Target-Guided Diffusion Models for Unpaired Cross-modality Medical Image Translation

Yimin Luo, Qinyu Yang, Ziyi Liu, Zenglin Shi, Weimin Huang, Guoyan Zheng, and Jun Cheng

***Abstract*— In a clinical setting, the acquisition of certain medical image modality is often unavailable due to various considerations such as cost, radiation, etc. Therefore, unpaired cross-modality translation techniques, which involve training on the unpaired data and synthesizing the target modality with the guidance of the acquired source modality, are of great interest. Previous methods for synthesizing target medical images are to establish one-shot mapping through generative adversarial networks (GANs). As promising alternatives to GANs, diffusion models have recently received wide interests in generative tasks. In this paper, we propose a target-guided diffusion model (TGDM) for unpaired cross-modality medical image translation. For training, to encourage our diffusion model to learn more visual concepts, we adopted a perception prioritized weight scheme (P2W) to the training objectives. For sampling, a pre-trained classifier is adopted in the reverse process to relieve modality-specific remnants from source data. Experiments on both brain MRI-CT and prostate MRI-US datasets demonstrate that the proposed method achieves a visually realistic result that mimics a vivid anatomical section of the target organ. In addition, we have also conducted a subjective assessment based on the synthesized samples to further validate the clinical value of TGDM.**

***Index Terms*— Unpaired Learning, Cross-Modality Translation, Medical Image Synthesis, Diffusion Model, Guidance Sampling.**

I. Introduction

Medical imaging has become an increasingly essential tool in modern healthcare, such as positron emission tomography (PET), magnetic resonance imaging (MRI), X-rays, computed tomography (CT), and ultrasound (US). Each modality uniquely quantifies distinct characteristics of the underlying anatomy [1]. Different imaging techniques have their own advantages and disadvantages, which often complement each other to provide clinicians with a more comprehensive understanding of patient conditions. However, the acquisition of certain modality is often restricted due to multiple considerations, such as cost and radiation [2]. Accordingly, there is a growing interest in cross-modality synthesis to obtain a missing modality from a given source modality [3], [4]. Thanks to the recent advancements of deep learning methods, we are getting closer to a promising path toward cross-modality medical image translation [5], [6].

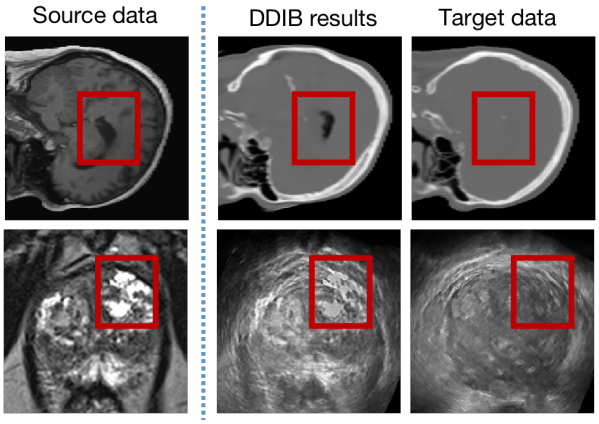

Fig. 1. Two representative examples for cross-modality medical image translation via DDIB. The patterns within the red bounding boxes are the remnants from the source MRI input. The objective of this paper is to optimize the diffusion model and reduce such patterns in translation results.

Current medical image translation methods are generally divided into two main types: paired image-to-image translation and unpaired image-to-image translation [7]. Paired approaches obtain a mapping that translates one image to another based on paired training data. However, it is often challenging and costly to obtained paired data clinically. Therefore, unpaired methods become more attractive. For example, we often need to register ultrasound images to preoperative prostate MRI in image-guided surgery. However, it is difficult to obtain paired MRI-US datasets to train a robust registration model. To address this issue, unpaired image translation techniques could be developed to obtain large-scale paired MRI-US data. Generative adversarial networks (GANs) have been extensively developed and applied to numerous medical translation tasks, including across MR scanner synthesis [2], multi-contrast MR synthesis [8]–[10], and cross-modality synthesis [11], [12]. Despite the remarkable ability to synthesize realistic medical samples, GAN-based methods indirectly characterize the distribution of the

Yimin Luo, Zenglin Shi, Weimin Huang, Jun Cheng are with the Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), 138632, Singapore. Yimin Luo is also with the School of Biomedical Engineering and Imaging Sciences, King's College London, St. Thomas' hospital, London, SE1 7EH, U.K. (e-mail: yimin.luo@kcl.ac.uk, cheng_jun@i2r.a-star.edu.sg).

Qingyu Yang is with the AI lab of Tencent, Shenzhen, Guangdong, 518054, China.

Ziyi Liu is with the School of Computer Sciences, Wuhan University, Wuhan, Hubei, 430071, China.

Guoyan Zheng is with the Institute of Medical Robotics, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China.

Corresponding author: Jun Cheng

target modality through a generator-discriminator interplay without considering the content fidelity [13]. This implicit characterization is often susceptible to learning biases, such as premature convergence and mode collapse. Moreover, GAN models usually rely on a rapid one-shot sampling process without intermediate iterations, which inherently restricts the reliability of the mapping. Consequently, these limitations often have a detrimental impact on the quality and diversity of the synthesized images [14].

To address the above limitations, diffusion models employ explicit likelihood characterization and a gradual sampling process to enhance the fidelity of the synthesized images. The originator of the diffusion is the denoising diffusion probability model (DDPM) [15], which consists of two Markov chains. One disturbs the data into a forward chain of noise and the other converts the noise back to the reverse chain of the data. Then, variational inference is used to gradually generate samples consistent with the original data distribution. Due to their stability in training and gradual generation mechanism, diffusion models have great advantages over previous GANs. Sasaki *et al*. [16] pioneered the method of Unit-DDPM in unpaired image translation tasks. Unit-DDPM is comprised of two parallel diffusion processes for the source and target modalities, where noisy samples from each modality were given as conditional input to reverse diffusion steps for the other modality. Recently, Su *et al*. [17] proposed dual diffusion implicit bridges (DDIB), which has become a scalable and rigorous addition to the family of unpaired image translation. Despite its effectiveness to achieve translation between RGB images, DDIB often retains some modality-specific information from the original domain when applied to medical image translation. Fig. 1 shows two representative results together with their paired ground truth in different tasks. The first example is from brain MRI-CT translation and the second is from prostate MRI-US translation. As we can see in the first example, the black patterns (lateral ventricle) in the red rectangular appear in the translated results. However, such patterns seldom appear in real CT data as CT cannot capture some tissues due to strong attenuation of bones. Moreover, a similar observation of white plaque can be found in the second example. To overcome this limitation, we propose a target guidance method to adopt a guidance representing global discrimination to sampling process. In addition, to promote the diffusion model to learn rich visual concept, we also introduce a weighting scheme to pay more attention to visual content learning at high noise level during training.

The main contributions of this work are summarized in the following aspects:

- We propose a target-guided diffusion model (TGDM) for unpaired medical image cross-modality translation. For training, to encourage diffusion model to learn more visual concepts, we adopt a perception prioritized weighting scheme (P2W) to the training objectives of diffusion. For sampling, we further introduce a target-guided method in its reverse process to reduce modality-specific remnants of source data.
- Experimental results on brain MRI-CT and prostate MRI-US translations demonstrated that the proposed method is able to achieve cross-modality translation effectively. Furthermore, a detailed analysis on guidance strength and subjective assessments are provided to validate the clinical value of the proposed method.

II. Related Works

A. Diffusion Models

Diffusion models [15] learn a Markov chain, gradually converting a Gaussian noise distribution into the data distribution on which the models are trained. The training procedure usually involves two phases: a forward process and a reverse process. Firstly, the forward process starts from the data distribution $q(x_0)$ and sequentially corrupts the data to $\mathcal{N}(\mathbf{0}, \mathbf{I})$ with Markov diffusion kernels $q\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}\right)$ defined by a fixed variance schedule $\{\beta_t\}_{t=1}^{T}$. The process can be expressed by:

$$q\left(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0\right) = \prod_{t=1}^{T} q\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}\right) \tag{1}$$

$$q\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}\right) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \mathbf{I}\right), \tag{2}$$

where the sample x_0 gradually loses its distinguishable features as the step t increases. Meanwhile, we sample x_t at any arbitrary time step t with $\mathcal{N}\left(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, \left(1-\bar{\alpha}_t\right)\mathbf{I}\right)$ in a closed form solution using reparameterization trick, where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. Subsequently, the reverse (generative) process is defined as another Markov chain parameterized by θ, denoising an arbitrary Gaussian noise to a clean data sample:

$$p_\theta\left(\boldsymbol{x}_{0:T}\right) = p\left(\boldsymbol{x}_T\right) \prod_{t=1}^{T} p_\theta\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right) \tag{3}$$

$$p_\theta\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right) = \mathcal{N}\left(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta\left(\boldsymbol{x}_t\right), \boldsymbol{\Sigma}_\theta\left(\boldsymbol{x}_t\right)\right), \tag{4}$$

where $p\left(\boldsymbol{x}_T\right) = \mathcal{N}\left(\boldsymbol{x}_T; \mathbf{0}, \mathbf{I}\right)$.

Training is performed to maximize the model log likelihood by minimizing the variational upper bound of the negative one. By doing this, given a random Gaussian noise sample, we synthesize a sample through this model. The classic objective is derived as same as in [15]:

$$\mathcal{L}_{\text{classic}} = \sum_t \mathbb{E}_{\boldsymbol{x}_0, t, \epsilon}\left[\left\|\epsilon - \boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t\right)\right\|^2\right] \tag{5}$$

Thanks to the exceptional quality and rich diversity of the generated samples, diffusion models have garnered remarkable success for a variety of image-related applications in computer vision fields, such as image generation [14], [15], [18]–[31] image restoration [32]–[35], image translation [17], [36], [37], and more. Capitalizing on the recent advances in computer vision, we observe a growing interest in diffusion models for medical imaging, especially the medical image synthesis. Due to the bottlenecks in medical image annotation, including the complexity of high-quality data collection, insufficient experts, and patient privacy concerns, obtaining large scale of annotated medical image datasets is very challenging. This makes generative models increasingly useful. Müller-Franzes

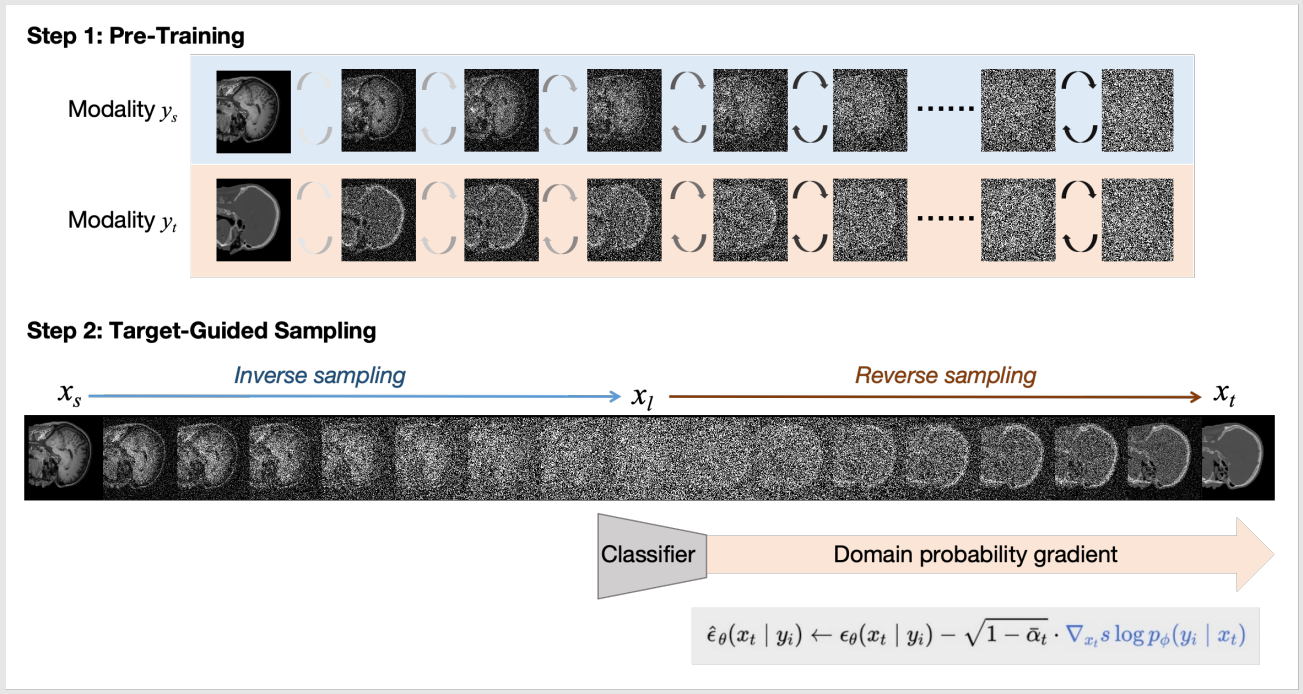

Fig. 2. The pipeline of the proposed target-guided diffusion models (TGDM) for unpaired cross-modality medical image translation. Our method is composed of two steps: pre-training and target-guided sampling. We pre-train a classifier on unpaired datasets from two different medical imaging domains, and its architecture is given in Fig. 3. The pre-trained classifier provides the pre-trained diffusion model with target guidance during its reverse sampling process. The target guidance is a probability gradient belonging to target domain.

et al. [38] presented a conditional latent DDPM for image synthesis in the medical domain, and excessive experiments showed that this method is a superior alternative to GANs for generative tasks involving optical coherence tomography (OCT), histology colorectal image, and chest x-ray image. Similar diffusion model-based generative works for MRI and PET can be found in [39] and [40] respectively. Moreover, to remove noise in retinal OCT, Hu *et al*. [41] presented a diffusion probabilistic model that is fully unsupervised to learn from noise instead of signal. The experimental results demonstrate that this model can significantly improve the image quality of retinal OCT with a small training cost.

B. Cross-Modality Translation

Cross-modality translation is an effective solution that involves synthesis of a missing target modality with the guidance of an existing source modality. It is useful in many clinical applications, such as CT or MRI based radiotherapy planning, sample synthesis, and fast MRI imaging [42]. Recent advancements in deep learning have brought researchers closer to a promising path to capture the conditional distributions from the training data. Zhu *et al*. [43] designed a forward and backward cycle model for unsupervised image-to-image translation and proposed a cycle consistent adversarial network (CycleGAN). CycleGAN has gained a lot of attention, and inspired many new variants and extensions for medical applications, particularly in MR-to-CT synthesis [11], [44]–[47]. Choi *et al*. [48], [49] proposed a more advanced translation network named StarGAN which translates the image with richer textures than CycleGAN. StarGAN has also been introduced for medical image translations on brain MRI translations [50]. Abu-Srhan *et al*. [51] further improved StarGAN via a new combination of non-adversarial losses for bidirectional brain MR-to-CT synthesis. Yi *et al*. [52] proposed DualGAN for image translation from two sets of unlabeled images from two domains. Zhang *et al*. [7] proposed NAGAN to achieve noise adaptation on both OCT and ultrasound. These GAN-based methods effectively transfer the intensity distributions and reduce the domain gaps.

As promising alternatives to GANs, diffusion models have received interest for image translation tasks. Saharia *et al*. [37] developed a unified framework for image-to-image translation based on conditional diffusion models, which greatly inspired further innovations [53]. Besides conditional diffusion models, Li *et al*. [54] proposed a Brownian bridge diffusion model (BBDM), which models image-to-image translation as a stochastic Brownian bridge process, and learns the translation between two domains directly through the bidirectional diffusion process. To improve diversity in medical imaging protocols, Ö zbey *et al*. [55] introduced a novel adversarial diffusion model, namely SynDiff, for medical image translation. SynDiff embodies a considerable enhancement with its innovative adversarial diffusion design. However, its translation performance is constrained due to the reliance on an independent CycleGAN model trained from paired data synthesis. Currently, diffusion models for cross domain translation from unpaired data remain under-explored in medical imaging fields, which is the focus of this paper.

III. METHODS

In this paper, our target is to design a novel diffusion model for unpaired image translation between two different medical modalities. The source domain typically provides the content of the images, while the target domain provides the domain-specific style of the images.

A. Overview

Recently, a diffusion model based unpaired image translation method, namely DDIB, is innovatively proposed [17] for image-to-image translation in natural images. Although DDIB is promising for natural images, its use for medical images faces some challenges due to large gaps between different modalities in medical imaging. As illustrated in the introduction, the depiction of human tissue exhibits significant changes from one medical imaging modality to another. DDIB shows limited effectiveness across different medical imaging modalities. Its translation results often retain modality-specific information from the original domain, which leads to unrealistic results. To address this limitation, we improve the training objective of diffusion with perception prioritized weighting (P2W) and design a two-stage sampling process in which strong guidance of target modality is introduced to provide gradients. The overall pipeline is illustrated in Fig. 2.

B. Training Objective

As shown in Fig. 2, we need to train diffusion models on the unpaired datasets. Su *et al*. [17] innovatively adopted the diffusion model to unpaired image translation through a special dual sampling process. Although it has lead to good progress, the classic training objective does not include a better content learning, which is important in image translation tasks. For each step t, denoising score matching loss $L_{\text{diffusion}}$ is a distance between two Gaussian distributions, which can be rewritten in terms of noise predictor ϵ_θ. At a low noise level, the image sample is only slightly corrupted and recovering it does not require prior knowledge of image contexts. However, when sample is seriously corrupted, its contents are unrecognizable, and we need to prioritize solving the pretext task of these important noise levels.

Since the data distribution between two medical imaging modalities is often different and we observe that the joint training of two modalities can be performed stably. Here, we design to embed the modality category information y in the joint training process to fit a conditional distribution of samples in specific modality. The entire training process is unpaired. Moreover, Choi *et al*. [56] found that restoring data corrupted with certain noise levels offers a proper pretext task for the diffusion model to learn richer visual concepts, and they proposed to assign higher weights to the loss at levels where the model learns perceptually rich contents while minimal weights to where the model learns imperceptible details. Since using this training strategy can significantly improves the performance of diffusion models in generative tasks, to further promote translation content-consistency, the diffusion model should pay more attention to visual content learning at high noise level during training. Motivated by this, we improve and redesign the weighting scheme of the objective function as

$$L = \sum_t \mathbb{E}_{x_0,\epsilon}\left[W \left\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, y, t\right)\right\|^2\right] \tag{6}$$

where W represents the P2W and is introduced in $L_{\text{diffusion}}$ to prioritize learning from more important noise levels. Here, W is calculated as

$$W = \frac{1}{\sqrt{1+\bar{\alpha}_t/(1-\bar{\alpha}_t)}} = \sqrt{1-\bar{\alpha}_t}. \tag{7}$$

On the other hand, to achieve our target guidance sampling, we pre-train a diffusion model $\epsilon_\theta\left(x_t \mid y\right)$ with category embedding as well as a two-class (source and target data) classifier $p_\phi\left(y \mid x_t\right)$ with time t information embedding, where y represents the target modality label. In this paper, we use a classifier architecture which is simply the down-sampling trunk of the U-Net model with an attention pool, as presented in Fig. 3, and the loss function is as follows:

$$L_{\text{pre-classifier}} = \sum_t \mathbb{E}_{x_0,\epsilon}\left[||y - p_\phi\left(x_t, t\right)||_c\right] \tag{8}$$

where $||\cdot||_c$ represents the cross-entropy loss function.

C. Target Guidance Sampling

After obtaining a diffusion model and a classifier trained on the original sample sets, we come to a two-stage sampling process. The overall sampling algorithm is in Algorithm 1. Since DDIB often retains modality-specific information in cross-modality translation tasks, strong modality information should be used in the sampling process for better translation results. Recently, Zhao *et al*. [58] found that embedding an energy function of cross-domain correlation to the sampling process of the diffusion models can obtain a conditional distribution between the two domains. Dhariwal and Nichol [14] found that the classifier guidance enables the diffusion model to generate a more class-consistent image from a random Gaussian noise. Motivated by these works, we propose a target-guided sampling process in this paper. Our sampling has two main steps: inverse sampling and target guidance sampling. We firstly obtain latent code from source data through an inverse sampling. Then, we synthesize its counterparts in target domain through a reverse sampling process.

Hereby, we exploit the reverse process of a DDIM sampling and source-target classifier guidance sampling to further synthesize the target samples. According to [59], we infer a stochastic latent code by running the deterministic generative process of DDIMs in reverse:

$$x_{t+1} = \sqrt{\bar{\alpha}_{t+1}}\left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta\left(x_t \mid y_i\right)}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\bar{\alpha}_{t+1}}\epsilon_\theta\left(x_t \mid y_i\right) \tag{9}$$

Finally, we translate the target samples based on the pre-trained diffusion model $\epsilon_\theta\left(x_t, t\right)$ by iteratively using the modality probability gradients $\nabla_{x_t} s \log p_\phi\left(y_i \mid x_t\right)$ from our

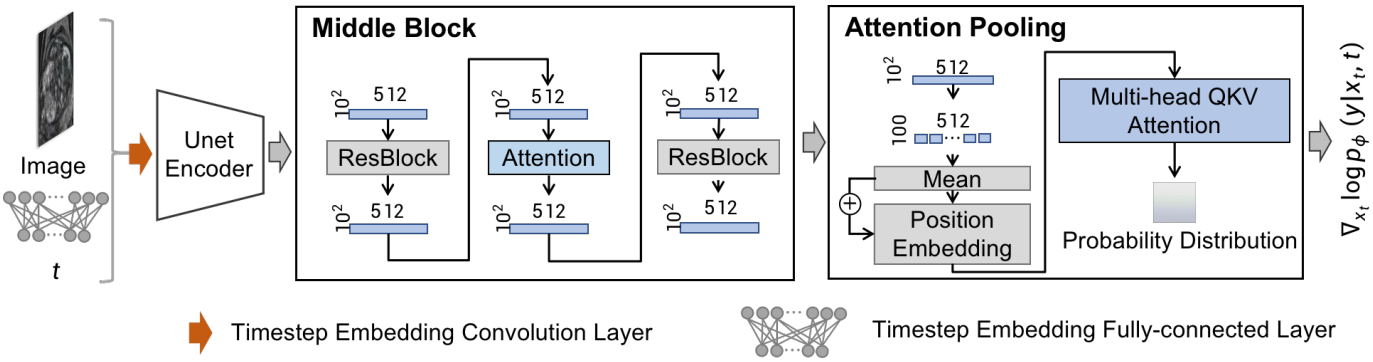

Fig. 3. The architecture of our pre-trained classifier. This classifier is the down-sampling trunk of the U-Net model with an attention pool [57] at the 8×8 layer. The output of this classifier is the probability gradient of the image input belonging to target domain.

pre-trained classifier $p_\phi(y \mid x_t)$. Our target-guided diffusion model performs this classifier guidance sampling by iteratively replacing $\epsilon_\theta(x_t \mid y_i)$ in each sampling step with

$$\hat{\epsilon}_\theta(x_t \mid y_i) = \epsilon_\theta(x_t \mid y_i) - \sqrt{1-\bar{\alpha}_t} \cdot \nabla_{x_t} s \log p_\phi(y_i \mid x_t), \tag{10}$$

where the coefficient s is called the guidance scale. Hereby, s represents how close the translated results to the target modality. In our tasks, we find that an appropriate guidance scale is beneficial to achieve high-quality medical image translation. A large scale may contribute to the loss in diversity.

Compared to DDIB, the establishment of $p_\theta(x_{t-1}|x_t)$ is changed to $p_\theta(x_{t-1}|x_t, y)$. Since,

$$p(x_{t-1}|y) = \frac{p(x_{t-1})p(y|x_{t-1})}{p(y)} \tag{11}$$

Adding condition x_t to each item, we obtain:

$$p_\theta(x_{t-1}|x_t, y) == \frac{p_\theta(x_{t-1}|x_t)p_\phi(y|x_{t-1}, x_t)}{p(y|x_t)}, \tag{12}$$

where x_t represents the resultant image after adding noise on x_{t-1}. Since x_t has extremely small impact on the classification, we have

$$p_\phi(y|x_{t-1}, x_t) \approx p_\phi(y|x_{t-1}) \tag{13}$$

Substituting Eq. (13) into Eq. (12), we obtain

$$p_\theta(x_{t-1}|x_t, y) = p_\theta(x_{t-1}|x_t)e^{(\log p_\phi(y|x_{t-1}) - \log p_\phi(y|x_t))} \tag{14}$$

Hence, we exploit the gradients $\nabla x_t log p_\varphi(y|x_t, t)$ to guide our diffusion model towards the target modality y. By doing this, our sampling process incorporates the global modality information for better translation results.

Algorithm 1 Target-guided Diffusion Model Sampling.

Input: Source $\tilde{x}$, Implicit diffusion model $\epsilon_\theta(x, t)$, Modality classifier $p_\varphi(y|x_t, t)$, label y, gradient scale s, Reverse sampling steps L, Sampling steps T

Stage 1: obtain latent code from source data

for $l = 1, 2, \cdots, L-1$ **do**

  $m \leftarrow \tilde{x}$ **if** $l = 1$ **else** $m \leftarrow m_l$ // m represents the intermediate variable of Gaussian noise interference

  $\epsilon_\theta^{(l)} \leftarrow \epsilon_\theta(m, l)$

  $m_{l+1} \leftarrow \sqrt{\bar{\alpha}_{l+1}} \cdot \epsilon_\theta^{(l)} + \sqrt{(1-\bar{\alpha}_{l+1})(\bar{\alpha}_l - 1)} \cdot \left(\frac{m_l}{\sqrt{\bar{\alpha}_l}} - \epsilon_\theta^{(l)}\right)$

end for

Stage 2: obtain target data from latent code

for $t = T, T-1, \cdots, 1$ **do**

  $x \leftarrow m_L$ **if** $t = T$ **else** $x \leftarrow x_t$

  $\epsilon_\theta^{(t)} \leftarrow \epsilon_\theta(x, t) - \sqrt{1-\bar{\alpha}_t} \cdot s \cdot \nabla_{x_t} \log p_\varphi(y|x_t, t)$

  $\sigma_t \leftarrow \sqrt{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}} \cdot \sqrt{\frac{1-\bar{\alpha}_t}{\bar{\alpha}_{t-1}}}$

  $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ **if** $t < T$ **else** $\epsilon \leftarrow 0$

  $x_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}} \left(\frac{x_t - \sqrt{1-\bar{\alpha}_t} \cdot \epsilon_\theta^{(t)}}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\bar{\alpha}_{t-1}\sigma_t^2} \cdot \epsilon_\theta^{(t)} + \sigma_t \cdot \epsilon$

end for

return x_0 // the translated result

D. Evaluations

Assessing the performance of cross-modality medical image translation techniques poses a significant challenge as acquiring paired samples can be arduous clinically. When paired reference images from the target modality are available for validation, the standard evaluation metrics used are the peak signal-to-noise ratio (PSNR), and structural similarity (SSIM). In addition, we also use Fréchet Inception Distance (FID) [60], which is used to quantify the discrepancy between the probability distributions of synthetic and real volumes. When paired reference images are unfortunately unavailable, as shown in [61], [62], only FID [60] is used. As FID tends to penalize non-GAN models, we also adopt subjective evaluation to further validate the clinical acceptance of the translated results. In this work, seven clinicians from the ultrasonography department have been invited to anonymously evaluate and compare the translated US results using a voting assessment approach.

IV. Experiment

In this section, we exploit the proposed method to two unpaired cross-modality medical image translation tasks. We show the details of our experimental results.

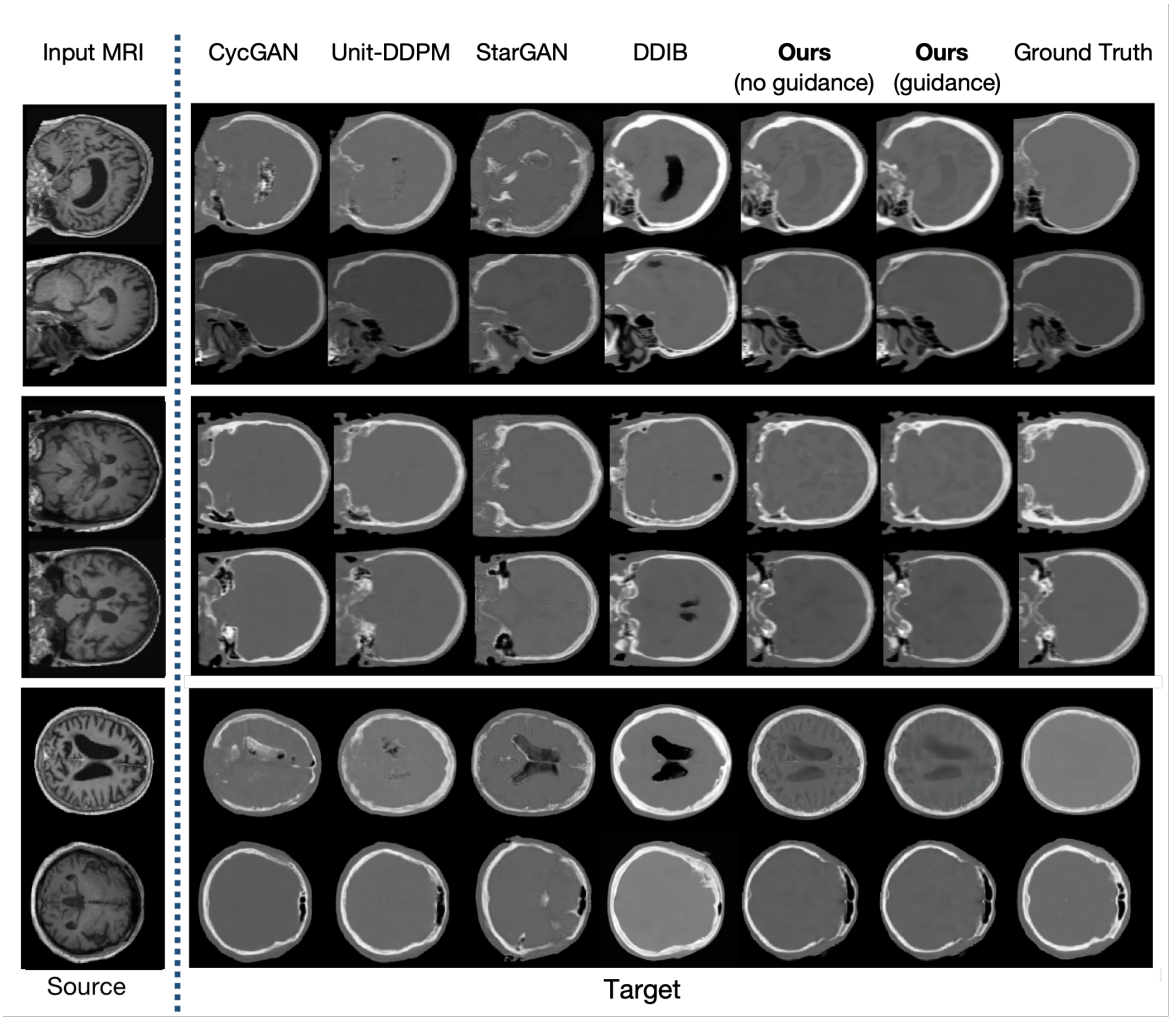

Fig. 4. The visual comparison of the brain MRI to CT translation results from three directions (X, Y, and Z). From left to right: the input brain MRI images, and the translation results of CycleGAN, Unit-DDPM, StarGAN, DDIB, our method without and with guidance (guidance scale $s = 5$).

A. Dataset and Implementation

MRI-CT brain is a new dataset consisting of 3D MRI-CT brain scans from 50 different patients, which is slightly higher than those in [63] and [64]. Our dataset comprises images acquired in the head region for clinical purposes such as dementia, epilepsy, and brain tumor grading. The MRI images in our dataset have a voxel spacing of $0.8 \times 0.8 \times 0.8\ mm^3$, while the CT images have a voxel spacing of $0.9 \times 0.9 \times 2.5\ mm^3$. To align the two modalities, we perform registration and sample the aligned images with a voxel spacing of $1.0 \times 1.0 \times 1.5\ mm^3$. The gray values of the CT images are uniformly distributed in the Hounsfield unit range of $[-1024, 2252.7]$. To ensure consistency, we resample all the training data to an isotropic resolution and normalize each MR image to have zero mean and unit variance. For the CT images, we normalize them to the range of $[0, 1]$, expecting that the synthetic values generated as output would also fall within the same range. The final synthetic CT image is obtained by multiplying the normalized output with the Hounsfield unit range observed in our training data. As we have ground truth (paired CT samples) for validation, we use PSNR, SSIM, and FID for translation performance evaluation. Experimentally, we analyzed brain MRI and CT images from 50 patient subjects and each subject has MRI-CT mage pairs from three imaging directions (X,Y, and Z). We use 4725 samples from 35 subjects for training and 1800 samples from 15 subjects for validation.

Prostate-MRI-US-Biopsy is a new dataset which is extracted from [65]. In the dataset, only the T2 weighted MRI is available and the 3D ultrasound is reconstructed from series of 2D ultrasound images. The original 2D ultrasound images are not included in the dataset. This dataset was derived from tracked biopsy sessions using the Artemis biopsy system and comprises 1,600 MRI images and 1,600 reconstructed ultrasound images selected and cropped from 160 subjects with the coronal section of the prostate in the centre of the patch, and their patch sizes are both 320×320 pixels. Most of these samples can clearly display the left and right lobes of the prostate. Moreover, we also select another 5 subjects there the ultrasound data is most close to the MRI for the performance evaluation of different methodologies. Since we don't have strictly paired ultrasound with golden ground truth for validation, we only use FID for translation performance evaluation. In addition, we have also invited clinicians to anonymously evaluate and compare the translated ultrasound

TABLE I
TRANSLATION RESULTS ON THE MRI-CT DATASET

Methods	PSNR↑	SSIM↑	FID↓
CycleGAN [43]	18.08	0.695	95.46
Unit-DDPM [16]	15.86	0.548	145.72
StarGAN [49]	16.41	0.559	152.01
DDIB [17]	15.55	0.524	122.45
Ours ($s = 5$)	**18.62**	0.612	**71.03**

TABLE II
AN ABLATION STUDY AND ANALYSIS ON P2W AND TARGET (CLASSIFIER) GUIDANCE ON THE MRI-CT DATASET.

	Guidance	P2W	PSNR	SSIM	FID
$s = 0$	×	×	15.55	0.524	122.45
$s = 0$	×	✓	18.53	0.633	73.18
$s = 0.5$	✓	✓	18.55	0.637	72.93
$s = 1$	✓	✓	18.58	**0.640**	72.36
$s = 3$	✓	✓	18.60	0.626	71.45
$s = 5$	✓	✓	18.62	0.612	71.03
$s = 10$	✓	✓	**18.63**	0.595	**70.72**

results using a voting assessment approach.

Experimentally, we implement the proposed TGDM using PyTorch 1.8.0. We conduct the diffusion training on a system with an NVIDIA QUADRO GV100 (32GB PCIE) GPU and the classifier training on an NVIDIA RTX 3090 (24GB PCIE) GPU.

B. MRI-CT Translation

For the brain MRI-CT dataset, we compare the proposed diffusion model-based method to four well-known unpaired image translation methods including CycleGAN [43], Unit-DDPM [16], StarGAN [49], and DDIB [17]. Fig. 5 illustrates the visual comparisons of our method and other methods on this dataset, and the translation CT results are displayed from three directions (X, Y, and Z). It can be witnessed that DDIB results preserved black patterns in the brain center region from its source MRI input and our method with target guidance (guidance scale $s = 5$) can obviously improve this drawback. Note that, the edge of DDIB results is much brighter than ground through and other methods, which seriously impacts the authenticity of the translation results. However, we observe notably poor and unstable performance for all examined methods. The MRI to CT results show superior performances for CycleGAN in some cases, where CycleGAN is able to completely hide the anatomical structures in the resulting CT scan. On the other hand, CycleGAN shows worse performance in other cases, where there are obvious residuals of the anatomy. The same situation can be witnessed in Unit-DDPM results. Unit-DDPM comprises two parallel diffusion processes for the source and target modalities, where noisy samples from each modality are given as conditional input to reverse diffusion steps for the other modality. In this work, the difference between the two modalities is too significant, the soft tissue information on MRI samples which is scarcely present in its CT counterparts should be fully removed. It is an extremely challenging task to current diffusion models, and it is also worth further research.

To sum up, according to our initial comparisons to previous DDIB, our diffusion model can preserve better style of the target CT modality and anatomical information and achieves relatively better visual results and stability. As we have ground truth for validation, moreover, we use PSNR, SSIM, and FID for performance evaluation. Table I compares average results of the mentioned methodologies. It can be witnessed that our TGDM shows better translation performance than previous DDIB by achieving higher PSNR and SSIM as well as lower FID. Although this method is not particularly advantageous over CycleGAN, it is still overall better than it. All differences are found to be statistically different using a paired t-test ($p < 0.01$).

Ablation Study and Analysis: We have also conducted a comprehensive ablation study to validate the effectiveness of the target guidance in the sampling process and P2W in the training process. The results of this study are presented in Table II. As displayed in Eq. (10), the coefficient s represents the guidance scale, and this parameter is significant to how closed the translated results to the target medical imaging modality. Accordingly, $s = 0$ represents the guidance scale in the sampling process is 0, which means no guidance. In this table, we firstly compare the performance of our method without guidance in the sampling process with the previous diffusion models that use classic objective in the training process. It can be witnessed that P2W promotes diffusion model training and achieves obviously better PSNR, SSIM, and FID results. Moreover, we make a further analysis on the selection of this parameter as well as its influences on the sampling process. Table II also compares the average evaluation results under different s. Along with the increase of the guidance scale s, the PSNR gradually increases, while the FID slightly decreases. For SSIM, note that, it initially increases and then decreases.

Overall, we determine that P2W has efficacy in training diffusion models for unpaired cross-modality medical image translation tasks. Nonetheless, there is only marginal differences in PSNR, SSIM and FID by using different scales in the sampling process. Accordingly, optimizing the training process of diffusion model is necessary and the domain probability gradients provided by a pre-trained classifier in the sampling process still have limited effect on removing modality-specific remnants of source data. Therefore, more specific and local guidance, such as contour and segmentation mask, is essential for improvement.

C. MRI-US Translation

Moreover, we also do a further validation on another MRI-US dataset. Fig. 5 illustrates the visual comparisons of our TGDM and other methods on this dataset. There are significant differences between different examined methods on these tasks. Visibly, previous DDIB provided much brighten results, which means that it can hardly learn the data distribution

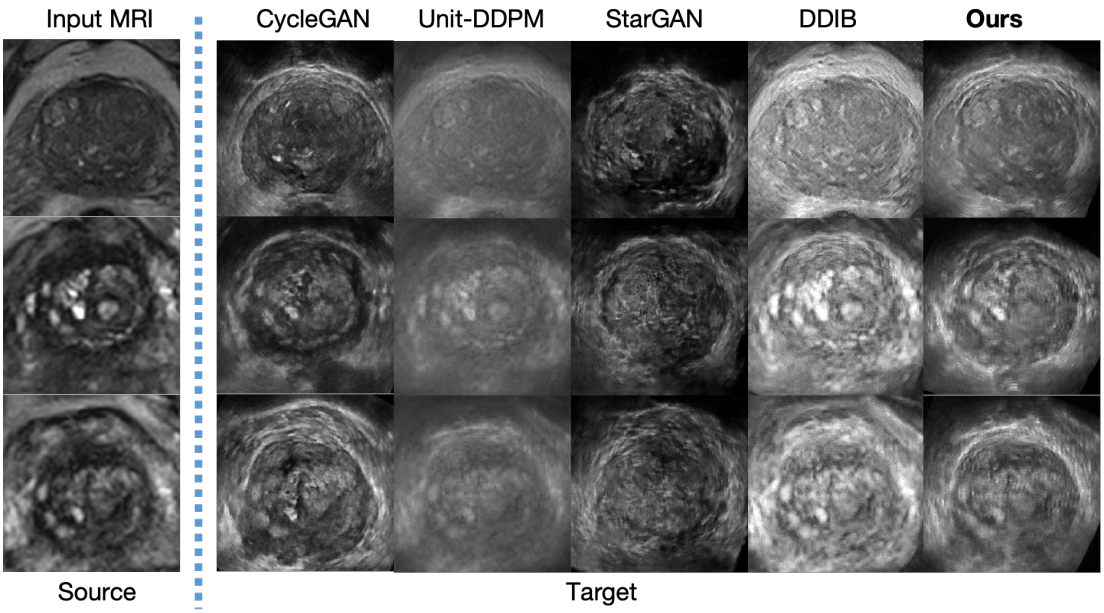

Fig. 5. The visual comparison of the prostate MRI to ultrasound translation results. From left to right are the input prostate MRI images, and the translation results of CycleGAN, Unit-DDPM, StarGAN, DDIB and our method.

TABLE III
TRANSLATION RESULTS ON THE MRI-US DATASET

Methods	FID↓
CycleGAN [43]	316.54
StarGAN [49]	371.32
Unit-DDPM [16]	284.75
DDIB [17]	355.28
Ours	**143.33**

during the diffusion model training. Our method reduces this drawback and modality-specific remnants of the input MRI data. The result of CycleGAN displays clear contour and texture information which does not usually occur in ultrasound samples. Moreover, Unit-DDPM produced very blurred translation results, but these results show less modality-specific remnants than our method. Finally, StarGAN produced dack and very blurred translation results.

Since we do not have ground truth, we can only use FID for evaluation. Table III compares average results of the mentioned methods. There are also significant differences between different examined methods in terms of FID, which may result from the small size of our validation dataset. All differences were found to be statistically different using a paired t-test ($p < 0.01$). Overall, it can be clearly witnessed that our TGDM show better translation performance than those methods by achieving lower FID.

Clinician Assessment: Due to the lack of strictly paired data acted as ground truth, we should seek other approaches to strengthen our validation. To further validate the clinical value of the MRI-US translation, in addition, seven clinicians from the ultrasonography department of two hospitals evaluated and compared translation results using a voting assessment approach. For each MRI image, we translate it to an ultrasound image using three different methods (CycleGAN [43], StarGAN [48], and ours). Then the three translated ultrasound images are shown together with the MRI image to the clinicians, who were asked to select the one that is most likely to be a real ultrasound image with acceptable quality based on their experiences. If none of the three images are acceptable, the vote will go to 'Give up'. Ten sets of images are used in our subjective test and the voting scores were tallied among seven clinicians and are compared in Fig. 6. Notably, according to our investigation, the proposed method also had the highest votes when comparing to GAN-based methods, which displays the potentials of diffusion to unpaired cross-modality medical image translation.

The inherent subjectivity among the seven observers renders the straightforward voting outcome somewhat less persuasive. Nonetheless, it is witnessed that clinicians are receptive to synthetic medical images, particularly those generated using diffusion models. Consequently, we are optimistic about the potential of this approach, and anticipate the possibility of extending validation to downstream applications, such as diagnosis and registration, in subsequent research.

D. Discussions

Despite desirable medical image translation results, our method requires a pre-trained classifier in the model training, which leads to some additional effort when the method is used for new data. Secondly, our current study is still at early stage and the size of the dataset is not large. Therefore, we only evaluated the method in a small validation set. Further evaluation in large dataset in downstream tasks would be helpful in future work. We hope this work can help to attract more researchers to work on medical image translation with more data.

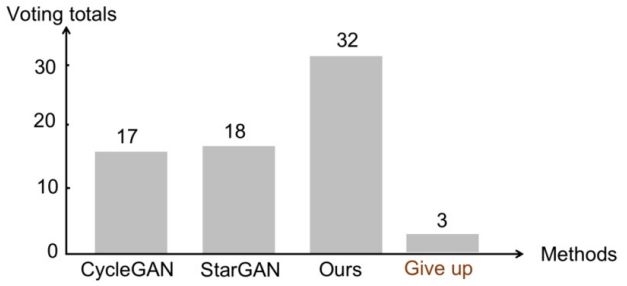

Fig. 6. Clinician voting results on the Prostate-MRI-US-Biopsy dataset via three translation methods: CycleGAN, StarGAN, and our method. We additionally set a choice 'give up' for the invited clinicians which represents 'no preference to these translation results'

V. Conclusion

Diffusion models have emerged as promising alternative to GANs, garnering considerable attention in recent image generation tasks. Despite their potential, diffusion models have not yet been extensively considered for unpaired cross-modality translation in medical imaging. Previous unpaired diffusion model-based image translation methods, such as DDIB, cannot work well to remove remnants in medical image translations. To address this issue, this paper introduces a target guidance diffusion model for unpaired cross-modality medical image translation. During training, our model incorporates a P2W to enrich the learning of diverse visual concepts. In the sampling process, a pre-trained classifier is employed within the reverse process to mitigate the modality-specific features from the source data. Evaluations conducted on both brain MRI-CT and prostate MRI-US datasets demonstrate that our target guidance diffusion model presents superior translation quality, with fewer remnants of the original medical imaging modality. This indicates that diffusion models hold promise for unpaired cross-modality medical image translation tasks. Furthermore, seven clinicians from the ultrasonography department were invited to provide an assessment based on the translated samples to validate the clinical value.

VI. Acknowledgement

This work is supported by the Agency for Science, Technology and Research under its AI^3 Horizontal Technology Coordinating Office Grant C231118001 and C211118006. The authors are grateful to all anonymous reviewers for their insightful comments on this work, especially the experimental section. Moreover, the author would like to thank Xiaoye Wang for his help and discussion of the experiments and all anonymous clinicians for their active participation in the voting process, which greatly improves our analysis subsection of this work.

References

[1] D. H. Ye, D. Zikic, B. Glocker, A. Criminisi, and E. Konukoglu, "Modality propagation: coherent synthesis of subject-specific scans with data-driven regularization," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part I 16*. Springer, 2013, pp. 606–613.

[2] D. Nie, R. Trullo, J. Lian, L. Wang, C. Petitjean, S. Ruan, Q. Wang, and D. Shen, "Medical image synthesis with deep convolutional adversarial networks," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 12, pp. 2720–2730, 2018.

[3] N. Cordier, H. Delingette, M. Lê, and N. Ayache, "Extended modality propagation: image synthesis of pathological cases," *IEEE transactions on medical imaging*, vol. 35, no. 12, pp. 2598–2608, 2016.

[4] Y. Huang, L. Shao, and A. F. Frangi, "Cross-modality image synthesis via weakly coupled and geometry co-regularized joint dictionary learning," *IEEE transactions on medical imaging*, vol. 37, no. 3, pp. 815–827, 2017.

[5] H. Van Nguyen, K. Zhou, and R. Vemulapalli, "Cross-domain synthesis of medical images using efficient location-sensitive deep network," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I 18*. Springer, 2015, pp. 677–684.

[6] V. Sevetlidis, M. V. Giuffrida, and S. A. Tsaftaris, "Whole image synthesis using a deep encoder-decoder network," in *Simulation and Synthesis in Medical Imaging: First International Workshop, SASHIMI 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 21, 2016, Proceedings 1*. Springer, 2016, pp. 127–137.

[7] T. Zhang, J. Cheng, H. Fu, Z. Gu, Y. Xiao, K. Zhou, S. Gao, R. Zheng, and J. Liu, "Noise adaptation generative adversarial network for medical image analysis," *IEEE transactions on medical imaging*, vol. 39, no. 4, pp. 1149–1159, 2019.

[8] S. U. Dar, M. Yurt, L. Karacan, A. Erdem, E. Erdem, and T. Cukur, "Image synthesis in multi-contrast mri with conditional generative adversarial networks," *IEEE transactions on medical imaging*, vol. 38, no. 10, pp. 2375–2388, 2019.

[9] D. Lee, J. Kim, W.-J. Moon, and J. C. Ye, "Collagan: Collaborative gan for missing image data imputation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2487–2496.

[10] M. Yurt, S. U. Dar, A. Erdem, E. Erdem, K. K. Oguz, and T. Çukur, "mustgan: multi-stream generative adversarial networks for mr image synthesis," *Medical image analysis*, vol. 70, p. 101944, 2021.

[11] C.-B. Jin, H. Kim, M. Liu, W. Jung, S. Joo, E. Park, Y. S. Ahn, I. H. Han, J. I. Lee, and X. Cui, "Deep ct to mr synthesis using paired and unpaired data," *Sensors*, vol. 19, no. 10, p. 2361, 2019.

[12] 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.

[13] 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.

[14] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, 2021.

[15] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[16] H. Sasaki, C. G. Willcocks, and T. P. Breckon, "Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models," *arXiv preprint arXiv:2104.05358*, 2021.

[17] X. Su, J. Song, C. Meng, and S. Ermon, "Dual diffusion implicit bridges for image-to-image translation," in *International Conference on Learning Representations*, 2022.

[18] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Advances in neural information processing systems*, vol. 32, 2019.

[19] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations*, 2020.

[20] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8162–8171.

[21] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, 2020.

[22] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[23] Y. Song and S. Ermon, "Improved techniques for training score-based generative models," *Advances in neural information processing systems*, vol. 33, pp. 12 438–12 448, 2020.

[24] A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. Mcgrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," in *International Conference on Machine Learning*. PMLR, 2022, pp. 16 784–16 804.

[25] Y. Song, C. Durkan, I. Murray, and S. Ermon, "Maximum likelihood training of score-based diffusion models," *Advances in Neural Information Processing Systems*, vol. 34, pp. 1415–1428, 2021.

[26] A. Sinha, J. Song, C. Meng, and S. Ermon, "D2c: Diffusion-decoding models for few-shot conditional generation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 533–12 548, 2021.

[27] A. Vahdat, K. Kreis, and J. Kautz, "Score-based generative modeling in latent space," *Advances in Neural Information Processing Systems*, vol. 34, pp. 11 287–11 302, 2021.

[28] F. Bao, C. Li, J. Zhu, and B. Zhang, "Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models," in *International Conference on Learning Representations*, 2021.

[29] T. Dockhorn, A. Vahdat, and K. Kreis, "Score-based generative modeling with critically-damped langevin diffusion," in *International Conference on Learning Representations*, 2021.

[30] N. Liu, S. Li, Y. Du, A. Torralba, and J. B. Tenenbaum, "Compositional visual generation with composable diffusion models," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*. Springer, 2022, pp. 423–439.

[31] Y. Jiang, S. Yang, H. Qiu, W. Wu, C. C. Loy, and Z. Liu, "Text2human: Text-driven controllable human image generation," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 4, pp. 1–11, 2022.

[32] M. Daniels, T. Maunu, and P. Hand, "Score-based generative neural networks for large-scale optimal transport," *Advances in neural information processing systems*, vol. 34, pp. 12 955–12 965, 2021.

[33] H. Chung, B. Sim, and J. C. Ye, "Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 413–12 422.

[34] B. Kawar, M. Elad, S. Ermon, and J. Song, "Denoising diffusion restoration models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23 593–23 606, 2022.

[35] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 461–11 471.

[36] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon, "Ilvr: Conditioning method for denoising diffusion probabilistic models," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2021, pp. 14 347–14 356.

[37] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *ACM SIGGRAPH 2022 Conference Proceedings*, 2022, pp. 1–10.

[38] G. Müller-Franzes, J. M. Niehues, F. Khader, S. T. Arasteh, C. Haarburger, C. Kuhl, T. Wang, T. Han, T. Nolte, S. Nebelung *et al.*, "A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis," *Scientific Reports*, vol. 13, no. 1, p. 12098, 2023.

[39] B. Kim and J. C. Ye, "Diffusion deformable model for 4d temporal medical image generation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part I*. Springer, 2022, pp. 539–548.

[40] Z. Dorjsembe, S. Odonchimed, and F. Xiao, "Three-dimensional medical image synthesis with denoising diffusion probabilistic models," in *Medical Imaging with Deep Learning*, 2022.

[41] D. Hu, Y. K. Tao, and I. Oguz, "Unsupervised denoising of retinal oct with diffusion probabilistic model," in *Medical Imaging 2022: Image Processing*, vol. 12032. SPIE, 2022, pp. 25–34.

[42] J. Chen, S. Chen, L. Wee, A. Dekker, and I. Bermejo, "Deep learning based unpaired image-to-image translation applications for medical physics: a systematic review," *Physics in Medicine & Biology*, 2023.

[43] 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.

[44] K. N. Brou Boni, J. Klein, A. Gulyban, N. Reynaert, and D. Pasquier, "Improving generalization in mr-to-ct synthesis in radiotherapy by using an augmented cycle generative adversarial network with unpaired data," *Medical physics*, vol. 48, no. 6, pp. 3003–3010, 2021.

[45] W. Li, Y. Li, W. Qin, X. Liang, J. Xu, J. Xiong, and Y. Xie, "Magnetic resonance image (mri) synthesis from brain computed tomography (ct) images based on deep learning methods for magnetic resonance (mr)-guided radiotherapy," *Quantitative imaging in medicine and surgery*, vol. 10, no. 6, p. 1223, 2020.

[46] C. Wang, G. Yang, G. Papanastasiou, S. A. Tsaftaris, D. E. Newby, C. Gray, G. Macnaught, and T. J. MacGillivray, "Dicyc: Gan-based deformation invariant cross-domain information fusion for medical image synthesis," *Information Fusion*, vol. 67, pp. 147–160, 2021.

[47] H. Yang, J. Sun, A. Carass, C. Zhao, J. Lee, J. L. Prince, and Z. Xu, "Unsupervised mr-to-ct synthesis using structure-constrained cyclegan," *IEEE transactions on medical imaging*, vol. 39, no. 12, pp. 4249–4261, 2020.

[48] 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.

[49] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "Stargan v2: Diverse image synthesis for multiple domains," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8188–8197.

[50] V. M. Bashyam, J. Doshi, G. Erus, D. Srinivasan, A. Abdulkadir, A. Singh, M. Habes, Y. Fan, C. L. Masters, P. Maruff *et al.*, "Deep generative medical image harmonization for improving cross-site generalization in deep learning predictors," *Journal of Magnetic Resonance Imaging*, vol. 55, no. 3, pp. 908–916, 2022.

[51] A. Abu-Srhan, I. Almallahi, M. A. Abushariah, W. Mahafza, and O. S. Al-Kadi, "Paired-unpaired unsupervised attention guided gan with transfer learning for bidirectional brain mr-ct synthesis," *Computers in Biology and Medicine*, vol. 136, p. 104763, 2021.

[52] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2849–2857.

[53] H. Lee, M. Kang, and B. Han, "Conditional score guidance for text-driven image-to-image translation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[54] B. Li, K. Xue, B. Liu, and Y.-K. Lai, "Bbdm: Image-to-image translation with brownian bridge diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, 2023, pp. 1952–1961.

[55] M. Özbey, S. U. Dar, H. A. Bedel, O. Dalmaz, Ş. Özturk, A. Güngör, and T. Çukur, "Unsupervised medical image translation with adversarial diffusion models," *IEEE Transactions on Medical Imaging*, vol. 42, no. 12, pp. 3524–3539, 2023.

[56] J. Choi, J. Lee, C. Shin, S. Kim, H. Kim, and S. Yoon, "Perception prioritized training of diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 472–11 481.

[57] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[58] M. Zhao, F. Bao, C. Li, and J. Zhu, "Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations," *Advances in Neural Information Processing Systems*, vol. 35, pp. 3609–3623, 2022.

[59] Z. Zhang, Z. Zhao, and Z. Lin, "Unsupervised representation learning from pre-trained diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 22 117–22 130, 2022.

[60] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[61] J. Li, Z. Qu, Y. Yang, F. Zhang, M. Li, and S. Hu, "Tcgan: a transformer-enhanced gan for pet synthetic ct," *Biomedical Optics Express*, vol. 13, no. 11, pp. 6003–6018, 2022.

[62] P. Paavilainen, S. U. Akram, and J. Kannala, "Bridging the gap between paired and unpaired medical image translation," in *MICCAI Workshop on Deep Generative Models*. Springer, 2021, pp. 35–44.

[63] D. Nie, R. Trullo, and et al., "Medical image synthesis with context-aware generative adversarial networks," in *MICCAI*, 2017, pp. 417–425.

[64] X. Han, "Mr-based synthetic ct generation using a deep convolutional neural network," *Med. Phys.*, vol. 44, no. 4, pp. 1408–1419, 2017.

[65] S. Natarajan, A. Priester, D. Margolis, J. Huang, and L. Marks, "Prostate mri and ultrasound with pathology and coordinates of tracked biopsy (prostate-mri-us-biopsy)," *Cancer Imaging Arch*, vol. 10, p. 7937, 2020.