SEMI-SUPERVISED SEGMENTATION-GUIDED TUMOR-AWARE GENERATIVE ADVERSARIAL NETWORK FOR MULTI-MODALITY BRAIN TUMOR TRANSLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-modality brain tumor images are widely used for clinical diagnosis since they can provide complementary information. Yet, due to considerations such as time, cost, and artifacts, it is difficult to get fully paired multi-modality images. Therefore, most of the brain tumor images are modality-missing in practice and only a few are labeled, due to a large amount of expert knowledge required. To tackle this problem, multi-modality brain tumor image translation has been extensively studied. However, existing works often lead to tumor deformation or distortion because they only focus on the whole image. In this paper, we propose a semi-supervised segmentation-guided tumor-aware generative adversarial network called S^3TAGAN, which utilizes unpaired brain tumor images with few paired and labeled ones to learn an end-to-end mapping from source modality to target modality. Specifically, we train a semi-supervised segmentation network to get pseudo labels, which aims to help the model focus on the local brain tumor areas. The model can synthesize more realistic images using pseudo tumor labels as additional information to help the global translation. Experiments show that our model achieves competitive results on both quantitative and qualitative evaluations. We also verify the effectiveness of the generated images via the downstream segmentation tasks.

1 INTRODUCTION

Multi-modality medical images are widely used in various tasks such as clinical detection. There are different kinds of imaging technologies in practice. For example, magnetic resonance imaging (MRI) is a common and noninvasive imaging technique. With the help of an additional magnetic field, MRI can determine the nucleus types of a certain part of the human body and then generate structural images with high resolution.

MRI is further divided into several modalities, such as T1-weighted (T1), T1-with-contrast-enhanced (T1ce), T2-weighted (T2), and T2-fluid-attenuated inversion recovery (Flair). Each modality of imaging can show complement lesion information from different angles. In Flair images, the cerebrospinal fluid shows hypointense signals while the lesions containing water appear as hyperintense signals. In T1 images, the cerebrospinal fluid is hypointense which tends to be black, while the gray matter is gray and the white matter is bright. Therefore, T1 images can present the anatomical structure, which is convenient for diagnoses. T1ce can show the structures and the edges of the tumors, which is convenient to observe the morphology of different types of tumors. T2 can better display the lesions because the brightness in the edema site is higher. Obviously, fully paired multi-modality images help doctors to make diagnoses more accurately.

The benefits of using multi-modality images to assist medical analysis have been widely recognized. However, due to the consideration of time, cost, artifacts, and other practical factors, physicians often get some of the modalities for examination in practice. In other words, most of the images are modality-missing, which has an adverse impact on the accuracy of physicians' diagnoses. If we can generate the corresponding missing modalities of given images by image translation, physicians can get more comprehensive information for diagnoses.

Many existing methods for multi-modality image translation based on deep learning have achieved good results in natural images. However, when applied to medical images, especially brain tumor images, the results are often unsatisfactory. Compared with two-dimensional natural images, three-dimensional medical images have more structural information. 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. In addition, the hierarchical structures of brain tumors are complex and irregular, which leads to blur or deformation in image translation. Therefore, the translation of brain tumor images has always been a challenge in the field of medical image translation.

To solve the problem of local distortion or blur in brain tumor image translation, we propose to use pseudo-labels generated by a segmentation network to guide the translation. The model contains a global branch and a local branch. For a given source image of arbitrary modality, we first put it into the segmentation network to get pseudo labels of three kinds of tumors, whole tumor, tumor core and enhancing tumor. Then the source image is inputted into the global branch and the dot product of it and the three pseudo labels are inputted into the local branch. In this way, the translation network can focus on the different parts of tumors. Since the training data are mostly unpaired and only few of them are labeled and paired, we train the segmentation network by the semi-supervised method proposed in CPS(Chen et al., 2021b). For paired images with Ground Truth, we use L1 loss for further constraint. The segmentation network and the translation network are trained at the same time to promote each other. Furthermore, in order to make our model applicable to images of arbitrary modality, similar to StarGAN(Choi et al., 2018), the discriminator tries to not only distinguish whether the images are real or fake, but also judge the modality which they belong to. In this way, we do not need to train a segmentation network for each modality, but only need to train a unified model to solve all cases.

In this way, we achieve an end-to-end translation, which means that given brain tumor images of arbitrary modality with both the source and target modality vectors, the model can directly output the final target images without any other manual intervention. We name our model as Semi-Supervised Segmentation-guided Tumor-Aware Generative Adversarial Network(S^3TAGAN).

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

- We propose a Semi-Supervised Segmentation-Guided Tumor-Aware Generative Adversarial Network, named S^3TAGAN, which is guided by different parts of tumors and improves the translation effectiveness using unpaired brain tumor images with few paired and labeled ones. We also propose a local consistency loss to preserve the anatomical structure of the tumors.
- We show qualitative and quantitative results in the multi-modality translation task on the BRATS 2020 dataset. Our model achieves better results compared with the state-of-the-art methods. We also verify the quality of the generated images through downstream segmentation tasks.

2 RELATED WORKS

Cross-modality image translation has been intensively studied in recent years. For instance, Pix2pix(Isola et al., 2017) provides a solution to generate images from the given source modality to the given target modality based on cGAN(Mirza & Osindero, 2014). However, it requires paired data for training, which is hard to realize. Therefore, how to achieve unsupervised image translation by utilizing unpaired data has attracted the interest of many researchers. CycleGAN(Zhu et al., 2017) and DiscoGAN(Kim et al., 2017) 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(Liu et al., 2017) believes that the essence of image translation tasks lies in calculating the joint distribution by utilizing the edge distribution of images in two known domains. Since there may be infinite joint distributions corresponding to two marginal distributions, some additional assumptions must be added. UNIT 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

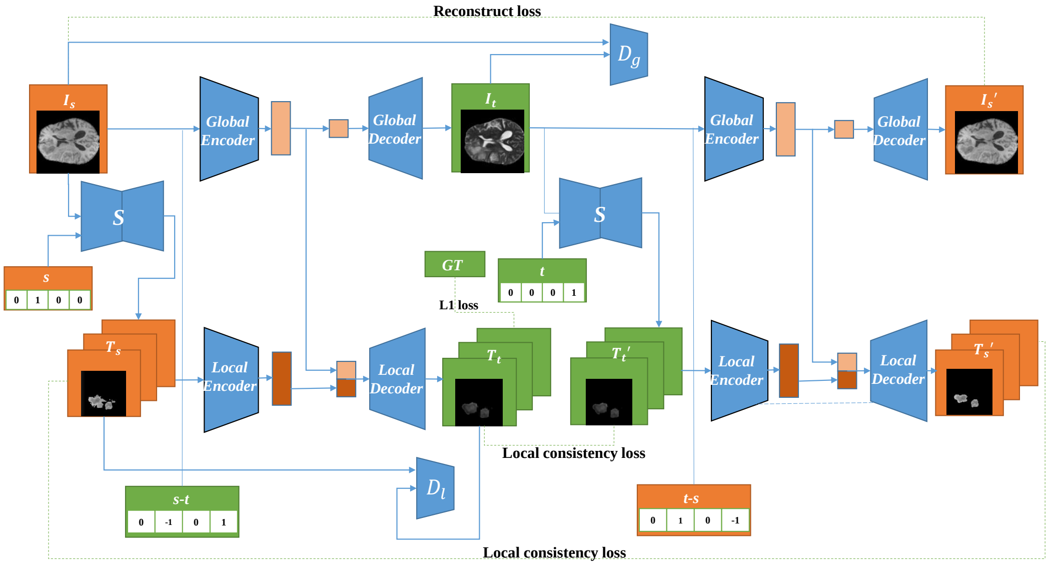

Figure 1: Overview of the proposed S^3TAGAN framework. Given the source image and its corresponding modality vector, we get three pseudo labels by the segmentation network and calculate the tumor images for the local branch. The generator contains two branches, one is the global branch which aims to generate the whole image from the source whole image and the given modality vector, while the other one is the local branch which aims to generate the local tumor images from the source tumor images and the given modality vector. Then the generator tries to reconstruct the whole image and tumor images. The global discriminator tries to determine whether the whole image is real or fake and classify its modality while the local discriminator responds to the local tumor images.

latent code back to the image domain. However, the images generated by the above models do not have style diversity. For a given image, the generated target modal image is unique. In order to solve this problem, MUNIT(Huang et al., 2018) and DRIT(Lee et al., 2018) 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. In this way, the style code can be obtained by style encoder or sampling, so that the image of the target modalities can be various.

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(Choi et al., 2018) 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. StarGAN v2(Choi et al., 2020) uses a variable style code to replace the mask vectors on this basis, and the generated target images of each modality have different styles. DRTI++(Lee et al., 2020) also adds domain code for translation so that any target modality images can be generated by a unified generator.

Although the above model can achieve multi-modality translation, it can not focus on local targets but only on the whole image. (Zhang et al., 2018b) propose that for unsupervised learning, cycle consistency loss will easily lead to local deformation of the image if there are no other constraints. InstaGAN(Mo et al., 2018) 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(Bhattacharjee et al., 2020) and INIT(Shen et al., 2019) respectively propose to use object detection and segmentation to assist translation. EaGANs(Yu et al., 2019) proposes to integrate edge maps that contain critical textural information to boost synthesis quality. TC-MGAN(Xin et al., 2020) 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(Chen et al., 2021a) can focus on the target area by using a segmentation network but it gets dissatisfactory results on brain tumor datasets. 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 S^3TAGAN 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 with few paired and labeled ones.

3 Method

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

3.1 Framework and pipeline

Given an image I_s from the source modality, we first put it into the segmentation network to get three pseudo labels, whole tumor, tumor core and enhancing tumor. Then we multiply the source image with three pseudo labels respectively to get the source tumor images T_s which only contains different tumor areas. Given the source modality vector s and an arbitrary target modality vector t, we aim to train a generator that can translate the source whole image I_s and the source tumor images T_s to the target whole image I_t and the target tumor images T_t. The mapping is denoted as: $(I_t, T_t) = G(I_s, T_s, s, t)$. Note that the segmentation network is only required during the training process, only I_s,s and t are used during the inference process, which is denoted as: $I_t = G(I_s, s, t)$. The framework of the model is shown in Figure 1.

Generator. The generator is comprised of two encoder-decoder pairs, one for the global branch and the other for the local branch. The global decoder receives the feature encoded by the global encoder and generates the target whole image I_t while the local decoder receives the features from both the global encoder and the local encoder to generate the target tumor images T_t. The generator translates the target whole image I_t and its corresponding tumor images T'_t to the reconstructed whole image I'_s and tumor images T'_s. In this way, a cycle training process is accomplished.

Discriminator. We use two discriminators to distinguish the reality of images and the modality they belong to. The discriminator D_g is responsible for the whole images in the global branch and the discriminator D_l is responsible for the tumor images in the local branch respectively.

Segmentation network. Given the image and its corresponding modality vector, the segmentation network generates three pseudo labels of the three kinds of tumors, which are binary masks that represent the foreground and background of the tumors. Then we calculate the tumor images by the dot product of the whole image and three pseudo labels. Taking source image I_s and its modality vector s for example, the mapping is denoted as: $T_s = I_s * S(I_s, s)$. On account of the poor proportion of labeled data, we train the segmentation network by a semi-supervised method proposed in CPS(Chen et al., 2021b). The generated target image I_t is also inputted into the segmentation network similar to I_s, which is denoted as $T'_t = I_t * S(I_t, t)$.

3.2 Training objective functions

Adversarial loss. Adversarial loss can make the images generated by the generator more realistic to confuse the discriminator. The traditional adversarial losses for the global branch and the local branch are defined as follows:

$$L_g^{adv} = E_{I_s}[logD_g^{src}(I_s)] + E_{I_t}[log(1 - D_g^{src}(I_t))], \tag{1}$$

$$L_l^{adv} = E_{T_s}[logD_l^{src}(T_s)] + E_{T_t}[log(1 - D_l^{src}(T_t))]. \tag{2}$$

Taking the global branch for example to explain, $D_g^{src}(I_s)$ represents the probability that the discriminator considers the images I_s as real, and $D_g^{src}(I_t)$ represents the probability that the discriminator considers the generated images as real. In order to correctly distinguish the reality of the images, the discriminator aims to minimize $D_g^{src}(I_t)$ and $D_l^{src}(T_t)$, while the generator, on the other hand, aims to maximize these terms to confuse the discriminator.

Since the traditional adversarial loss may lead to unstable adversarial learning, WGAN(Arjovsky et al., 2017) proposes a new adversarial loss, which solves the problem of instability of the training process in GAN, reduces the problem of mode collapse to a large extent, and ensures the diversity of generated samples. WGAN-GP(Gulrajani et al., 2017) proposes to use a gradient penalty strategy instead of the weight clipping strategy in WGAN, which makes the training process in GAN more stable and improves the quality of generated images. The final adversarial losses are shown as follows:

$$\begin{aligned} L_{D_g}^{adv} =& \lambda_{gp} E_{\alpha,s,I_s}[(\|\nabla D_g^{src}(\alpha I_s + (1-\alpha)I_t)\|_2 - 1)^2] \\ & - E_{I_s}[D_g^{src}(I_s)], \end{aligned} \tag{3}$$

$$\begin{aligned} L_{D_l}^{adv} =& \lambda_{gp} E_{\alpha,s,T_s}[(\|\nabla D_l^{src}(\alpha T_s + (1-\alpha)T_t)\|_2 - 1)^2] \\ & - E_{T_s}[D_l^{src}(T_s)], \end{aligned} \tag{4}$$

$$L_{G_g}^{adv} = E_{I_s,s}[D_g^{src}(I_t)], \tag{5}$$

$$L_{G_l}^{adv} = E_{T_s,s}[D_l^{src}(T_t)], \tag{6}$$

where λ_{gp} is set as 1, α is set as a random number whose range is between [0, 1] and subject to a uniform distribution in this paper.

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 discriminator aims to judge the modality they belong to. The difference is we add an extra discriminator for the local branch. For real images, we define the modality classification loss as follows:

$$L_{D_g}^{r\_cls} = E_{I_s}[-log D_g^{cls}(s|I_s)], \tag{7}$$

$$L_{D_l}^{r\_cls} = E_{T_s}[-log D_l^{cls}(s|T_s)], \tag{8}$$

where the term $D_g^{cls}(s|I_s)$ represents a probability distribution over the modality vector for the whole images and the term $D_l^{cls}(s|T_s)$ represents the one for the tumor images. Similarly, we define the modality classification loss for fake images as follows:

$$L_{D_g}^{f\_cls} = E_{I_s,s}[-log D_g^{cls}(t|I_t)], \tag{9}$$

$$L_{D_l}^{f\_cls} = E_{T_s,s}[-log D_l^{cls}(t|T_t)]. \tag{10}$$

Local consistency loss. T_t represents the generated tumor images and T_t' represents the tumor areas of the generated whole image I_t. Since we hope the segmentation network can better guide the translation of the global branch, we constrain the similarity of these two to alleviate the problem of distortion in brain tumor image translation. Similarly, the reconstructed tumor image T_s' and the source tumor images T_s are supposed to be similar. We propose a local consistency loss as an extra constraint to improve the translation effect, which is defined as follows:

$$L_{local} = E([\|T_t - T_t'\|_1]) + E([\|T_s - T_s'\|_1]). \tag{11}$$

Reconstruct loss. The model can translate the source image I_s to the image I_t of any modality. However, this does not guarantee that the generated image I_t just simply changes the image style information and still contains all the content information of the source image I_s. To solve this problem, we input I_t into the translation network for cycle translation to obtain the reconstructed image I_s'. If I_s' is consistent with I_s, the content information is not missed during translation. The reconstruct loss is defined as follows:

$$L_{rec} = E[\|I_s - I_s'\|_1]. \tag{12}$$

Identity mapping loss. Given an image of arbitrary modality, if the target modality happens to be its source modality, we denote the mapping as $I_{idt}, T_{idt} = G(I_s, T_s, s, s)$. We hope the generated images to be as consistent as possible with the source images. We use identity mapping loss to enforce the generated image not to lose origin information, which is defined as follows:

$$L_{idt} = E[\|I_s - I_{idt}\|_1 + E[\|T_s - T_{idt}\|_1]. \tag{13}$$

Semi-supervised loss. For images that are paired and labeled, we further use ground truth to constrain the generated tumor images to alleviate the problem of local image deformation. We defined the semi-supervised loss as follows:

$$L_{ss} = E[\|T_t - GT\|_1], \tag{14}$$

where GT represents the ground truth of the target tumor images.

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

$$\begin{aligned} L_D = &\lambda_{D_g}^{adv} L_{D_g}^{adv} + \lambda_{D_l}^{adv} L_{D_l}^{adv} + \lambda_{D_g}^{cls} L_{D_g}^{r\_cls} \\ &+ \lambda_{D_l}^{cls} L_{D_l}^{r\_cls}, \end{aligned} \tag{15}$$

$$\begin{aligned} L_G = &\lambda_{G_g}^{adv} L_{G_g}^{adv} + \lambda_{G_l}^{adv} L_{G_l}^{adv} + \lambda_{D_g}^{cls} L_{D_g}^{f_{cls}} \\ &+ \lambda_{D_l}^{cls} L_{D_l}^{f_{cls}} + \lambda_{rec} L_{rec} + \lambda_{local} L_{local} \\ &+ \lambda_{idt} L_{idt} + \lambda_{ss} L_{ss}, \end{aligned} \tag{16}$$

where $\lambda_{D_g}^{adv}$,$\lambda_{D_l}^{adv}$,$\lambda_{D_g}^{cls}$,$\lambda_{D_l}^{cls}$,λ_{rec}, λ_{local}, λ_{idt},λ_{ss} are hyper-parameters to balance to losses. We set $\lambda_{D_g}^{adv}$,$\lambda_{D_l}^{adv}$,$\lambda_{D_g}^{cls}$,$\lambda_{D_l}^{cls}$ to be 1.0 and λ_{rec}, λ_{local}, λ_{idt},λ_{ss} to be 10.0 in this paper.

4 Experiments

4.1 Settings

Datasets. We conduct all our experiments on BRATS2020(Menze et al., 2014)(Bakas et al., 2017)(Bakas et al., 2018) dataset. BRATS2020 provides brain tumor images of four modalities: T1-weighted (T1), T1-with-contrast-enhanced (T1ce), T2-weighted (T2) and T2-fluid-attenuated inversion recovery (Flair). Three kinds of tumors which are named Whole Tumor(WT), Tumor Core(TC) and Enhancing Tumor(ET), are labeled for segmentation. We use 150 patients' images as the training samples and 20 percent of them are treated as labeled and paired ones. We resize the images to 128*128. Details are shown in the supplementary material.

Evaluation metrics. For the translation task, we use structural similarity index measure(SSIM), peak-signal-noise ratio(PSNR) and learned perceptual image patch similarity(LPIPS)(Zhang et al., 2018a) to measure the similarity between the generated images and ground truth. For the downstream segmentation task, we use DICE to measure the integrity of the predicted pseudo labels which are generated by nnU-Net(Isensee et al., 2021). That's because nnU-Net is an acknowledged method that achieves state-of-the-art performances for medical image segmentation.

Baselines. We compare our translation results with StarGAN(Choi et al., 2018), DRIT++(Lee et al., 2018), Targan(Chen et al., 2021a) and ReMIC(Shen et al., 2020). StarGAN proposes to use a unified model to translate images to arbitrary modalities. DRIT++ disentangles an image to the content code and the attribute code during the training process and generates images by using the content code extracted from the input images and the attribute code sampled from the standard normal distribution. TarGAN alleviates the problem of image deformation on the target area by utilizing an extra shape controller. ReMIC generates images by using multi-modality paired images. Note that we implement a semi-supervised ReMIC for comparison.

Implementation details. We implement PatchGAN(Isola et al., 2017) as the backbone for both the global discriminator and local discriminator. And we use U-net as the backbone for the generator and the segmentation network. We train our model for 100 epochs with a learning rate of 10^{-4} for the generator and both the discriminators for the first 50 epochs and linearly decay the learning rate to 10^{-6} at the final epoch. Adam(Kingma & Ba, 2014) 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. Details are shown in the supplementary material. All the experiments are conducted on PyTorch with NVIDIA RTX 3090.

4.2 Results

In this section, we demonstrate our translation results compared with other baseline methods. Then we verify the effectiveness of the generated images via the downstream segmentation tasks.

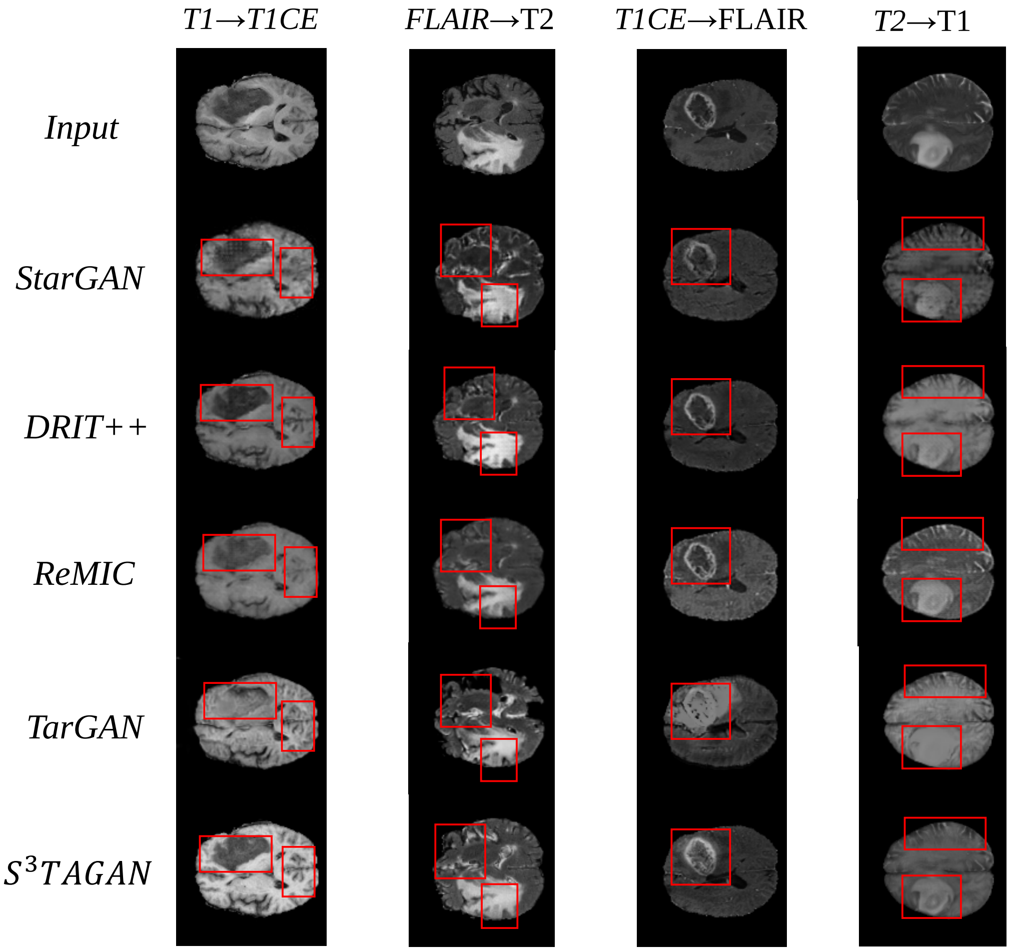

Figure 2: qualitative results of our model and the other baselines on the BRATS2020 dataset. The source image is denoted as Input. The four columns show the images of various mappings and the rows represent different methods. Red boxes show that our method generates images with clearer textures and more structural information. The results demonstrate that our method achieves better translation effectiveness.

4.2.1 Translation results

Qualitative evaluation. Figure 2 shows the qualitative results of our model and the other baselines. StarGAN and DRIT++ generate images with checkerboard artifacts in some cases, while ReMIC and TarGAN may lead to blur or deformation in the tumor areas. Our method generates images with clearer textures and more structural information.

Quantitative evaluation. We use SSIM, PSNR and LPIPS to measure the similarity between the generated images and ground truth.

As shown in Table 1, our method gets higher SSIM and PSNR than the other baselines. The value is the average of all the cases for the mapping of any source modality to an arbitrary target modality. We also use the smallest rectangle to frame the tumor areas for every generated image and calculate the SSIM and PSNR of the framed images with their corresponding ground truth, which is denoted as local SSIM and local PSNR. These two metrics represent the translation effectiveness in the tumor areas. S^3TAGAN get higher scores of them, which means that our method preserves more information in the tumor areas. LPIPS is also a metric to measure perceptual similarity. The

Table 1: Quantitative evaluation of the images generated by our method and the baselines. We report the mean value for translation between any two modalities here. The symbol ↑ denotes higher is better while the symbol ↓ denotes lower is better.

Method	SSIM↑	PSNR↑	local SSIM↑	local PSNR↑	LPIPS↓
StarGAN	83.35	25.99	56.26	18.55	9.79
DRIT++	86.57	30.86	60.94	19.76	8.66
TarGAN	84.66	28.74	58.79	18.36	9.76
ReMIC	84.59	28.70	57.68	18.28	9.12
S^{3}TAGAN	**88.37**	**34.36**	**61.51**	**22.51**	**5.96**

Table 2: The quantitative evaluation for downstream segmentation task conducted by nnU-Net. We use DICE as the metric. Our method achieves better results than the baselines on all three kinds of tumors.

Method	DICE (%)		
	WT	TC	ETC
StarGAN	73.56	59.76	35.65
DRIT++	77.96	65.81	42.36
Targan	77.37	64.41	42.84
Remic	77.10	63.56	42.12
S^{3}TAGAN	**79.07**	**66.69**	**43.37**

lower value of this metric means higher similarity which represents that our model achieves better translation effectiveness.

4.2.2 Downstream segmentation results

Given an image from an arbitrary modality, we translate it to the other three modalities by our model and all the baselines respectively. Then we put the fully multi-modality images generated by the above methods into nnU-Net to compare their segmentation effectiveness. We use DICE of whole tumor(WT), tumor core(TC) and enhancing tumor(ET) to measure the results. As shown in Table 2, our method achieves better performance than all the baselines, which also represents that we generate more accurate information on the tumor areas.

4.3 Effectiveness of local branch

In this section, we conduct an ablation study to validate the effectiveness of the local branch which is guided by the segmentation network in our proposed S^3TAGAN. We replace the predicted pseudo labels with the following three situations: (a)ground truth labels. (b) full-zeros maps. (c) random maps that each value is either zero or one. As shown in Table 3, the performance of S^3TAGAN **with labels** is the upper bound in theory, which represents the best guidance for local tumor translation. Note that the performance of segmentation guidance is close to this situation, which demonstrates the effectiveness of our model. While S^3TAGAN with zeros maps is the lower bound in theory, which represents no segmentation-guided learning. S^3TAGAN with random maps represents learning with guidance for random areas but not the tumor areas. Translation effectiveness is improved slightly in this situation. The results show the robustness of our model.

4.4 Ratio of paired and labeled data

In order to test the effect of the ratio of paired and labeled data for semi-supervised learning on our model, we adjust the ratio to 10 percent and 100 percent. As shown in Table 4, more paired and labeled data for supervision can improve translation effectiveness. Note that the effectiveness of 20 percent supervision which is our default setting is close to 100 percent supervision.

Table 3: Ablation study for the effectiveness of local branch. The results demonstrate the effectiveness and robustness of our model.

	SSIM	PSNR	local SSIM	local PSNR	LPIPS
S^{3}TAGAN	88.37	34.36	61.51	22.51	5.96
S^{3}TAGAN w ground truth labels	88.42	34.56	61.71	22.58	5.91
S^{3}TAGAN w zeros maps	88.03	33.10	61.26	22.19	6.02
S^{3}TAGAN w random maps	88.15	33.85	61.38	22.40	6.01

Table 4: Sensitive analysis for the ratio of paired and labeled data for training. The results show that few paired multi-modality data can also benefit the translation.

	SSIM	PSNR	local SSIM	local PSNR	LPIPS
S^{3}TAGAN(10%)	88.25	32.96	61.73	22.28	6.01
S^{3}TAGAN(20% by default)	88.37	34.36	61.51	22.51	5.96
S^{3}TAGAN(100%)	88.46	34.39	61.60	22.71	5.94

5 Conclusion

In this paper, we propose a semi-supervised segmentation-guided method called S^3TAGAN to translate brain tumor images, which learns a mapping between any two modalities. We use unpaired images for training with only few paired and labeled ones, which is in agreement with the practical situation. With the guidance of the segmentation network, the local branch focuses on the brain tumor areas and alleviates the problem of deformation in the tumor areas, which benefits the quality of both the generated whole images and tumor images. Experiments demonstrate that our model achieves better translation effectiveness with strong robustness. The results of the downstream segmentation task also verify the effectiveness of our model.

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4 (1):1–13, 2017.

Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin 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.

Deblina Bhattacharjee, Seungryong Kim, Guillaume Vizier, and Mathieu Salzmann. Dunit: Detection-based unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4787–4796, 2020.

Junxiao Chen, Jia Wei, and Rui Li. Targan: Target-aware generative adversarial networks for multi-modality medical image translation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 24–33. Springer, 2021a.

Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2613–2622, 2021b.

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul 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*, pp. 8789–8797, 2018.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8188–8197, 2020.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 172–189, 2018.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International conference on machine learning*, pp. 1857–1865. PMLR, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 35–51, 2018.

Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. Drit++: Diverse image-to-image translation via disentangled representations. *International Journal of Computer Vision*, 128(10):2402–2417, 2020.

Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30, 2017.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34 (10):1993–2024, 2014.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Instagan: Instance-aware image-to-image translation. *arXiv preprint arXiv:1812.10889*, 2018.

Liyue Shen, Wentao Zhu, Xiaosong Wang, Lei Xing, John M Pauly, Baris Turkbey, Stephanie Anne Harmon, Thomas Hogue Sanford, Sherif Mehralivand, Peter L Choyke, et al. Multi-domain image completion for random missing input data. *IEEE transactions on medical imaging*, 40(4): 1113–1122, 2020.

Zhiqiang Shen, Mingyang Huang, Jianping Shi, Xiangyang Xue, and Thomas S Huang. Towards instance-level image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3683–3692, 2019.

Bingyu Xin, Yifan Hu, Yefeng Zheng, and Hongen 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)*, pp. 1803–1807. IEEE, 2020.

Biting Yu, Luping Zhou, Lei Wang, Yinghuan Shi, Jurgen Fripp, and Pierrick Bourgeat. Ea-gans: edge-aware generative adversarial networks for cross-modality mr image synthesis. *IEEE transactions on medical imaging*, 38(7):1750–1762, 2019.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018a.

Zizhao Zhang, Lin Yang, and Yefeng 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*, pp. 9242–9251, 2018b.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.