net(TC)	0.8753$\pm$0.0004	31.34 $\pm$ 1.30	0.6156$\pm$0.0012	19.96 $\pm$ 1.26	0.0661$\pm$0.0255
	UTAD-Net(ET)	0.8723 $\pm$ 0.0005	30.88 $\pm$ 1.15	0.6143 $\pm$ 0.0008	19.93 $\pm$ 1.17	0.0716 $\pm$ 0.0268
Ablation	UTAD-Net(zeros map)	0.8535 $\pm$ 0.0007	29.26 $\pm$ 1.48	0.6013 $\pm$ 0.0005	18.75 $\pm$ 1.59	0.0869 $\pm$ 0.0264
	UTAD-Net(random map)	0.8509 $\pm$ 0.0007	28.47 $\pm$ 1.40	0.5994 $\pm$ 0.0009	18.26 $\pm$ 1.31	0.0884 $\pm$ 0.0281

TABLE II

THE QUANTITATIVE EVALUATION FOR DOWNSTREAM SEGMENTATION TASK CONDUCTED BY NNU-NET.

Method		DSC			ASSD(mm)			HD95(mm)		
		WT	TC	ET	WT	TC	ET	WT	TC	ET
Baseline	SingleGT	0.8012 $\pm$ 0.0217	0.6944 $\pm$ 0.0256	0.4775 $\pm$ 0.0273	1.69 $\pm$ 0.52	2.44 $\pm$ 0.65	2.92 $\pm$ 0.61	5.69 $\pm$ 1.48	6.72 $\pm$ 1.65	7.73 $\pm$ 1.59
	StarGAN[23]	0.7520 $\pm$ 0.0243	0.6357 $\pm$ 0.0284	0.4018 $\pm$ 0.0310	3.77 $\pm$ 0.72	4.01 $\pm$ 0.76	3.70 $\pm$ 0.88	10.57 $\pm$ 3.02	10.89 $\pm$ 3.16	11.48 $\pm$ 3.28
	DRIT++[24]	0.7867 $\pm$ 0.0226	0.6749 $\pm$ 0.0271	0.4403 $\pm$ 0.0263	2.53 $\pm$ 0.56	2.62 $\pm$ 0.84	2.81 $\pm$ 0.63	7.79 $\pm$ 2.54	8.12 $\pm$ 2.10	8.63 $\pm$ 2.26
	TarGAN[11]	0.7782 $\pm$ 0.0283	0.6670 $\pm$ 0.0265	0.4361 $\pm$ 0.0301	2.51 $\pm$ 0.75	2.89 $\pm$ 0.72	2.96 $\pm$ 0.80	7.93 $\pm$ 2.67	8.24 $\pm$ 2.41	8.81 $\pm$ 2.79
	ReMIC[33]	0.7737 $\pm$ 0.0294	0.6618 $\pm$ 0.0322	0.4328 $\pm$ 0.0306	2.63 $\pm$ 0.63	2.95 $\pm$ 0.68	2.78 $\pm$ 0.81	8.03 $\pm$ 2.60	8.39 $\pm$ 2.23	8.75 $\pm$ 2.59
	ResViT[25]	0.7983 $\pm$ 0.0233	0.7037 $\pm$ 0.0270	0.4756 $\pm$ 0.0286	2.83 $\pm$ 0.69	2.29 $\pm$ 0.78	2.68 $\pm$ 0.74	7.43 $\pm$ 2.49	7.03 $\pm$ 2.28	7.56 $\pm$ 2.64
Teacher	UTAD-Net(WT)	0.8190$\pm$0.0224	0.6987 $\pm$ 0.0243	0.4573 $\pm$ 0.0290	1.53$\pm$0.56	2.14 $\pm$ 0.60	2.32 $\pm$ 0.78	3.89$\pm$1.79	6.17 $\pm$ 1.50	7.80 $\pm$ 2.46
	UTAD-Net(TC)	0.8040 $\pm$ 0.0256	0.7432$\pm$0.0284	0.5176 $\pm$ 0.0293	1.79 $\pm$ 0.63	1.81$\pm$0.70	1.75 $\pm$ 0.74	6.09 $\pm$ 1.84	5.78$\pm$2.06	5.71 $\pm$ 2.36
	UTAD-Net(ET)	0.8004 $\pm$ 0.0263	0.7097 $\pm$ 0.0305	0.5452$\pm$0.0303	1.92 $\pm$ 0.65	2.17 $\pm$ 0.72	1.68$\pm$0.79	6.50 $\pm$ 2.42	6.40 $\pm$ 2.52	5.46$\pm$2.27
Student	UTAD-Net(WT)	0.8084$\pm$0.0258	0.6937 $\pm$ 0.0262	0.4483 $\pm$ 0.0294	1.62$\pm$0.60	2.56 $\pm$ 0.68	2.52 $\pm$ 0.71	5.17$\pm$2.15	7.10 $\pm$ 2.36	8.05 $\pm$ 2.63
	UTAD-Net(TC)	0.7992 $\pm$ 0.0241	0.7184$\pm$0.0285	0.5011 $\pm$ 0.0321	1.82 $\pm$ 0.60	2.05$\pm$0.63	2.49 $\pm$ 0.69	6.17 $\pm$ 2.48	6.49$\pm$2.35	6.76 $\pm$ 2.65
	UTAD-Net(ET)	0.7953 $\pm$ 0.0262	0.6840 $\pm$ 0.0246	0.5255$\pm$0.0255	2.13 $\pm$ 0.67	2.59 $\pm$ 0.66	2.42$\pm$0.75	6.62 $\pm$ 2.62	7.96 $\pm$ 2.58	6.26$\pm$2.56
Ablation	UTAD-Net(zeros map)	0.7772 $\pm$ 0.0265	0.6652 $\pm$ 0.0294	0.4308 $\pm$ 0.0283	2.67 $\pm$ 0.62	2.91 $\pm$ 0.76	2.93 $\pm$ 0.75	8.01 $\pm$ 2.76	8.44 $\pm$ 2.66	8.92 $\pm$ 2.68
	UTAD-Net(random map)	0.7756 $\pm$ 0.0248	0.6631 $\pm$ 0.0315	0.4306 $\pm$ 0.0287	2.75 $\pm$ 0.65	2.96 $\pm$ 0.73	2.86 $\pm$ 0.86	8.10 $\pm$ 2.80	8.52 $\pm$ 2.78	8.84 $\pm$ 2.77

generates more realistic images with clearer textures and richer structural details.

2) *Quantitative evaluation:* As shown in Table I, the teacher networks of our models guided by different masks significantly outperform the baseline on various metrics, and the performance of the student network is slightly worse than the teacher network but still better than all baselines. The value is the average of all the possible mappings. In addition to SSIM, PSNR and LPIPS, we use the smallest rectangle to frame areas of Whole Tumor for every generated image and calculate SSIM and PSNR of the framed areas with their corresponding ground truth, which is denoted as local SSIM and local PSNR. These two metrics can measure the translation effect of the tumor areas. UTAD-Net performs better on the two metrics, which suggests that our method generates more realistic tumor areas.

C. Downstream segmentation results

Given an image from an arbitrary modality, we translate it to the other three modalities by our method and all the baselines respectively. Then we use nnU-Net to compare the segmentation effectiveness of the fully multi-modality images generated by each of the above methods. As shown in Table II, the teacher network achieves optimal results, which suggests that guiding with the tumor masks leads to more realistic tumor areas. Note that the SingleGT refers to randomly selecting one

modality of the ground truth for each sample in the training set for segmentation. Although the student network's performance is slightly lower than that of the teacher network, it still outperforms all baselines and singleGT on the corresponding downstream segmentation tasks, indicating that the images generated by our model have practical clinical significance.

D. Ablation study

In this section, we conduct two parts of the ablation study, one to validate the effectiveness of the local branch in the teacher network, and the other one to confirm the efficacy of the student network structure.

1) *Effectiveness of the local branch in the teacher network:* We conduct an ablation study to validate the effectiveness of the local branch which is guided by the mask. We replace the tumor mask in the teacher network with two settings: (a) a map with all zero values. (b) a random map that each value is either zero or one. As shown in Table I and Table II, the performance of UTAD-Net with a map of zeros (i.e., translation without the guidance of a tumor mask) is degenerated in terms of image quality and downstream segmentation. UTAD-Net with a random map performs even worse, because the noise images without semantic information have a negative effect on the model's performance.

2) *The structure of the student network:* The purpose of the student network is to perceive the tumor areas like the

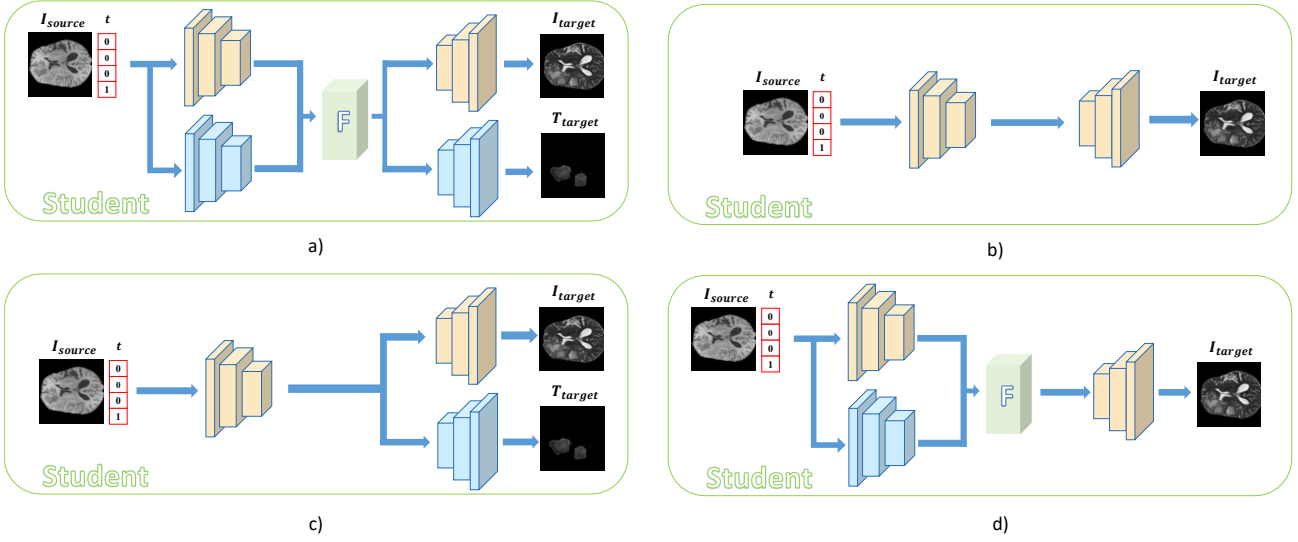


Fig. 4. Four architecture schemes of the student network. a) Complete global branch and local branch (UTAD-Net). b) Only global branch. c) Without local tumor encoder. d) Without local tumor decoder.

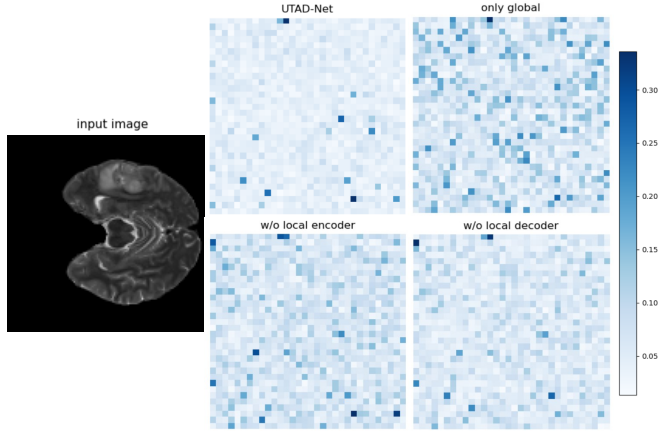


Fig. 5. The feature error map of the teacher network and the student network. Each pixel represents a channel.

teacher network without tumor masks. In order to learn the features extracted by the teacher network from the tumor areas better, we designed the following four different architecture schemes of the student network for comparison, as illustrated in Figure 4. We compute the error between the features obtained from the fusion block of the teacher network and the corresponding layer features of student network, averaging across different pixels to obtain the feature error map along the channel dimension. Note that for student networks in structures a) and d), we select the features obtained from the fusion block, whereas for structures b) and c), we choose the features from the last layer of the encoder. The error map is shown in Figure 5. The absence of either the local encoder or local decoder structure results in a reduction in the accuracy of the student network. The student network with only the global branch performs the worst, while maintaining

TABLE III
THE QUANTITATIVE EVALUATION OF THE FEATURE ERROR FOR THE TEACHER NETWORK AND THE FOUR DIFFERENT ARCHITECTURE SCHEMES OF THE STUDENT NETWORK.

	MAE	MSE
UTAD-Net	0.0500	0.0124
only global	0.2514	0.2229
w/o local encoder	0.1912	0.1588
w/o local decoder	0.1790	0.1423

the same structure as the teacher network yields the most similar features. We use Mean Absolute Error (MAE) and Mean Squared Error (MSE) to assess the similarity of features between the teacher network and the student network, where lower values indicate higher similarity. The quantified results are presented in Table III, demonstrating that the structure combining both global and local branches ultimately yields the best performance.

V. CONCLUSION

We propose UTAD-Net to translate brain tumor images, in order to capture tumor areas precisely and generate more realistic images between any two modalities. With the guidance of the tumor masks, the teacher network can focus on the brain tumor areas and alleviate the problem of deformation in the generated images. The student network gains the ability to perceive the tumor areas and generate target images without masks through distillation. Experiments demonstrate that our model achieves superior translation results and generates more realistic images with practical clinical significance, which results in improved downstream segmentation.

VI. ACKNOWLEDGMENTS

This work is supported in part by the Guangdong Provincial Natural Science Foundation (2023A1515011431),

the Guangzhou Science and Technology Planning Project (202201010092), the National Natural Science Foundation of China (72074105), NSF-1850492 and NSF-2045804.

REFERENCES

- [1] A. Işın, C. Direkoğlu, and M. Şah, "Review of mri-based brain tumor image segmentation using deep learning methods," *Procedia Computer Science*, vol. 102, pp. 317–324, 2016.
- [2] Y. Hou, T. Li, Q. Zhang, H. Yu, and H. Ge, "Brain tumor segmentation based on knowledge distillation and adversarial training," in *2021 international joint conference on neural networks (IJCNN)*. IEEE, 2021, pp. 1–7.
- [3] X. Liu, A. Yu, X. Wei, Z. Pan, and J. Tang, "Multimodal mr image synthesis using gradient prior and adversarial learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 6, pp. 1176–1188, 2020.
- [4] Z. Liu, J. Wei, R. Li, and J. Zhou, "Learning multi-modal brain tumor segmentation from privileged semi-paired mri images with curriculum disentanglement learning," *Computers in Biology and Medicine*, vol. 159, p. 106927, 2023.
- [5] B. Xin, Y. Hu, Y. Zheng, and H. Liao, "Multi-modality generative adversarial networks with tumor consistency loss for brain mr image synthesis," in *2020 IEEE 17th international symposium on biomedical imaging (ISBI)*. IEEE, 2020, pp. 1803–1807.
- [6] W. Yuan, J. Wei, J. Wang, Q. Ma, and T. Tasdizen, "Unified generative adversarial networks for multimodal segmentation from unpaired 3d medical images," *Medical Image Analysis*, vol. 64, p. 101731, 2020.
- [7] T. Zhou, S. Canu, P. Vera, and S. Ruan, "Brain tumor segmentation with missing modalities via latent multi-source correlation representation," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23*. Springer, 2020, pp. 533–541.
- [8] —, "Latent correlation representation learning for brain tumor segmentation with missing mri modalities," *IEEE Transactions on Image Processing*, vol. 30, pp. 4263–4274, 2021.
- [9] Z. Zhou, V. Sodha, J. Pang, M. B. Gotway, and J. Liang, "Models genesis," *Medical image analysis*, vol. 67, p. 101840, 2021.
- [10] Z. Shen, M. Huang, J. Shi, X. Xue, and T. S. Huang, "Towards instance-level image-to-image translation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3683–3692.
- [11] J. Chen, J. Wei, and R. Li, "Targan: Target-aware generative adversarial networks for multi-modality medical image translation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 24–33.
- [12] D. Bhattacharjee, S. Kim, G. Vizier, and M. Salzmann, "Dunit: Detection-based unsupervised image-to-image translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4787–4796.
- [13] Z. Zhang, L. Yang, and Y. Zheng, "Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9242–9251.
- [14] Z. Huang, L. Lin, P. Cheng, K. Pan, and X. Tang, "Ds 3-net: Difficulty-perceived common-to-t1ce semi-supervised multimodal mri synthesis network," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VI*. Springer, 2022, pp. 571–581.
- [15] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki *et al.*, "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*, 2018.
- [16] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.
- [17] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.
- [18] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.
- [19] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *International conference on machine learning*. PMLR, 2017, pp. 1857–1865.
- [20] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," *Advances in neural information processing systems*, vol. 30, 2017.
- [21] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 172–189.
- [22] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 35–51.
- [23] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8789–8797.
- [24] H.-Y. Lee, H.-Y. Tseng, Q. Mao, J.-B. Huang, Y.-D. Lu, M. Singh, and M.-H. Yang, "Drit++: Diverse image-to-image translation via disentangled representations," *International Journal of Computer Vision*, vol. 128, no. 10, pp. 2402–2417, 2020.
- [25] O. Dalmaz, M. Yurt, and T. Çukur, "Resvit: Residual vision transformers for multimodal medical image synthesis," *IEEE Transactions on Medical Imaging*, vol. 41, no. 10, pp. 2598–2614, 2022.
- [26] S. Mo, M. Cho, and J. Shin, "Instagan: Instance-aware image-to-image translation," *arXiv preprint arXiv:1812.10889*, 2018.
- [27] B. Yu, L. Zhou, L. Wang, Y. Shi, J. Fripp, and P. Bourgeat, "Ea-gans: edge-aware generative adversarial networks for cross-modality mr image synthesis," *IEEE transactions on medical imaging*, vol. 38, no. 7, pp. 1750–1762, 2019.
- [28] T. Zhou, H. Fu, G. Chen, J. Shen, and L. Shao, "Hi-net: hybrid-fusion network for multi-modal mr image synthesis," *IEEE transactions on medical imaging*, vol. 39, no. 9, pp. 2772–2781, 2020.
- [29] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," *Advances in neural information processing systems*, vol. 30, 2017.
- [30] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos, "Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features," *Scientific data*, vol. 4, no. 1, pp. 1–13, 2017.
- [31] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.
- [32] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.
- [33] L. Shen, W. Zhu, X. Wang, L. Xing, J. M. Pauly, B. Turkbey, S. A. Harmon, T. H. Sanford, S. Mehrabian, P. L. Choyke *et al.*, "Multi-domain image completion for random missing input data," *IEEE transactions on medical imaging*, vol. 40, no. 4, pp. 1113–1122, 2020.
- [34] A. Hore and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *2010 20th international conference on pattern recognition*. IEEE, 2010, pp. 2366–2369.
- [35] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.
- [36] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.
- [37] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.
- [38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.