Fine-grained Medical Image Synthesis with Dual-Attention Adversarial Learning

Qiuyu Xiao[1], Dong Nie[1,]

[1]Department of Computer Science, University of North Carolina at Chapel Hill, USA

Abstract. Medical imaging plays a critical role in various clinical applications. However, due to considerations such as cost and risk, the acquisition of certain image modalities can be limited. To address this issue, many cross-modality medical image synthesis methods have been proposed. Nevertheless, current methods struggle to accurately model hard-to-synthesize regions (e.g., tumor or lesion regions). To overcome this challenge, we propose a simple yet effective strategy: a dual-discriminator (dual-D) adversarial learning system. In this system, 1) a global discriminator (global-D) provides an overall evaluation of the synthetic image, and 2) a local discriminator (local-D) performs a dense evaluation of the synthetic image's local regions. Additionally, we introduce a difficult-region-aware attention mechanism that enhances the modeling of hard-to-synthesize regions (e.g., tumor or lesion regions) based on the local-D. Experimental results demonstrate the robustness and accuracy of our proposed method in synthesizing target images from corresponding source images. Specifically, we evaluated our method on two datasets: i.e., 1) generating T2 MRI from T1 MRI for brain tumor images, and 2) generating CT from MRI. Our proposed method outperforms state-of-the-art techniques in both datasets and tasks. Furthermore, our proposed difficult-region-aware attention mechanism proves effective in generating more realistic images, particularly in the hard-to-synthesize regions.

1 Introduction

The importance of medical imaging for clinical diagnosis, disease treatment, and medical research has steadily increased over the last few decades. Multiple imaging modalities, such as magnetic resonance imaging (MRI) and computed tomography (CT), provide complementary information, which is essential for the comprehensive assessment of complex diseases, whether in diagnostic examinations or as part of medical research trials. Different imaging modalities are often required at various stages of disease diagnosis and treatment. However, in practice, it is not always feasible to obtain all necessary modalities. Therefore, it is highly beneficial to explore solutions for synthesizing the modality of interest (or target) from the available source modalities.

In the past, many researchers have attempted to directly synthesize high-quality medical modality images. However, this task is often challenging due to several issues. First, the mapping from the source modality to the target modality

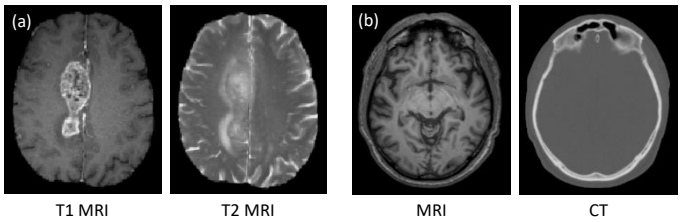

Fig. 1. Two pairs of corresponding source and target images from the same subjects. (a) shows T1 MRI/T2 MRI brain tumor images; (b) shows MRI/CT brain images.

(or its inverse) is typically complex and ill-posed. Second, different modalities can exhibit significantly different image appearances, e.g., MRI and CT as shown in Fig. 1(b). Third, certain regions in the image (such as tumor in Fig. 1(a))may have completely different image contrasts and appearances. Nonetheless, there often exist potential connections, or even highly nonlinear relationships, between different modalities that can be leveraged for synthesizing one modality from another.

Convolutional neural network (CNN) offers a new approach for learning highly nonlinear relationships by employing multiple-layer mapping [1, 3–5, 7, 8, 10–15]. For example, Huang et al. [4] proposed a method to simultaneously perform super-resolution and cross-modality medical image synthesis using weakly-supervised joint convolutional sparse coding. Similarly, Nie et al. [7] introduced a supervised adversarial learning framework with gradient difference loss to synthesize CT images from MRI scans. Although the training of the aforementioned image synthesis methods can achieve good performance in most cases, they often fail to produce reasonable results in certain situations, such as in tumor regions (as shown in Fig. 1(a)). This is because the training process tends to be dominated by the majority of samples or regions that are easier to synthesize, such as normal tissue regions, while neglecting the minority of tumor or lesion regions, which are crucial biomarkers in clinical diagnosis. Therefore, it is essential to develop a method that can better model tumor and lesion regions in medical image synthesis.

In this work, we propose a dual-discriminator adversarial learning framework with a difficult-region-aware attention mechanism to address the aforementioned issues. Specifically, in addition to the regular CNN-based discriminator, we introduce a dense fully convolutional network (FCN) as the local discriminator to assess the difficulty level of each local region in image synthesis. More importantly, we further propose a difficult-region-aware attention mechanism to better model the hard-to-synthesize regions (i.e., tumor regions). Experimental results demonstrate that the proposed method can effectively synthesize target images with significantly improved modeling capacity for the hard-to-synthesize regions. To the best of our knowledge, this is the first work to address the challenge of hard-to-synthesize regions in cross-modality image synthesis tasks.

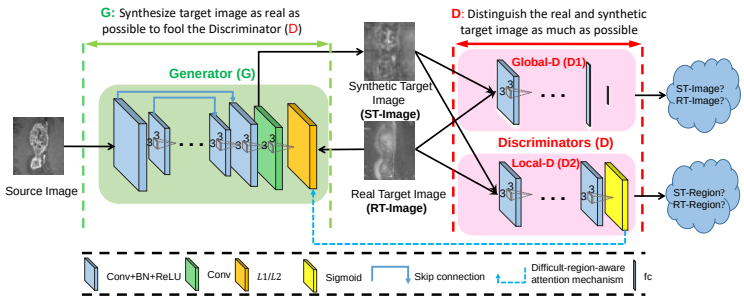

Fig. 2. Architecture used in the deep supervised generative adversarial setting to synthesize target image. This framework contains one generator and two discriminators. A difficult-region-aware attention mechanism is also included in the framework.

2 Methods

To address the aforementioned issues and challenges, we propose a deep convolutional adversarial network framework that involves adversarial training of the generator (UNet) using dual discriminators, namely, a CNN as the global discriminator and an FCN as the local discriminator. Fig. 2 illustrates the entire framework.

2.1 UNet for Medical Image Synthesis

UNet [9], an evolutionary version of FCN, incorporates both high-resolution and rich-semantic feature maps to enhance localization accuracy, making it widely used for segmentation and reconstruction in medical image analysis. In this paper, we adopt UNet as the image generator for medical image synthesis because it can alleviate the loss of spatial details and recover fine-grained details in dense predictions, compared to FCN.

As mentioned in the Introduction section, typically an L_1/L_2 loss is used to train the network as below,

$$L_G(X, Y) = \|Y - G(X)\|^p \tag{1}$$

where Y is the ground-truth target image, and $G(X)$ is the generated target image from the source image X by the Generator network G and p is 1 or 2.

2.2 Adversarial Learning for Medical Image Synthesis

To enhance the perceptual quality of the generated target images, we propose using adversarial learning to improve the performance of UNet within local regions of synthetic images. Our adversarial network comprises a global discriminator (denoted as $D1$ in Fig. 2), typically implemented with a CNN, which distinguishes between the real target image and the generated one as a whole, and also a local discriminator (denoted as D2 in Fig. 2), typically implemented with an FCN, which discerns between synthetic and real images at the voxel level.

It's worth noting that in D2 each element of the dense outputs corresponds to the quality of the synthesized local region around the respective voxel.

Global Adversarial Learning: In our study, we follow the WGAN-GP [2] to form the global discriminator ($D1$) due to its powerful adversarial learning capacity. Concretely, the loss function for $D1$ can be defined as:

$$L_{D1}(X,Y) = E_x[D1(G(X))] - E_Y[D1(Y)] + \lambda E_{\hat{X}}[(\left\| \bigtriangledown_{\hat{X}} D1(\hat{X}) \right\|_2 - 1)^2] \quad (2)$$

where X is the source input image, Y is the corresponding target image, $G(X)$ is the estimated image by the generator, and $\hat{X}$ is uniformly sampled along straight lines among synthetic and real samples. And λ is a constant weighting hyper-parameter.

The global adversarial loss term used to train G is defined as below.

$$L_{G\_ADV1}(X,Y) = -E_x[D1(G(X))] \quad (3)$$

With the above definitions, $D1$ can globally distinguish the real target image from the synthetic target data generated by G. At the same time, G aims to produce more realistic target images for confusing $D1$. The details of $D1$ follows the suggestions in [2].

Local Adversarial Learning: To obtain the local confidence information of how well each local region is synthesized, we formulate the training objective of the local discriminator as the summation of binary cross-entropy loss over the image domain, as given in Eq. 4. Note that the reason we choose to cross entropy loss instead of W-distance, is to apply sigmoid function to the dense outputs for obtaining the probability that can be used as confidence value. Here, we use G and $D2$ to denote the generator and local-D networks, respectively.

$$L_{D2}(\mathbf{X},\mathbf{Y};\theta_{\mathbf{D2}}) = L_{BCE}(D2(\mathbf{Y},\theta_{\mathbf{D2}}),\mathbf{1}) + L_{BCE}(D2(G(\mathbf{X}),\theta_{\mathbf{D2}}),\mathbf{0}), \quad (4)$$

where L_{BCE} is binary cross-entropy loss. $\mathbf{X}$ and $\mathbf{Y}$ represent the input image and its corresponding real target image, respectively. $\theta_{\mathbf{D2}}$ is network parameters for the local-D network.

For training the generator network, besides the L_1/L_2 loss defined in Eq. 1 and the global adversarial learning loss defined in Eq. 5, the local adversarial loss ("ADV") to improve G and fool $D2$ can be defined below:

$$L_{ADV2}(\mathbf{X},\theta_{\mathbf{G}}) = L_{BCE}(D2(G(\mathbf{X};\theta_{\mathbf{G}})),\mathbf{1}) \quad (5)$$

The training of the two networks is performed in an alternating fashion. First, D is updated by taking a mini-batch of real target image and a mini-batch of generated target image (corresponding to the output of G). Then, G is updated by using another mini-batch of samples including sources and their corresponding ground-truth target images.

2.3 Region-attention based Adversarial Difficulty Learning

Due to the inhomogeneous characteristics and irregular distribution of medical images, certain regions within the images are typically more challenging to synthesize effectively. Consequently, there is a strong need to develop a model that

can accurately depict these hard-to-synthesize regions. As the local discriminator can provide confidence information regarding the synthesis quality of each local region, we can leverage this information to prioritize attention towards the hard-to-synthesize regions , such as tumors and lesions, ensuring they are better modeled. To achieve this, we propose a difficulty-region-aware attention mechanism designed to accurately measure the difficulty level of synthesizing different regions. Specifically, we design a difficulty-region-aware L_1/L_2 loss by using region-level attentions from the adversarial local confidence map.

The voxel-level difficulty-region-aware attention from the confidence map (M) is formulated (based on Eq. 1) as below:

$$L_{AttG}(X,Y) = F \odot \|Y - G(X)\|^p \tag{6}$$

where $\odot$ is the element-wise multiplication and $F = (1-M)^\beta$, and β is the voxel-level attention parameter. Note, F here works as a scaling factor, which is used to largely suppress the contribution of easy-to-synthesize regions to the training loss and emphasize the hard-to-synthesize regions.

With the difficult-region-aware L_1/L_2 loss in Eq. 6, we can pay more attention in the less confident (i.e., hard-to-synthesize) regions and thus better model them (e.g., tumor and lesion). Although our proposed idea is simple, the adversarial difficulty-region-aware attention mechanism presents an opportunity to effectively utilize voxel-wise focal loss in a regression context. It's important to note that, thus far, there have been few effective works proposed for attention mechanisms in regression networks.

Total Loss for Training Generator: To this end, the total loss for training generator includes the attention based L_1/L_2 loss, the global adversarial loss, and the local adversarial loss, which can be summarized below Eq. 7.

$$L_G = L_{AttG} + \lambda_1 L_{G\_ADV1} + \lambda_2 L_{G\_ADV2} \tag{7}$$

The above training loss could encourage G to generate target images with voxel-wise correspondence to real target image, and also best fooling the discriminators both globally and locally.

2.4 Training Details

The discriminator $D1$ is a typical CNN including three stages of convolution, BN, ReLU and max pooling, followed by one convolutional layer and three fully connected layers (where the first two use ReLU as activation functions). The filter size is 3×3, the numbers of the filters are 32, 64, 128, respecitvely, and 256 for the convolutional layers, and the numbers of the output nodes in the fully connected layers are 512, 128 and 1, respectively. The dense discriminator $D2$ is a typical FCN with three down-sampling. All networks were trained using the Adam optimizer. The code is implemented using the pytorch library[1], and it will be publicly released upon acceptance of this paper.

[1] https://github.com/pytorch/pytorch

3 Experiments and Results

We selected the BRATS dataset for evaluating our proposed method, which is a publicly available dataset comprising MRI scans from brain tumor patients [6]. The dataset consists of a total of 354 pairs of T1 MRI and T2 MRI scans. Among these, 200 subjects were allocated for training, 60 for validation, and the remaining 94 for testing purposes.

To demonstrate the advantage of our proposed method in terms of synthesis accuracy, we compare it with three widely-used approaches: atlas-based, FCN, UNet, and sGAN.

3.1 Impact of Proposed Dual-Discriminator Strategy

To demonstrate the effectiveness of our proposed dual-discriminator strategy, we conducted experiments comparing three methods on the BRATS dataset: sGAN with a global discriminator (sGAN-1), sGAN with a local discriminator (sGAN-2), and our proposed dual-discriminator strategy (sGAN-dual). The PSNR values obtained were 27.3 dB, 27.6 dB, and 28.3 dB for sGAN-1, sGAN-2, and sGAN-dual, respectively. It's important to note that these results were achieved using the ordinary L_1 loss for the generator. Furthermore, besides interpreting our proposed dual-discriminator strategy as local and global adversarial constraints, we can view it from another perspective. Specifically, the use of our dual-discriminator strategy can effectively mitigate the issue of adversarial gradient vanishing, thereby enhancing the stability and efficacy of adversarial learning.

3.2 Impact of Difficult-Region-Aware Attention Mechanism

To demonstrate the impact of our proposed difficult-region-aware attention mechanism, we conduct experiments to compare the performance with and without this mechanism on the BRATS dataset. The experimental results indicate an improvement of 0.2 dB in terms of PSNR when employing our proposed attention mechanism. Furthermore, to delve deeper into the effect of our proposed mechanism, we focus on evaluating the synthesis performance solely on tumor regions. By utilizing manually segmented tumor regions from this dataset, we compute the PSNR specifically on tumor regions within the testing set, achieving an average improvement of 0.6 dB.

We also visualize results in Fig. 3. We can clearly see that the generated image by using our proposed difficult-region-aware attention mechanism (i.e., 'dual-D+attention') could recover much more details, compared to the results without using our proposed mechanism (i.e., 'dual-D'), especially for the tumor regions.

In order to gain a deeper understanding of why the difficult-region-aware mechanism is effective, we analyze the confidence map generated by the local discriminator (i.e. $D2$). Our analysis shows that initially, tumor regions are evaluated as poorly synthesized, as indicated by the low local confidence scores. Consequently, more attention is directed towards these tumor regions during

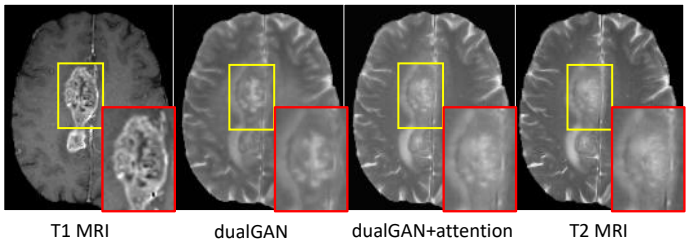

Fig. 3. Visual evaluation of our proposed difficult-region-aware attention mechanism. Using our proposed mechanism, the respective results (third column) is more similar to the real target T2 MRI (fourth column), compared to the case without using our proposed mechanism (second column).

the subsequent training stages of the generator network. As a result, by the end of the training process, the generator network learns to better synthesize tumor regions as well.

3.3 Comparison with Other Methods

For a qualitative comparison of the image synthesis results obtained by different methods, we present synthetic target images alongside their corresponding real target images in Fig. 4. It is evident that the proposed algorithm excels in preserving continuity, coherence, and smoothness in the synthetic results, owing to the utilization of both global and local adversarial learning constraints across the image patches. Notably, the tumor region of generated T1 MRI can recover much more details than other methods, closely resembling the real T2 MRI. We attribute this improvement to the difficult-region-aware attention mechanism, which prioritizes regions recognized as challenging to synthesize, such as tumor regions.

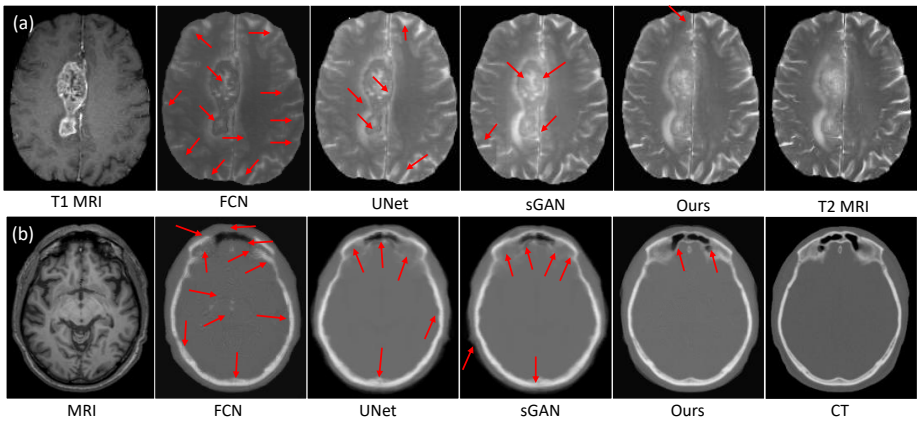

Fig. 4. Visual comparison of results by different methods for two cases of application: (a) T1 MRI to T2 MRI synthesis and (b) MRI to CT synthesis. Red arrows indicate poorly-synthesized regions.

We also quantitatively compare the predicted results in Table 1, in terms of both PSNR and MAE. Our proposed method outperforms all other competing methods in both metrics. Fig. 4(a) shows synthesis results on BRATS dataset (with brain tumors) by different methods. It can be seen that our result is more consistent with the real T2 MRI (right).

Table 1. Average MAE and PSNR on 94 testing subjects from the BRATS dataset.

Method	MAE	PSNR
FCN	34.5(8.6)	25.0(2.3)
UNet	28.8(**6.9**)	26.2(**1.8**)
sGAN	27.0(5.7)	26.0(1.5)
Ours	**25.8**(5.2)	**27.5(1.4)**

Table 2. Average MAE and PSNR on 16 subjects from the brain dataset.

Method	MAE	PSNR
FCN	24.4(15.1)	22.7(3.2)
UNet	21.8(12.8)	26.7(2.1)
sGAN	20.4(11.2)	27.3(**1.7**)
Ours	**18.4(10.3)**	**28.6**(1.8)

To show the generalization ability of our proposed method, we also evaluate it on another brain dataset for synthesizing CT from MRI. Fig. 4(b) shows CT synthesis results by different methods, and Table 2 gives quantitative comparison results. It is clear that our proposed method can work better than the state-of-the-art methods, demonstrating the good generalization of our proposed method to other datasets for other image synthesis tasks.

4 Conclusions

We have introduced dual discriminators within an adversarial learning framework, comprising a global discriminator for overall evaluation and a local discriminator for region-wise evaluation, to address critical challenges in medical image synthesis. Additionally, we have proposed a difficult-region-aware attention mechanism to effectively handle hard-to-synthesize regions (i.e., tumor and lesion). Our proposed model has been applied to two tasks: 1) synthesizing T2 MRI from corresponding T1 MRI and 2) synthesizing brain CT images from their corresponding MR images. Experimental results demonstrate that our method outperforms three state-of-the-art methods. Furthermore, our proposed difficult-region-aware attention mechanism enhances the synthesis of hard-to-synthesize regions. Lastly, we tested the generalization of our proposed method across different image synthesis tasks using the same framework, achieving good performance consistently.

Bibliography

[1] Sanuwani Dayarathna et al. Deep learning based synthesis of mri, ct and pet: Review and analysis. *Medical Image Analysis*, page 103046, 2023.

[2] Ishaan Gulrajani et al. Improved training of wasserstein gans. In *NIPS*, pages 5767–5777, 2017.

[3] Xiao Han. Mr-based synthetic ct generation using a deep convolutional neural network method. *Medical Physics*, 44(4):1408–1419, 2017.

[4] Yawen Huang et al. Simultaneous super-resolution and cross-modality synthesis of 3d medical images using weakly-supervised joint convolutional sparse coding. *arXiv preprint arXiv:1705.02596*, 2017.

[5] Huabing Liu et al. Multimodal brain tumor segmentation boosted by monomodal normal brain images. *IEEE Transactions on Image Processing*, 2024.

[6] Bjoern H Menze et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE TMI*, 34(10):1993, 2015.

[7] Dong Nie et al. Medical image synthesis with context-aware generative adversarial networks. In *MICCAI*, 2017.

[8] Shaoyan Pan et al. Synthetic ct generation from mri using 3d transformer-based denoising diffusion model. *Medical Physics*, 51(4):2538–2548, 2024.

[9] Olaf Ronneberger et al. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015.

[10] G Szalkowski et al. Image synthesis for planning and target tracking of mr-based stereotactic radiation therapy. In *MEDICAL PHYSICS*, volume 48. WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA, 2021.

[11] Gregory Szalkowski et al. Synthetic digital reconstructed radiographs for mr-only robotic stereotactic radiation therapy: A proof of concept. *Computers in biology and medicine*, 138:104917, 2021.

[12] Jelmer M Wolterink et al. Generative adversarial networks for noise reduction in low-dose ct. *TMI*, 36(12):2536–2545, 2017.

[13] Can Zhao et al. A deep learning based anti-aliasing self super-resolution algorithm for mri. In *MICCAI*, pages 100–108. Springer, 2018.

[14] Xuanru Zhou et al. Multimodality mri synchronous construction based deep learning framework for mri-guided radiotherapy synthetic ct generation. *Computers in Biology and Medicine*, 162:107054, 2023.

[15] Jun-Yan Zhu et al. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *CVPR*, pages 2223–2232, 2017.