Unsupervised Medical Image Translation with Adversarial Diffusion Models

Muzaffer Özbey, Onat Dalmaz, Salman UH Dar, Hasan A Bedel, Şaban Özturk, Alper Güngör, and Tolga Çukur* *Senior Member*

***Abstract*— Imputation of missing images via source-to-target modality translation can improve diversity in medical imaging protocols. A pervasive approach for synthesizing target images involves one-shot mapping through generative adversarial networks (GAN). Yet, GAN models that implicitly characterize the image distribution can suffer from limited sample fidelity. Here, we propose a novel method based on adversarial diffusion modeling, SynDiff, for improved performance in medical image translation. To capture a direct correlate of the image distribution, SynDiff leverages a conditional diffusion process that progressively maps noise and source images onto the target image. For fast and accurate image sampling during inference, large diffusion steps are taken with adversarial projections in the reverse diffusion direction. To enable training on unpaired datasets, a cycle-consistent architecture is devised with coupled diffusive and non-diffusive modules that bilaterally translate between two modalities. Extensive assessments are reported on the utility of SynDiff against competing GAN and diffusion models in multi-contrast MRI and MRI-CT translation. Our demonstrations indicate that SynDiff offers quantitatively and qualitatively superior performance against competing baselines.**

***Index Terms*— medical image translation, synthesis, unsupervised, unpaired, adversarial, diffusion, generative**

I. Introduction

Multi-modal imaging is key for comprehensive assessment of anatomy and function in the human body [1]. Complementary tissue information captured by individual modalities serve to improve diagnostic accuracy and performance in downstream imaging tasks [2]. Unfortunately, broad adoption of multi-modal protocols is fraud with challenges due to economic and labor costs [3]–[6]. Medical image translation is a powerful solution that involves synthesis of a missing target modality under guidance from an acquired source modality [7]–[10]. This recovery is an ill-conditioned problem given nonlinear variations in tissue signals across modalities [11]–[13]. At this juncture, learning-based methods are offering performance leaps by incorporating nonlinear data-driven priors to improve problem conditioning [14]–[17].

Learning-based image translation involves network models trained to capture a prior on the conditional distribution of target given source images [18]–[20]. In recent years, generative adversarial network (GAN) models have been broadly adopted for translation tasks, given their exceptional realism in image synthesis [21]–[26]. A discriminator that captures information regarding the target distribution concurrently guides a generator to perform one-shot mapping from the source onto the target image [27]–[31]. Based on this adversarial mechanism, state-of-the-art results have been reported with GANs in numerous translation tasks including synthesis across MR scanners [23], multi-contrast MR synthesis [21], [25], [27], [32], and cross-modal synthesis [33]–[35].

While powerful, GAN models indirectly characterize the distribution of the target modality through a generator-discriminator interplay without evaluating likelihood [36]. Such implicit characterization is potentially amenable to learning biases, including premature convergence and mode collapse. Moreover, GAN models commonly employ a rapid one-shot sampling process without intermediate steps, inherently limiting the reliability of the mapping performed by the network. In turn, these issues can limit the quality and diversity of synthesized images [37]. As a promising alternative, recent computer vision studies have adopted diffusion models based on explicit likelihood characterization and a gradual sampling process to improve sample fidelity in unconditional generative modeling tasks [37], [38]. However, the potential of diffusion methods in medical image translation remains largely unexplored, partly owing to the computational burden of image sampling and difficulties in unpaired training of regular diffusion models [38].

Here, we propose a novel adversarial diffusion model for medical image synthesis, SynDiff, to perform efficient and high-fidelity modality translation (Fig. 1). Given the source image, SynDiff leverages conditional diffusion to generate the target image. Unlike regular diffusion models, SynDiff adopts a fast diffusion process with large step size for efficiency. Accurate sampling in reverse diffusion steps is achieved by a novel source-conditional adversarial projector that denoises the target image sample with guidance from the source image. To enable unsupervised learning, a cycle-consistent architecture is devised with bilaterally coupled diffusive and non-diffusive processes between the two modalities (Fig. 2). Comprehensive demonstrations are performed for multi-contrast MRI and MRI-CT translation. Our results clearly indicate the superiority of SynDiff against competing GAN and diffusion models. Code for SynDiff is publicly available at https://github.com/icon-lab/SynDiff.

Contributions

- We introduce the first adversarial diffusion model in the

This study was supported in part by TUBITAK BIDEB scholarships awarded to A. Bedel, O. Dalmaz, A. Gungor, and by TUBA GEBIP 2015 and BAGEP 2017 fellowships awarded to T. Çukur (Corresponding author: Tolga Çukur). M. Ozbey and O. Dalmaz contributed equally to this study. M. Ozbey, O. Dalmaz, S.UH. Dar, H.A. Bedel, S. Ozturk, A. Gungor and T. Çukur are with the Department of Electrical and Electronics Engineering, and the National Magnetic Resonance Research Center (UMRAM), Bilkent University, Ankara, Turkey (e-mails: {muzaffer, onat, salman, abedel, sozturk, agungor, cukur}@ee.bilkent.edu.tr). S. Ozturk is also with Amasya University, Turkey. A. Gungor is also with ASELSAN Research Center, Turkey. The authors thank Mahmut Yurt for discussions on diffusion models in early phases of the study.

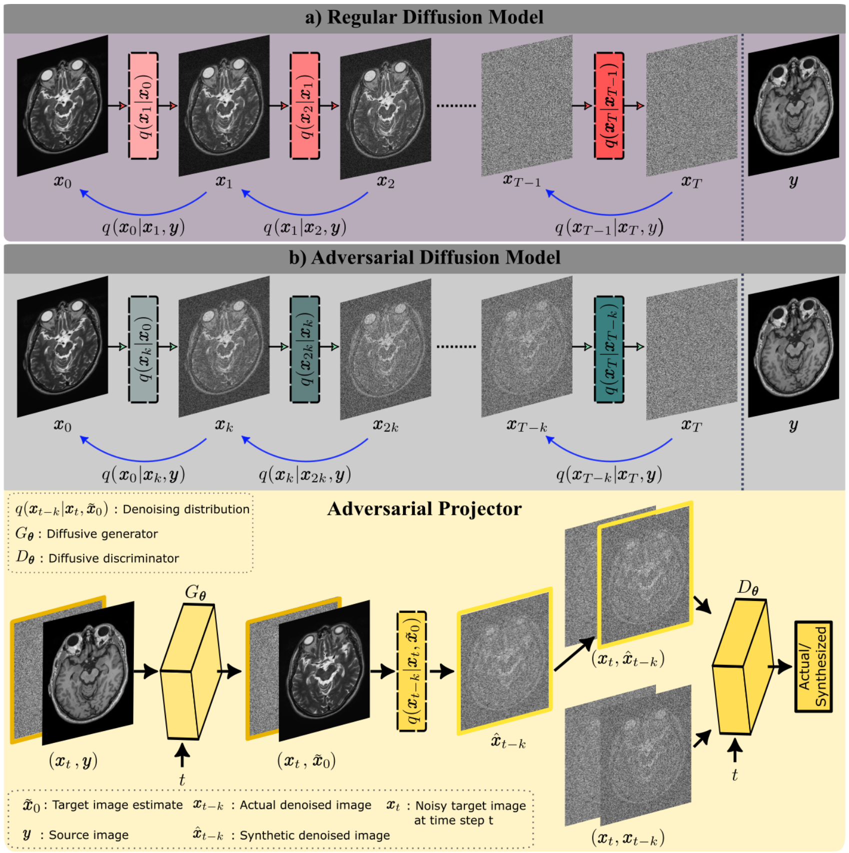
a) Regular Diffusion Model
$q(\boldsymbol{x}_1|\boldsymbol{x}_0)$
$q(\boldsymbol{x}_2|\boldsymbol{x}_1)$
$q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1})$
$\boldsymbol{x}_0$
$\boldsymbol{x}_1$
$\boldsymbol{x}_2$
$\boldsymbol{x}_{T-1}$
$\boldsymbol{x}_T$
$\boldsymbol{y}$
$q(\boldsymbol{x}_0|\boldsymbol{x}_1,\boldsymbol{y})$
$q(\boldsymbol{x}_1|\boldsymbol{x}_2,\boldsymbol{y})$
$q(\boldsymbol{x}_{T-1}|\boldsymbol{x}_T,\boldsymbol{y})$
b) Adversarial Diffusion Model
$q(\boldsymbol{x}_k|\boldsymbol{x}_0)$
$q(\boldsymbol{x}_{2k}|\boldsymbol{x}_k)$
$q(\boldsymbol{x}_T|\boldsymbol{x}_{T-k})$
$\boldsymbol{x}_0$
$\boldsymbol{x}_k$
$\boldsymbol{x}_{2k}$
$\boldsymbol{x}_{T-k}$
$\boldsymbol{x}_T$
$\boldsymbol{y}$
$q(\boldsymbol{x}_0|\boldsymbol{x}_k,\boldsymbol{y})$
$q(\boldsymbol{x}_k|\boldsymbol{x}_{2k},\boldsymbol{y})$
$q(\boldsymbol{x}_{T-k}|\boldsymbol{x}_T,\boldsymbol{y})$
Adversarial Projector
$q(\boldsymbol{x}_{t-k}|\boldsymbol{x}_t,\tilde{\boldsymbol{x}}_0)$: Denoising distribution
$G_{\boldsymbol{\theta}}$: Diffusive generator
$D_{\boldsymbol{\theta}}$: Diffusive discriminator
$G_{\boldsymbol{\theta}}$
$(\boldsymbol{x}_t\,,\boldsymbol{y})$
t
$(\boldsymbol{x}_t\,,\,\tilde{\boldsymbol{x}}_0)$
$q(\boldsymbol{x}_{t-k}|\boldsymbol{x}_t,\tilde{\boldsymbol{x}}_0)$
$\hat{\boldsymbol{x}}_{t-k}$
$(\boldsymbol{x}_t,\hat{\boldsymbol{x}}_{t-k})$
$(\boldsymbol{x}_t,\boldsymbol{x}_{t-k})$
$D_{\boldsymbol{\theta}}$
t
Actual/ Synthesized
$\tilde{\boldsymbol{x}}_0$: Target image estimate
$\boldsymbol{x}_{t-k}$: Actual denoised image
$\boldsymbol{x}_t$: Noisy target image at time step t
$\boldsymbol{y}$: Source image
$\hat{\boldsymbol{x}}_{t-k}$: Synthetic denoised image

Fig. 1: **a)** Regular diffusion models gradually transform between actual image samples for the target modality ($\boldsymbol{x}_0$) and isotropic Gaussian noise ($\boldsymbol{x}_T$) in T steps, with T on the order of thousands. Each forward step (right arrows) adds noise to the current sample to create a noisier sample with forward transition probability $q\left(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t\right)$. Each reverse step (left arrows) suppresses the added noise to create a denoised sample. For image translation, a source modality ($\boldsymbol{y}$) can also be provided as conditioning input to the reverse steps resulting in a reverse transition probability $q\left(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{y}\right)$ assumed to be Gaussian, and operationalized via a neural network $p_{\boldsymbol{\theta}}\left(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{y}\right)$ that estimates its mean. **b)** The proposed adversarial diffusion model performs fast transformation between $\boldsymbol{x}_0$ and $\boldsymbol{x}_T$ in T/k steps, with step size $k \gg 1$. Each forward step adds a greater amount to noise to compensate for large k, breaking apart the normality assumption for reverse transition probabilities $q\left(\boldsymbol{x}_{t-k}|\boldsymbol{x}_t,\boldsymbol{y}\right)$. To improve accuracy, reverse diffusion steps are operationalized via a novel adversarial projector that uses a generator $G_{\boldsymbol{\theta}}$ and a discriminator $D_{\boldsymbol{\theta}}$. $G_{\boldsymbol{\theta}}$ first produces an estimate of the target image $\tilde{\boldsymbol{x}}_0$ given $\boldsymbol{x}_t$ and $\boldsymbol{y}$, and a denoised image sample $\hat{\boldsymbol{x}}_{t-k}$ is then synthesized from the denoising distribution $q\left(\boldsymbol{x}_{t-k}|\boldsymbol{x}_t,\tilde{\boldsymbol{x}}_0\right)$. Meanwhile, $D_{\boldsymbol{\theta}}$ distinguishes between actual ($\boldsymbol{x}_{t-k}$) and synthetic samples ($\hat{\boldsymbol{x}}_{t-k}$) for the denoised image.

literature for high-fidelity medical image synthesis.

- We introduce the first diffusion-based method for unsupervised medical image translation that enables training on unpaired datasets of source-target modalities.
- We propose a novel source-conditional adversarial projector to capture reverse transition probabilities over large step sizes for efficient image sampling.

II. Related Work

To translate medical images, conditional GANs perform one-shot source-to-target mapping via a generator trained using an adversarial loss [23]. Adversarial loss terms are known to improve sensitivity to high-frequency details in tissue structure over canonical pixel-wise losses [21]. As such, GAN-based translation has been broadly adopted in many applications. Augmenting adversarial with pixel-wise losses, a first group of studies considered supervised training on paired sets of source-target images matched across subjects [24], [26]–[30]. For improved flexibility, other studies proposed cycle-consistency or mutual information losses to enable unsupervised learning from unpaired data [21], [33], [39]–[44]. In general, enhanced spatial acuity and realism have been reported in target images synthesized with GANs when compared to vanilla convolutional models [21]. That said, several problems can arise in GAN models, including lower mapping reliability for the one-shot sampling process [37], premature convergence of the discriminator before the generator is properly trained [31], and poor representational diversity due to mode collapse [36]. In turn, these problems can lower sample quality and diversity, limiting generalization performance of GAN-based image translation.

As a recent alternative to GANs, deep diffusion models have received interest for generative modeling tasks in computer vision [37], [38]. Starting from a pure noise sample, diffusion models generate image samples from a desired distribution through repetitive denoising. Denoising is performed via a neural network architecture trained to maximize a correlate on data likelihood. Due to the gradual stochastic sampling process and explicit likelihood characterization, diffusion models can improve the reliability of the network mapping to offer enhanced sample quality and diversity. Given this potential, diffusion-based methods have recently been adopted for unimodal imaging tasks such as image reconstruction [45]–[49], unconditional image generation [50], and anomaly detection [51], [52]. Nevertheless, these methods are typically based on unconditional diffusion processes devised to process single-modality images. Furthermore, current methods often involve vanilla diffusion models that rely on a large number of inference steps for accurate image generation. This prolonged sampling process introduces computational challenges in adoption of diffusion models.

Here, we propose a novel adversarial diffusion model for improved efficiency and performance in medical image translation. Note that translation involves a nonlinear intensity mapping from source- onto target-modality images of a given subject's anatomy [11]. Since the underlying tissue structure is common between modalities, the source image contains critical information to constrain the structure depicted in the synthesized target image [10]. To improve anatomical accuracy in target images, SynDiff leverages a conditional diffusion process where high-quality structural information from actual source images are provided as conditioning input to reverse diffusion steps. For this purpose, a novel source-conditional adversarial projector is employed that provides efficient and accurate image sampling over few large diffusion steps. Note also that supervised training of reliable translation models requires access to paired source-target images acquired from a large number of subjects [23]. Yet, multi-modal imaging of a large cohort is challenging due to economic and time costs [21]. While a cycle-consistent architecture can be formed via bilateral diffusive processes between source and target modalities, relatively slow training and poor efficiency of regular diffusion models can limit performance [38]. To improve efficacy in unsupervised learning, SynDiff leverages a non-diffusive module within a novel cycle-consistent architecture to produce source-image estimates paired with target images in the training set. To our knowledge, SynDiff is the first adversarial diffusion model for medical image synthesis, and the first diffusion-based method for unsupervised medical image translation in the literature. Based on these unique advances, we provide the first demonstrations of unsupervised translation in multi-contrast MRI and multi-modal MRI-CT based on diffusion modeling.

Few recent studies have considered improvements on vanilla diffusion models with partially related aims to our proposed method. A study on natural image generation has used an adversarial diffusion model, DDGAN, to improve efficiency in reverse diffusion steps [53]. DDGAN is an unconditional diffusion model that generates random images starting from noise; and it uses an adversarial generator for reverse diffusion without guidance from a source image. In contrast, SynDiff is a conditional diffusion model that translates between source- and target-images of an anatomy. It uses a source-conditional adversarial projector for reverse diffusion to synthesize target images with anatomical correspondence to a guiding source image. Besides the diffusive module, SynDiff also embodies a non-diffusive module to permit unsupervised training. A study on unsupervised translation of natural images has proposed a non-adversarial diffusion model, UNIT-DDPM [54]. Based on the notion that source-target modalities share a latent space, UNIT-DDPM uses parallel diffusion processes to simultaneously generate samples for both modalities in a large number of reverse steps; and the noisy source-image samples drawn from the source diffusion process are used to condition the generation of target images in the target diffusion process. In contrast, SynDiff uses an adversarial projector for efficient sampling in few steps; and it leverages source-image estimates that are produced by a non-diffusive module to provide high-quality anatomical guidance for synthesis. A recent study has independently considered a conditional score-based method, UMM-CGSM, for imputation of missing sequences in a multi-contrast MRI protocol [55]. UMM-CGSM uses a non-adversarial model with relatively large number of inference steps; and it performs supervised training on paired datasets of source-target images. In contrast, SynDiff adopts an adversarial diffusion model for efficient sampling over few steps; and it can perform unsupervised learning.

TABLE I: Description of variables related to images, diffusion processes, networks and probability distributions. Throughout the manuscript, vectorial quantities are annotated in bold font.

Images	
$\boldsymbol{x}_0$	Actual target-image sample
$\boldsymbol{x}_t$	Noisy target-image sample at time step t
$\boldsymbol{x}_T$	Noisy target-image sample at time step T, (i.e., drawn from isotropic Gaussian distribution)
$\boldsymbol{y}$	Guiding source image
$\boldsymbol{x}_{t-k}$	Actual target-image sample at time step $t-k$
$\hat{\boldsymbol{x}}_{t-k}$	Synthesized target-image sample at time step $t-k$
$\boldsymbol{x}_0^A$, $\boldsymbol{x}_0^B$	Unpaired training images from modalities A and B
$\tilde{\boldsymbol{y}}^A, \tilde{\boldsymbol{y}}^B$	Source images estimated by non-diffusive module
$\breve{\boldsymbol{x}}_0^A, \breve{\boldsymbol{x}}_0^B$	Target images synthesized by non-diffusive module
$\hat{\boldsymbol{x}}_0^A, \hat{\boldsymbol{x}}_0^B$	Target images synthesized by diffusive module
Regular diffusion	
β_t	Noise variance for regular diffusion at time step t
$\boldsymbol{\epsilon}$	Standard normal random vector
$\boldsymbol{\mu}(\boldsymbol{x}_t, t), \boldsymbol{\Sigma}(\boldsymbol{x}_t, t)$	Network estimates for mean and covariance of the conditional distribution of $\boldsymbol{x}_{t-1}$ given $\boldsymbol{x}_t$
ψ_t	$1-\beta_t$
$\overline{\psi}_t$	$\prod_{r=[0,1,..,t]} \psi_r$
$\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$	Network estimate for the added noise at time step t
Adversarial diffusion	
k	Step size for fast diffusion
γ_t	Noise variance for fast diffusion at time step t
$\overline{\beta}_{\text{min}}, \overline{\beta}_{\text{max}}$	Parameters that control the progression of noise variance
$\boldsymbol{f}_i$	Feature maps in the ith subblock of the diffusive generator
$\boldsymbol{m}$	Learnable temporal embedding added onto feature maps to encode the time step t
α_t	$1-\gamma_t$
$\overline{\alpha}_t$	$\prod_{r=[0,k,..,t]} \alpha_r$
Networks	
$G_{\boldsymbol{\phi}}^A, D_{\boldsymbol{\phi}}^A$	Non-diffusive generator-discriminator pair for learning to estimate a source image $\tilde{\boldsymbol{y}}^A$ given $\boldsymbol{x}_0^B$
$G_{\boldsymbol{\phi}}^B, D_{\boldsymbol{\phi}}^B$	Non-diffusive generator-discriminator pair for learning to estimate a source image $\tilde{\boldsymbol{y}}^B$ given $\boldsymbol{x}_0^A$
$G_{\boldsymbol{\theta}}^A, D_{\boldsymbol{\theta}}^A$	Diffusive generator-discriminator pair for learning to synthesize a target image $\hat{\boldsymbol{x}}_{t-k}^A$ given $\boldsymbol{x}_t^A$
$G_{\boldsymbol{\theta}}^B, D_{\boldsymbol{\theta}}^B$	Diffusive generator-discriminator pair for learning to synthesize a target image $\hat{\boldsymbol{x}}_{t-k}^B$ given $\boldsymbol{x}_t^B$
Probability distributions	
$q(\boldsymbol{x}_0)$	Actual image distribution
$q(\boldsymbol{x}_t \vert \boldsymbol{x}_{t-1})$	Forward transition probability
$q(\boldsymbol{x}_{t-1} \vert \boldsymbol{x}_t)$	Reverse transition probability
$p_\theta(\boldsymbol{x}_{t-1} \vert \boldsymbol{x}_t)$	Network estimate for reverse transition probability
$q(\boldsymbol{x}_{t-k} \vert \boldsymbol{x}_t, \boldsymbol{y})$	Reverse transition probability for fast conditional diffusion with step size $k \gg 1$
$p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-k} \vert \boldsymbol{x}_t, \boldsymbol{y})$	Network estimate for reverse transition probability in fast conditional diffusion with step size $k \gg 1$

III. Theory

A. Denoising Diffusion Models

Regular diffusion models map between pure noise samples and actual images through a gradual process over T time steps (Fig. 1a). In the forward direction, a small amount of Gaussian noise is added repeatedly onto an input image $\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0)$ to obtain a sample $\boldsymbol{x}_T$ from an isotropic Gaussian distribution for sufficiently large T. Forward diffusion forms a Markov chain where the mapping from $\boldsymbol{x}_{t-1}$ to $\boldsymbol{x}_t$ and the respective forward transition probability are:

$$\boldsymbol{x}_t = \sqrt{1-\beta_t}\boldsymbol{x}_{t-1} + \sqrt{\beta_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \tag{1}$$

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

where β_t is noise variance, $\boldsymbol{\epsilon}$ is added noise, $\mathcal{N}$ is a Gaussian distribution, $\boldsymbol{I}$ is an identity covariance matrix. Reverse diffusion also forms a Markov chain from $\boldsymbol{x}_T$ onto $\boldsymbol{x}_0$, albeit each step aims to gradually denoise the samples. Under large T and small β_t, the reverse transition probability between $\boldsymbol{x}_{t-1}$ and $\boldsymbol{x}_t$ can be approximated as a Gaussian distribution [56], [57]:

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) := \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}(\boldsymbol{x}_t, t), \boldsymbol{\Sigma}(\boldsymbol{x}_t, t)) \tag{3}$$

Diffusion models typically operationalize each reverse diffusion step as mapping through a neural network that provides estimates for $\boldsymbol{\mu}$ and/or $\boldsymbol{\Sigma}$. Training is then performed by minimizing a variational bound on log-likelihood:

$$L_{vb} = \mathbb{E}_{q(\boldsymbol{x}_{0:T})}\left[log\,\frac{p_\theta(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right] \leq \mathbb{E}_{q(\boldsymbol{x}_0)}\left[log\,p_{\boldsymbol{\theta}}(\boldsymbol{x}_0)\right] \tag{4}$$

where $\mathbb{E}_q$ denotes expectation over q, $p_{\boldsymbol{\theta}}$ is the network parametrization of the joint distribution of input variables, $\boldsymbol{\theta}$ are network parameters, $\boldsymbol{x}_{0:T}$ denote the collection of image samples between time steps 0 and T, and $\boldsymbol{x}_{1:T}|\boldsymbol{x}_0$ denote image samples between time steps 1 and T conditioned on the sample at time step 0. The bound can be decomposed as:

$$\begin{aligned} L_{vb} \;\; &= \;\; log\,p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \\ &\quad - \sum_{t=1}^{T} KL(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)\,||\,p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)) \end{aligned} \tag{5}$$

where KL denotes Kullback-Leibler divergence, and $KL(q(\boldsymbol{x}_T|\boldsymbol{x}_0)\,||\,p(\boldsymbol{x}_T))$ is omitted as it does not depend on $\boldsymbol{\theta}$. A common parametrization omits $\boldsymbol{\Sigma}$ to focus on $\boldsymbol{\mu}$:

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) = \frac{1}{\sqrt{\psi_t}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\overline{\psi}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\right) \tag{6}$$

where $\psi_t = 1 - \beta_t$ and $\overline{\psi}_t = \prod_{r=[0,1,..,t]} \psi_r$. In Eq. 6, $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ can be derived if the network is used to estimate the added noise $\boldsymbol{\epsilon}$ by minimizing the following loss [58]:

$$L_{err} = \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\sqrt{\overline{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\epsilon}, t)\|_2^2\right] \tag{7}$$

where t, $\boldsymbol{x}_0$ and $\boldsymbol{\epsilon}$ are sampled from the discrete uniform distribution $\mathcal{U}(0,T)$, $q(\boldsymbol{x}_0)$ and $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$, respectively. During inference, reverse diffusion steps are performed starting from a random sample $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. For each step $t \in T...1$, $\boldsymbol{\mu}$ is derived using Eq. 6 based on the network estimate $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$, and $\boldsymbol{x}_{t-1}$ is sampled based on Eq. 3.

B. SynDiff

Here, we introduce a novel diffusion model for efficient, high-fidelity translation between source and target modalities of a given anatomy. SynDiff uses a diffusive module equipped with a source-conditional adversarial projector for fast and accurate reverse diffusion sampling (Fig. 1b). It also employs a non-diffusive module for estimating source images paired with corresponding target images, so as to enable unsupervised learning (Fig. 2). The adversarial diffusion process that forms the basis of the diffusive module, the network architecture, and the learning procedures for SynDiff are detailed below.

1) Adversarial Diffusion Process: Regular diffusion models prescribe relatively large T such that the step size is sufficiently small to satisfy the normality assumption in Eq. 3, but this limits efficiency in image generation. Here, we instead propose fast diffusion with the following forward steps:

$$\boldsymbol{x}_t = \sqrt{1-\gamma_t}\boldsymbol{x}_{t-k} + \sqrt{\gamma_t}\boldsymbol{\epsilon} \tag{8}$$

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-k}) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{1-\gamma_t}\boldsymbol{x}_{t-k}, \gamma_t\boldsymbol{I}\right) \tag{9}$$

where $k \gg 1$ is step size. The noise variance γ_t is set as:

$$\gamma_t = 1 - \mathrm{e}^{\overline{\beta}_{\text{min}}\frac{k}{T} - (\overline{\beta}_{\text{max}} - \overline{\beta}_{\text{min}})\frac{2tk-k^2}{2T^2}} \tag{10}$$

$\overline{\beta}_{\text{min}}$ and $\overline{\beta}_{\text{max}}$ control the progression of noise variance in an exponential schedule [59].

Guidance from a source image ($\boldsymbol{y}$) is available during medical image translation, so a conditional process is proposed in the reverse diffusion direction. Note that, for $k \gg 1$, there is no closed form expression for $q(\boldsymbol{x}_{t-k}|\boldsymbol{x}_t, \boldsymbol{y})$ and the normality assumption used to compute Eq. 4 breaks down [38]. Here we introduce a novel source-conditional adversarial projector to capture the complex transition probability $q(\boldsymbol{x}_{t-k}|\boldsymbol{x}_t, \boldsymbol{y})$ for large k in our conditional diffusion model, as inspired by a recent report on unconditional generation of natural images using adversarial learning to capture $q(\boldsymbol{x}_{t-k}|\boldsymbol{x}_t)$ [53]. In SynDiff, a conditional generator $G_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \boldsymbol{y}, t)$ performs gradual denoising in each reverse step to synthesize $\hat{\boldsymbol{x}}_{t-k} \sim p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-k}|\boldsymbol{x}_t, \boldsymbol{y})$. $G_{\boldsymbol{\theta}}$ receives the image pair ($\boldsymbol{x}_t$,$\boldsymbol{y}$) as a two-channel input, and it extracts intermediate feature maps $\boldsymbol{f}_i$ where $i \in [1, ..., N]$ is the subblock index in an encoder-decoder structure [59]. A learnable temporal embedding $\boldsymbol{m}$ is computed given t, and this embedding is added as a channel-specific bias term onto the feature maps in each subblock [59]: $\boldsymbol{f}_i' = \boldsymbol{f}_i + \boldsymbol{m}$. Meanwhile, a discriminator $D_{\boldsymbol{\theta}}(\{\hat{\boldsymbol{x}}_{t-k} \text{ or } \boldsymbol{x}_{t-k}\}, \boldsymbol{x}_t, t)$ distinguishes samples drawn from estimated versus true denoising distributions ($p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-k}|\boldsymbol{x}_t, \boldsymbol{y})$ vs. $q(\boldsymbol{x}_{t-k}|\boldsymbol{x}_t, \boldsymbol{y})$). D_θ receives either ($\boldsymbol{x}_t$,$\hat{\boldsymbol{x}}_{t-k}$) or ($\boldsymbol{x}_t$,$\boldsymbol{x}_{t-k}$) as a two-channel input. The temporal embedding $\boldsymbol{m}$ is also added as a bias term onto the feature maps across $D_{\boldsymbol{\theta}}$. A non-saturating adversarial loss is adopted for $G_{\boldsymbol{\theta}}$ [60]:

$$L_{G_{\boldsymbol{\theta}}} = \mathbb{E}_{t,q(\boldsymbol{x}_t|\boldsymbol{x}_0,\boldsymbol{y}),p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-k}|\boldsymbol{x}_t,\boldsymbol{y})}[-log(D_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_{t-k}))] \tag{11}$$

where $t \sim \mathcal{U}(\{0, k, ..., T\})$, and the discriminator arguments are abbreviated for brevity. $D_{\boldsymbol{\theta}}$ also adopts a non-saturating adversarial loss with gradient penalty [61]:

$$\begin{aligned} L_{D_{\boldsymbol{\theta}}} = \mathbb{E}_{t,q(\boldsymbol{x}_t|\boldsymbol{x}_0,\boldsymbol{y})}\Big[&\mathbb{E}_{q(\boldsymbol{x}_{t-k}|\boldsymbol{x}_t,\boldsymbol{y})}\left[-log(D_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-k}))\right] \\ &+\mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-k}|\boldsymbol{x}_t,\boldsymbol{y})}[-log(1 - D_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_{t-k}))] \\ &+\eta\mathbb{E}_{q(\boldsymbol{x}_{t-k}|\boldsymbol{x}_t,\boldsymbol{y})}\left\|\nabla_{\boldsymbol{x}_{t-k}}D_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-k})\right\|_2^2\Big] \end{aligned} \tag{12}$$

where η is the weight for the gradient penalty.

Evaluation of Eqs. 11-12 require sampling from $q(\boldsymbol{x}_{t-k}|\boldsymbol{x}_t, \boldsymbol{y})$ that is unknown. Yet, $\boldsymbol{x}_0$ and $\boldsymbol{y}$ are non-linearly related images of the same anatomy and $\boldsymbol{x}_t$ is conditionally independent of $\boldsymbol{y}$ given $\boldsymbol{x}_0$. Thus, if the non-linear mapping from a particular $\boldsymbol{y}$ onto $\boldsymbol{x}_0$ is injective (i.e., one-to-one), the reverse transition probability can be expressed as $q(\boldsymbol{x}_{t-k}|\boldsymbol{x}_t, \boldsymbol{x}_0, \boldsymbol{y}) = q(\boldsymbol{x}_{t-k}|\boldsymbol{x}_t, \boldsymbol{x}_0)$ [38]. Bayes' rule can then be used to express the denoising distribution in terms of

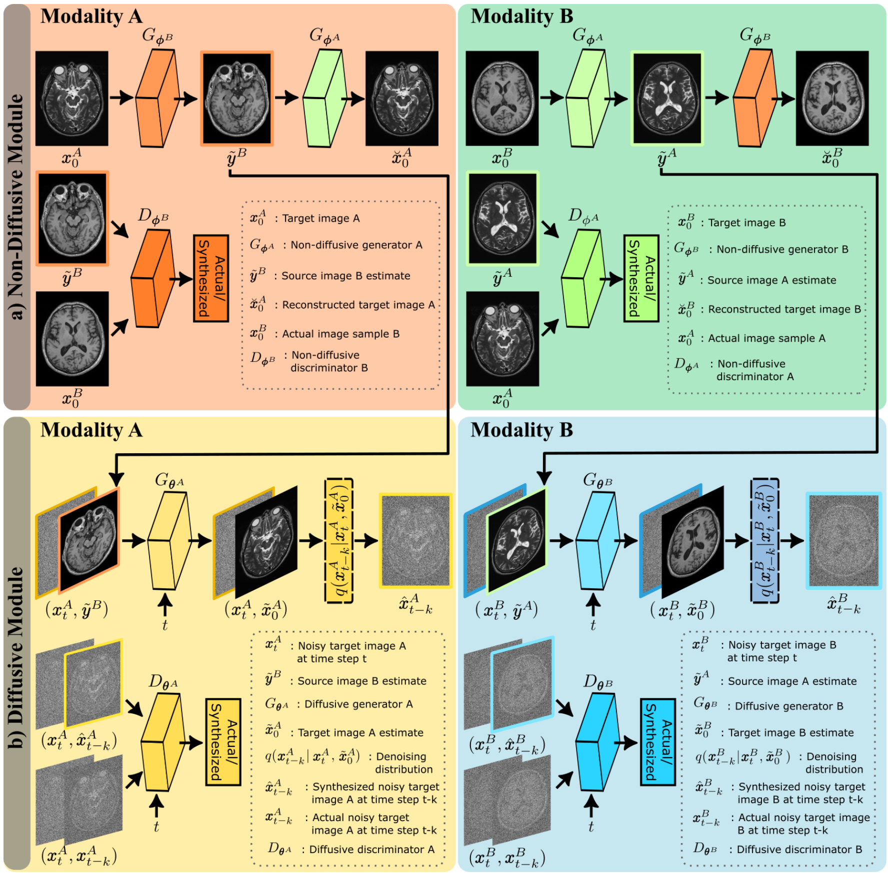
a) Non-Diffusive Module
Modality A
G_{ϕ^B}
G_{ϕ^A}
x_0^A
$\tilde{y}^B$
$\breve{x}_0^A$
D_{ϕ^B}
$\tilde{y}^B$
x_0^B
Actual/ Synthesized
x_0^A : Target image A
G_{ϕ^A} : Non-diffusive generator A
$\tilde{y}^B$: Source image B estimate
$\breve{x}_0^A$: Reconstructed target image A
x_0^B : Actual image sample B
D_{ϕ^B} : Non-diffusive discriminator B
Modality B
G_{ϕ^A}
G_{ϕ^B}
x_0^B
$\tilde{y}^A$
$\breve{x}_0^B$
D_{ϕ^A}
$\tilde{y}^A$
x_0^A
Actual/ Synthesized
x_0^B : Target image B
G_{ϕ^B} : Non-diffusive generator B
$\tilde{y}^A$: Source image A estimate
$\breve{x}_0^B$: Reconstructed target image B
x_0^A : Actual image sample A
D_{ϕ^A} : Non-diffusive discriminator A
b) Diffusive Module
Modality A
G_{θ^A}
$(x_t^A, \tilde{y}^B)$
t
$(x_t^A, \tilde{x}_0^A)$
$q(x_{t-k}^A|x_t^A, \tilde{x}_0^A)$
$\hat{x}_{t-k}^A$
D_{θ^A}
$(x_t^A, \hat{x}_{t-k}^A)$
(x_t^A, x_{t-k}^A)
Actual/ Synthesized
x_t^A : Noisy target image A at time step t
$\tilde{y}^B$: Source image B estimate
G_{θ^A} : Diffusive generator A
$\tilde{x}_0^A$: Target image A estimate
$q(x_{t-k}^A|x_t^A, \tilde{x}_0^A)$: Denoising distribution
$\hat{x}_{t-k}^A$: Synthesized noisy target image A at time step t-k
x_{t-k}^A : Actual noisy target image A at time step t-k
D_{θ^A} : Diffusive discriminator A
Modality B
G_{θ^B}
$(x_t^B, \tilde{y}^A)$
t
$(x_t^B, \tilde{x}_0^B)$
$q(x_{t-k}^B|x_t^B, \tilde{x}_0^B)$
$\hat{x}_{t-k}^B$
D_{θ^B}
$(x_t^B, \hat{x}_{t-k}^B)$
(x_t^B, x_{t-k}^B)
Actual/ Synthesized
x_t^B : Noisy target image B at time step t
$\tilde{y}^A$: Source image A estimate
G_{θ^B} : Diffusive generator B
$\tilde{x}_0^B$: Target image B estimate
$q(x_{t-k}^B|x_t^B, \tilde{x}_0^B)$: Denoising distribution
$\hat{x}_{t-k}^B$: Synthesized noisy target image B at time step t-k
x_{t-k}^B : Actual noisy target image B at time step t-k
D_{θ^B} : Diffusive discriminator B

Fig. 2: For unsupervised learning, SynDiff leverages a cycle-consistent architecture that bilaterally translates between two modalities (A, B). For synthesizing a target image $\hat{\boldsymbol{x}}_0^A$ of modality A, the diffusive module in Fig. 1b requires guidance from a source image $\boldsymbol{y}^B$ of modality B for the same anatomy. However, a paired source image of the same anatomy might be unavailable in the training set. To enable training on unpaired images, SynDiff uses a non-diffusive module to first estimate a paired source image $\tilde{\boldsymbol{y}}^B$ from $\boldsymbol{x}_0^A$. Similarly, for synthesizing a target image $\hat{\boldsymbol{x}}_0^B$ of modality B with the diffusive module, the non-diffusive module first estimates a paired source image $\tilde{\boldsymbol{y}}^A$ from $\boldsymbol{x}_0^B$. **a)** To do this, the non-diffusive module comprises two generator-discriminator pairs ($G_{\boldsymbol{\phi}^{A,B}}$, $D_{\boldsymbol{\phi}^{A,B}}$) that generate initial translation estimates for $\boldsymbol{x}_0^A \rightarrow \tilde{\boldsymbol{y}}^B$ (orange) and $\boldsymbol{x}_0^B \rightarrow \tilde{\boldsymbol{y}}^A$ (green). **b)** These initial translation estimates $\tilde{\boldsymbol{y}}^{A,B}$ are then used as guiding source-modality images in the diffusive module. For cycle-consistent learning, the diffusive module also comprises two generator-discriminator pairs ($G_{\boldsymbol{\theta}^{A,B}}$, $D_{\boldsymbol{\theta}^{A,B}}$) to generate denoised image estimates for $(\boldsymbol{x}_t^A, \tilde{\boldsymbol{y}}^B, t) \rightarrow \hat{\boldsymbol{x}}_{t-k}^A$ (yellow) and $(\boldsymbol{x}_t^B, \tilde{\boldsymbol{y}}^A, t) \rightarrow \hat{\boldsymbol{x}}_{t-k}^B$ (blue).

forward transition probabilities:

$$q(\boldsymbol{x}_{t-k}|\boldsymbol{x}_t, \boldsymbol{x}_0) = q(\boldsymbol{x}_t|\boldsymbol{x}_{t-k}, \boldsymbol{x}_0)\frac{q(\boldsymbol{x}_{t-k}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} \tag{13}$$

Using Eq. 8, it can then be shown that $q(\boldsymbol{x}_{t-k}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_{t-k}; \overline{\boldsymbol{\mu}}(\boldsymbol{x}_t, \boldsymbol{x}_0), \overline{\gamma}\boldsymbol{I})$ with the following parameters:

$$\overline{\boldsymbol{\mu}} = \frac{\sqrt{\overline{\alpha}_{t-k}}\gamma_t}{1-\overline{\alpha}_t}\boldsymbol{x}_0 + \frac{\sqrt{\alpha_t}\,(1-\overline{\alpha}_{t-k})}{1-\overline{\alpha}_t}\boldsymbol{x}_t, \overline{\gamma} = \frac{1-\overline{\alpha}_{t-k}}{1-\overline{\alpha}_t}\gamma_t \tag{14}$$

where $\alpha_t = 1 - \gamma_t$ and $\overline{\alpha}_t = \prod_{r=[0,k,..,t]} \alpha_r$.

Eqs. 11-12 also require sampling from the network-parameterized denoising distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-k}|\boldsymbol{x}_t, \boldsymbol{y})$. A trivial albeit deterministic sample would be the generator output, i.e. $\hat{\boldsymbol{x}}_{t-k} \sim \delta(\boldsymbol{x}_{t-k} - G_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \boldsymbol{y}, t))$. To maintain stochasticity, we instead operationalize the generator distribution as follows:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-k}|\boldsymbol{x}_t, \boldsymbol{y}) := q(\boldsymbol{x}_{t-k}|\boldsymbol{x}_t, \tilde{\boldsymbol{x}}_0 = G_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \boldsymbol{y}, t)) \tag{15}$$

where $G_{\boldsymbol{\theta}}$ predicts $\tilde{\boldsymbol{x}}_0$ that is t/k steps away from $\boldsymbol{x}_t$. Following a total of T/k reverse diffusion steps, the eventual denoised image will be obtained via sampling $\hat{\boldsymbol{x}}_0 \sim p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_k, \boldsymbol{y})$.

2) Network Architecture: To synthesize a target-modality image, the reverse diffusion steps parametrized in Eq.15 require guidance from a source-modality image of the same anatomy. However, the training set might include only unpaired images $\boldsymbol{x}_0^A$, $\boldsymbol{x}_0^B$ for the modalities A, B, respectively. To learn from unpaired training sets, we introduce a cycle-consistent architecture based on non-diffusive and diffusive modules that bilaterally translate between the two modalities.

Non-diffusive module. SynDiff leverages a non-diffusive module to estimate a source image paired with each target image in the training set. A source-image estimate $\tilde{\boldsymbol{y}}^B$ of modality B is produced given a target image $\boldsymbol{x}_0^A$ of modality A; and a source-image estimate $\tilde{\boldsymbol{y}}^A$ is produced given a target image $\boldsymbol{x}_0^B$. To do this, two generator-discriminator pairs ($G_{\boldsymbol{\phi}^A}$,$D_{\boldsymbol{\phi}^A}$) and ($G_{\boldsymbol{\phi}^B}$,$D_{\boldsymbol{\phi}^B}$) with parameters $\boldsymbol{\phi}^{A,B}$ are employed [21]. The generators produce the estimates $\tilde{\boldsymbol{y}}^{A,B}$ as:

$$\tilde{\boldsymbol{y}}^B = G_{\boldsymbol{\phi}^B}(\boldsymbol{x}_0^A)$$
$$\tilde{\boldsymbol{y}}^A = G_{\boldsymbol{\phi}^A}(\boldsymbol{x}_0^B) \tag{16}$$

A non-saturating adversarial loss is adopted for $G_{\boldsymbol{\phi}^{A,B}}$:

$$L_{G_{\boldsymbol{\phi}}} = \mathbb{E}_{p_{\boldsymbol{\phi}}(\boldsymbol{y}|\boldsymbol{x}_0)}[-log(D_{\boldsymbol{\phi}}(\tilde{\boldsymbol{y}}))] \tag{17}$$

where $p_{\boldsymbol{\phi}}(y|x_0)$ denotes the network parametrization of the conditional distribution of the source given the target image, and the conditioning input x_0 to the discriminator is omitted for brevity. Meanwhile, the discriminators distinguish samples of estimated versus true source images by adopting a non-saturating adversarial loss:

$$L_{D_{\boldsymbol{\phi}}} = \mathbb{E}_{q(\boldsymbol{y}|\boldsymbol{x}_0)}[-log(D_{\boldsymbol{\phi}}(\boldsymbol{y}))] + \mathbb{E}_{p_{\boldsymbol{\phi}}(\boldsymbol{y}|\boldsymbol{x}_0)}[-log(1 - D_{\boldsymbol{\phi}}(\tilde{\boldsymbol{y}}))] \tag{18}$$

where $q(\boldsymbol{y}|\boldsymbol{x}_0)$ is the true conditional distribution of the source given the target image. Note that, for $D_{\boldsymbol{\phi}^B}$, $\boldsymbol{y}$ corresponds to $\boldsymbol{x}_0^B$ and the conditioning input is $\boldsymbol{x}_0^A$; whereas for $D_{\boldsymbol{\phi}^A}$, $\boldsymbol{y}$ corresponds to $\boldsymbol{x}_0^A$ and the conditioning input is $\boldsymbol{x}_0^B$.

Diffusive module. SynDiff then leverages a diffusive module to synthesize target images given source-image estimates from the non-diffusive module as guidance. A synthetic target image $\hat{\boldsymbol{x}}^A$ is produced given $\tilde{\boldsymbol{y}}^B$; and a synthetic target image $\hat{\boldsymbol{x}}^B$ is produced given $\tilde{\boldsymbol{y}}^A$. To do this, two adversarial diffusion processes are used with respective generator-discriminator pairs ($G_{\boldsymbol{\theta}^A}$,$D_{\boldsymbol{\theta}^A}$) and ($G_{\boldsymbol{\theta}^B}$,$D_{\boldsymbol{\theta}^B}$) of parameters $\boldsymbol{\theta}^{A,B}$. Starting with Gaussian noise images $\boldsymbol{x}_T^{A,B}$ at time step T, target images are synthesized in T/k reverse diffusion steps. In each step, the generators first produce deterministic estimates of denoised target images as noted in Sec. III-B.1:

$$\begin{aligned}\tilde{\boldsymbol{x}}_0^A &= G_{\boldsymbol{\theta}^A}(\boldsymbol{x}_t^A, \boldsymbol{y} = \tilde{\boldsymbol{y}}^B, t)\\ \tilde{\boldsymbol{x}}_0^B &= G_{\boldsymbol{\theta}^B}(\boldsymbol{x}_t^B, \boldsymbol{y} = \tilde{\boldsymbol{y}}^A, t)\end{aligned} \tag{19}$$

Afterwards, the denoising distribution for each modality as described in Eq. 15 is used to synthesize target images:

$$\begin{aligned}\hat{\boldsymbol{x}}_{t-k}^A &\sim q(\boldsymbol{x}_{t-k}^A | \boldsymbol{x}_t^A, \tilde{\boldsymbol{x}}_0^A)\\ \hat{\boldsymbol{x}}_{t-k}^B &\sim q(\boldsymbol{x}_{t-k}^B | \boldsymbol{x}_t^B, \tilde{\boldsymbol{x}}_0^B)\end{aligned} \tag{20}$$

3) Learning Procedures: To achieve unsupervised learning, SynDiff leverages a cycle-consistency loss by comparing true target images against their reconstructions. In the diffusive module, reconstructions are taken as synthetic target images $\hat{\boldsymbol{x}}_0^{A,B}$. In the non-diffusive module, source-image estimates are projected to the target domain via the generators:

$$\begin{aligned}\breve{\boldsymbol{x}}_0^A &= G_{\boldsymbol{\phi}^A}(\tilde{\boldsymbol{y}}^B)\\ \breve{\boldsymbol{x}}_0^B &= G_{\boldsymbol{\phi}^B}(\tilde{\boldsymbol{y}}^A)\end{aligned} \tag{21}$$

where $\breve{\boldsymbol{x}}_0^{A,B}$ denote the corresponding reconstructions. Afterwards, the cycle-consistency loss is defined as:

$$\begin{aligned}L_{\text{cyc}} = \mathbb{E}_{t,q(\boldsymbol{x}_0^{A,B}),q(\boldsymbol{x}_t^{A,B}|\boldsymbol{x}_0^{A,B})} \Big[\lambda_{1\phi}(\|\boldsymbol{x}_0^A - \breve{\boldsymbol{x}}_0^A\|_1 +\\ \|\boldsymbol{x}_0^B - \breve{\boldsymbol{x}}_0^B\|_1) + \lambda_{1\theta}(\|\boldsymbol{x}_0^A - \hat{\boldsymbol{x}}_0^A\|_1 + \|\boldsymbol{x}_0^B - \hat{\boldsymbol{x}}_0^B\|_1)\Big]\end{aligned} \tag{22}$$

where $\lambda_{1\phi,1\theta}$ are the weights for cycle-consistency loss terms from the non-diffusive and diffusive modules respectively, and ℓ_1-norm of the difference between two images is taken as a consistency measure [21]. The diffusive and non-diffusive modules are trained jointly without any pretraining procedures. Accordingly, the overall generator loss is:

$$L_G^{\text{total}} = \lambda_{2\phi}(L_{G_{\boldsymbol{\phi}}^A} + L_{G_{\boldsymbol{\phi}}^B}) + \lambda_{2\theta}(L_{G_{\boldsymbol{\theta}}^A} + L_{G_{\boldsymbol{\theta}}^B}) + L_{\text{cyc}} \tag{23}$$

where $\lambda_{2\phi,2\theta}$ are the weights for adversarial loss terms from the non-diffusive and diffusive modules respectively, and for each modality $L_{G_{\boldsymbol{\phi}}}$ is defined as in Eq. 17 and $L_{G_{\boldsymbol{\theta}}}$ is defined as in Eq. 11. The overall discriminator loss is given as:

$$L_D^{\text{total}} = \lambda_{2\phi}(L_{D_{\boldsymbol{\phi}}^A} + L_{D_{\boldsymbol{\phi}}^B}) + \lambda_{2\theta}(L_{D_{\boldsymbol{\theta}}^A} + L_{D_{\boldsymbol{\theta}}^B}) \tag{24}$$

with $L_{D_{\boldsymbol{\phi}}}$ defined as in Eq. 18 and $L_{D_{\boldsymbol{\theta}}}$ defined as in Eq. 12.

During training, the non-diffusive module must be used to produce estimates of source images paired with given target images. During inference, however, the task is to synthesize an unacquired target image given the acquired source image of an anatomy, so only the respective generator within the diffusive module that performs the desired task is needed. For instance, to perform the mapping $A \rightarrow B$ (i.e., source→target), $G_{\boldsymbol{\theta}^B}(\boldsymbol{x}_t^B, \boldsymbol{y}^A, t)$ is used where $\boldsymbol{x}_t^B$ is the target-image sample of modality B at time step t and $\boldsymbol{y}^A$ is the acquired source image of modality A provided as input. Inference starts at time step T with a Gaussian noise sample $\boldsymbol{x}_T^B$ drawn from $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$, and the noisy target-image sample produced at the end of each reverse diffusion step is taken as the input target-image sample in the following step. A total of T/k reverse diffusion steps are taken as outlined in Eqs. 19-20 to attain $\hat{\boldsymbol{x}}_0^B$ at time step 0 as the synthetic target image.

IV. Methods

A. Datasets

We demonstrated SynDiff on two multi-contrast brain MRI datasets (IXI[1], BRATS [62]), and a multi-modal pelvic MRI-CT dataset [63]. In each dataset, a three-way split was performed to create training, validation and test sets with no subject overlap. While all unsupervised medical image translation models were trained on unpaired images, performance assessments necessitate the presence of paired and registered source-target volumes. Thus, in the validation and test sets, separate volumes of a given subject were spatially registered to enable calculation of quantitative metrics. Registrations were implemented in FSL via affine transformation and mutual information loss [64]. In each subject, each imaging volume was separately normalized to a mean intensity of 1. The maximum voxel intensity across subjects was then normalized to 1 to ensure an intensity range of [0,1]. Cross-sectional images were zero-padded as necessary to attain a consistent 256×256 image size in all datasets prior to modeling.

1) IXI Dataset: T_1-, T_2-, PD-weighted images from 40 healthy subjects were analyzed, with (25,5,10) subjects reserved for (training,validation,test). T_2 and PD volumes were registered onto T_1 volumes in validation/test sets. In each subject, 100 axial cross-sections with brain tissue were selected. Scan parameters were TE=4.6ms, TR=9.81ms for T_1; TE=100ms, TR=8178.34ms for T_2; TE=8ms, TR=8178.34ms for PD images; and a common spatial resolution=0.94×0.94×1.2mm^3.

2) BRATS Dataset: T_1-, T_2-, Fluid Attenuation Inversion Recovery (FLAIR) weighted brain MR images from 55 glioma patients were analyzed, with a (training, validation, test) split of (25,10,20) subjects. T_2 and FLAIR volumes were registered onto T_1 volumes in validation/test sets. In each subject, 100 axial cross-sections containing brain tissue were selected. Diverse scan protocols were used at multiple institutions.

3) Pelvic MRI-CT Dataset: Pelvic T_1-, T_2-weighted MRI, and CT images from 15 subjects were analyzed, with a (training, validation, test) split of (9,2,4) subjects. T_1 and CT volumes were registered onto T_2 volumes in validation/test sets. In each subject, 90 axial cross-sections were selected. For T_1 scans, TE=7.2ms, TR=500-600ms, 0.88×0.88×3mm^3 resolution, or TE=4.77ms, TR=7.46ms, 1.10×1.10×2mm^3 resolution were prescribed. For T_2 scans, TE=97ms, TR=6000-6600ms, 0.88×0.88×2.50mm^3 resolution, or TE=91-102ms, TR = 12000-16000ms, 0.88-1.10×0.88-1.10×2.50mm^3 resolution were prescribed. For CT scans, 0.10×0.10×3mm^3 resolution, Kernel=B30f, or 0.10×0.10×2mm^3 resolution, Kernel=FC17 were prescribed. To implement synthesis tasks from accelerated MRI scans [65], [66], fully-sampled MRI data were retrospectively undersampled 4-fold in two dimensions to attain low-resolution images at a 16x acceleration rate [65].

B. Competing Methods

We demonstrated SynDiff against several state-of-the-art non-attentional GAN, attentional GAN, and diffusion models.

[1] https://brain-development.org/ixi-dataset/

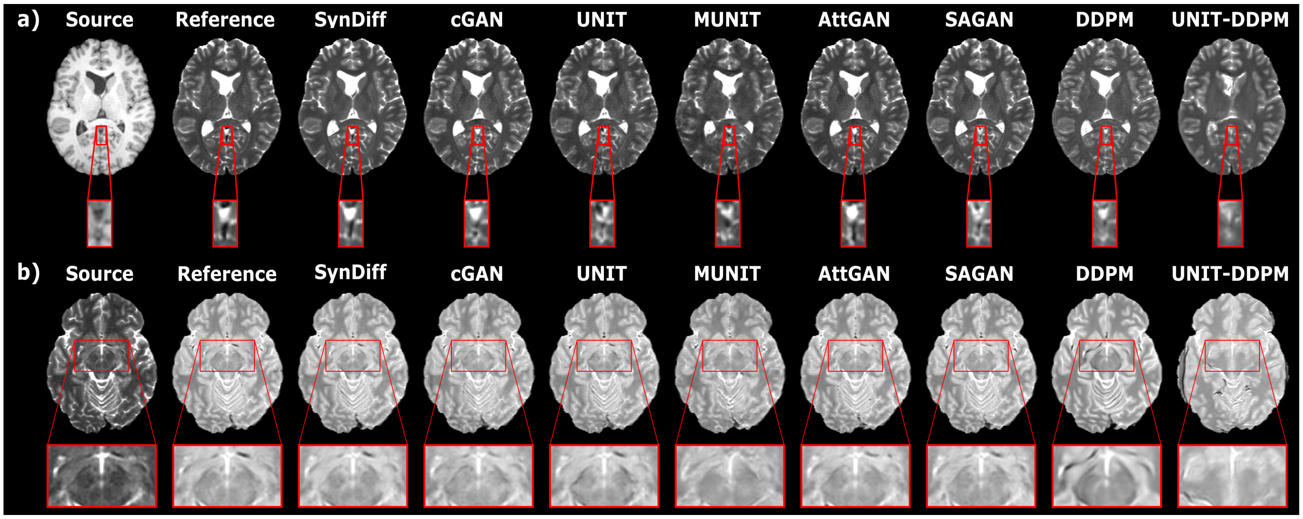

Fig. 3: SynDiff was demonstrated on IXI for translation between MRI contrasts. Synthesized images from competing methods are displayed along with the source and the ground-truth target (reference) images for representative a) $T_1 \rightarrow T_2$, b) $T_2 \rightarrow$PD tasks. Display windows of a) [0 0.65], b) [0 0.80] are used. Compared to baselines, SynDiff yields lower noise and artifacts, and maintains higher anatomical fidelity.

TABLE II: Performance for multi-contrast MRI translation tasks in IXI. PSNR (dB) and SSIM (%) are listed as mean±std across the test set. Boldface marks the top-performing model in each task.

	$T_2 \rightarrow T_1$		$T_1 \rightarrow T_2$		PD$\rightarrow T_1$		$T_1 \rightarrow$PD		PD$\rightarrow T_2$		$T_2 \rightarrow$PD	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SynDiff	**30.42±1.40**	**94.77±1.26**	**30.32±1.46**	**94.28±1.32**	**30.09±1.36**	**94.99±1.17**	**30.85±1.56**	**94.03±1.12**	**33.64±0.86**	**96.58±0.36**	**35.47±1.15**	**96.98±0.36**
cGAN	29.22±1.20	93.46±1.33	29.24±1.26	93.01±1.44	28.42±1.03	93.38±1.19	29.92±1.45	93.19±1.21	33.58±0.75	96.46±0.39	34.24±1.00	96.09±0.47
UNIT	29.17±1.15	93.54±1.34	28.34±0.98	92.02±1.45	28.10±0.99	92.97±1.21	29.29±1.08	92.36±1.24	32.57±0.65	96.22±0.38	34.74±1.07	96.66±0.39
MUNIT	26.35±0.88	89.78±1.78	26.61±0.86	88.28±1.90	25.99±0.89	89.73±1.70	27.59±1.02	88.71±1.60	29.17±0.71	92.01±1.01	29.80±0.82	91.61±1.00
AttGAN	29.27±1.38	93.74±1.36	28.37±1.08	92.21±1.45	28.02±1.07	92.83±1.19	29.65±1.42	92.98±1.23	32.15±0.67	95.93±0.44	35.11±1.11	96.76±0.40
SAGAN	28.85±1.26	93.38±1.40	29.01±1.32	92.87±1.43	27.93±1.22	93.04±1.29	29.58±1.51	92.76±1.25	32.44±0.71	95.91±0.46	34.75±0.83	96.64±0.38
DDPM	24.93±0.69	89.49±1.69	28.04±1.03	91.14±1.58	24.95±0.74	89.08±1.67	27.16±0.95	90.45±1.33	30.49±0.84	94.74±0.69	29.67±0.71	93.18±0.83
UNIT-DDPM	24.01±0.72	86.59±2.16	22.44±1.26	81.64±3.06	23.81±0.97	86.62±2.44	26.81±1.35	88.57±2.04	25.43±0.49	88.08±1.09	25.13±1.42	84.47±2.53

All competing methods performed unsupervised learning on unpaired source and target modalities. For each model, hyperparameter selection was performed to maximize performance on the validation set. A common set of parameters that offered near-optimal quantitative performance while maintaining high spatial acuity was selected across translation tasks. The selected parameters included number of training epochs, learning rate for the optimizer, and loss-term weightings for each model. Additionally, the step size was selected for diffusion models.

1) SynDiff: In the non-diffusive module, generators used a ResNet backbone with three encoding, six residual, and three decoding blocks [67]; and discriminators used six blocks with two convolutional layers followed by two-fold spatial downsampling. In the diffusive module, generators used a UNet backbone with six encoding and decoding blocks [68]. Each block had two residual subblocks followed by a convolutional layer. For encoding, the convolutional layer halved feature map resolution and channel dimensionality was doubled every other block. For decoding, the convolutional layer doubled resolution and channel dimensionality was halved every other block. Residual sublocks received a temporal embedding derived by projecting a 32-dimensional sinusoidal position encoding through a two-layer multi-layer perceptron (MLP) [59]. They also received 256-dimensional random latents from a three-layer MLP to modulate feature maps via adaptive normalization [69]. Discriminators used six blocks with two convolutional layers followed by two-fold downsampling, and the temporal embedding was added onto feature maps in each block. Cross-validated hyperparameters were: 50 epochs, 10^{-4} learning rate, μ=0.5, T=1000, a step size of k=250, and T/k=4 diffusion steps. Weights for cycle-consistency and adversarial loss terms were $\lambda_{1\phi,1\theta}$=0.5 and $\lambda_{2\phi,2\theta}$=1, respectively. Lower and upper bounds on the noise variance schedule were set according to $\overline{\beta}_{\text{min}}$=0.1, $\overline{\beta}_{\text{max}}$=20.

2) cGAN: A cycle-consistent GAN model was considered with architecture and loss functions adopted from [21]. cGAN comprised two generators with ResNet backbones, and two discriminators with a cascade of convolutional blocks followed by instance normalization. Cross-validated hyperparameters were 100 epochs, 2x10^{-4} learning rate linearly decayed to 0 in the last 50 epochs. Weights for cycle-consistency and adversarial losses were 100 and 1.

3) UNIT: An unsupervised GAN model that assumes a shared latent space between source-target modalities was considered, with architecture and loss functions adopted from [70]. UNIT comprised two discriminators and two translators with ResNet backbones in a cyclic setup. The translators contained parallel-connected domain image encoders and generators with a shared latent space. The discriminators contained a cascade of downsampling convolutional blocks. Cross-validated hyperparameters were 100 epochs, 10^{-4} learning

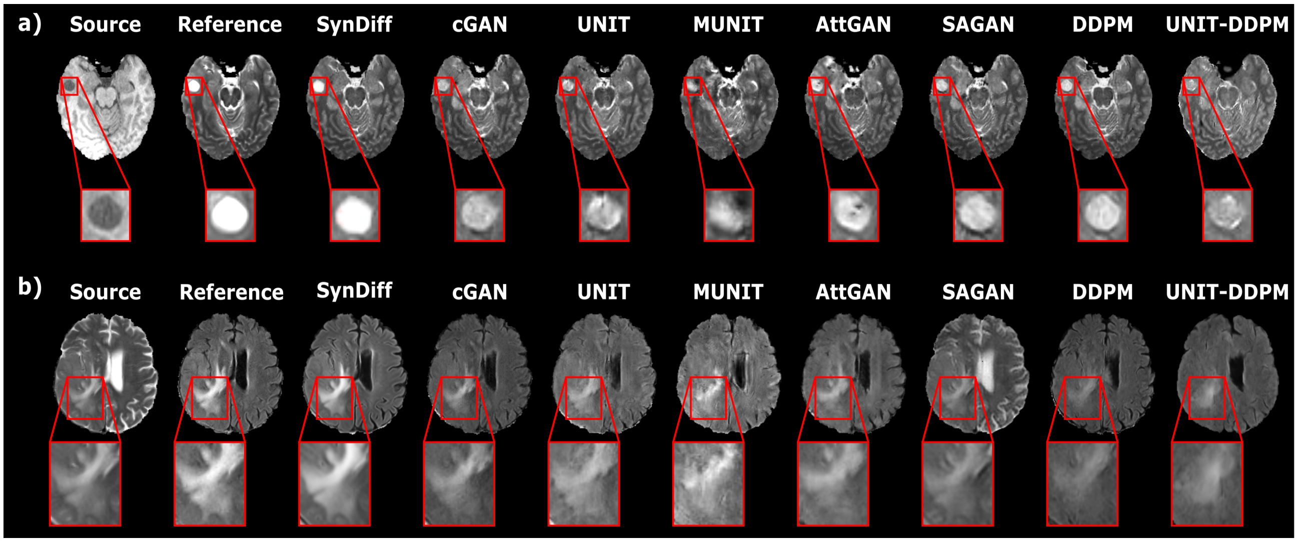

Fig. 4: SynDiff was demonstrated on BRATS for translation between MRI contrasts. Synthesized images are displayed along with the source and the ground-truth target (reference) images for representative a) $T_1 \rightarrow T_2$, b) $T_2 \rightarrow$FLAIR tasks. Display windows of a) [0 0.75], b) [0 0.80] are used. SynDiff lowers noise/artifact levels and more accurately depicts detailed structure compared to baselines.

TABLE III: Performance for multi-contrast MRI translation tasks in BRATS. PSNR (dB) and SSIM (%) listed as mean±std across the test set.

	$T_2 \rightarrow T_1$		$T_1 \rightarrow T_2$		FLAIR$\rightarrow T_1$		$T_1 \rightarrow$FLAIR		FLAIR$\rightarrow T_2$		$T_2 \rightarrow$FLAIR	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SynDiff	**28.90±0.73**	**93.79±0.95**	**27.10±1.26**	**92.35±1.27**	**26.47±0.69**	**89.37±1.50**	**26.45±1.02**	**87.79±1.67**	**26.75±1.18**	**91.69±1.50**	**28.17±0.90**	**90.44±1.48**
cGAN	27.41±0.45	92.07±0.92	27.00±1.11	91.90±1.13	26.35±0.77	89.03±1.51	26.44±0.73	85.98±1.51	25.99±1.30	90.02±1.67	27.41±0.78	88.48±1.45
UNIT	25.76±0.69	87.99±1.08	23.72±1.15	86.62±1.34	26.28±0.75	88.40±1.46	26.41±0.75	86.12±1.43	25.29±1.34	88.41±1.73	26.92±0.69	86.84±1.43
MUNIT	25.88±0.73	88.16±1.17	23.70±1.12	86.03±1.34	25.08±0.64	86.38±1.42	24.91±0.76	82.73±1.49	24.22±1.11	85.78±1.39	25.26±0.65	83.19±1.42
AttGAN	27.22±0.47	91.87±0.89	26.05±1.16	91.11±1.36	25.59±0.60	87.37±1.32	23.71±1.13	82.12±2.04	24.36±1.14	87.19±1.52	26.56±0.73	86.44±1.38
SAGAN	26.94±0.54	91.70±0.96	26.60±1.10	91.55±1.19	21.70±1.02	79.82±3.03	23.95±1.19	81.40±2.44	20.33±1.49	79.72±2.00	22.52±1.02	81.02±1.76
DDPM	27.36±0.58	91.94±0.96	26.34±1.17	91.50±1.27	23.41±0.64	81.55±2.43	24.49±1.12	82.12±1.97	21.23±1.50	82.38±2.45	25.49±0.60	84.71±1.40
UNIT-DDPM	19.84±1.54	85.92±2.28	23.71±1.50	88.75±2.49	20.31±0.84	79.30±2.08	21.33±1.18	81.80±1.99	20.03±1.61	77.21±2.03	24.15±1.03	82.07±1.84

rate. Weights for cycle-consistency, adversarial, reconstruction losses were 10, 1, and 10.

4) MUNIT: An unsupervised GAN model that assumes a shared content space albeit distinct style distributions for source-target modalities was considered, with architecture and loss functions adopted from [71]. MUNIT comprised pairs of discriminators, content encoders with ResNet backbones, MLP style encoders, and decoders with ResNet backbones. Cross-validated hyperparameters were 100 epochs, 10^{-4} learning rate. Weights for image, content, style reconstruction, adversarial losses were 10, 1, 1, and 1.

5) AttGAN: A cycle-consistent GAN model with attentional generators [72] was adopted for unsupervised translation. AttGAN comprised two convolutional attention UNet generators and two patch discriminators [72]. Cross-validated hyperparameters were 100 epochs, 2x10^{-4} learning rate linearly decayed to 0 in the last 50 epochs. Weights for cycle-consistency and adversarial losses were 100 and 1.

6) SAGAN: A cycle-consistent GAN model with self-attention generators [73] was adopted for unsupervised translation. SAGAN comprised two generators based on a ResNet backbone with self-attention layers in the last two residual blocks, and two patch discriminators [73]. Cross-validated hyperparameters were 100 epochs, 2x10^{-4} learning rate linearly decayed to 0 in the last 50 epochs. Weights for cycle-consistency and adversarial losses were 100 and 1.

7) DDPM: A recent diffusion model with improved sampling efficiency was considered, with architecture and loss functions adopted from [74]. The source modality was given as a conditioning input to reverse diffusion steps, and cycle-consistent learning was achieved by including non-diffusive modules as in SynDiff. Cross-validated hyperparameters were 50 epochs, 10^{-4} learning rate, T=1000, k=1, and 1000 diffusion steps. A cosine noise schedule was used as in [74]. Weight for cycle-consistency loss was 1.

8) UNIT-DDPM: A recent diffusion model allowing unsupervised training was considered, with architecture and loss functions adopted from [54]. UNIT-DDPM comprised two parallel diffusion processes for the source and target modalities, where noisy samples from each modality were given as conditioning input to reverse diffusion steps for the other modality. Cross-validated hyperparameters were 50 epochs, 10^{-4} learning rate, T=1000, k=1, and 1000 diffusion steps. A cosine noise schedule was used [74]. Weight for cycle-consistency loss was 1, and the release time was 1 as in [54].

C. Modeling Procedures

All models were implemented in Python using the PyTorch framework. Models were trained using Adam optimizer with β_1=0.5, β_2=0.9. Models were executed on a workstation equipped with Nvidia RTX 3090 GPUs. Model performance was evaluated on the test set within each dataset. For fair

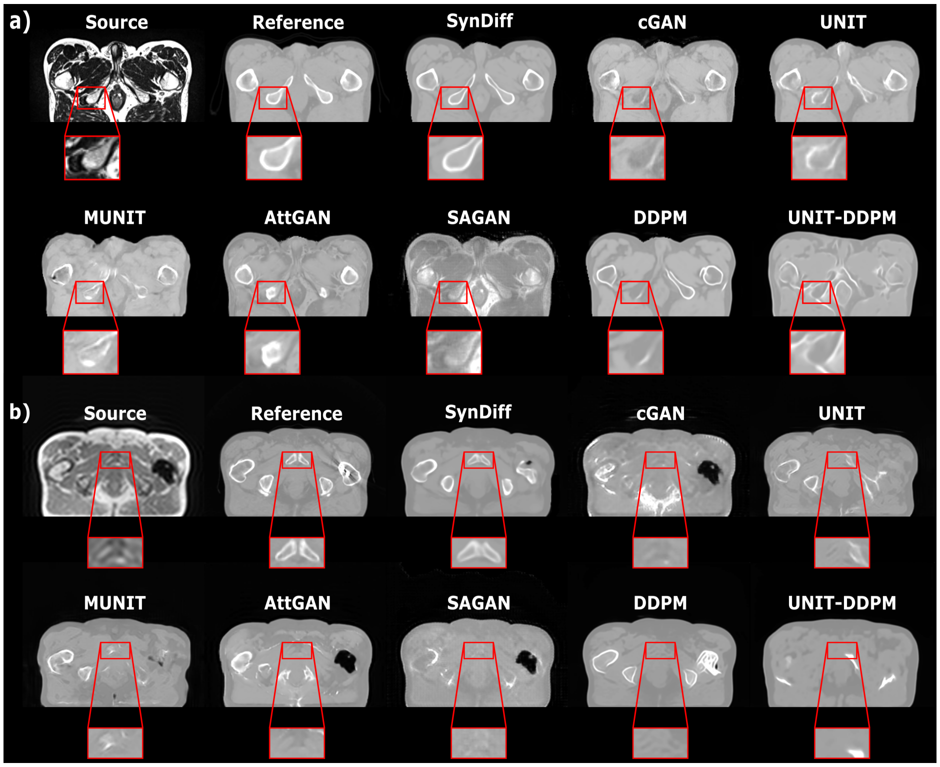

Fig. 5: SynDiff was demonstrated on the pelvic dataset for multi-modal MRI-CT translation. Synthesized images are displayed along with the source and the ground-truth target (reference) images for representative a) $T_2 \rightarrow$CT, b) accelerated $T_1 \rightarrow$CT tasks. Display windows of a) [-1000 1050] HU, and b) [-1000 1000] HU are used. Compared to diffusion and GAN baselines, SynDiff achieves lower artifact levels, and more accurately estimates anatomical structure near diagnostically-relevant regions.

comparison, evaluations of both deterministic and stochastic methods were performed based on a single target image synthesized at each cross section given the respective source image. Performance was assessed via peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) metrics in conditional synthesis tasks where a ground-truth reference is available. For unconditional tasks, Fréchet inception distance (FID) score was utilized to assess the perceptual quality of the generated random synthetic images by comparing their overall distribution to that of actual images. Prior to assessment, all images were normalized by their mean, and all examined images in a given cross-section were then normalized by the maximum intensity in the reference image. Significance of performance differences between competing methods were assessed via non-parametric Wilcoxon signed-rank tests ($p<0.05$).

V. Results

A. Multi-Contrast MRI Translation

We demonstrated SynDiff for unsupervised MRI contrast translation against state-of-the-art non-attentional GAN (cGAN, UNIT, MUNIT), attentional GAN (AttGAN, SAGAN), and regular diffusion (DDPM, UNIT-DDPM) models. First, experiments were performed on brain images from healthy subjects in IXI. Table II lists performance metrics for $T_2 \rightarrow T_1$, $T_1 \rightarrow T_2$, PD$\rightarrow T_1$, $T_1 \rightarrow$PD, PD$\rightarrow T_2$, and $T_2 \rightarrow$PD synthesis tasks. SynDiff yields the highest performance in all tasks ($p<0.05$), except for PD$\rightarrow T_2$ where cGAN performs similarly. On average, SynDiff outperforms non-attentional GANs by 2.2dB PSNR and 2.5% SSIM, attentional GANs by 1.4dB PSNR and 1.2% SSIM, and regular diffusion models by 5.7dB PSNR and 6.6% SSIM ($p<0.05$). Representative images are displayed in Fig. 3. GANs show noise or local inaccuracies in tissue contrast. Regular diffusion models suffer from a degree of spatial warping and blurring. UNIT-DDPM shows relatively lower anatomical accuracy, with occasional losses in tissue features. In comparison, SynDiff yields lower noise and artifacts, and higher accuracy in tissue depiction.

Next, experiments were conducted on brain images from

TABLE IV: Performance for multi-modal MRI-CT translation tasks in the pelvic dataset. PSNR (dB) and SSIM (%) listed as mean±std across the test set. 'acc.' stands for accelerated.

	$T_2 \rightarrow$CT		$T_1 \rightarrow$CT		acc. $T_2 \rightarrow$CT		acc. $T_1 \rightarrow$CT	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SynDiff	**26.86** **±0.51**	**87.94** **±2.53**	**25.16** **±1.53**	**86.02** **±2.05**	**26.71** **±0.63**	**87.32** **±2.84**	**25.47** **±1.09**	**85.00** **±2.10**
cGAN	25.07 ±0.17	84.91 ±1.84	24.11 ±1.00	77.81 ±1.84	21.24 ±0.51	69.62 ±0.85	20.35 ±0.32	64.73 ±1.47
UNIT	26.10 ±0.49	86.40 ±2.71	25.04 ±0.39	82.62 ±1.52	25.20 ±0.37	84.83 ±1.43	24.92 ±0.39	81.44 ±1.13
MUNIT	22.90 ±1.05	77.42 ±2.17	24.76 ±0.62	79.81 ±1.20	23.44 ±0.77	77.88 ±2.04	24.42 ±0.34	79.64 ±1.05
AttGAN	23.81 ±0.18	74.35 ±0.84	24.76 ±1.06	82.48 ±2.49	23.91 ±0.29	76.47 ±0.66	21.34 ±0.51	67.24 ±1.52
SAGAN	21.03 ±0.33	67.77 ±0.86	23.89 ±1.02	77.05 ±2.87	19.61 ±0.78	61.92 ±0.32	23.28 ±0.96	70.02 ±2.85
DDPM	24.66 ±0.19	83.24 ±2.62	24.92 ±0.81	82.63 ±3.64	24.35 ±0.47	83.25 ±1.70	24.62 ±0.59	83.04 ±2.40
UNIT-DDPM	21.49 ±0.72	80.23 ±2.69	20.26 ±1.17	76.79 ±1.37	21.89 ±0.77	77.69 ±3.06	21.45 ±0.23	77.10 ±2.83

TABLE V: Average training times per cross-section (sec), inference times per cross-section (sec) and memory load (gigabytes).

	SynDiff	cGAN	UNIT	MUNIT	AttGAN	SAGAN	DDPM	UNIT-DDPM
Training	2.35	0.14	0.26	0.24	0.47	0.22	1.34	1.16
Inference	0.182	0.060	0.041	0.040	0.083	0.076	85.773	52.225
Memory	2.12	0.77	1.31	2.26	0.86	1.08	2.95	2.77

glioma patients in BRATS. Table III lists performance metrics for $T_2 \rightarrow T_1$, $T_1 \rightarrow T_2$, FLAIR→T_1, T_1→FLAIR, FLAIR→T_2, and T_2→FLAIR tasks. SynDiff again achieves the highest synthesis performance in all tasks ($p<0.05$), except for cGAN that yields similar PSNR in T_1→FLAIR, and performs similarly in FLAIR→T_1. On average, SynDiff outperforms non-attentional GANs by 1.5dB PSNR and 3.5% SSIM, attentional GANs by 2.7dB PSNR and 5.0% SSIM, and diffusion models by 4.2dB PSNR and 6.8% SSIM ($p<0.05$). Representative images are displayed in Fig. 4. Non-attentional GANs show elevated noise and artifact levels. Attentional GANs occasionally suffer from leakage of contrast features from the source image (e.g., hallucination of regions with notably brighter or darker signal levels). Regular diffusion models show a degree of blurring and feature losses. Instead, SynDiff generates high-fidelity target images with low noise and artifacts.

B. Multi-Modal Translation

We also demonstrated SynDiff for unsupervised translation between separate modalities. In particular, experiments were performed using SynDiff, non-attentional GAN, attentional GAN, and regular diffusion models on the pelvic dataset for MRI-CT translation. Table IV lists performance metrics for T_2→CT, T_1→CT, accelerated T_2→CT, and accelerated T_1→CT synthesis tasks. SynDiff achieves the highest performance in all tasks ($p<0.05$). On average, SynDiff outperforms non-attentional GANs by 2.1dB PSNR and 7.6% SSIM, attentional GANs by 3.3dB PSNR and 14.4% SSIM, and diffusion models by 2.8dB PSNR and 6.1% SSIM ($p<0.05$). Representative images are displayed in Fig. 5. Non-attentional GANs and AttGAN show local contrast losses and artifacts, SAGAN suffers from contrast leakage, and regular diffusion models yield over-smoothing that can cause loss of fine features. While UNIT offers higher synthesis performance for some segments near tissue boundaries, particularly around the peripheral body-background boundary, SynDiff has generally higher performance across the image. Overall, SynDiff synthesizes target images with high anatomical fidelity. Note that the reference CT image in Fig. 5b has metal-induced streak artifacts that are generally absent from synthetic CT images. Implanted metals lead to reduced signal intensity in MRI, whereas they elicit streak artifacts in CT that diverge from regular tissue appearance. Since the training and validation subjects in the pelvic dataset did not carry any implants, the trained models learned to associate dark regions in T_1-weighted MR images with regular tissues that elicit low signal such as outer bone layers [51]. In turn, the trained models synthesize CT images with regular tissue appearance as opposed to artifacts near metal.

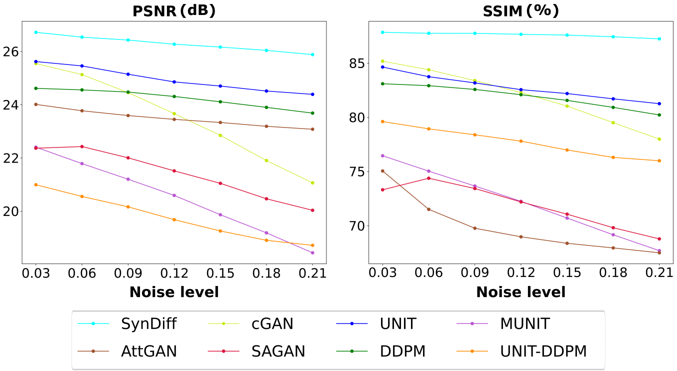

Fig. 6: Performance of competing methods as a function of added noise level on source-modality images. Results shown for the representative T_2→CT task in terms of PSNR (left), SSIM (right).

C. Model Complexity

A practical concern for medical image translation is the computational complexity of the applied models. Table V lists the training time, inference time and memory use of competing methods. As expected, one-shot GAN models have notably fast training and inference compared to diffusion models. While SynDiff has relatively comparable training times to other diffusion models, its fast diffusion process improves inference efficiency above two-orders-of-magnitude over DDPM and UNIT-DDPM. In terms of memory utilization, SynDiff has higher demand than cGAN, attentional GANs and UNIT, comparable demand to MUNIT, albeit notably lower demand than DDPM and UNIT-DDPM. Overall, SynDiff offers a more favorable compromise between image fidelity and computational complexity than regular diffusion models.

D. Image Variability

Image translation models involving random noise variables produce stochastic outputs, which can induce variability in target images independently synthesized for a given source image. To assess image variability, we examined target-image samples from competing stochastic methods, SynDiff, MUNIT, DDPM and UNIT-DDPM. For each task, a random selection of 50 cross sections was considered from the test set. For each cross section, 10 target-image samples were synthesized independently given the respective source image. Mean and standard deviation (std.) of performance metrics were computed across 10 samples. On average across cross sections, the std. across samples is less than 0.02dB in PSNR and 0.07% in SSIM for all methods, except for UNIT-DDPM with std. less than 0.27dB in PSNR and 0.31% in SSIM. Thus, all stochastic methods have minimal std. values relative to mean values, suggesting limited variability in synthesized target images.

E. Reliability against Noise

An important concern for translation methods is their reliability against distributional shifts in the noise level between training and test sets. To examine this issue, varying levels of noise were added onto source images in the test set for the multi-modal T_2→CT task. Zero-mean bivariate Gaussian

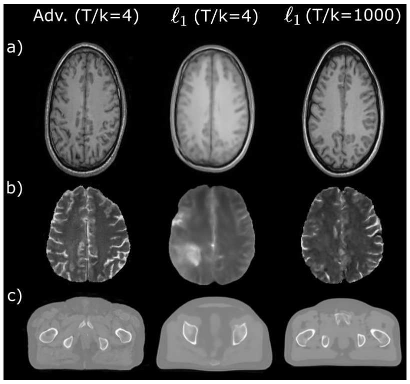

Fig. 7: The adversarial projector in SynDiff with T/k=4 steps was compared against a variant model using an ℓ_1-loss based projector with T/k=4 and T/k=1000. Image samples are shown for the unconditional synthesis tasks: a) T_1 in IXI, b) T_2 in BRATS and c) CT in pelvic datasets. Display windows of a) [0 0.90], b) [0 0.80] for MRI images, and c) [-1000 1300] HU for CT images are used.

TABLE VI: Performance of variant models in unconditional synthesis tasks. FID is listed across the training set.

	T_1(IXI)	T_2(BRATS)	CT(Pelvic)
Adv. proj. (T/k=4)	**30.75**	75.04	58.21
ℓ_1 proj. (T/k=4)	141.22	96.96	107.57
ℓ_1 proj. (T/k=1000)	52.78	**64.66**	**54.11**

white noise was added onto each cross-section at std. values ranging in [0.03 0.21] relative to the mean pixel intensity [49]. Fig. 6 plots performance of models trained on images without added noise when tested on noise-added images. Naturally, all methods show performance losses with increasing noise level. Compared with the performance on the original images, performance losses at the highest noise level (0.21 std) are 3.4dB PSNR, 7.6% SSIM for non-attentional GANs, 0.9dB PSNR, 3.9% SSIM for attentional GANs, and 1.9dB PSNR, 3.6% SSIM for regular diffusion models. In contrast, SynDiff shows relatively modest performance differences of 1.0dB PSNR, 0.7% SSIM. These results suggest that SynDiff maintains a degree of reliability against noise.

F. Ablation Studies

We conducted a set of ablation studies to systematically evaluate the importance of the main elements in SynDiff. To demonstrate the importance of the adversarial diffusion process, we compared the diffusive module in SynDiff based on an adversarial projector against a variant diffusive module based on an ℓ_1-loss based projector for reverse diffusion. The variant module shared the same overall loss function, albeit it ablated the adversarial loss terms for the diffusive generators and discriminators. As such, the remaining loss terms for the diffusive module were based on pixel-wise ℓ_1-loss similar to regular diffusion models. For focused assessment of the diffusive module, demonstrations were performed in unconditional synthesis tasks where guidance from the non-diffusive module was removed from all models. Synthetic images in representative tasks are displayed in Fig. 7, and FID scores are listed in Table VI. Compared to the ℓ_1 projector at T/k=4, the adversarial projector at T/k=4 substantially improves visual image quality and FID scores over the ℓ_1 projector at T/k=4, while performing competitively with the ℓ_1 projector at T/k=1000. These results demonstrate the utility of adversarial projections for efficient and accurate image sampling during reverse diffusion.

TABLE VII: Performance of variant models ablated of adversarial loss, cycle-consistency loss and the diffusive module. PSNR and SSIM listed as mean±std across the test set.

	PD→T_1		T_1→T_2		T_2→CT	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SynDiff	**30.09**	**94.99**	**27.10**	**92.35**	**26.86**	**87.94**
	±1.36	**±1.17**	**±1.26**	**±1.27**	**±0.51**	**±2.53**
w/o adv. loss	17.69	57.97	17.87	67.87	12.48	52.36
	±0.62	±3.45	±1.20	±2.30	±2.09	±3.76
w/o cyc. loss	26.18	91.88	24.70	89.84	22.81	76.84
	±0.73	±1.39	±1.51	±2.21	±0.30	±2.29
Non-diff. module	28.53	93.30	26.67	90.80	22.09	80.40
	±1.02	±1.20	±1.05	±1.21	±1.98	±0.32

TABLE VIII: Performance of variant models for varying number of steps T/k and varying loss-term weights ($\lambda_{1\phi}, \lambda_{1\theta}, \lambda_{2\phi}, \lambda_{2\theta}$). PSNR and SSIM listed as mean±std across the test set.

	PD→T_1		T_1→T_2		T_2→CT	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
T/k=2	29.47±1.35	94.46±1.24	27.11±1.31	92.48±1.27	26.97±0.53	87.76±2.56
T/k=4	30.09±1.36	94.99±1.17	27.10±1.26	92.35±1.27	26.86±0.51	87.94±2.53
T/k=8	29.83±1.28	94.85±1.18	27.24±1.24	92.42±1.24	26.95±0.59	87.95±2.78
$\lambda_{1\phi}$=0.25	30.04±1.33	94.96±1.14	26.67±1.20	91.89±1.26	26.81±0.56	87.69±2.66
$\lambda_{1\phi}$=0.5	30.09±1.36	94.99±1.17	27.10±1.26	92.35±1.27	26.86±0.51	87.94±2.53
$\lambda_{1\phi}$=1	30.11±1.29	94.97±1.14	27.33±1.19	92.72±1.17	27.07±0.63	88.22±2.74
$\lambda_{1\theta}$=0.25	29.85±1.23	94.81±1.14	27.18±1.21	92.28±1.20	27.09±0.58	88.01±2.81
$\lambda_{1\theta}$=0.5	30.09±1.36	94.99±1.17	27.10±1.26	92.35±1.27	26.86±0.51	87.94±2.53
$\lambda_{1\theta}$=1	30.06±1.32	94.98±1.15	27.90±1.25	92.03±1.27	27.06±0.44	88.23±2.52
$\lambda_{2\phi}$=0.5	30.12±1.29	95.04±1.14	27.82±1.20	93.12±1.18	26.80±0.45	88.10±2.46
$\lambda_{2\phi}$=1	30.09±1.36	94.99±1.17	27.10±1.26	92.35±1.27	26.86±0.51	87.94±2.53
$\lambda_{2\phi}$=2	29.66±1.22	94.69±1.15	26.99±1.24	91.82±1.34	26.72±0.53	87.69±2.64
$\lambda_{2\theta}$=0.5	29.97±1.22	94.91±1.14	27.29±1.24	92.59±1.22	26.76±0.50	87.34±2.55
$\lambda_{2\theta}$=1	30.09±1.36	94.99±1.17	27.10±1.26	92.35±1.27	26.86±0.51	87.94±2.53
$\lambda_{2\theta}$=2	29.60±1.21	94.77±1.14	27.20±1.26	92.48±1.24	26.83±0.48	87.97±2.55

We then examined the contributions of adversarial, cycle-consistent and diffusive learning in SynDiff. A first variant model was constructed by ablating adversarial loss; a second variant model was constructed by ablating cycle-consistency loss; and a third variant model was constructed by ablating the diffusive module to synthesize target images directly using the non-diffusive module. As listed in Table VII, SynDiff achieves substantially higher performance than all variants, indicating the importance of each learning strategy. We also assessed the test performance of SynDiff as a function of the number of diffusion steps (T/k), and as a function of weights that control the balance between separate loss terms ($\lambda_{1\phi}, \lambda_{1\theta}, \lambda_{2\phi}, \lambda_{2\theta}$). In each case, models were trained across a range of values centered around the parameters selected based on validation performance. As seen in Table VIII, there are generally minute differences in image quality among variants based on different parameter values. On average across tasks, we find less than 0.2dB PSNR, 0.2% SSIM difference between the selected and remaining T/k values, and less than 0.3dB PSNR, 0.4% SSIM difference between the selected and remaining loss-term

TABLE IX: Performance of variant models as mean±std across the test set. The non-diffusive module was pretrained in variant models. In pretrained-frozen, the non-diffusive module was not updated while training the diffusive module. In pretrained-trained, the non-diffusive module was also updated while training the diffusive module.

	PD→T$_1$		T$_1$→T$_2$		T$_2$→CT	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SynDiff	**30.09**	**94.99**	27.10	92.35	26.86	87.94
	±1.36	**±1.17**	±1.26	±1.27	±0.51	±2.53
Pretrained-frozen	29.19	94.22	27.23	92.50	26.77	87.65
	±1.28	±1.23	±1.30	±1.31	±0.82	±2.81
Pretrained-trained	29.24	94.24	**27.44**	**92.75**	**26.97**	**88.40**
	±1.17	±1.21	**±1.23**	**±1.24**	**±0.7**	**±2.74**

TABLE X: Performance of variant models as mean±std across the test set. In variant models, the non-diffusive module was only trained for n_{ND} epochs while the diffusive module was fully trained.

	PD→T$_1$		T$_1$→T$_2$		T$_2$→CT	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
$n_{ND} = 5$	19.88	70.31	25.26	89.99	22.59	74.59
	±0.60	±2.70	±1.08	±1.13	±0.39	±2.44
$n_{ND} = 10$	27.68	92.12	26.04	91.12	24.29	81.21
	±0.70	±1.17	±1.13	±1.26	±0.19	±2.71
$n_{ND} = 25$	29.51	94.45	26.47	91.41	26.03	86.14
	±1.14	±1.12	±1.33	±1.48	±0.35	±2.52
$n_{ND} = 50$	30.09	94.99	27.10	92.35	26.86	87.94
	±1.36	±1.17	±1.26	±1.27	±0.51	±2.53

weights. Overall, these results suggest that SynDiff shows a degree of reliability against parameter variations.

Next, we questioned whether SynDiff would benefit from pretraining of the non-diffusive module to improve stability. To address this question, SynDiff was compared against variant models that pretrained the non-diffusive module for 50 epochs to optimize its translation performance, and later combined the pretrained non-diffusive module with a randomly initialized diffusive module. A pretrained-frozen variant trained the combined model while the non-diffusive module was frozen. A pretrained-trained variant trained the combined model while both diffusive and non-diffusive modules were updated. As listed in Table IX, there are marginal performance changes between SynDiff and variants, with differences less than 0.3dB PSNR and 0.3% SSIM on average across tasks. This result suggests that the two modules can be jointly trained without notable stability issues. Compared to SynDiff, pretraining moderately reduces the performance of variant models in the easier PD→T$_1$ task, albeit it generally increases their performance in the relatively hard T$_1$→T$_2$ and T$_2$→CT tasks. To assess the underlying reason for this pattern, we compared the translation performance of non-diffusive modules in SynDiff versus the pretrained-trained variant. On average, pretraining yields 1.5dB lower PSNR, 2.6% lower SSIM for the non-diffusive module in PD→T$_1$, whereas it yields on average 1.3dB higher PSNR, 0.7% lower SSIM in remaining tasks. This finding suggests that pretraining the non-diffusive module can lead to overfitting in easier translation tasks, while it can help enhance performance in relatively hard tasks.

Finally, we assessed the dependence of the diffusive module on the quality of the source-image estimates provided by the non-diffusive module. For this purpose, we trained variant models in which the non-diffusive module was intentionally undertrained to produce suboptimal source-image estimates. Accordingly, the training of the non-diffusive module was stopped early by freezing its weights after a certain number of epochs (n_{ND}), while the training of the diffusive module was continued for the full 50 epochs. Table X lists performance of variant models across a range of n_{ND} values. Compared to SynDiff at n_{ND}=50, we find relatively modest performance differences of 0.7dB PSNR, 1.1% SSIM at n_{ND}=25, and more notable differences of 2.0dB PSNR, 3.6% SSIM starting at n_{ND}=10. These results indicate that while training of the diffusive module shows a degree of reliability against suboptimal source-image estimates, a well-functioning non-diffusive module is key for the performance of the diffusive module in unsupervised medical image translation.

VI. Discussion

A. Diffusion versus GAN Models

In unconstrained image generation tasks, regular diffusion models have been reported to offer benefits over GAN models that can suffer from limited training stability and sample diversity despite their high image quality [38], [74]. While SynDiff significantly outperforms all competing methods, here we observe that regular diffusion models such as DDPM are less competitive against GAN models in anatomically-constrained medical image translation, particularly in multi-contrast MRI tasks. Note that unconditional models for image generation are typically trained on large datasets with highly heterogeneous samples. In contrast, the conditional translation models considered here are trained on datasets of relatively limited size and heterogeneity [21], [24]. Furthermore, medical images carry higher intrinsic noise than natural images. This can limit the spatial acuity of regular diffusion models trained with pixel-wise losses that show lower sensitivity than adversarial losses to fine-grained features such as noise [21], [75]. Given these differences, benefits of diffusion models in terms of stability and sample diversity might be less discernible in medical image translation. Further work is warranted to systematically explore the relative performance of diffusion models against GANs as a function of the size, heterogeneity, and noise levels of medical imaging datasets.

Another difference between diffusion and GAN models for medical image translation concerns the variability of independent target images synthesized from a given source image. Both model classes draw samples from the conditional distribution of the target given the source modality, but the target images can be deterministic or stochastic depending on the use of random variables. Among competing methods, all GAN models receive only source images to produce deterministic images, except for MUNIT that receives random noise variables at intermediate stages. Meanwhile, all diffusion models produce stochastic images as they initiate sampling of target images from a random noise image. Here, we observed that all diffusion models including SynDiff show limited variability across independent target samples synthesized from the same source image, likely because the influence of the random noise image diminishes across diffusion steps. Still, future studies are warranted for an in-depth assessment of the variability of translation estimates and their utility in characterizing uncertainty in diffusion models.

B. Limitations

SynDiff is a diffusion-based method that adopts adversarial loss in its diffusive module to accelerate image sampling, and in its non-diffusive module to enable unsupervised training. In theory, these losses might introduce vulnerability against training instabilities, typically manifested as oscillatory patterns and suboptimal convergence in model performance [61]. To rule out this potential issue, we inspected the validation performance of SynDiff across training epochs. We do not find any notable sign of instabilities as model performance across epochs progresses smoothly towards a convergent point, without abrupt jumps (not reported). We also observe that pretraining of the non-diffusive module does not yield a notable benefit, suggesting that the joint training of diffusive and non-diffusive modules can be performed stably. In cases where instability is suspected during training of SynDiff, stabilization of adversarial components can be achieved via spectral normalization or feature matching [61].

The non-diffusive module in SynDiff computes source-image estimates paired with target images in the training set, and the diffusive module is trained based on these estimates. To assess the reliance of the diffusive module on the non-diffusive module, we systematically undertrained the non-diffusive module to produce suboptimal source-image estimates. Note that although the diffusive module was trained with low quality source-image estimates, it was still tested with acquired source images during inference. This creates discrepancy between the distribution of source-image inputs to the diffusive module between the training and test sets. While the diffusive module shows a degree of reliability against moderate discrepancies, its performance degrades under significant discrepancies towards more aggressive levels of undertraining. Thus, a well-functioning non-diffusive module is key for training of the diffusive module.

Here, high-quality images were synthesized while translating between MRI contrasts (e.g., T_1, T_2) and MRI to CT. Yet, we observed notably poor performance in CT-to-MRI translation for all examined methods (not reported). Note that CT primarily yields strong contrast for the dense outer bone layers based on X-ray attenuation, whereas MRI shows strong contrast among soft tissues and bone based on tissue magnetization. As such, the primary information on soft tissues needed to synthesize MRI images is scarcely present in CT images, resulting in a one-to-many mapping from CT to MRI and compromising model performance. In particular, SynDiff expresses the denoising distribution based on the theoretical assumption that the source-to-target mapping is an injective function, so accuracy of the computed reverse transition probabilities can be compromised during CT-to-MRI translation. For such ill-posed tasks, image quality might be improved by employing traditional or learning-based regularization priors on the target modality [23], [27], [76].

C. Future Work

Several technical developments can be pursued to improve SynDiff. Here, we considered synthesis tasks in which source and target modalities were unpaired across subjects. When paired source-target images are available, SynDiff can be adapted for supervised training by substituting a pixel-wise in place of cycle-consistency loss and providing actual source images as conditioning input [21], [77]. Performance improvements might also be viable by expanding the size of training datasets based on a collection of undersampled source- and target-modality acquisitions [66], or a combination of paired and unpaired source-target modality data [34].

Architectural developments might also help improve translation performance. The diffusive and non-diffusive generators in SynDiff were implemented based on convolutional backbones. Recent studies have reported that transformer-based architectures can improve contextual sensitivity in medical imaging tasks compared to convolutional architectures [35], [78]. The importance of contextual representations in implementing reverse diffusion steps remains to be demonstrated, yet attention mechanisms in transformers might help enhance the generalization performance to atypical anatomy [79].

Finally, developments on computational efficiency might be considered to further improve practicality. Unlike regular diffusion models with slow inference, SynDiff offers a more competitive inference time with GAN models. Yet, it may be possible to attain further speed benefits by combining the adversarial projector in SynDiff with alternative acceleration approaches, such as initiating sampling with an intermediate image [80] or running the diffusion process in a compact latent space [50]. SynDiff's training time is notably higher than GANs, and moderately longer than regular diffusion models due to the computation of added adversarial components and losses. When needed, training efficiency might be improved by parallel execution on multiple GPUs [53].

D. Potential Applications

A primary application of SynDiff is imputation of missing scans in multi-contrast MRI and multi-modal imaging. In clinical protocols, a subset of scans are typically omitted due to time constraints, or due to motion artifacts in uncooperative patients [21]. To maintain the original protocol, omitted scans can then be imputed from acquired scans. While successful results have been demonstrated here for mapping between native MRI contrasts and mapping MRI to CT, information required to synthesize the target image may not be sufficiently encoded in the source image in other cases. For instance, MRI contrasts enhanced with exogenous agents carry distinct information from native contrasts, so it is relatively difficult to synthesize contrast-enhanced MRI images from native MRI contrasts [25]. In such cases, translation performance can be improved by incorporating multiple source modalities that capture more diverse tissue information [26], [28], [29].

Another potential application for SynDiff is unsupervised adaptation of learning-based models for downstream tasks such as segmentation and classification across separate domains (e.g., scanners, imaging sites, modalities). When the amount of labeled data is limited in a primary domain, a model adequately trained in a secondary domain with a large labeled dataset might be transferred [81], [82]. However, blind model transfer will incur substantial performance loss given inherent shifts in the data distribution across domains. Assuming that a sufficiently large set of unlabeled images are available in the primary domain, SynDiff can be used to translate between primary and secondary domains [83]. Performance of the transferred model can improve when translated images are given as input, since their distribution is more closely aligned with secondary-domain images. That said, similar to the case of scan imputation, success in domain adaptation is bounded by the extent of information shared between domains. Downstream models can show suboptimal performance on translated images when information on the secondary domain is not sufficiently encoded in the primary domain.

VII. Conclusion

In this study, we introduced a novel adversarial diffusion model for medical image translation between source and target modalities. SynDiff leverages a fast diffusion process to efficiently synthesize target images, and a conditional adversarial projector for accurate reserve diffusion sampling. Unsupervised learning is achieved via a cycle-consistent architecture that embodies coupled diffusion processes between the two modalities. SynDiff achieves superior quality compared to state-of-the-art GAN and diffusion models, and it holds great promise for high-fidelity medical image translation. The fast conditional diffusion process in SynDiff might also offer performance benefits over GANs in other applications such as denoising and super-resolution [49], [84], [85].

References

[1] J. E. Iglesias, E. Konukoglu, D. Zikic, B. Glocker, K. Van Leemput, and B. Fischl, "Is synthesizing MRI contrast useful for inter-modality analysis?" in *Med Image Comput Comput Assist Interv*, 2013, pp. 631–638.

[2] J. Lee, A. Carass, A. Jog, C. Zhao, and J. Prince, "Multi-atlas-based CT synthesis from conventional MRI with patch-based refinement for MRI-based radiotherapy planning," in *SPIE Med Imag.*, vol. 10133, 2017, p. 101331I.

[3] 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 *Med Image Comput Comput Assist Interv*, 2013, pp. 606–613.

[4] T. Huynh *et al.*, "Estimating CT image from MRI data using structured random forest and auto-context model," *IEEE Trans Med Imag*, vol. 35, no. 1, pp. 174–183, 2016.

[5] A. Jog, A. Carass, S. Roy, D. L. Pham, and J. L. Prince, "Random forest regression for magnetic resonance image synthesis," *Med Image Anal*, vol. 35, pp. 475–488, 2017.

[6] T. Joyce, A. Chartsias, and S. A. Tsaftaris, "Robust multi-modal MR image synthesis," in *Med Image Comput Comput Assist Interv*, 2017, pp. 347–355.

[7] N. Cordier, H. Delingette, M. Le, and N. Ayache, "Extended modality propagation: Image synthesis of pathological cases," *IEEE Trans Med Imag*, vol. 35, pp. 2598–2608, 2016.

[8] Y. Wu *et al.*, "Prediction of CT substitutes from MR images based on local diffeomorphic mapping for brain PET attenuation correction," *J Nucl Med*, vol. 57, no. 10, pp. 1635–1641, 2016.

[9] C. Zhao, A. Carass, J. Lee, Y. He, and J. L. Prince, "Whole brain segmentation and labeling from CT using synthetic MR images," in *Mach Learn Med Imaging*, 2017, pp. 291–298.

[10] Y. Huang, L. Shao, and A. F. Frangi, "Cross-modality image synthesis via weakly coupled and geometry co-regularized joint dictionary learning," *IEEE Trans Med Imag*, vol. 37, no. 3, pp. 815–827, 2018.

[11] S. Roy, A. Jog, A. Carass, and J. L. Prince, "Atlas based intensity transformation of brain MR images," in *Multimodal Brain Image Anal.*, 2013, pp. 51–62.

[12] D. C. Alexander, D. Zikic, J. Zhang, H. Zhang, and A. Criminisi, "Image quality transfer via random forest regression: Applications in diffusion MRI," in *Med Image Comput Comput Assist Interv*, 2014, pp. 225–232.

[13] Y. Huang, L. Shao, and A. F. Frangi, "Simultaneous super-resolution and cross-modality synthesis of 3D medical images using weakly-supervised joint convolutional sparse coding," *Comput Vis Pattern Recognit*, pp. 5787–5796, 2017.

[14] H. Van Nguyen, K. Zhou, and R. Vemulapalli, "Cross-domain synthesis of medical images using efficient location-sensitive deep network," in *Med Image Comput Comput Assist Interv*, 2015, pp. 677–684.

[15] R. Vemulapalli, H. Van Nguyen, and S. K. Zhou, "Unsupervised cross-modal synthesis of subject-specific scans," in *Int Conf Comput Vis*, 2015, pp. 630–638.

[16] V. Sevetlidis, M. V. Giuffrida, and S. A. Tsaftaris, "Whole image synthesis using a deep encoder-decoder network," in *Simul Synth Med Imaging*, 2016, pp. 127–137.

[17] D. Nie, X. Cao, Y. Gao, L. Wang, and D. Shen, "Estimating CT image from MRI data using 3D fully convolutional networks," in *Deep Learn Data Label Med Appl*, 2016, pp. 170–178.

[18] C. Bowles *et al.*, "Pseudo-healthy image synthesis for white matter lesion segmentation," in *Simul Synth Med Imaging*, 2016, pp. 87–96.

[19] A. Chartsias, T. Joyce, M. V. Giuffrida, and S. A. Tsaftaris, "Multimodal MR synthesis via modality-invariant latent representation," *IEEE Trans Med Imag*, vol. 37, no. 3, pp. 803–814, 2018.

[20] W. Wei *et al.*, "Fluid-attenuated inversion recovery MRI synthesis from multisequence MRI using three-dimensional fully convolutional networks for multiple sclerosis," *J Med Imaging*, vol. 6, no. 1, p. 014005, 2019.

[21] 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 Trans Med Imaging*, vol. 38, no. 10, pp. 2375–2388, 2019.

[22] B. Yu, L. Zhou, L. Wang, J. Fripp, and P. Bourgeat, "3D cGAN based cross-modality MR image synthesis for brain tumor segmentation," *Int. Symp. Biomed. Imaging*, pp. 626–630, 2018.

[23] D. Nie *et al.*, "Medical image synthesis with deep convolutional adversarial networks," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 12, pp. 2720–2730, 2018.

[24] K. Armanious *et al.*, "MedGAN: Medical image translation using GANs," *Comput Med Imaging Grap*, vol. 79, p. 101684, 2019.

[25] D. Lee, J. Kim, W.-J. Moon, and J. C. Ye, "CollaGAN: Collaborative GAN for missing image data imputation," in *Comput Vis Pattern Recognit*, 2019, pp. 2487–2496.

[26] H. Li *et al.*, "DiamondGAN: Unified multi-modal generative adversarial networks for MRI sequences synthesis," in *Med. Image Comput Comput Assist Interv*, 2019, pp. 795–803.

[27] B. Yu, L. Zhou, L. Wang, Y. Shi, J. Fripp, and P. Bourgeat, "Ea-GANs: Edge-aware generative adversarial networks for cross-modality MR image synthesis," *IEEE Trans Med Imag*, vol. 38, no. 7, pp. 1750–1762, 2019.

[28] A. Sharma and G. Hamarneh, "Missing MRI pulse sequence synthesis using multi-modal generative adversarial network," *IEEE Trans Med Imag*, vol. 39, pp. 1170–1183, 2020.

[29] G. Wang *et al.*, "Synthesize high-quality multi-contrast magnetic resonance imaging from multi-echo acquisition using multi-task deep generative model," *IEEE Trans Med Imag*, vol. 39, no. 10, pp. 3089–3099, 2020.

[30] T. Zhou, H. Fu, G. Chen, J. Shen, and L. Shao, "Hi-Net: Hybrid-fusion network for multi-modal MR image synthesis," *IEEE Trans Med Imag*, vol. 39, no. 9, pp. 2772–2781, 2020.

[31] H. Lan, A. Toga, and F. Sepehrband, "SC-GAN: 3D self-attention conditional GAN with spectral normalization for multi-modal neuroimaging synthesis," *bioRxiv:2020.06.09.143297*, 2020.

[32] 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," *Med Image Anal*, vol. 70, p. 101944, 2021.

[33] H. Yang *et al.*, "Unpaired brain MR-to-CT synthesis using a structure-constrained cycleGAN," *arXiv:1809.04536*, 2018.

[34] C.-B. Jin *et al.*, "Deep CT to MR synthesis using paired and unpaired data," *Sensors*, vol. 19, no. 10, p. 2361, 2019.

[35] O. Dalmaz, M. Yurt, and T. Çukur, "ResViT: Residual vision transformers for multi-modal medical image synthesis," *IEEE Trans Med Imaging*, vol. 44, no. 10, pp. 2598–2614, 2022.

[36] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *Comput Vis Pattern Recognit*, pp. 1125–1134, 2017.

[37] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *Adv Neural Inf Process Syst*, vol. 34, 2021, pp. 8780–8794.

[38] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Adv Neural Inf Process Syst*, vol. 33, 2020, pp. 6840–6851.

[39] A. Chartsias, T. Joyce, R. Dharmakumar, and S. A. Tsaftaris, "Adversarial image synthesis for unpaired multi-modal cardiac data," in *Simul Synth Med Imaging*, 2017, pp. 3–13.

[40] J. M. Wolterink, A. M. Dinkla, M. H. F. Savenije, P. R. Seevinck, C. A. T. van den Berg, and I. Išgum, "Deep MR to CT synthesis using unpaired data," in *Simul Synth Med Imaging*, Cham, 2017, pp. 14–23.

[41] Y. Hiasa *et al.*, "Cross-modality image synthesis from unpaired data using cycleGAN: Effects of gradient consistency loss and training data size," in *Simul Synth Med Imaging*, 2018, pp. 31–41.

[42] M. Sohail, M. N. Riaz, J. Wu, C. Long, and S. Li, "Unpaired multi-contrast MR image synthesis using generative adversarial networks," in *Simul Synth Med Imaging*, Cham, 2019, pp. 22–31.

[43] Y. Ge *et al.*, "Unpaired MR to CT synthesis with explicit structural constrained adversarial learning," in *Int. Symp. Biomed. Imaging*, 2019, pp. 1096–1099.

[44] X. Dong *et al.*, "Synthetic CT generation from non-attenuation corrected PET images for whole-body PET imaging," *Phys Med Biol*, vol. 64, no. 21, p. 215016, 2019.

[45] A. Jalal, M. Arvinte, G. Daras, E. Price, A. G. Dimakis, and J. Tamir, "Robust compressed sensing mri with deep generative priors," in *Adv Neural Inf Process Syst*, vol. 34, 2021, pp. 14 938–14 954.

[46] H. Chung and J. C. Ye, "Score-based diffusion models for accelerated mri," *Med Image Anal*, vol. 80, p. 102479, 2022.

[47] Y. Song, L. Shen, L. Xing, and S. Ermon, "Solving inverse problems in medical imaging with score-based generative models," in *Int Conf Learn Represent*, 2022.

[48] A. Güngör *et al.*, "Adaptive diffusion priors for accelerated mri reconstruction," *Med Image Anal*, p. 102872, 2023. [Online]. Available: https://doi.org/10.1016/j.media.2023.102872

[49] H. Chung, E. S. Lee, and J. C. Ye, "Mr image denoising and super-resolution using regularized reverse diffusion," *arXiv:2203.12621*, 2022.

[50] W. H. L. Pinaya *et al.*, "Brain imaging generation with latent diffusion models," *arXiv:2209.07162*, 2022.

[51] J. Wolleb, F. Bieder, R. Sandkühler, and P. C. Cattin, "Diffusion models for medical anomaly detection," in *Med Image Comput Comput Assist Inter*, vol. 13438. Springer, 2022, pp. 35–45.

[52] W. H. L. Pinaya *et al.*, "Fast Unsupervised Brain Anomaly Detection and Segmentation with Diffusion Models," *arXiv:2206.03461*, 2022.

[53] Z. Xiao, K. Kreis, and A. Vahdat, "Tackling the generative learning trilemma with denoising diffusion GANs," in *Int Conf Learn Represent*, 2022.

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

[55] X. Meng *et al.*, "A novel unified conditional score-based generative framework for multi-modal medical image completion," *arXiv:2207.03430*, 2022.

[56] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Int Conf Mach Learn*, 2015, pp. 2256–2265.

[57] W. Feller, "On the theory of stochastic processes, with particular reference to applications," in *Proc Berkeley Symp Math Stat Probab*, vol. 1. University of California Press, 1949, pp. 403–433.

[58] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Adv Neural Inf Process Syst*, vol. 33, pp. 6840–6851, 2020.

[59] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv:2011.13456*, 2020.

[60] I. Goodfellow *et al.*, "Generative adversarial nets," *Adv Neural Inf Process Syst*, vol. 27, 2014.

[61] L. Mescheder, A. Geiger, and S. Nowozin, "Which training methods for gans do actually converge?" in *Int Conf Mach Learn*, 2018, pp. 3481–3490.

[62] B. H. Menze *et al.*, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans Med Imag*, vol. 34, no. 10, pp. 1993–2024, 2015.

[63] T. Nyholm *et al.*, "MR and CT data with multiobserver delineations of organs in the pelvic area—part of the gold atlas project," *Med Phys*, vol. 45, no. 3, pp. 1295–1300, 2018.

[64] M. Jenkinson and S. Smith, "A global optimisation method for robust affine registration of brain images," *Med Image Anal*, vol. 5, pp. 143–156, 2001.

[65] K. H. Kim, W.-J. Do, and S.-H. Park, "Improving resolution of mr images with an adversarial network incorporating images with different contrast," *Med Phys*, vol. 45, no. 7, pp. 3120–3131, 2018.

[66] M. Yurt *et al.*, "Semi-supervised learning of MRI synthesis without fully-sampled ground truths," *IEEE Trans Med Imaging*, vol. 41, no. 12, pp. 3895–3906, 2022.

[67] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf Comput Vis Pattern Recognit*, 2016, pp. 770–778.

[68] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Med Image Comput Comput Assist Inter*. Springer, 2015, pp. 234–241.

[69] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *IEEE Comp Vis Pattern Recognit*, 2019, pp. 4401–4410.

[70] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Adv Neural Inf Process Syst*, vol. 30, 2017.

[71] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *European Conf Comput Vis*, 2018, pp. 172–189.

[72] O. Oktay *et al.*, "Attention U-Net: Learning where to look for the pancreas," *arXiv:1804.03999*, 2018.

[73] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Int Conf Mach Learn*, vol. 97, 2019, pp. 7354–7363.

[74] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Int Conf Mach Learn*, 2021, pp. 8162–8171.

[75] T. Xiang, M. Yurt, A. B. Syed, K. Setsompop, and A. Chaudhari, "DDM^2: Self-Supervised Diffusion MRI Denoising with Generative Diffusion Models," *arXiv:2302.03018*, 2023.

[76] G. Elmas *et al.*, "Federated Learning of Generative Image Priors for MRI Reconstruction," *IEEE Trans Med Imaging*, pp. 1–1, 2022. [Online]. Available: https://doi.org/10.1109/TMI.2022.3220757

[77] Q. Lyu and G. Wang, "Conversion between CT and MRI images using diffusion and score-matching models," *arXiv:2209.12104*, 2022.

[78] Y. Luo *et al.*, "3D Transformer-GAN for high-quality PET reconstruction," in *Med Image Comput Comput Assist Interv*, 2021, pp. 276–285.

[79] Y. Korkmaz, S. U. H. Dar, M. Yurt, M. Ozbey, and T. Cukur, "Unsupervised MRI reconstruction via zero-shot learned adversarial transformers," *IEEE Trans Med Imaging*, vol. 41, no. 7, pp. 1747–1763, 2022.

[80] H. Chung, B. Sim, and J. C. Ye, "Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction," in *IEEE Conf Comput Vis Pattern Recognit*, 2022, pp. 12 413–12 422.

[81] F. Wu and X. Zhuang, "Unsupervised domain adaptation with variational approximation for cardiac segmentation," *IEEE Trans Med Imag*, vol. 40, no. 12, pp. 3555–3567, 2021.

[82] F. Wu, L. Li, and X. Zhuang, "Multi-modality cardiac segmentation via mixing domains for unsupervised adaptation," in *STACOM: Int Work Stat Atl Comp Mod Heart*, 2022, pp. 179–188.

[83] J. Gao, J. Zhang, X. Liu, T. Darrell, E. Shelhamer, and D. Wang, "Back to the source: Diffusion-driven test-time adaptation," *arXiv:2207.03442*, 2022.

[84] A. Gungor, B. Askin, D. A. Soydan, E. U. Saritas, C. B. Top, and T. Çukur, "TranSMS: Transformers for super-resolution calibration in magnetic particle imaging," *IEEE Trans Med Imaging*, vol. 41, no. 12, pp. 3562–3574, 2022.

[85] Q. Yang *et al.*, "Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss," *IEEE Trans Med Imag*, vol. 37, no. 6, pp. 1348–1357, 2018.