

NOVEL DEEP LEARNING ALGORITHMS FOR MULTI-MODAL MEDICAL IMAGE SYNTHESIS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
MASTER OF SCIENCE
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

By
Onat Dalmaz
August 2023

Novel Deep Learning Algorithms For Multi-Modal Medical Image
Synthesis

By Onat Dalmaz

August 2023

We certify that we have read this thesis and that in our opinion it is fully adequate,
in scope and in quality, as a thesis for the degree of Master of Science.

Tolga Çukur(Advisor)

Emine Ülkü Sarıtaş Çukur

Mehmet Erkut Erdem

Approved for the Graduate School of Engineering and Science:

Orhan Arıkan
Director of the Graduate School

ABSTRACT

NOVEL DEEP LEARNING ALGORITHMS FOR MULTI-MODAL MEDICAL IMAGE SYNTHESIS

Onat Dalmaz

M.S. in Electrical and Electronics Engineering

Advisor: Tolga Çukur

August 2023

Multi-modal medical imaging is a powerful tool for diagnosis and treatment of various diseases, as it provides complementary information about tissue morphology and function. However, acquiring multiple images from different modalities or contrasts is often impractical or impossible due to various factors such as scan time, cost, and patient comfort. Medical image translation has emerged as a promising solution to synthesize target-modality images given source-modality images. Ability to synthesize unavailable images enhance the ubiquity and utility of multi-modal protocols while decreasing examination costs and toxicity exposure such as ionizing radiation and contrast agents. Existing medical image translation methods prominently rely on generative adversarial networks (GANs) with convolutional neural networks (CNNs) backbones. CNNs are designed to perform local processing with compact filters, and this inductive bias is prone to limited contextual sensitivity. Meanwhile, GANs suffer from limited sample fidelity and diversity due to one-shot sampling and implicit characterization of the image distribution. To overcome the challenges with CNN based GAN models, in this thesis, first ResViT was introduced that leverages novel aggregated residual transformer (ART) blocks that synergistically fuse representations from convolutional and transformer modules. Then SynDiff is introduced, a conditional diffusion model that progressively maps noise and source images onto the target image via large diffusion steps and adversarial projections, capturing a direct correlate of the image distribution and improving sample quality and speed. ResViT provides a unified implementation to avoid the need to rebuild separate synthesis models for varying source-target modality configurations, whereas SynDiff enables unsupervised training on unpaired datasets via a cycle-consistent architecture. ResViT and SynDiff was demonstrated on synthesizing missing sequences in multi-contrast MRI, and CT images from MRI, and their state-of-the-art performance in medical image translation was shown.

Keywords: Multi-modal, Medical image synthesis, deep learning, transformer, diffusion models.



ÖZET

ÇOK-KIPLI TIBBİ GÖRÜNTÜ SENTEZİ İÇİN YENİ DERİN ÖĞRENME ALGORİTMALARI

Onat Dalmaz

Elektrik ve Elektronik Mühendisliği, Yüksek Lisans

Tez Danışmanı: Tolga Çukur

Haziran 2023

Çoklu kip tıbbi görüntüleme, doku morfolojisi ve işlevi hakkında tamamlayıcı bilgiler sağlayarak çeşitli hastalıkların tanı ve tedavisinde güçlü bir araçtır. Ancak, farklı kip veya kontrastlardan çoklu görüntüler elde etmek çoğu zaman tarama süresi, maliyet ve hasta konforu gibi çeşitli faktörler nedeniyle tatbiki olmayabilir, veya tümüyle imkansız olabilmektedir. Tıbbi görüntü çevirimi, kaynak kip görüntülerine dayanarak hedef kip görüntülerini sentezlemek için umut verici bir çözüm olarak ortaya çıkmıştır. Mevcut olmayan görüntüleri sentezleme yeteneği, çok kipli protokollerinin yaygınlığını ve kullanılabilirliğini artırırken muayene maliyetlerini ve iyonize radyasyon ve kontrast ajanları gibi toksisite maruziyetini azaltır. Mevcut tıbbi görüntü çevirimi yöntemleri öne çıkan şekilde evrişimli sinir ağları (CNN) omurgalı üretici çekişmeli ağlar (GAN) üzerine kurulmuştur. CNN'ler kompakt filtrelerle yerel işlem yapmak için tasarlanmıştır ve bu model varsayımı sınırlı bağlamsal duyarlılığa eğilimlidir. Bu arada, GAN'ler tek atışlı örnekleme ve görüntü dağılımının dolaylı karakterizasyonu nedeniyle sınırlı örnek sadakati ve çeşitliliğinden muzdariptir. CNN tabanlı GAN modellerinin zorluklarını aşmak için, bu tezde ilk olarak evrişimli ve dönüştürücü modüllerden sinerjik olarak temsilleri birleştiren yenilikçi birikimli artık dönüştürücü (ART) bloklarından yararlanan ResViT'i tanıtıyoruz. Ardından SynDiff'i tanıtıyoruz; büyük difüzyon adımları ve çekişmeli izdüşümler yoluyla gürültüyü ve kaynak görüntülerini hedef görüntünün üzerine ilerleyici bir şekilde eşleyen, görüntü dağılımının doğrudan bir bağıntısını yakalayan ve örnek kalitesini ve hızını iyileştiren koşullu bir difüzyon modeli. ResViT, değişen kaynak-hedef kip yapılandırmaları için ayrı sentez modelleri oluşturma ihtiyacını ortadan kaldırmak için birleştirilmiş bir uygulama sağlarken, SynDiff eşleştirilmemiş veri kümelerinde denetimsiz eğitimi döngüsel tutarlı bir mimari aracılığıyla mümkün kılar. ResViT ve SynDiff'i çok kontrastlı MRI'da eksik dizilerin sentezlenmesinde ve MRG'den BT görüntülerinin

sentezlenmesinde gösteriyoruz ve tıbbi görüntü çevirimindeki son teknoloji performanslarını gösteriyoruz.



Anahtar sözcükler: Çok-modlu, tıbbi görüntü sentezi, derin öğrenme, dönüştürücü, difüzyon modelleri.

Acknowledgement

First of all, I would like to express my deepest gratitude to my dear parents Belgin and Ali. They have a huge role in me choosing Bilkent for my undergrad. They always encouraged me to follow an academic path, which I truly enjoy and feel like I am reaching my full potential. I am really proud to be their child. They were always there for me in good times and bad times, helping and supporting me no matter what. Without them, I could not get even close to where I am right now. I love you guys. I would also like to give a special shout-out to my uncle, Dr. Koyuturk. Your encouragement and guidance have been a constant source of inspiration, and I could not have made it this far without you.

I thank my closest friend Baturay for being with me throughout the entire Bilkent process. Yet, this is not a goodbye and I know that we are going to find yasin ercire at sirinyer pazaryeri someday. Even though you're going to be in Connecticut and I will be in California, we're still in the same country so I know that our friendship and research collaboration will not end sometime soon. Also, since you're going earlier, prepare for a lot of "yardım reyiz" that I am going to do to you. Kaan, I am never going to forget our project deadlines, how we completed those 1 minute "before" the deadlines, "reds", deep conversations, and always supporting each other no matter what. I love you and am going to feel your absence. Hande, who am I going to tell all my crazy stories at No29? Who is going to prepare us dinner and play some movies while we cry? OK, if I continue I am going to cry so let's stop. Dede, aka the leader of Foca crew, I still want our deep conversations about almost every aspect of life during "before"s, and also our weekly calls as well. Let's meet at Belfast during some random time, and you say the keyword and piss me off. I think you might be the greatest person that ever lived. And lastly, Begiüs. You were with me throughout most of my Master's journey and supported me whenever I need you the most no matter what. I can always act as a little child with you since you adore kids. I know that you're not gonna make yourself missed since we will do christmas in London, and I will be coming with a custom Onat delivery.

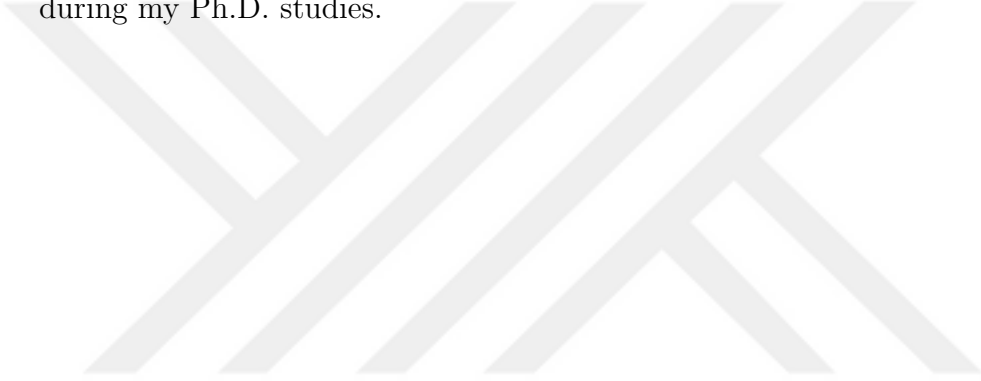
Sanem, I know that we have not talked very frequently throughout last year, yet I am pretty sure that we could keep up easily if we ever meet again. You are the one that can make me laugh without a slight effort, with merely some magical words: "ERU, ezo, bizim oglan". You are with me no matter what and made me feel your unconditional support. You were my comfort zone that I can completely be myself. I am pretty sure that our paths will converge some day.

I would like to follow the tradition at ICON and express my gratitude to all my labmates. Upon joining the lab, I learned everything from scratch by asking questions to Salman and Mahmut, and sometimes irritating them. They always patiently answered my dumbest questions. But don't think that you can run away Mahmut, I am coming to Stanford this Fall. Muzaffer, I will remember two thing for sure: how we were writing papers/dealing with revisions in the middle of the summer while everyone was on vacation. I couldn't do it without our deep conversations. Also, trying to produce results when we have 1 hours before the meeting with Dr. Tolga. Gökberk and Atakan, will never forget Toronto ISMRM with you guys. I believe you guys will get offers from excellent places next year. Lastly, Irmak and I always got along very well, yet I would never expect us to become roommates for the upcoming Autumn quarter.

I would also like to thank my jury members. I met with Dr. Sarıtaş during my earlier years at Bilkent, and she has always been very polite and supportive towards me. Thanks to her courses in undergrad and Master's, I have a strong grasp of fundamental concepts in medical imaging, and electrical engineering in a broader sense. I thank Prof. Aykut Erdem for our collaborations during my first year of Master's. Even though that project did not end up as a journal article, it thought me a lot of about machine learning. I sincerely thank you for your precious time that you dedicate to this thesis as juries.

Lastly, I would like to thank my advisor, Dr. Çukur, for his invaluable guidance and support during my Master's studies. He has been a constant source of inspiration and encouragement for me, and has helped me overcome many challenges and difficulties along the way. His mentorship has played an instrumental role in shaping my academic and professional growth, and I have learned a lot from his

expertise and experience in medical imaging. He has always been generous with his time and resources, and has provided me with constructive feedback and useful advice. I also appreciate his relentless pursuit of excellence in research, which has inspired me and influenced me to become a passionate researcher. When I first started my studies, writing a paper seem impossible to me. After 3 years, he simply rendered writing papers easy and simple for me. He has been the best advisor I could have asked for in my Master's studies, and I will miss him dearly during my Ph.D. studies.



Contents

1	Introduction	1
1.1	Multi-modal Medical Imaging	1
1.2	Contributions	2
1.3	Outline	4
2	Fundamentals of Multi-Modal Imaging	5
2.1	Basics of MRI Contrast Mechanisms	5
2.2	Basics of CT Imaging	7
2.3	Multi-Modal MRI-CT	8
3	ResViT: Residual vision transformers for multi-modal medical image synthesis	11
3.1	Introduction	12
3.2	Related Work	14
3.3	Theory and Methods	18

3.3.1	Residual Vision Transformers	18
3.3.2	Datasets	25
3.3.3	Competing Methods	26
3.3.4	Architectural Details	29
3.3.5	Modeling Procedures	31
3.3.6	Experiments	32
3.4	Results	37
3.4.1	Multi-Contrast MRI Synthesis	37
3.4.2	Across-Modality Synthesis	41
3.4.3	Ablation Studies	41
3.5	Discussion	47
4	Unsupervised Medical Image Translation with Adversarial Diffusion Models	50
4.1	Introduction	51
4.2	Related Work	54
4.3	Theory	58
4.3.1	Denoising Diffusion Models	58
4.3.2	SynDiff	59
4.4	Methods	66

4.4.1	Datasets	66
4.4.2	Competing Methods	67
4.4.3	Modeling Procedures	72
4.5	Results	74
4.5.1	Multi-Contrast MRI Translation	74
4.5.2	Multi-Modal Translation	76
4.5.3	Model Complexity	76
4.5.4	Image Variability	77
4.5.5	Reliability against Noise	77
4.5.6	Ablation Studies	79
4.6	Discussion	83
4.6.1	Diffusion versus GAN Models	83
4.6.2	Limitations	85
4.6.3	Future Work	86
4.6.4	Potential Applications	87
5	Conclusion	89

List of Figures

2.1	Sample 2D cross sections of T_1 , T_2 -weighted MR and CT images of a male pelvis.	9
2.2	Synthesis methods can reduce the radiation exposure and contrast agent injection for the patient by imputing CT scans or contrast-enhanced MRI scans with synthetic images.	10
3.1	The generator in ResViT follows an encoder-decoder architecture bridged with a central information bottleneck to distill task-specific information. The encoder and decoder comprise convolutional layers to maintain local precision and inductive bias in learned structural representations. Meanwhile, the information bottleneck comprises a stack of novel aggregated residual transformer (ART) blocks. ART blocks learn contextual representations via vision transformers, and synergistically fuse CNN-based local and transformer-based global representations.	16

3.2	ResViT is a conditional image synthesis model that can unify various source-target modality configurations into a single model for improved practicality. a) During training, ResViT takes as input the entire set of images within the multi-modal protocol, including both source and target modalities. For model consolidation across multiple synthesis tasks, various configurations of source-target modalities are expressed in terms of availability conditions in ResViT. b) During inference, the specific source-target configuration is determined via the availability conditions in each given test subject.	20
3.3	ResViT was demonstrated on the IXI dataset for two representative many-to-one synthesis tasks: a) $T_1, T_2 \rightarrow PD$, b) $T_2, PD \rightarrow T_1$. Synthesized images from all competing methods are shown along with the source images and the reference target image. ResViT improves synthesis performance in regions that are depicted sub-optimally in competing methods. Overall, ResViT generates images with lower artifact and noise levels and sharper tissue depiction.	33
3.4	ResViT was demonstrated on the BRATS dataset for two representative many-to-one synthesis tasks: a) $T_1, T_2 \rightarrow FLAIR$, b) $T_2, FLAIR \rightarrow T_1$. Synthesized images from all competing methods are shown along with the source images and the reference image. ResViT improves synthesis performance, especially in pathological regions (e.g., tumors, lesions) in comparison to competing methods. Overall, ResViT images have better-delineated tissue boundaries and lower artifact/noise levels.	36

- 3.5 ResViT_{uni} was demonstrated against other unified models on brain MRI datasets for two representative tasks: a) $T_1, PD \rightarrow T_2$ in IXI, b) $T_1, FLAIR \rightarrow T_2$ in BRATS. Synthesized images from all competing methods are shown along with the source images and the reference target image. ResViT_{uni} improves synthesis performance especially in pathological regions (tumors, lesions) in comparison to competing methods. Overall, ResViT_{uni} generates images with lower artifact and noise levels and more accurate tissue depiction for tasks in both datasets. 38
- 3.6 ResViT was demonstrated on the pelvic MRI-CT dataset for the T_2 -weighted MRI $\rightarrow$ CT task. Synthesized images from all competing methods are shown along with the source and reference images. ResViT enhances synthesis of relevant morphology in the CT domain as evidenced by the elevated accuracy near bone structures. 41
- 3.7 a) Representative synthesis results from ResViT and pGAN are shown along with the reference images for $T_2, FLAIR \rightarrow T_1$, $T_1, FLAIR \rightarrow T_2$, and $T_1, T_2 \rightarrow FLAIR$ in BRATS; and MRI $\rightarrow$ CT in the pelvic dataset. Error maps between the synthetic and reference images for each method are displayed, along with the attention map within the first transformer module of ResViT. Here, the attention maps were overlaid onto the reference image for improved visualization. Attention maps focus on image regions where ResViT substantially reduces synthesis errors compared to pGAN. b) ResViT was demonstrated against an ablated variant obtained by removing the adversarial term from the loss function (i.e., ablating the discriminator subnetwork). Representative results are shown for $T_1, T_2 \rightarrow PD$ in IXI, $T_1, T_2 \rightarrow FLAIR$ in BRATS, and MRI $\rightarrow$ CT in the pelvic dataset. Synthesized images from the models are shown along with the reference images. Adversarial loss improves the acuity of synthesized images. . . . 46

- 4.1 **a)** Regular diffusion models gradually transform between actual image samples for the target modality ($\mathbf{x}_0$) and isotropic Gaussian noise ($\mathbf{x}_T$) 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 $q(\mathbf{x}_{t+1}|\mathbf{x}_t)$. Each reverse step (left arrows) suppresses the added noise to create a denoised sample. For image translation, a source modality ($\mathbf{y}$) can also be provided as conditioning input to the reverse steps resulting in a reverse transition probability $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{y})$ assumed to be Gaussian, and operationalized via a neural network $p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{y})$ that estimates its mean. **b)** The proposed adversarial diffusion model performs fast transformation between $\mathbf{x}_0$ and $\mathbf{x}_T$ 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 $q(\mathbf{x}_{t-k}|\mathbf{x}_t, \mathbf{y})$. To improve accuracy, reverse diffusion steps are operationalized via a novel adversarial projector that uses a generator G_{θ} and a discriminator D_{θ} . G_{θ} first produces an estimate of the target image $\tilde{\mathbf{x}}_0$ given $\mathbf{x}_t$ and $\mathbf{y}$, and a denoised image sample $\hat{\mathbf{x}}_{t-k}$ is then synthesized from the denoising distribution $q(\mathbf{x}_{t-k}|\mathbf{x}_t, \tilde{\mathbf{x}}_0)$. Meanwhile, D_{θ} distinguishes between actual ($\mathbf{x}_{t-k}$) and synthetic samples ($\hat{\mathbf{x}}_{t-k}$) for the denoised image. 53

- 4.2 For unsupervised learning, SynDiff leverages a cycle-consistent architecture that bilaterally translates between two modalities (A , B). For synthesizing a target image $\hat{\mathbf{x}}_0^A$ of modality A , the diffusive module in Fig. 4.1b requires guidance from a source image $\mathbf{y}^B$ 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 $\tilde{\mathbf{y}}^B$ from $\mathbf{x}_0^A$. Similarly, for synthesizing a target image $\hat{\mathbf{x}}_0^B$ of modality B with the diffusive module, the non-diffusive module first estimates a paired source image $\tilde{\mathbf{y}}^A$ from $\mathbf{x}_0^B$. **a)** To do this, the non-diffusive module comprises two generator-discriminator pairs ($G_{\phi^{A,B}}$, $D_{\phi^{A,B}}$) that generate initial translation estimates for $\mathbf{x}_0^A \rightarrow \tilde{\mathbf{y}}^B$ (orange) and $\mathbf{x}_0^B \rightarrow \tilde{\mathbf{y}}^A$ (green). **b)** These initial translation estimates $\tilde{\mathbf{y}}^{A,B}$ 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 ($G_{\theta^{A,B}}$, $D_{\theta^{A,B}}$) to generate denoised image estimates for $(\mathbf{x}_t^A, \tilde{\mathbf{y}}^B, t) \rightarrow \hat{\mathbf{x}}_{t-k}^A$ (yellow) and $(\mathbf{x}_t^B, \tilde{\mathbf{y}}^A, t) \rightarrow \hat{\mathbf{x}}_{t-k}^B$ (blue). 60
- 4.3 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) $T_2 \rightarrow \text{PD}$ 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. 68
- 4.4 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) $T_2 \rightarrow \text{FLAIR}$ 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. . . . 70

- 4.5 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. 73
- 4.6 Performance of competing methods as a function of added noise level on source-modality images. Results shown for the representative $T_2 \rightarrow CT$ task in terms of PSNR (left), SSIM (right). 78
- 4.7 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. 79

List of Tables

3.1	Validation performance of candidate ResViT configurations in representative synthesis tasks. Performance is taken as PSNR (dB) between synthesized and reference target images. A_i denotes the presence of a transformer module in the i th ART block.	30
3.2	Validation performance of ResViT models with varying sizes of transformer modules in representative synthesis tasks.	30
3.3	Average inference times (msec) of synthesis models per single cross-section.	31
3.4	Performance of task-specific synthesis models in many-to-one ($T_1, T_2 \rightarrow PD$, $T_1, PD \rightarrow T_2$, and $T_2, PD \rightarrow T_1$) and one-to-one ($T_2 \rightarrow PD$ and $PD \rightarrow T_2$) tasks in the IXI dataset. PSNR (dB) and SSIM (%) are listed as mean $\pm$ std across test subjects. Boldface indicates the top-performing model for each task.	34
3.5	Performance of task-specific synthesis models in many-to-one tasks ($T_1, T_2 \rightarrow FLAIR$, $T_1, FLAIR \rightarrow T_2$, and $T_2, FLAIR \rightarrow T_1$) and one-to-one tasks ($T_2 \rightarrow FLAIR$ and $FLAIR \rightarrow T_2$) across test subjects in the BRATS dataset. Boldface indicates the top-performing model for each task.	35

3.6	Performance of unified synthesis models in many-to-one tasks ($T_1, T_2 \rightarrow PD$, $T_1, PD \rightarrow T_2$, and $T_2, PD \rightarrow T_1$) across test subjects in the IXI dataset. Boldface indicates the top-performing model for each task.	40
3.7	Performance of unified synthesis models in many-to-one tasks ($T_1, T_2 \rightarrow FLAIR$, $T_1, FLAIR \rightarrow T_2$, and $T_2, FLAIR \rightarrow T_1$) across test subjects in the BRATS dataset. Boldface indicates the top-performing model for each task.	40
3.8	Performance for the across-modality synthesis task (T_2 -weighted MRI $\rightarrow$ CT) across test subjects in the pelvic MRI-CT dataset. Boldface indicates the top-performing model for each task.	42
3.9	Test performance of ResViT and variants ablated of transformer modules, convolutional modules or adversarial loss. FID is a summary metric across the entire test set. Boldface indicates the top-performing model for each task.	43
3.10	Test performance of ResViT ($A_1 - A_6$) and variants ablated of weight tying and individual transformer modules. Boldface indicates the top-performing model for each task.	43
3.11	Test performance of ResViT and variants ablated of pre-training and delayed insertion procedures for transformers. Boldface indicates the top-performing model for each task.	44
3.12	Test performance of ResViT and variants built by: adding encoder-decoder skip connections, removing skip connections in convolutional modules, removing skip connections in transformer modules, using unlearned down/upsampling blocks in ART, removing down/upsampling blocks in ART via a higher degree of down/upsampling in the encoder/decoder. Boldface indicates the top-performing model for each task.	45

3.13	Input feature maps and corresponding combination weights for the channel compression (CC) module were inspected in trained ResViT models. L_2 -norm of feature tensors from the transformer module (g_j), feature tensors from the previous ART block (f_j), combination weights for g_j , and combination weights for f_j are listed.	45
4.1	Description of variables related to images, diffusion processes, networks and probability distributions. Throughout the manuscript, vectorial quantities are annotated in bold font.	57
4.2	Performance for multi-contrast MRI translation tasks in IXI. PSNR (dB) and SSIM (%) are listed as mean $\pm$ std across the test set. Boldface marks the top-performing model in each task.	68
4.3	Performance for multi-contrast MRI translation tasks in BRATS. PSNR (dB) and SSIM (%) listed as mean $\pm$ std across the test set.	70
4.4	Performance for multi-modal MRI-CT translation tasks in the pelvic dataset. PSNR (dB) and SSIM (%) listed as mean $\pm$ std across the test set. ‘acc.’ stands for accelerated.	74
4.5	Average training times per cross-section (sec), inference times per cross-section (sec) and memory load (gigabytes).	75
4.6	Performance of variant models in unconditional synthesis tasks. FID is listed across the training set.	78
4.7	Performance of variant models ablated of adversarial loss, cycle-consistency loss and the diffusive module. PSNR and SSIM listed as mean $\pm$ std across the test set.	80
4.8	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 $\pm$ std across the test set.	81

4.9	Performance of variant models as $\text{mean} \pm \text{std}$ 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.	82
4.10	Performance of variant models as $\text{mean} \pm \text{std}$ 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.	82

Chapter 1

Introduction

1.1 Multi-modal Medical Imaging

Multi-modal medical imaging (MMI) is a powerful tool for in vivo assessment of anatomy and function in the human body [1]. MMI involves acquiring images from different types of scanners (e.g., CT, MRI, PET, ultrasound) [2] or different contrasts from a single scanner (e.g., multi-contrast MRI) [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14], which provide complementary information about tissue morphology, function, and metabolism. In turn, MMI can enhance diagnostic accuracy and performance in downstream imaging tasks [15]. Yet, MMI often faces several practical challenges, such as high costs of scanners and maintenance, long scan times that increase patient discomfort and motion artifacts [16], and patient safety issues such as radiation exposure and contrast agent toxicity [17, 18]. One way to reduce the scan time in medical imaging is to use image reconstruction techniques that can recover high-quality images from incomplete or noisy measurements [19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36]. These techniques have shown significant improvements in accelerating MRI acquisitions. However, they often are not applicable in some clinical retrospective scenarios where measurements of the target image is not available.

Medical image translation is a promising solution that aims to synthesize missing target modalities from available source modalities, reducing the need for additional scans [37, 38, 39, 40, 41, 42, 43, 44] (Note that medical image synthesis and translation is used interchangeably throughout this thesis). This is an ill-posed inverse problem due to nonlinear variations in tissue signals across modalities. Learning-based methods have shown great potential in solving this problem by incorporating data-driven priors to improve problem conditioning. These methods involve deep neural network models trained to capture the conditional distribution of target given source images [45, 46, 47]. Convolutional neural networks (CNNs) have been widely used for medical image translation, achieving significant improvements over traditional methods [45, 46, 48, 49, 50, 47]. Generative adversarial networks (GANs) with CNN backbones were later introduced to enhance the realism and detail of the generated images by using an adversarial loss [51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66]. However, CNNs rely on local image features extracted by compact filters, which may limit their ability to capture long-range spatial dependencies and contextual information [67, 68, 69, 70, 71, 72]. Furthermore, GANs have their own drawbacks, such as implicit distribution modeling, learning biases, mode collapse, and unreliable one-shot sampling [73, 74, 75, 76]. In this thesis, two novel approaches were presented to address the shortcomings of CNN-based GAN models for improved reliability and performance in medical image translation.

1.2 Contributions

This thesis first aims to address limitations of existing CNN-based adversarial medical image synthesis methods. Existing approaches employ convolutions to exploit correlations among small neighborhoods of image pixels, and this successfully reduces the number of model parameters to facilitate learning and manifest feature locality. However, CNNs have limited ability to capture contextual features due to their local processing with compact filters. To address this issue, ResViT was proposed, a novel generative adversarial model that combines the

contextual sensitivity of vision transformers with the precision of convolution operators and the realism of adversarial learning. ResViT’s generator consists of novel aggregated residual transformer (ART) blocks that fuse residual convolutional and transformer modules in a central bottleneck. ART blocks use residual connections and channel compression to enhance the diversity and relevance of the learned features. A weight sharing strategy is applied among ART blocks to reduce the computational cost. ResViT has a unified implementation that can handle different source-target modality combinations without rebuilding separate models. Effectiveness of the proposed approach demonstrated for synthesizing missing sequences in multi-contrast MRI and CT images from MRI, and show that it outperforms competing CNN- and transformer-based methods in terms of qualitative and quantitative evaluations.

Next, this thesis presents a novel diffusion-based medical image translation method to alleviate shortcomings of adversarial modeling. Most medical image synthesis methods use generative adversarial networks (GAN) that map source images to target images in one step. However, GAN models that try to learn the image distribution implicitly may suffer from limited sample fidelity. Here in this thesis, we introduce a novel method based on adversarial diffusion modeling, SynDiff, was introduced 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. SynDiff also enables unsupervised training with unpaired datasets via a cycle-consistent architecture with coupled diffusive and non-diffusive modules that bilaterally translate between two modalities. SynDiff was compared against competing GAN and diffusion models in multi-contrast MRI and MRI-CT translation. Extensive experimental results indicate the quantitative and qualitative superiority of SynDiff against competing baselines.

1.3 Outline

The rest of this thesis is organized as follows. Chapter 2 provides relevant background on multi-modal MR and CT imaging. Chapter 3 introduces the residual vision transformer model, and Chapter 4 describes the adversarial diffusion method presents for multi-modal medical image synthesis and translation. Lastly, Chapter 5 summarizes and concludes the main findings and contributions presented in this thesis.

Chapter 2

Fundamentals of Multi-Modal Imaging

2.1 Basics of MRI Contrast Mechanisms

Magnetic resonance imaging (MRI) is based on hydrogen nuclei ^1H that possess both magnetic and angular momentum. The hydrogen nuclei adjust their alignment according to the magnetic field. In the absence of any external magnetic field B_0 , the hydrogen nuclei are arranged randomly and the net magnetization is nearly zero. However, with the application of an external B_0 , the hydrogen nuclei begin to precess around the longitudinal direction of the applied B_0 with so called the Larmor frequency, $\omega_0 = B_0\gamma$, where γ is the gyro-magnetic ratio of underlying material. Followingly, an external radio-frequency (RF) field pulse, denoted as B_1 , can tilt the net magnetization of the nuclei to the transverse plane by an arbitrary flip angle. After turning off external RF field, the net magnetization tries to return to its equilibrium state. During this relaxation period, the transverse part of magnetization decays with a time constant T_2 , meanwhile longitudinal magnetization recovering with a time constant T_1 . In MRI, the receiver coils can detect the magnetic flux changes along transverse direction and therefore the signal in the transverse plane can be easily picked up. Note that, for

spatial encoding of the signal, additional magnetic field gradients G are required. That said, the received signal by the coil is given as:

$$s(t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} M_0(x, y, z) e^{-t/T_2(r)} e^{-i\gamma \int_0^t G(\tau) \cdot r d\tau} dx dy dz \quad (2.1)$$

where $M_0(x, y, z)$ is transverse magnetization at point (x, y, z) , $r = (x, y, z)$ and $G = (G_x, G_y, G_z)$. Frequency (Fourier) domain in MRI is usually referred to as k-space. To achieve a k-space formulation, equation 2.1 can be rearranged by expressing the gradient terms with spatial frequency variables:

$$k_x(t) = \frac{\gamma}{2\pi} \int_0^t G_x(\tau) d\tau \quad (2.2)$$

$$k_y(t) = \frac{\gamma}{2\pi} \int_0^t G_y(\tau) d\tau \quad (2.3)$$

$$k_z(t) = \frac{\gamma}{2\pi} \int_0^t G_z(\tau) d\tau \quad (2.4)$$

$$s(t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} M_0(x, y, z) e^{-t/T_2(r)} e^{-i2\pi(k_x(t)+k_y(t)+k_z(t))} dx dy dz \quad (2.5)$$

Note that transverse magnetization can be manipulated via the applied RF pulse and gradient fields based on the tissue-specific relaxation parameters. With this manipulation, desired properties of the tissues can be imaged via various pulse sequences which enables the multi-contrast MRI acquisition. For instance, MRI signal for the common spin-echo sequences can be represented as:

$$S_0 = K_0 \rho(r) [1 - e^{-TR/T_1(r)}] e^{TE/T_2(r)} \quad (2.6)$$

where K_0 is a constant, $\rho(r)$ is proton density (PD), TR is the repetition time, and TE are the echo time for the RF pulse, and $T_1(r)$ and $T_2(r)$ are relaxation parameters of longitudinal and transverse magnetization, respectively.

This equation is an example of how to alter the contrast in acquired images by simple modification of TR and TE. T_1 -weighted contrast images, for example, can be produced by a low TE and a moderate TR. Similarly, T_2 -weighted contrast images can be acquired by a moderate TE and long TR duration.

2.2 Basics of CT Imaging

Computed tomography (CT) imaging uses a combination of X-rays and computer technology to produce cross-sectional images of the body or objects. It can be used for various diagnostic, therapeutic, or industrial purposes. It depicts in detail areas inside the body, such as bones, muscles, organs, and blood vessels. To produce images of thin cross-sections within an anatomy, CT scanners work by rotating an X-ray source and a line of sensors around the body and computing how much X-ray beams are weakened by different tissues inside the body. A computer then leverages reconstruction algorithms to form images from the multiple X-ray measurements taken from various angles.

The attenuation coefficient of a tissue at a given X-ray energy is given by:

$$\mu(E) = \mu_p(E) + \mu_c(E) \quad (2.7)$$

where $\mu_p(E)$ is the photoelectric absorption coefficient and $\mu_c(E)$ is the Compton scattering coefficient.

The intensity of the X-ray beam after passing through a tissue of thickness x is given by:

$$I(x) = I_0 e^{-\mu x} \quad (2.8)$$

where I_0 is the initial intensity and μ is the attenuation coefficient.

The projection of the attenuation coefficient along a ray path s is given by:

$$p(s) = \int_{-\infty}^{\infty} \mu(x, y) ds \quad (2.9)$$

where (x, y) are the coordinates of the tissue along the ray path.

The reconstructed image of the attenuation coefficient is obtained by applying an inverse Radon transform to the projections:

$$f(x, y) = \mathcal{R}^{-1}[p(s, \theta)] \quad (2.10)$$

$$f(x, y) = \mathcal{R}^{-1}[p(s, \theta)] = \frac{1}{4\pi^2} \int_0^\pi \int_{-\infty}^{\infty} p(s, \theta) \frac{\partial}{\partial s} \left[\frac{1}{s} \int_{-\infty}^s e^{i\xi f} H(\xi) d\xi \right] ds d\theta \quad (2.11)$$

where $f(x, y)$ is the reconstructed image, $p(s, \theta)$ is the projection at angle θ , $h(t)$ is a filter function, and $H(\xi)$ is its Fourier transform, and $\mathcal{R}^{-1}$ is the inverse Radon transform operator.

2.3 Multi-Modal MRI-CT

MRI and CT scans are both imaging techniques that can produce detailed images of the internal structures of the body. However, they have different advantages and disadvantages that may affect the choice of imaging modality depending on the clinical scenario.

Figure 2.1 displays T_1 , T_2 -weighted MR and CT images of a male pelvis. MRI scans create images of the soft tissues using strong magnetic fields and radio waves, and thereby reveal more information about the tissue composition and function than CT scans. MR imaging also safer because they do not use ionizing

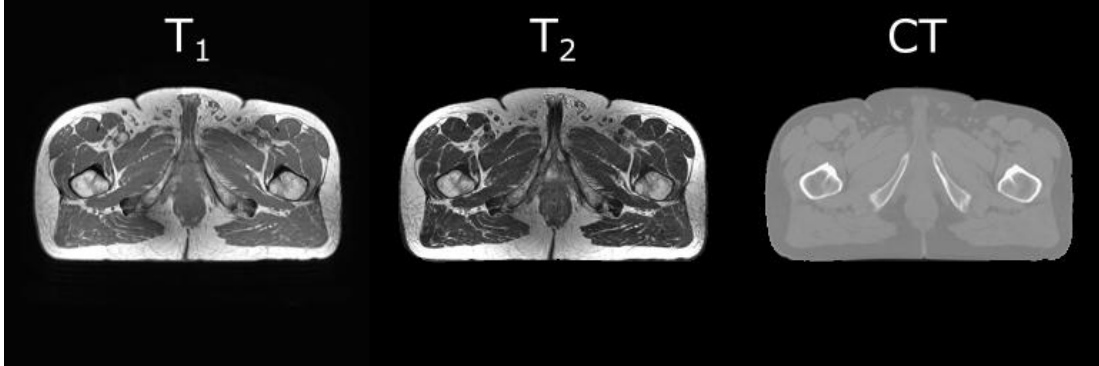


Figure 2.1: Sample 2D cross sections of T_1 , T_2 -weighted MR and CT images of a male pelvis.

radiation, which can harm the cells and increase cancer risk of patients. They are especially useful for diagnosing diseases that affect the soft tissues, such as tumors, infections, inflammation, stroke, and multiple sclerosis. On the other hand, CT scans create images of the bones, organs, and blood vessels using X-rays and computer technology. They can reveal more information about the shape and size of the structures than MRI scans. CT exams are faster, more widely available than MRI scans, and cheaper compared to MRI. They are especially useful for diagnosing diseases that affect the bones or cause bleeding, such as fractures, trauma, cancer, aneurysms, and appendicitis.

The main trade-off between MRI and CT scans is the balance between exam time, image quality, and radiation exposure. MRI scans can provide better image quality than CT scans for certain conditions, but they also take longer and require more cooperation from the patient. CT scans can provide faster and more convenient imaging than MRI scans for certain conditions, but they also expose the patient to radiation that can have harmful effects in the long term. Therefore, the choice of imaging modality depends on the clinical question, the urgency of the situation, the availability of resources, and the patient's preferences.

Medical image translation methods are techniques that can synthesize a missing or desired image modality from an available or acquired image modality. For example, medical image translation can generate CT images from MR images, or vice versa. These approaches can also recover missing MRI contrast or CT scans

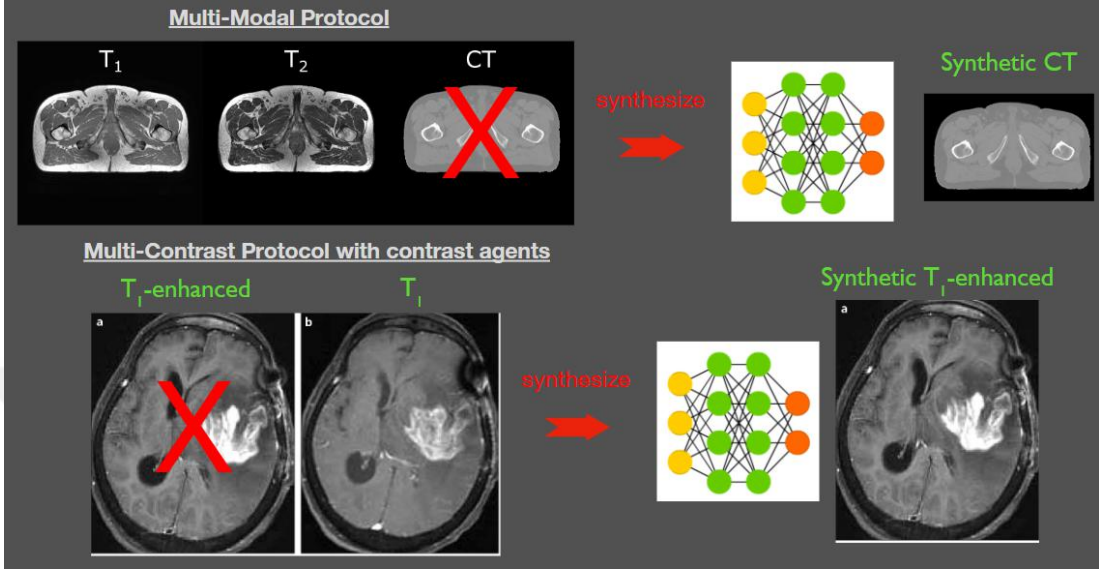


Figure 2.2: Synthesis methods can reduce the radiation exposure and contrast agent injection for the patient by imputing CT scans or contrast-enhanced MRI scans with synthetic images.

by translating between different modalities or sequences. Thus, they hold several potential benefits for clinical practice and research. First, they can reduce the imaging workload and cost by eliminating the need to acquire multiple modalities or sequences for the same patient or region of interest [77]. Second, they can reduce the radiation exposure and contrast agent injection for the patient by replacing CT scans or contrast-enhanced MRI scans with synthetic images (See Fig. 2.2). Furthermore, they can improve the diagnosis and treatment planning by providing complementary information from different modalities or sequences. Lastly, they can facilitate the development and evaluation of machine learning methods for downstream tasks such as classification and segmentation [78, 79, 80, 81], by augmenting the available data and labels. Taken together, medical image synthesis methods can facilitate ubiquitous multi-modal medical imaging.

Chapter 3

ResViT: Residual vision transformers for multi-modal medical image synthesis

The content of this chapter reflects the studies described in the following publications:

- **O. Dalmaz**, M. Yurt, and T. Çukur, "ResViT: Residual Vision Transformers for Multimodal Medical Image Synthesis," in *IEEE Transactions on Medical Imaging*, vol. 41, no. 10, pp. 2598-2614, Apr 2022. [Online]. Available: <https://ieeexplore.ieee.org/document/9758823>
- **O. Dalmaz**, M. Yurt, S. U. H. Dar, and T. Cukur, "Cycle-Consistent Adversarial Transformers for Unpaired MR Image Translation," in *30th annual meeting of International Society for Magnetic Resonance Imaging (ISMRM)*, London, May 2022.
- **O. Dalmaz**, M. Yurt, and T. Cukur, "Adversarial Residual Transformers For Multi-Modal Medical Image synthesis," in *IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, Virtual Conference, Mar. 2022. (Virtual Conference)

- **O. Dalmaz**, M. Yurt, and T. Cukur, “Medical Image Synthesis with Residual Vision Transformers,” *NeurIPS Medical Imaging Meets*, Virtual Conference, Dec. 2021. (Virtual Conference)

3.1 Introduction

Medical imaging plays a pivotal role in modern healthcare by enabling in vivo examination of pathology in the human body. In many clinical scenarios, multi-modal protocols are desirable that involve a diverse collection of images from multiple scanners (e.g., CT, MRI) [2], or multiple acquisitions from a single scanner (multi-contrast MRI) [3]. Complementary information about tissue morphology, in turn, empower physicians to diagnose with higher accuracy and confidence. Unfortunately, numerous factors including uncooperative patients and excessive scan times prohibit ubiquitous multi-modal imaging [17, 18]. As a result, there has been ever-growing interest in synthesizing unacquired images in multi-modal protocols from the subset of available images, bypassing costs associated with additional scans [1, 82].

Medical image synthesis aims to predict target-modality images for a subject given source-modality images acquired under a limited scan budget [37]. This is an ill-posed inverse problem since medical images are high dimensional, target-modality data are absent during inference, and there exist nonlinear differences in tissue contrast across modalities [83, 84, 15, 85, 86, 87]. Unsurprisingly, recent adoption of deep learning methods for solving this difficult problem has enabled major performance leaps [88, 89, 90, 91, 92, 93, 94, 95]. In learning-based synthesis, network models effectively capture a prior on the joint distribution of source-target images [49, 46, 53]. Earlier studies using CNNs for this purpose reported significant improvements over traditional approaches [45, 46, 48, 49, 50, 47]. Generative adversarial networks (GANs) were later introduced that leverage an adversarial loss to increase capture of detailed tissue structure [51, 52, 53, 54, 55, 56, 57, 58]. Further improvements were attained

by leveraging enhanced architectural designs [59, 60, 61, 62], and learning strategies [63, 64, 65]. Despite their prowess, prior learning-based synthesis models are fundamentally based on convolutional architectures that use compact filters to extract local image features [73, 74]. Exploiting correlations among small neighborhoods of image pixels, this inductive bias reduces the number of model parameters to facilitate learning. However, it also limits expressiveness for contextual features that reflect long-range spatial dependencies [67, 68].

Indeed, spatially distant voxels in medical images can show strong correlations for both healthy and pathological tissue. For instance, bone tissue in the skull or CSF in the ventricles broadly distribute over large brain regions, whether spatially contiguous or segregated. While pathology such as brain lesions might have less regular anatomical priors, their spatial distribution (e.g., location, quantity, shape) still shows disease-specific patterns [96]. For instance, multiple diffusely-distributed lesions are prominent even in early stages of multiple sclerosis (MS) or Alzheimer’s (AD); with characteristic lesions located near periventricular and juxtacortical regions in MS, and near hippocampus, entorhinal cortex and isocortex in AD [97]. Meanwhile, relatively few lesions manifest as spatially-contiguous clumps spanning across large volumes in high-grade tumors; with characteristic lesions located near the cerebrum and cerebellum in gliomas, and near the skull in meningiomas [97]. Thus, the spatial distribution of pathology also involves long-range interactions regarding the relative position and composition of lesions with respect to healthy tissue. In principle, synthesis performance can be enhanced by priors that capture long-range interactions related to healthy and abnormal tissues. While transformer architectures in computer vision use global attention operators to improve contextual sensitivity [98], adoption of vanilla transformers in tasks with pixel-level outputs is challenging due to computational burden and limited localization abilities [99].

Here, we propose a novel deep learning model for medical image synthesis, ResViT, that combines the sensitivity of vision transformers to global context, the localization power of CNNs, and the realism of adversarial learning. ResViT’s generator follows an encoder-decoder architecture with a central bottleneck to distill task-critical information. The encoder and decoder contain CNN

blocks to leverage local precision of convolution operators [100]. The bottleneck comprises novel aggregated residual transformer (ART) blocks to synergistically preserve local and global context, with a weight-sharing strategy to minimize model complexity. To improve practical utility, a unified ResViT implementation is introduced that consolidates models for numerous source-target configurations. Demonstrations are performed for synthesizing missing sequences in multi-contrast MRI, and CT from MRI. Comprehensive experiments on imaging datasets from healthy subjects and patients clearly indicate the superiority of the proposed method against competing methods. Code to implement the ResViT model is publicly available at <https://github.com/icon-lab/ResViT>.

Contributions

- We introduce the first adversarial model for medical image synthesis with a transformer-based generator to translate multi-modal imaging data.
- We introduce novel aggregated residual transformer (ART) blocks to synergistically preserve localization and context.
- We introduce a weight sharing strategy among ART blocks to lower model complexity and mitigate computational burden.
- We introduce a unified synthesis model that generalizes across multiple configurations of source-target modalities.

3.2 Related Work

The immense success of deep learning in inverse problems has motivated its rapid adoption in medical imaging [101, 102]. Medical image synthesis is a particularly ill-posed problem since target images are predicted without any target-modality data [55]. Earlier studies in this domain have proposed local networks based on patch-level processing [90, 103, 104]. While local networks offer benefits over

traditional approaches, they can show limited sensitivity to broader context across images [49]. Later studies adopted deep CNNs for image-level processing with increasing availability of large imaging databases. CNN-based synthesis has been successfully demonstrated in various applications including synthesis across MR scanners [105, 106, 55, 107], multi-contrast MR synthesis [45, 46, 48, 49, 50, 47], and CT synthesis [108, 109, 110, 111]. Despite significant improvements they enable, CNNs trained with pixel-wise loss terms tend to suffer from undesirable loss of detailed structure [73, 74, 53].

To improve capture of structural details, GANs [51] were proposed to learn the distribution of target modalities conditioned on source modalities [112]. Adversarial losses empower GANs to capture an improved prior for recovery of high-spatial-resolution information [73, 74, 53]. In recent years, GAN-based methods were demonstrated to offer state-of-the-art performance in numerous synthesis tasks, including data augmentation as well as multi-modal synthesis [113, 114, 53, 57]. Important applications of GAN models include CT to PET [115, 116], MR to CT [117, 118, 119], unpaired cross-modality [120, 121, 122, 123], 3T-to-7T [124, 125], and multi-contrast MRI synthesis [56, 52, 53, 60, 57, 58, 64, 65, 62, 54, 63, 61, 59, 55].

While GAN models have arguably emerged as a gold standard in recent years, they are not without limitation. In particular, GANs are based on purely convolutional operators known to suffer from poor across-subject generalization to atypical anatomy and sub-optimal learning of long-range spatial dependencies [67, 68]. Recent studies have incorporated spatial or channel attention mechanisms to modulate CNN-derived feature maps [126, 127, 128, 60, 129, 130, 131]. Such modulation motivates the network to give greater focus to regions that may suffer from greater errors [127, 128]. While attention maps might be distributed across image regions, multiplicative gating of local CNN features offers limited expressiveness in modeling of global context [132, 99, 133].

To incorporate contextual representations, transformer-based methods have received recent interest in medical imaging tasks such as segmentation [99, 132, 134], reconstruction [135, 136, 137], and synthesis [138, 139, 140]. Among relevant

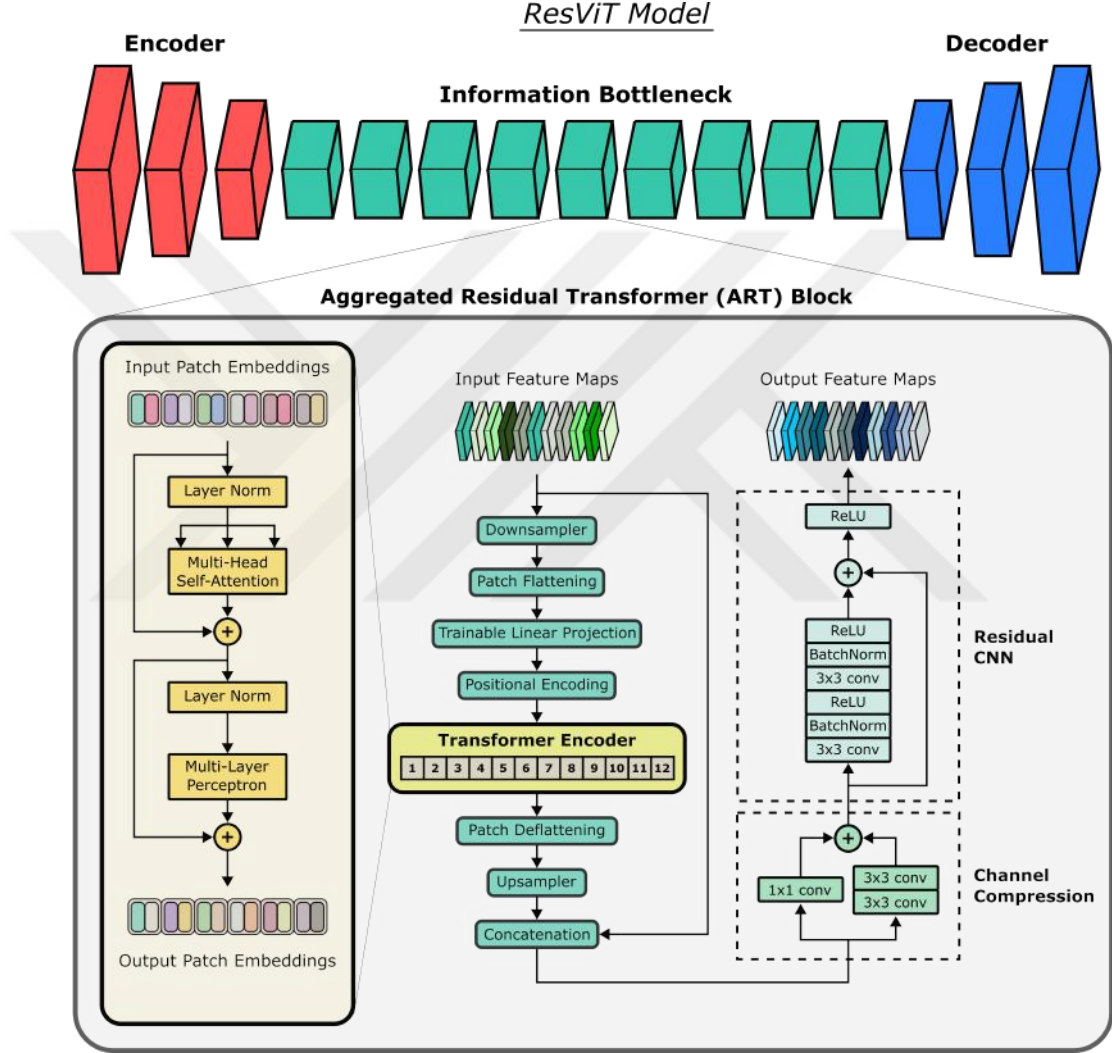


Figure 3.1: The generator in ResViT follows an encoder-decoder architecture bridged with a central information bottleneck to distill task-specific information. The encoder and decoder comprise convolutional layers to maintain local precision and inductive bias in learned structural representations. Meanwhile, the information bottleneck comprises a stack of novel aggregated residual transformer (ART) blocks. ART blocks learn contextual representations via vision transformers, and synergistically fuse CNN-based local and transformer-based global representations.

methods are Transformer GAN that suppresses noise in low-dose PET images [135], TransCT that suppresses noise in low-dose CT images [136], and SLATER that recovers MR images from undersampled k-space acquisitions [137]. While these methods reconstruct images for single-modality data, ResViT translates imaging data across separate modalities. Furthermore, Transformer GAN is an adversarial model with convolutional encoder-decoder and a bottleneck that contains a transformer module without external residual connections. TransCT is a non-adversarial model where CNN blocks first learn textural components of low-frequency (LF) and high-frequency (HF) image parts; and a transformer block without external residual connections then combines the encoded HF and textural LF maps. In comparison, ResViT is an adversarial model that employs a hybrid architecture in the bottleneck of its generator comprising a cascade of residual transformer and residual CNN modules. Unlike SLATER based on an unconditional model that maps latent variable to images via cross-attention transformers, ResViT is a conditional model based on self-attention transformers.

Few recent studies have independently introduced transformer-based methods for medical image synthesis. VTGAN generates retinal angiograms from fundus photographs [138] and GANBERT performs MR-to-PET synthesis [139], whereas ResViT performs multi-contrast MRI and MR-to-CT synthesis. Both VTGAN and GANBERT use entirely convolutional generators and only include transformers in their discriminators. In contrast, ResViT incorporates transformers in its generator to explicitly leverage long-range context. The closest study to our work is PTNet that performs one-to-one translation between T_1 - and T_2 -weighted images in infant MRI [140]. However, PTNet is a non-adversarial model without a discriminator, and it follows a convolution-free architecture. In contrast, ResViT is an adversarial model with a hybrid CNN-transformer architecture to achieve high localization and contextual sensitivity along with a high degree of realism in synthesized images. Furthermore, a broader set of tasks are considered for ResViT including one-to-one and many-to-one translation for multi-contrast MRI and CT.

A unique component of ResViT is the novel ART blocks in its generator that

contain a cascade of transformer and CNN modules equipped with skip connections. These residual paths enable effective aggregation contextual and convolutional representations. Based on this powerful component, we provide the first demonstrations of a transformer architecture for many-to-one synthesis tasks and a unified synthesis model for advancing practicality over task-specific methods.

3.3 Theory and Methods

3.3.1 Residual Vision Transformers

Here we propose a novel adversarial method for medical image synthesis named residual vision transformers, ResViT, that can unify various source-target modality configurations into a single model for improved practicality. ResViT leverages a hybrid architecture of deep convolutional operators and transformer blocks to simultaneously learn high-resolution structural and global contextual features (Fig. 3.1). The generator subnetwork follows an encoder - information bottleneck - decoder pathway, and the discriminator subnetwork is composed of convolutional operators. The generator’s bottleneck contains a stack of novel aggregated residual transformer (ART) blocks. Each ART block is organized as the cascade of a transformer module that extracts hidden contextual features, and a CNN module that extracts hidden local features of input feature maps. Importantly, skip connections are inserted around both modules to create multiple paths of information flow through the block. These paths propagate multiple sets of features to the output: (a) Input features from the previous network layer passing through skip connections of transformer and CNN modules; (b) Contextual features computed by the transformer module passing through the skip connection of the CNN module; (c) Local features computed by the CNN module based on input features reaching through the skip connection of the transformer module; (d) Hybrid local-contextual features computed by the transformer-CNN cascade. Therefore, the main motivation for use of residual CNN and transformer modules is to learn an aggregated representation that synergistically combines lower-level input features

along with their contextual, local, and hybrid local-contextual features.

The central segment of ResViT containing ART blocks acts as an information bottleneck for spatial and feature dimensions of medical image representations. On the one hand, the central segment processes feature maps that have been spatially downsampled by the encoder. This increases the relative emphasis on mid- to high-level spatial information over lower-level information [100]. On the other hand, ART blocks contain channel-compression (CC) modules that process concatenated feature maps from the previous ART block and the transformer module. CC modules downsample the concatenated maps in the feature dimension to distill a task-relevant set of convolutional and contextual features.

Given the computational efficiency of convolutional layers, CNNs pervasively process feature maps at relatively high spatial resolutions to improve spatial sensitivity for local features [100]. In contrast, vision transformers include computationally exhaustive self-attention layers, and so they typically process feature maps at notably lower spatial resolutions [98]. To ensure that both the residual CNNs and transformers in ART blocks receive input feature maps at their expected resolutions, we incorporated down and upsampling blocks respectively at the input and output of transformer modules. this particular design ensures compatibility between the resolutions of feature maps extracted from CNN and transformer modules in return for a modest increase in model complexity. In the remainder of this section, we explain the detailed composition of each architectural component, and we describe the loss functions to train ResViT.

3.3.1.1 Encoder

The first component of ResViT is a deep encoder network that contains a series of convolutional layers to capture a hierarchy of localized features of source images. Note that ResViT can serve a unified synthesis model, so its encoder receives as input the full set of modalities within the imaging protocol, both source and target modalities (Fig. 3.2). Source modalities are input via an identity mapping,

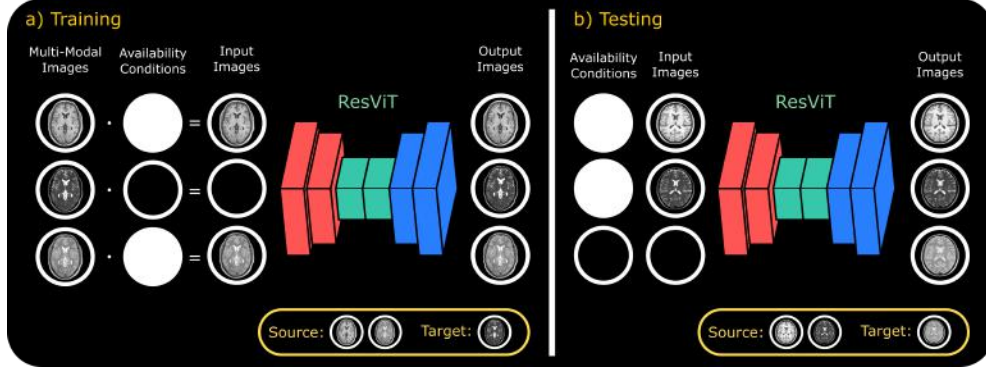


Figure 3.2: ResViT is a conditional image synthesis model that can unify various source-target modality configurations into a single model for improved practicality. a) During training, ResViT takes as input the entire set of images within the multi-modal protocol, including both source and target modalities. For model consolidation across multiple synthesis tasks, various configurations of source-target modalities are expressed in terms of availability conditions in ResViT. b) During inference, the specific source-target configuration is determined via the availability conditions in each given test subject.

whereas unavailable target modalities are masked out:

$$X_i^G = a_i \cdot m_i \quad (3.1)$$

where i denotes the channel index of the encoder input $i \in \{1, 2, \dots, I\}$, m_i is the image for the i th modality. In Eq. (1), a_i denotes the availability of the i th modality:

$$a_i = \begin{cases} 1 & \text{if } m_i \text{ is a source modality} \\ 0 & \text{if } m_i \text{ is a target modality} \end{cases} \quad (3.2)$$

During training, various different configurations of source-target modalities are considered within the multi-modal protocol (e.g., $T_1, T_2 \rightarrow PD$; $T_2, PD \rightarrow T_1$; $T_1, PD \rightarrow T_2$ for a three-contrast MRI protocol). During inference, the specific source-target configuration is determined via the availability conditions in individual test subjects. Given the availability-masked multi-channel input, the encoder uses convolutional operators to learn latent structural representations shared across the consolidated synthesis tasks. Encoder maps the multi-channel input X^G to the embedded latent feature map $f_{ne} \in \mathbb{R}^{N_C, H, W}$ via convolutional filters, where N_C is the number channels, H is the height and W is the width of the feature map. These representations are then fed to the information bottleneck.

3.3.1.2 Information Bottleneck

Next, ResViT employs a residual bottleneck to distill task-relevant information in the encoded features. Note that convolution operators have greater power in capturing localized features, whereas attention operators are more sensitive to context-driven features. To simultaneously maintain localization power and contextual sensitivity, we introduce ART blocks that aggregate the information from residual convolutional and transformer branches (Fig. 3.1). Receiving as input the j th layer feature maps $f_j \in \mathbb{R}^{N_C, H, W}$, an ART block first processes the feature maps via a vision transformer. Due to computational constraints, the transformer expects feature maps at smaller resolutions compared to convolutional layers. Thus, the spatial dimensions (H, W) of $f_j \in \mathbb{R}^{N_C, H, W}$ are lowered by a downsampling block (**DS**):

$$f'_j \in \mathbb{R}^{N'_C, H', W'} = \mathbf{DS}(f_j) \quad (3.3)$$

where **DS** is implemented as a stack of strided convolutional layers, $f'_j \in \mathbb{R}^{N'_C, H', W'}$ are downsampled feature maps with $W' = W/M$, $H' = H/M$, M denoting the downsampling factor. A transformer branch then processes f'_j to extract contextual information. Accordingly, f'_j is first split into non-overlapping patches of size (P, P) , and the patches are then flattened. Hereby, $N_P = W'H'/P^2$ patches with dimensionality $N'_C P^2$ are formed. The transformer embeds patches onto an N_D -dimensional space via trainable linear projections, supplemented with learnable positional encoding:

$$z_0 = [f_j^1 P_E; f_j^2 P_E; \dots; f_j^{N_P} P_E] + P_E^{pos} \quad (3.4)$$

where $z_0 \in \mathbb{R}^{N_P, N_D}$ are the input patch embeddings, $f_j^p \in \mathbb{R}^{N'_C P^2}$ is the p th patch, P_E is the embedding projection, and P_E^{pos} is the learnable positional encoding.

Next, the transformer encoder processes patch embeddings via a cascade of L layers of multi-head self-attention (**MSA**) [141] and multi-layer perceptrons (**MLP**) [142]. The output of the l th layer in the transformer encoder is given as:

$$z'_l = \mathbf{MSA}(\mathbf{LN}(z_{l-1})) + z_{l-1} \quad (3.5)$$

$$z_l = \mathbf{MLP}(\mathbf{LN}(z'_l)) + z'_l \quad (3.6)$$

MSA layers in Eq. 7 employ S separate self-attention heads:

$$\mathbf{MSA}(z) = [\mathbf{SA}_1(z); \mathbf{SA}_2(z); \dots; \mathbf{SA}_S(z)]U_{msa} \quad (3.7)$$

where $\mathbf{SA}_s$ stands for the s th attention head with $s \in \{1, 2, \dots, S\}$ and U_{msa} denotes the learnable tensor projecting attention head outputs. $\mathbf{SA}$ layers compute a weighted combination of all elements of the input sequence z : $\mathbf{SA}(z) = Av$ where attention weights $A_{a,b}$ are taken as pairwise similarity between the query q and key k :

$$A_{a,b} = \text{softmax}(q_a k_b^T / N_D^{0.5}) \quad (3.8)$$

where

The output of the transformer encoder z_L is then deflattened to form $g'_j \in \mathbb{R}^{N_D, H', W'}$. Resolution of g'_j is increased to match the size of input feature maps via an upsampling block **US** based on transposed convolutions:

$$g_j \in \mathbb{R}^{N_C, H, W} = \mathbf{US}(g'_j) \quad (3.9)$$

where $g_j \in \mathbb{R}^{N_C, H, W}$ are upsampled feature maps output by the transformer module. The global context learned via the transformer is then fused with the localized features captured with convolutional operators by channelwise concatenation.

To distill learned structural and contextual representations, the channels of the concatenated feature maps are then compressed via a channel compression (**CC**) module:

$$h_j \in \mathbb{R}^{N_C, H, W} = \mathbf{CC}(\text{concat}(f_j, g_j)) \quad (3.10)$$

where h_j are compressed feature maps. **CC** uses two parallel convolutional branches of varying kernel size. Finally, the feature maps are processed via a residual CNN (**ResCNN**) [100]:

$$f_{j+1} \in \mathbb{R}^{N_C, H, W} = \mathbf{ResCNN}(h_j) \quad (3.11)$$

where f_{j+1} denotes the output of the ART block at the j th network layer.

3.3.1.3 Decoder

The last component of the generator is a deep decoder based on transposed convolutional layers. Because ResViT can serve as a unified model, its decoder can synthesize all contrasts within the multi-modal protocol regardless of the specific source-target configuration (Fig. 3.2). The decoder receives as input the feature maps f_A distilled by the bottleneck and produces multi-modality images $\hat{Y}_i^G \in \hat{Y}^G$ in separate channels, where A is the total number of ART blocks, and $\hat{Y}_i^G$ denotes the i th synthesized modality.

3.3.1.4 Parameter Sharing Transformers

Multiple ART blocks are used in the information bottleneck to increase the capacity of ResViT in learning contextual representations. That said, multiple independent transformer blocks would inevitably elevate memory demand and risk of overfitting due to an excessive number of parameters. To prevent these risks, a weight-sharing strategy is adopted where the model weights for the transformer encoder are tied across separate ART blocks. The tied parameters in each transformer layer include the projection matrices T_q, T_k, T_v for query, key, value and projection tensors for attention heads U_{msa} in MSA layers, and weight matrices in MLP layers. Remaining parameters in transformer modules including down/up-sampling blocks, patch embeddings and positional encodings are kept independent. During backpropagation, updates for tied weights are computed based on the summed error gradient across ART blocks.

3.3.1.5 Discriminator

The discriminator in ResViT is based on a conditional PatchGAN architecture [73]. The discriminator performs patch-level differentiation between acquired and synthetic images. This implementation increases sensitivity to localized details related to high-spatial-frequency information. As ResViT can serve as a unified

model by generating all modalities in the multi-modal protocol including sources, an availability-guided selective discriminator is employed:

$$X_i^D(source) = X_i^G = a_i \cdot m_i \quad (3.12)$$

$$X_i^D(syn\ target) = (1 - a_i) \cdot Y_i^G \quad (3.13)$$

$$X_i^D(acq\ target) = (1 - a_i) \cdot m_i \quad (3.14)$$

where $X_i^D(source)$ are source images, $X_i^D(syn\ target)$ are synthesized target images, and $X_i^D(acq\ target)$ are acquired target images. The conditional discriminator receives as input the concatenation of source and target images:

$$X^D(synthetic) = concat(X_i^D(source), X_i^D(syn\ target)) \quad (3.15)$$

$$X^D(acquired) = concat(X_i^D(source), X_i^D(acq\ target)) \quad (3.16)$$

where $X^D(synthetic)$ is the concatenation of source and synthetic target images, and $X^D(acquired)$ is the concatenation of the source and acquired target images.

3.3.1.6 Loss Function

The first term in the loss function is a pixel-wise L_1 loss defined between the acquired and synthesized target modalities:

$$L_{pix} = \sum_{i=1}^I (1 - a_i) E[||G(X^G)_i - m_i||_1] \quad (3.17)$$

where E denotes expectation, and G denotes the generator submodule in ResViT. ResViT takes as input source modalities to reconstruct them at the output. Thus, the second term is a pixel-wise consistency loss between acquired and reconstructed source modalities based on an L_1 distance:

$$L_{rec} = \sum_{i=1}^I a_i E[||G(X^G)_i - m_i||_1] \quad (3.18)$$

The last term is an adversarial loss defined via the conditional discriminator (D):

$$\begin{aligned} L_{adv} = & - E[D(X^D(acquired))^2] \\ & - E[(D(X^D(synthetic)) - 1)^2] \end{aligned} \quad (3.19)$$

The three terms are linearly combined to form the overall objective:

$$L_{ResViT} = \lambda_{pix}L_{pix} + \lambda_{rec}L_{rec} + \lambda_{adv}L_{adv} \quad (3.20)$$

where λ_{pix} , λ_{rec} , and λ_{adv} are the weightings of the pixel-wise, reconstruction, and adversarial losses, respectively.

3.3.2 Datasets

We demonstrated the proposed ResViT model on two multi-contrast brain MRI datasets (IXI: <https://brain-development.org/ixi-dataset/>, BRATS [143, 144, 145]) and a multi-modal pelvic MRI-CT dataset [146].

3.3.2.1 IXI Dataset

T₁-weighted, T₂-weighted, and PD-weighted brain MR images from 53 healthy subjects were analyzed. 25 subjects were reserved for training, 10 were reserved for validation, and 18 were reserved for testing. From each subject, 100 axial cross-sections containing brain tissues were selected. Acquisition parameters were as follows. T₁-weighted images: TE = 4.603ms, TR = 9.813ms, spatial resolution = 0.94×0.94×1.2mm³. T₂-weighted images: TE = 100ms, TR = 8178.34ms, spatial resolution = 0.94×0.94×1.2mm³. PD-weighted images: TE = 8ms, TR = 8178.34ms, spatial resolution = 0.94×0.94×1.2mm³. The multi-contrast images in this dataset were unregistered. Hence, T₂- and PD-weighted images were spatially registered onto T₁-weighted images prior to modelling. Registration was performed via an affine transformation in FSL [147] based on mutual information.

3.3.2.2 BRATS Dataset

T₁-weighted, T₂-weighted, post-contrast T₂-weighted, and T₂ Fluid Attenuation Inversion Recovery (FLAIR) brain MR images from 55 subjects were analyzed. 25 subjects were reserved for training, 10 were reserved for validation, and 20

were reserved for testing. From each subject, 100 axial cross-sections containing brain tissues were selected. Please note that the BRATS dataset contains images collected under various clinical protocols and scanners at multiple institutions. As publicly shared, multi-contrast images are co-registered to the same anatomical template, interpolated to $1\times1\times1\text{mm}^3$ resolution and skull-stripped.

3.3.2.3 MRI-CT Dataset

T₂-weighted MR and CT images of the male pelvis from 15 subjects were used. 9 subjects were reserved for training, 2 were reserved for validation, and 4 were reserved for testing. From each subject, 90 axial cross-sections were analysed. Acquisition parameters were as follows. T₂-weighted images: Group 1, TE = 97ms, TR = 6000-6600ms, spatial resolution = $0.875\times0.875\times2.5\text{mm}^3$. Group 2, TE = 91-102ms, TR = 12000-16000ms, spatial resolution = $0.875\times1.1\times0.875\times2.5\text{mm}^3$. CT images: Group 1, spatial resolution = $0.98\times0.98\times3\text{mm}^3$, Kernel = B30f. Group 2: spatial resolution = $0.1\times0.1\times2\text{mm}^3$, Kernel = FC17. This dataset contains images collected under various protocols and scanners for each modality. As publicly shared, multi-modal images are co-registered onto T₂-weighted MR scans.

3.3.3 Competing Methods

We demonstrated the proposed ResViT model against several state-of-the-art image synthesis methods. The baseline methods included convolutional models (task-specific models: pGAN [53], pix2pix [73], medSynth [55]; unified models: MM-GAN [64], pGAN_{uni}), attention-augmented convolutional models (A-UNet [128], SAGAN [127]), and transformer models (task-specific models: TransUNet [99], PTNet [140]; unified: TransUNet_{uni}). Hyperparameters of each competing method were optimized via identical cross-validation procedures.

3.3.3.1 Convolutional models

pGAN A convolutional GAN model with ResNet backbone was considered [53]. pGAN comprises CNN-based generator and discriminator networks. Its generator follows an encoder-bottleneck-decoder pathway, where the encoder and decoder are identical to those in ResViT. The bottleneck contains a cascade of residual CNN blocks.

pix2pix A convolutional GAN model with U-Net backbone was considered [73]. pix2pix has a CNN-based generator with an encoder-decoder structure tied with skip connections.

medSynth A convolutional GAN model with residual U-Net backbone was considered as provided at <https://github.com/ginobilinie/medSynthesisV1> [55]. The generator of medSynth contains a long-range skip connection to connect the input to the network output.

MM-GAN A unified synthesis model based on a convolutional GAN was considered [64]. MM-GAN comprises CNN-based generator and discriminator networks, where the generator is based on U-Net. MM-GAN trains a single network under various source-target modality configurations. The original MM-GAN architecture was directly adopted, except for curriculum learning to ensure standard sample selection for all competing methods. The unification strategy in MM-GAN matches the unification strategy in ResViT.

pGAN_{uni} A unified version of the pGAN model was trained to consolidate multiple synthesis tasks. The unification procedure was identical to that of ResViT.

3.3.3.2 Attention-augmented convolutional models

Attention U-Net (A-UNet) A CNN-based U-Net architecture with additive attention gates was considered [128]. Here we adopted the original A-UNet model as the generator of a conditional GAN model, where the discriminator was identical to that in ResViT.

Self-Attention GAN (SAGAN) A CNN-based GAN model with self-attention modules incorporated into the generator was considered [127]. Here we adapted the original SAGAN model designed for unconditional mapping by inserting the self-attention modules into the pGAN model as described in [148]. For fair comparison, the number and position of attention modules in SAGAN were matched to those of transformer modules in ResViT.

3.3.3.3 Transformer models

TransUNet A recent hybrid CNN-transformer architecture was considered [99]. Here, we adopted the original TransUNet model as the generator of a conditional GAN architecture with an identical discriminator to ResViT. We further replaced the segmentation head with a convolutional layer for synthesis.

PTNet A recent convolution-free transformer architecture was considered [140]. Here we adopted the original PTNet model based on a non-adversarial architecture without a discriminator, and trained via an L_2 -norm pixel-wise loss (i.e., mean-squared error).

TransUNet_{uni} The TransUNet model was unified to consolidate multiple synthesis tasks. The unification procedure was identical to that of ResViT.

3.3.4 Architectural Details

The encoder in the ResViT model contained three convolutional layers of kernel size 7, 3, 3 respectively. The feature map in the encoder output was of size $\mathbb{R}^{256,64,64}$, and this dimensionality was retained across the information bottleneck. The decoder contained three convolutional layers of kernel size 3, 3, 7 respectively. The information bottleneck contained nine ART blocks. The downsampling blocks preceding transformers contained two convolutional layers with stride 2 and kernel size 3. The upsampling blocks succeeding transformers contained two transposed convolutional layers with stride 2 and kernel size 3. Down and upsampling factors were set to $M = 4$. Channel compression lowered the number of channels from 512 to 256. The transformer encoder was adopted by extracting the transformer component of the ImageNet-pretrained model R50+ViT-B/16 (https://github.com/google-research/vision_transformer). The transformer encoder expected an input map of 16×16 spatial resolution. Patch flattening was performed with size $P = 1$ yielding a sequence length of 256 [98]. Note that transformer modules contain substantially higher number of parameters compared to convolutional modules. Thus, retaining a transformer in each ART block results in significant model complexity, inducing computational burden and suboptimal learning. To alleviate these issues, transformer modules in ART blocks utilized tied weights, and they were only retained in a subset of ART blocks while remaining blocks reduced to residual CNNs.

The configuration of transformer modules, i.e. their total number and position, was selected via cross-validation experiments. Due to the extensive number of potential configurations, a pre-selection process was implemented. Accordingly, performance for a transformer module inserted in a single ART block ($A_1, A_2, \dots, A_9$) was measured, and the top half of positions was pre-selected. Composite configurations with multiple transformer modules were then formed based on the pre-selected blocks ($A_1 - A_5, A_1 - A_6 - A_9$ etc.). We observed that retaining more than 2 modules elevated complexity without any performance benefits. Validation performance for the best performing configurations

Configuration	$T_1, T_2 \rightarrow PD$	$T_1, T_2 \rightarrow FLAIR$	$MRI \rightarrow CT$
	PSNR	PSNR	PSNR
$A_1 - A_5$	33.23	24.82	26.40
$A_1 - A_6$	33.34	24.88	26.56
$A_1 - A_9$	33.27	24.77	26.58
$A_4 - A_9$	33.11	24.63	26.19
$A_5 - A_9$	33.05	24.65	26.27
$A_1 - A_6 - A_9$	32.89	24.82	26.20

Table 3.1: Validation performance of candidate ResViT configurations in representative synthesis tasks. Performance is taken as PSNR (dB) between synthesized and reference target images. A_i denotes the presence of a transformer module in the i th ART block.

Transformer size	$T_1, T_2 \rightarrow PD$	$T_1, T_2 \rightarrow FLAIR$	$MRI \rightarrow CT$
	PSNR	PSNR	PSNR
Base	33.34	24.88	26.56
Large	33.14	24.60	26.46

Table 3.2: Validation performance of ResViT models with varying sizes of transformer modules in representative synthesis tasks.

($A_1 - A_5$, $A_1 - A_6$, $A_1 - A_9$, $A_5 - A_9$, $A_4 - A_9$, $A_5 - A_9$, $A_1 - A_6 - A_9$) are listed in Table 3.1 for three representative tasks ($T_1, T_2 \rightarrow PD$ in IXI, $T_1, T_2 \rightarrow FLAIR$ in BRATS, and $MRI \rightarrow CT$ in MRI-CT). Consistently across tasks, the ($A_1 - A_6$) configuration yielded near-optimal performance and so it was selected for all experiments thereafter.

We also tuned the intrinsic complexity of transformer modules. To do this, two variant modules were examined: "base" and "large". The "base" module contained 12 layers with latent dimensionality $N_d = 768$, 12 attention heads, and 3073 hidden units in each layer of the *MLP*. Meanwhile, the "large" module contained 24 layers with latent dimensionality $N_d = 1024$, 16 attention heads and 4096 hidden units in each layer of the *MLP*. Validation performances based on the two variant modules are listed in Table 3.2. The "base" module that offers higher performance for lower computational complexity was selected for consequent experiments.

	ResViT	pGAN	pix2pix	medSynth	A-UNet	SAGAN	TransUNet	PTNet
Time	98	60	60	81	70	63	78	224

Table 3.3: Average inference times (msec) of synthesis models per single cross-section.

3.3.5 Modeling Procedures

All adversarial models were trained with the same PatchGAN discriminator. Task-specific models used adversarial and pixel-wise losses, whereas unified models used adversarial, pixel-wise, and reconstruction losses. The only exceptions were PTNet that excluded the adversarial loss, and medSynth that includes an additional gradient loss. For fair comparisons, identical optimization procedures and loss-term weighting were employed for all competing methods. Learning rate schedules, number of epochs, and loss-term weighting were selected via cross-validation. We trained multiple models for different number of epochs (in the set $\{50, 100, 150, 200\}$), different values of learning rate (in the set $\{10^{-5}, 10^{-4}, 2 \times 10^{-4}, 5 \times 10^{-4}, 10^{-3}\}$) and weightings of the loss functions (λ_{pix} and λ_{rec} in the set $\{20, 50, 100, 150\}$). We looked at the synthesis performance of the models trained with different hyperparameters on the validation set. Following selected parameters consistently yielded near-optimal performance in all methods. For all competing methods, Training was performed via the Adam optimizer [149] with $\beta_1 = 0.5$, $\beta_2 = 0.999$. Models were trained for 100 epochs, as further training did not yield any performance improvement. The learning rate was 2×10^{-4} for the first 50 epochs and linearly decayed to 0 in the remaining epochs. Loss-term weights were $\lambda_{adv} = 1$, $\lambda_{pix} = 100$ for task-specific models, and $\lambda_{adv} = 1$, $\lambda_{rec} = 100$, $\lambda_{pix} = 100$ for unified models. Transformer modules in both TransUNet and ResViT were initiated with pre-trained versions for object classification on ImageNet [150]. ART blocks were initiated without transformer modules and then fine-tuned for 50 epochs following insertion of transformers. Modelling was performed via the PyTorch framework on Nvidia RTX A4000 GPUs. Inference times are listed in Table 4.5.

Synthesis quality was assessed via Peak Signal to Noise Ratio (PSNR), and

Structural Similarity Index (SSIM) [151]. Metrics were calculated between ground truth and synthesized target images. Mean and standard deviations of metrics were reported across an independent test set, non-overlapping with training-validation sets. Significance of performance differences were evaluated with signed-rank tests ($p < 0.05$). Tests were conducted on subject-average metrics, except MRI $\rightarrow$ CT where cross-sectional metrics were tested in each subject due to limited number of test subjects.

3.3.6 Experiments

3.3.6.1 Multi-Contrast MRI Synthesis

Experiments were conducted on the IXI and BRATS datasets to demonstrate synthesis performance in multi-modal MRI. In the IXI dataset, one-to-one tasks of $T_2 \rightarrow PD$; $PD \rightarrow T_2$ and many-to-one tasks of $T_1, T_2 \rightarrow PD$; $T_1, PD \rightarrow T_2$; $T_2, PD \rightarrow T_1$ were considered. In the BRATS dataset, one-to-one tasks of $T_2 \rightarrow FLAIR$; $FLAIR \rightarrow T_2$, many-to-one tasks of $T_1, T_2 \rightarrow FLAIR$; $T_1, FLAIR \rightarrow T_2$; $T_2, FLAIR \rightarrow T_1$ were considered. In both datasets, task-specific ResViT models were compared against pGAN, pix2pix, medSynth, A-UNet, SAGAN, TransUNet, and PTNet. Meanwhile, unified ResViT models were demonstrated against $pGAN_{uni}$, MM-GAN, and $TransUNet_{uni}$.

3.3.6.2 MRI to CT Synthesis

Experiments were performed on the MRI-CT dataset to demonstrate across-modality synthesis performance. A one-to-one synthesis task of deriving target CT images from source MR images was considered. The task-specific ResViT model was compared against pGAN, pix2pix, medSynth, A-UNet, SAGAN, TransUNet, and PTNet.

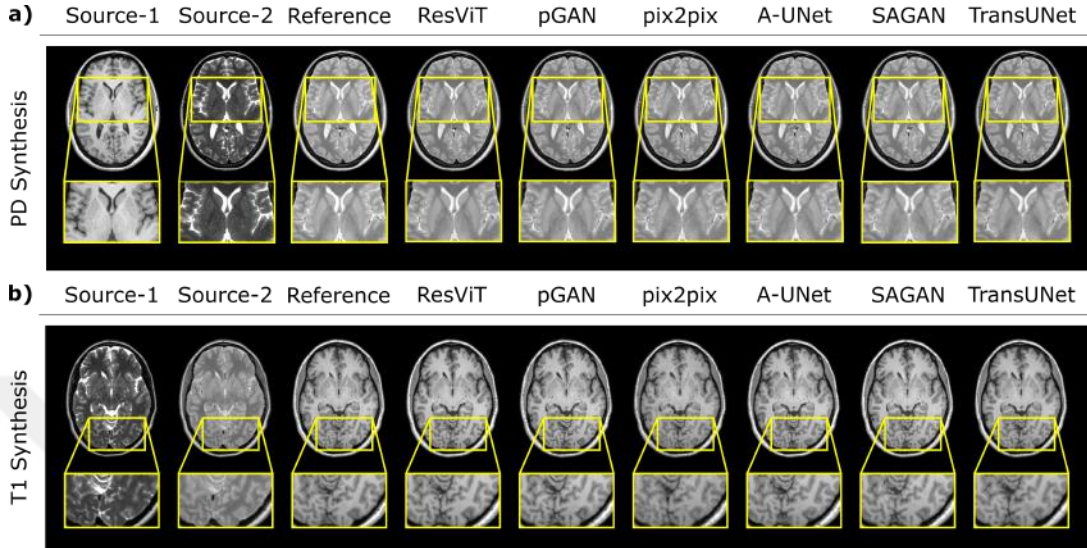


Figure 3.3: ResViT was demonstrated on the IXI dataset for two representative many-to-one synthesis tasks: a) $T_1, T_2 \rightarrow \text{PD}$, b) $T_2, \text{PD} \rightarrow T_1$. Synthesized images from all competing methods are shown along with the source images and the reference target image. ResViT improves synthesis performance in regions that are depicted sub-optimally in competing methods. Overall, ResViT generates images with lower artifact and noise levels and sharper tissue depiction.

3.3.6.3 Ablation Studies

Several lines of ablation experiments were conducted to demonstrate the value of the individual components of the ResViT model, including both architectural design elements and training strategies. Experiments were performed on three representative tasks: namely $T_1, T_2 \rightarrow \text{PD}$ in IXI, $T_1, T_2 \rightarrow \text{FLAIR}$ in BRATS, and $\text{MRI} \rightarrow \text{CT}$.

First, we assessed the performance contribution of the three main components in ResViT: transformer modules, convolutional modules and adversarial learning. Variant models were trained when transformer modules were ablated from ART blocks, when residual CNNs were ablated from transformer-retaining ART blocks, and when the adversarial loss term and the discriminator were ablated. In addition to PSNR and SSIM, we additionally measured the Fréchet inception distance (FID) [152] between the synthesized and ground truth images to evaluate the importance of adversarial learning.

	$T_1, T_2 \rightarrow PD$		$T_1, PD \rightarrow T_2$		$T_2, PD \rightarrow T_1$		$T_2 \rightarrow PD$		$PD \rightarrow T_2$	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
ResViT	33.92	97.72	35.71	97.73	29.58	95.21	32.90	97.19	34.24	97.24
	± 1.44	± 0.41	± 1.20	± 0.005	± 1.37	± 0.011	± 1.20	± 0.005	± 1.09	± 0.005
pGAN	32.91	0.966	33.95	0.965	28.71	0.941	32.20	0.963	33.05	0.963
	± 0.94	± 0.005	± 1.06	± 0.006	± 1.08	± 0.013	± 1.00	± 0.005	± 0.95	± 0.007
pix2pix	32.25	0.974	33.62	0.973	28.35	0.949	30.72	0.956	30.74	0.950
	± 1.24	± 0.006	± 1.31	± 0.009	± 1.24	± 0.016	± 1.28	± 0.007	± 1.63	± 0.012
medSynth	33.23	0.967	32.66	0.963	28.43	0.938	32.20	0.964	30.41	0.956
	± 1.09	± 0.005	± 1.30	± 0.007	± 1.01	± 0.013	± 1.10	± 0.006	± 3.98	± 0.025
A-UNet	32.24	96.28	32.43	95.89	28.95	91.64	32.05	96.99	33.32	95.08
	± 0.92	± 1.35	± 1.36	± 0.74	± 1.21	± 1.29	± 1.04	± 0.91	± 1.08	± 0.72
SAGAN	32.50	96.43	33.71	96.49	28.62	94.19	32.07	96.31	32.96	96.22
	± 0.93	± 0.47	± 1.00	± 0.56	± 1.10	± 1.27	± 0.98	± 0.62	± 1.01	± 0.65
TransUNet	32.53	95.76	32.49	96.02	28.21	94.11	30.90	96.03	31.73	95.81
	± 0.97	± 0.52	± 1.18	± 0.79	± 1.30	± 1.33	± 1.35	± 0.57	± 1.44	± 0.75
PTNet	32.57	96.82	33.23	96.22	28.15	94.75	31.46	96.36	31.79	95.90
	± 1.18	± 0.51	± 1.85	± 1.61	± 2.22	± 1.63	± 1.50	± 0.67	± 2.25	± 1.88

Table 3.4: Performance of task-specific synthesis models in many-to-one ($T_1, T_2 \rightarrow PD$, $T_1, PD \rightarrow T_2$, and $T_2, PD \rightarrow T_1$) and one-to-one ($T_2 \rightarrow PD$ and $PD \rightarrow T_2$) tasks in the IXI dataset. PSNR (dB) and SSIM (%) are listed as mean $\pm$ std across test subjects. Boldface indicates the top-performing model for each task.

Second, we probed the design and training procedures of ART blocks. We assessed the utility of tied weights across transformer modules, and multiple transformer-retaining ART blocks. Variant models were trained separately using untied weights in transformers, and based on a single transformer-retraining module at either first or sixth ART blocks. We also examined the importance of model initiation with ImageNet pre-trained transformer modules, and delayed insertion of transformer modules during training. Variant models were built by using randomly initialized transformer modules, by inserting pre-trained transformer modules into ART blocks at the beginning of training, and by inserting randomly initialized transformer modules at the beginning of training.

Third, we investigated the design of skip connections and down/upsampling modules. We considered benefits of additional long-range skip connections that bypass the bottleneck to propagate shallow features, and existing skip connections

	$T_1, T_2 \rightarrow \text{FLAIR}$		$T_1, \text{FLAIR} \rightarrow T_2$		$T_2, \text{FLAIR} \rightarrow T_1$		$T_2 \rightarrow \text{FLAIR}$		$\text{FLAIR} \rightarrow T_2$	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
ResViT	25.84 ± 1.13	0.886 ± 0.014	26.90 ± 1.20	0.938 ± 0.011	26.20 ± 1.31	0.924 ± 0.009	24.97 ± 1.07	0.870 ± 0.014	25.78 ± 0.92	0.908 ± 0.015
pGAN	24.89 ± 1.10	0.867 ± 0.015	26.51 ± 1.13	0.922 ± 0.012	25.72 ± 1.54	0.918 ± 0.011	24.01 ± 1.15	0.864 ± 0.014	25.09 ± 1.52	0.894 ± 0.015
pix2pix	24.31 ± 1.21	0.862 ± 0.015	26.12 ± 1.53	0.920 ± 0.012	25.80 ± 1.72	0.918 ± 0.011	23.15 ± 1.93	0.869 ± 0.016	24.52 ± 0.88	0.883 ± 0.014
medSynth	23.93 ± 1.45	0.863 ± 0.016	26.44 ± 0.76	0.921 ± 0.011	25.72 ± 1.62	0.914 ± 0.012	23.36 ± 1.88	0.864 ± 0.017	24.41 ± 0.82	0.888 ± 0.014
A-UNet	24.36 ± 1.24	0.857 ± 0.017	26.48 ± 1.21	0.924 ± 0.012	25.67 ± 1.35	0.918 ± 0.010	23.69 ± 1.57	0.873 ± 0.015	24.56 ± 0.94	0.891 ± 0.014
SAGAN	24.62 ± 1.17	0.869 ± 0.014	26.41 ± 1.22	0.919 ± 0.012	25.91 ± 1.42	0.918 ± 0.011	24.02 ± 1.35	0.860 ± 0.015	25.10 ± 0.88	0.893 ± 0.014
TransUNet	24.34 ± 1.26	0.872 ± 0.014	26.51 ± 0.92	0.920 ± 0.010	25.76 ± 1.69	0.921 ± 0.011	23.70 ± 1.75	0.864 ± 0.015	24.62 ± 0.81	0.891 ± 0.015
PTNet	22.18 ± 2.18	0.864 ± 0.028	25.16 ± 2.05	0.930 ± 0.020	22.41 ± 1.95	0.920 ± 0.014	19.12 ± 1.69	0.844 ± 0.018	23.99 ± 1.03	0.901 ± 0.016

Table 3.5: Performance of task-specific synthesis models in many-to-one tasks ($T_1, T_2 \rightarrow \text{FLAIR}$, $T_1, \text{FLAIR} \rightarrow T_2$, and $T_2, \text{FLAIR} \rightarrow T_1$) and one-to-one tasks ($T_2 \rightarrow \text{FLAIR}$ and $\text{FLAIR} \rightarrow T_2$) across test subjects in the BRATS dataset. Boldface indicates the top-performing model for each task.

in ART blocks for residual learning. Variant models were trained by introducing skip connections between respective layers of the encoder and decoder, and by removing skip connections around either the transformer or convolution modules in ART. We also assessed alternative designs for down/upsampling modules in ART to mitigate added model complexity. In a first variant, original down/upsampling modules were replaced with unlearned maxpooling modules for downsampling and bilinear interpolation modules for upsampling. In a second variant, additional downsampling layers in the encoder and upsampling layers in the decoder were included in order to remove down/upsampling modules in ART blocks.

Next, we inspected the relative strength of contextual features in the distilled task-relevant representations in ART blocks. For a quantitative assessment, we compared the L_2 -norm of the contextual feature map derived by the transformer module against that of the input feature map to the ART block relayed through the transformer’s skip connection. Note that these two maps are distilled via

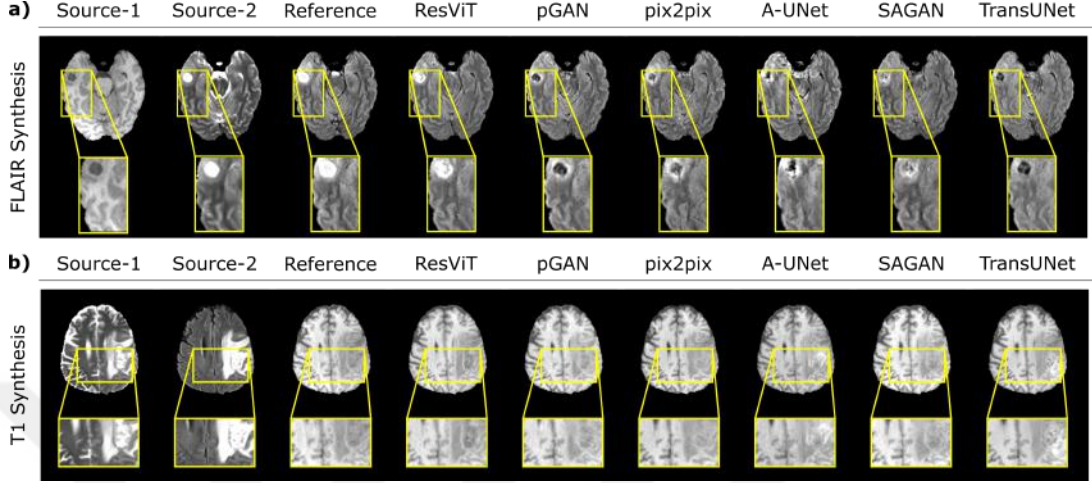


Figure 3.4: ResViT was demonstrated on the BRATS dataset for two representative many-to-one synthesis tasks: a) $T_1, T_2 \rightarrow \text{FLAIR}$, b) $T_2, \text{FLAIR} \rightarrow T_1$. Synthesized images from all competing methods are shown along with the source images and the reference image. ResViT improves synthesis performance, especially in pathological regions (e.g., tumors, lesions) in comparison to competing methods. Overall, ResViT images have better-delineated tissue boundaries and lower artifact/noise levels.

the channel compression (CC) module following concatenation. Thus, we also compared the L_2 -norm of the combination weights in the CC module for the contextual versus input features.

To interpret the information that self-attention mechanisms focus on during synthesis tasks, we computed and visualized the attention maps as captured by the transformer modules in ResViT. Attention maps were calculated based on the Attention Rollout technique, and a single average map was extracted for a given transformer module [153].

3.4 Results

3.4.1 Multi-Contrast MRI Synthesis

3.4.1.1 Task-Specific Synthesis Models

We demonstrated the performance of ResViT in learning task-specific synthesis models for multi-contrast MRI. ResViT was compared against convolutional models (pGAN, pix2pix, medSynth), attention-augmented CNNs (A-UNet, SAGAN), and recent transformer architectures (TransUNet, PTNet). First, brain images of healthy subjects in the IXI dataset were considered. PSNR and SSIM metrics are listed in Table 3.4 for many-to-one and one-to-one tasks. ResViT achieves the highest performance in both many-to-one ($p < 0.05$) and one-to-one tasks ($p < 0.05$). On average, ResViT outperforms convolutional models by 1.71dB PSNR and 1.08% SSIM, attention-augmented models by 1.40dB PSNR and 1.45% SSIM, and transformer models by 1.96dB PSNR and 1.13% SSIM ($p < 0.05$). Representative images for $T_1, T_2 \rightarrow \text{PD}$ and $T_2, \text{PD} \rightarrow T_1$ are displayed in Fig. 4.3a,b. Compared to baselines, ResViT synthesizes target images with lower artifact levels and sharper tissue depiction.

We then demonstrated task-specific ResViT models on the BRATS dataset containing images of glioma patients. PSNR and SSIM metrics are listed in Table 3.5 for many-to-one and one-to-one tasks. ResViT again achieves the highest performance in many-to-one ($p < 0.05$) and one-to-one tasks ($p < 0.05$), except $T_2 \rightarrow \text{FLAIR}$ where A-UNet has slightly higher SSIM. On average, ResViT outperforms convolutional models by 1.01dB PSNR and 1.41% SSIM, attention-augmented models by 0.84dB PSNR and 1.24% SSIM, and transformer models by 2.16dB PSNR and 1.25% SSIM ($p < 0.05$). Note that the BRATS dataset contains pathology with large across-subject variability. As expected, attention-augmented models show relative benefits against pure convolutional models, yet ResViT that explicitly models contextual relationships still outperforms all baselines. Representative target images for $T_1, T_2 \rightarrow \text{FLAIR}$ and $T_2, \text{FLAIR} \rightarrow T_1$ are

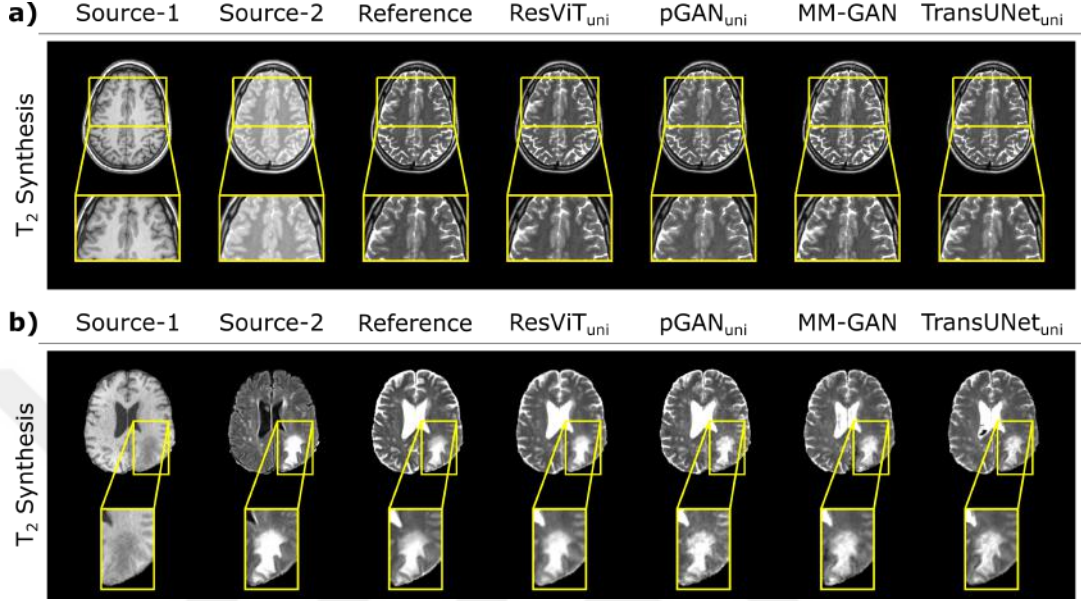


Figure 3.5: ResViT_{uni} was demonstrated against other unified models on brain MRI datasets for two representative tasks: a) T₁, PD $\rightarrow$ T₂ in IXI, b) T₁, FLAIR $\rightarrow$ T₂ in BRATS. Synthesized images from all competing methods are shown along with the source images and the reference target image. ResViT_{uni} improves synthesis performance especially in pathological regions (tumors, lesions) in comparison to competing methods. Overall, ResViT_{uni} generates images with lower artifact and noise levels and more accurate tissue depiction for tasks in both datasets.

displayed in Fig. 3.4a,b, respectively. Compared to baselines, ResViT synthesizes target images with lower artifact levels and sharper tissue depiction. Importantly, ResViT reliably captures brain lesions in patients in contrast to competing methods with inaccurate depictions including TransUNet.

Superior depiction of pathology in ResViT signals the importance of ART blocks in simultaneously maintaining local precision and contextual consistency in medical image synthesis. In comparison, transformer-based TransUNet and PTNet yield relatively limited synthesis quality which might be attributed to several fundamental differences between the models. TransUNet uses only a transformer in its bottleneck while propagating shallow convolutional features via encoder-decoder skip connections, and its decoder increases spatial resolution via bilinear upsampling that might be ineffective in suppressing high-frequency

artifacts [154]. In contrast, ResViT continues encoding and propagating convolutional features across the information bottleneck to create a deeper feature representation, and it employs transposed convolutions within upsampling modules to mitigate potential artifacts. PTNet is a convolution-free architecture that relies solely on self-attention operators that have limited localization ability [98], and it is a non-adversarial model trained via MSE loss that can lead to characteristic loss of structural details [73]. Instead, ResViT is a hybrid CNN-transformer architecture sensitive to local and contextual features, and it employs adversarial learning to improve sensitivity to detailed tissue structure.

3.4.1.2 Unified Synthesis Models

Task-specific models are trained and tested to perform a single synthesis task to improve performance, but a separate model has to be built for each task. Next, we demonstrated ResViT in learning unified synthesis models for multi-contrast MRI. A unified ResViT (ResViT_{uni}) was compared against unified convolutional (pGAN_{uni}, MM-GAN) and transformer models (TransUNet_{uni}). Performance of unified models were evaluated at test time on many-to-one tasks in IXI (Table 3.6) and BRATS (Table 3.7). ResViT_{uni} maintains the highest performance in many-to-one tasks in both IXI ($p < 0.05$) and BRATS ($p < 0.05$). In IXI, ResViT_{uni} outperforms pGAN_{uni} by 1.12dB PSNR and 0.70% SSIM, MM-GAN by 2.37dB PSNR and 1.80% SSIM, and TransUNet_{uni} by 2.69dB PSNR and 1.67% SSIM ($p < 0.05$). In BRATS, ResViT outperforms pGAN_{uni} by 0.74dB PSNR and 0.93% SSIM, MM-GAN by 0.77dB PSNR and 0.90% SSIM, and TransUNet_{uni} by 1.08dB PSNR and 1.43% SSIM ($p < 0.05$). Representative target images are displayed in Fig. 3.5. ResViT synthesizes target images with lower artifacts and sharper depiction than baselines. These results suggest that a unified ResViT model can successfully consolidate models for varying source-target configurations.

	$T_1, T_2 \rightarrow PD$		$T_1, PD \rightarrow T_2$		$T_2, PD \rightarrow T_1$	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
ResViT _{uni}	33.22	0.971	33.97	0.968	28.80	0.946
	± 1.21	± 0.005	± 1.04	± 0.006	± 1.20	± 0.013
pGAN _{uni}	31.86	0.965	32.90	0.962	27.86	0.937
	± 1.09	± 0.005	± 0.91	± 0.006	± 1.04	± 0.014
MM-GAN	30.73	0.955	30.91	0.951	27.23	0.925
	± 1.16	± 0.006	± 1.61	± 0.013	± 1.24	± 0.015
TransUNet _{uni}	30.30	0.956	30.77	0.949	26.86	0.930
	± 1.44	± 0.007	± 1.10	± 0.014	± 1.16	± 0.013

Table 3.6: Performance of unified synthesis models in many-to-one tasks ($T_1, T_2 \rightarrow PD$, $T_1, PD \rightarrow T_2$, and $T_2, PD \rightarrow T_1$) across test subjects in the IXI dataset. Boldface indicates the top-performing model for each task.

	$T_1, T_2 \rightarrow FLAIR$		$T_1, FLAIR \rightarrow T_2$		$T_2, FLAIR \rightarrow T_1$	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
ResViT _{uni}	25.32	0.876	26.81	0.921	26.24	0.922
	± 0.91	± 0.015	± 1.04	± 0.012	± 1.65	± 0.010
pGAN _{uni}	24.46	0.865	26.23	0.914	25.46	0.912
	± 0.99	± 0.014	± 1.08	± 0.012	± 1.20	± 0.009
MM-GAN	24.20	0.861	26.10	0.915	25.75	0.916
	± 1.34	± 0.015	± 1.48	± 0.014	± 1.64	± 0.011
TransUNet _{uni}	24.11	0.863	26.05	0.912	24.96	0.901
	± 1.19	± 0.014	± 1.46	± 0.013	± 1.24	± 0.012

Table 3.7: Performance of unified synthesis models in many-to-one tasks ($T_1, T_2 \rightarrow FLAIR$, $T_1, FLAIR \rightarrow T_2$, and $T_2, FLAIR \rightarrow T_1$) across test subjects in the BRATS dataset. Boldface indicates the top-performing model for each task.

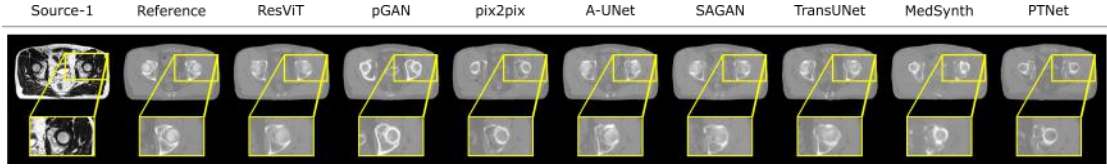


Figure 3.6: ResViT was demonstrated on the pelvic MRI-CT dataset for the T_2 -weighted MRI $\rightarrow$ CT task. Synthesized images from all competing methods are shown along with the source and reference images. ResViT enhances synthesis of relevant morphology in the CT domain as evidenced by the elevated accuracy near bone structures.

3.4.2 Across-Modality Synthesis

We also demonstrated ResViT in across-modality synthesis. T_2 -weighted MRI and CT images in the pelvic dataset were considered. ResViT was compared against pGAN, pix2pix, medSynth, A-UNet, SAGAN, TransUNet, and PTNet. PSNR and SSIM metrics are listed in Table 4.4. ResViT yields the highest performance in each subject ($p < 0.05$). On average, ResViT outperforms convolutional models by 1.89dB PSNR and 3.20% SSIM, attention-augmented models by 0.75dB PSNR and 1.95% SSIM, and transformer models by 1.76dB PSNR and 1.5% SSIM ($p < 0.05$). Representative target images are displayed in Fig. 4.5. Compared to baselines, ResViT synthesizes target images with lower artifacts and more accurate tissue depiction. Differently from multi-contrast MRI, attention-augmented models and TransUNet offer more noticeable performance benefits over convolutional models. That said, ResViT still maintains further elevated performance, particularly near bone structures in CT images. This finding suggests that the relative importance of contextual representations is higher in MRI-CT synthesis. With the help of its residual transformer blocks, ResViT offers reliable performance with accurate tissue depiction in this task.

3.4.3 Ablation Studies

We performed a systematic set of experiments to demonstrate the added value of the main components and training strategies used in ResViT. First, we compared ResViT against ablated variants where the convolutional modules in ART blocks, transformer modules in ART blocks, or the adversarial term in training

loss were separately removed. Table 3.9 lists performance metrics in the test set for three representative synthesis tasks. Consistently across tasks, ResViT yields optimal or near-optimal performance. ResViT achieves higher PSNR and SSIM in representative tasks compared to variants without transformer or convolutional modules ($p < 0.05$). It also yields lower FID than these variants, except in MRI $\rightarrow$ CT where ablation of the convolutional module slightly decreases FID. Importantly, ResViT maintains notably lower FID compared to the variant without adversarial loss (albeit slightly lower SSIM in $T_1, T_2 \rightarrow \text{FLAIR}$ and PSNR in MRI $\rightarrow$ CT). This is expected since FID is generally considered as a more suited metric to examine the perceptual benefits of adversarial learning than PSNR or SSIM that reflect heavier influence from relatively lower frequencies [152]. Representative synthesized images are also displayed in Fig. 3.7a. ResViT images more closely mimic the reference images, and show greater spatial acuity compared against the variant without adversarial loss. Taken together, these results indicate that adversarial learning enables ResViT to more closely capture the distributional properties of target-modality images.

Second, we compared ResViT against ablated variants where the weight tying procedure across transformer modules was neglected, or transformer modules in one of the two retaining ART blocks were removed. Table 3.10 lists performance metrics in the test set. ResViT yields higher performance than all variants ($p < 0.05$). These results demonstrate the added value of the weight tying procedure and the transformer configuration in ResViT. We also compared ResViT against ablated variants where the pre-training of transformer modules or their delayed insertion during training were selectively neglected, as listed in Table 3.11.

		ResViT	pGAN	pix2pix	medSynth	A-UNet	SAGAN	TransUNet	PTNet
MRI $\rightarrow$ CT	PSNR	28.45	26.80	26.53	26.36	27.80	27.61	27.76	25.63
		± 1.35	± 0.90	± 0.45	± 0.63	± 0.63	± 1.02	± 1.03	± 1.40
	SSIM	0.931	0.905	0.898	0.894	0.913	0.910	0.914	0.918
		± 0.009	± 0.008	± 0.004	± 0.009	± 0.004	± 0.006	± 0.009	± 0.009

Table 3.8: Performance for the across-modality synthesis task (T_2 -weighted MRI $\rightarrow$ CT) across test subjects in the pelvic MRI-CT dataset. Boldface indicates the top-performing model for each task.

	T ₁ , T ₂ → PD			T ₁ , T ₂ → FLAIR			MRI → CT		
	PSNR	SSIM	FID	PSNR	SSIM	FID	PSNR	SSIM	FID
ResViT	33.92	0.977	14.47	25.84	0.886	18.58	28.45	0.932	60.28
	±1.44	±0.004		±1.13	±0.014		±1.35	±0.009	
w/o trans.	32.91	0.966	14.56	24.96	0.868	19.21	26.73	0.899	95.38
modules	±0.96	±0.005		±1.10	±0.005		±0.91	±0.008	
w/o conv.	33.49	0.971	14.84	25.11	0.874	20.30	28.19	0.922	60.16
modules	±1.34	±0.005		±1.02	±0.014		±1.15	±0.009	
w/o adv.	33.75	0.977	15.80	22.95	0.891	40.68	28.58	0.932	65.49
loss	±1.45	±0.005		±1.93	±0.015		±1.13	±0.007	

Table 3.9: Test performance of ResViT and variants ablated of transformer modules, convolutional modules or adversarial loss. FID is a summary metric across the entire test set. Boldface indicates the top-performing model for each task.

	T ₁ , T ₂ → PD		T ₁ , T ₂ → FLAIR		MRI → CT	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
$A_1 - A_6$	33.92	0.977	25.84	0.886	28.45	0.931
	±1.44	±0.004	±1.13	±0.014	±1.35	±0.009
$A_1 - A_6$ (untied weights)	33.72	0.973	25.19	0.879	28.16	0.923
	±1.23	±0.005	±1.18	±0.014	±1.11	±0.007
A_1	33.51	0.971	24.98	0.883	28.06	0.921
	±1.15	±0.005	±1.60	±0.015	±1.31	±0.008
A_6	33.78	0.977	25.25	0.880	27.95	0.921
	±1.34	±0.004	±1.20	±0.014	±1.22	±0.008

Table 3.10: Test performance of ResViT ($A_1 - A_6$) and variants ablated of weight tying and individual transformer modules. Boldface indicates the top-performing model for each task.

Our results indicate that both pre-training and delayed insertion of transformers improve synthesis performance in representative tasks ($p < 0.05$).

Third, we examined the utility of the skip connections and down/upsampling blocks in the proposed architecture. We compared ResViT against variants built by incorporating additional skip connections between corresponding layers of the encoder and decoder, by removing the skip connection around the transformer module in ART blocks, and by removing the skip connection around the CNN module in ART blocks. Table 3.12 lists performance metrics in the test set. ResViT yields higher performance than all variants ($p < 0.05$). Our

T ₁ , T ₂ → PD	T ₁ , T ₂ → PD		T ₁ , T ₂ → FLAIR		MRI → CT	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM height
	T ₁ , T ₂ → FLAIR		MRI → CT			
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
ResViT	33.92	0.977	25.84	0.886	28.45	0.931
	±1.44	±0.004	±1.13	±0.014	±1.35	±0.009
w/o transformer pre-training	33.55	0.971	24.86	0.881	27.94	0.922
	±1.25	±0.005	±1.28	± 0.016	±1.25	±0.009
w/o delayed insertion	33.35	0.977	24.89	0.873	28.01	0.924
	±1.13	±0.004	±1.18	±0.015	±1.27	±0.008
w/o transformer pre-training or delayed insertion	33.58	0.971	24.74	0.869	27.66	0.913
	±1.16	±0.005	±1.30	±0.016	±0.78	±0.006

Table 3.11: Test performance of ResViT and variants ablated of pre-training and delayed insertion procedures for transformers. Boldface indicates the top-performing model for each task.

results indicate that ResViT does not benefit from encoder-decoder skip connections, yet benefits substantially from residual learning in ART blocks. We also compared ResViT against variants built by replacing down/upsampling modules in ART blocks with unlearned maxpooling/bilinear interpolation modules, and by increasing encoder downsampling and decoder upsampling rates to remove down/upsampling modules in ART blocks entirely. ResViT outperforms all variants as listed in Table 3.12 ($p < 0.05$), except for MRI→CT where the variant with unlearned down/upsampling and ResViT yield similar SSIM. These results demonstrate the benefits of the proposed down/upsampling scheme in ResViT.

Next, we inspected the relative strength of transformer-derived contextual features in distilled representations within ART blocks. To do this, we computed the L_2 -norms of contextual feature maps output by the transformer module, and input feature maps from the previous ART block relayed through the skip connection of the transformer module. We also computed the relative weighting of the two feature maps as the L_2 -norms of respective combination weights in the channel compression (CC) module. The resulting measurements for ResViT models trained in representative tasks are listed in Table 3.13. We find that contextual and input feature maps, and the respective combination weights in CC blocks have comparable strength.

	T ₁ , T ₂ → PD		T ₁ , T ₂ → FLAIR		MRI → CT	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
ResViT	33.92	0.977	25.84	0.886	28.45	0.931
	±1.44	±0.43	±1.13	±1.42	±1.35	±0.88
with enc.-dec.	33.74	97.32	25.24	88.43	28.06	92.62
skip connections	±1.24	±0.52	±1.16	± 1.52	±0.93	±0.65
w/o skip around	28.24	94.23	25.02	86.37	26.94	90.59
conv. modules	±1.27	±0.91	±0.98	±1.53	±0.73	±0.72
w/o skip around	31.53	96.18	24.06	86.78	27.08	90.82
trans. modules	±1.26	± 0.61	±1.28	±1.44	±0.80	±0.63
ART with unlearned	33.73	96.89	25.33	88.43	28.16	93.12
down/upsampling	±1.19	±0.54	±1.11	±1.39	±1.04	±0.71
ART w/o	31.51	96.13	23.61	86.68	26.79	91.54
down/upsampling	±1.27	±0.63	±1.53	±1.52	±0.62	±0.63

Table 3.12: Test performance of ResViT and variants built by: adding encoder-decoder skip connections, removing skip connections in convolutional modules, removing skip connections in transformer modules, using unlearned down/upsampling blocks in ART, removing down/upsampling blocks in ART via a higher degree of down/upsampling in the encoder/decoder. Boldface indicates the top-performing model for each task.

	T ₁ , T ₂ → PD	T ₁ , T ₂ → FLAIR	MRI → CT
g_j	272.10	388.92	418.69
f_j	528.54	565.26	605.72
CC weights for g_j	135.02	114.12	97.11
CC weights for f_j	221.29	165.61	114.78

Table 3.13: Input feature maps and corresponding combination weights for the channel compression (CC) module were inspected in trained ResViT models. L_2 -norm of feature tensors from the transformer module (g_j), feature tensors from the previous ART block (f_j), combination weights for g_j , and combination weights for f_j are listed.

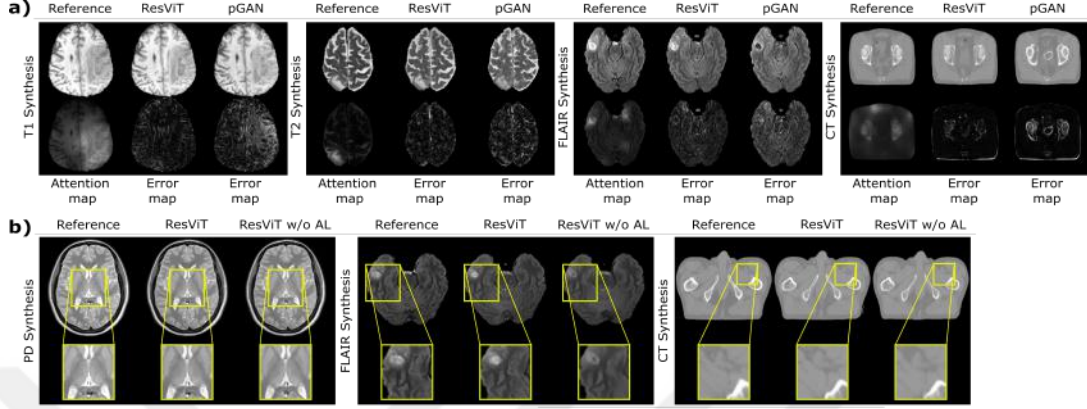


Figure 3.7: a) Representative synthesis results from ResViT and pGAN are shown along with the reference images for T₂, FLAIR → T₁, T₁, FLAIR → T₂, and T₁, T₂ → FLAIR in BRATS; and MRI → CT in the pelvic dataset. Error maps between the synthetic and reference images for each method are displayed, along with the attention map within the first transformer module of ResViT. Here, the attention maps were overlaid onto the reference image for improved visualization. Attention maps focus on image regions where ResViT substantially reduces synthesis errors compared to pGAN. b) ResViT was demonstrated against an ablated variant obtained by removing the adversarial term from the loss function (i.e., ablating the discriminator subnetwork). Representative results are shown for T₁, T₂ → PD in IXI, T₁, T₂ → FLAIR in BRATS, and MRI → CT in the pelvic dataset. Synthesized images from the models are shown along with the reference images. Adversarial loss improves the acuity of synthesized images.

Lastly, we wanted to visually interpret the benefits of the self-attention mechanisms in ResViT towards synthesis performance. Fig. 3.7b displays representative attention maps in ResViT. Synthetic images and error maps are also shown for ResViT as well as pGAN, which generally offered the closest performance to ResViT in our experiments. There is a high degree of correlation among the broad image regions that the attention maps focus on and the image regions where ResViT yields substantially reduced error levels compared to pGAN. This pattern is observed both for healthy tissues and lesions associated with disease pathology. Therefore, our results suggest that the self-attention mechanisms in ResViT strengthen contextual representations to yield lower synthesis errors in healthy subjects and patients.

3.5 Discussion

In this study, we proposed a novel adversarial model for image translation between separate modalities. Traditional GANs employ convolutional operators that have limited ability to capture long-range relationships among distant regions [68]. The proposed model aggregates convolutional and transformer branches within a residual bottleneck to preserve both local precision and contextual sensitivity. To our knowledge, this is the first adversarial model for medical image synthesis with a transformer-based generator. We further introduced a weight-sharing strategy among transformer modules to lower model complexity. Finally, a unification strategy was implemented to learn an aggregate model that copes with numerous source-target configurations without training separate models.

We demonstrated ResViT for missing sequence synthesis tasks in multi-contrast MRI and missing modality synthesis in multi-modal MRI-CT imaging. Comparisons were provided against state-of-the-art convolutional models with UNet and ResNet backbones and a recent transformer architecture with UNet backbone. ResViT outperforms competing methods in one-to-one and many-to-one synthesis tasks, and our results demonstrate the efficiency of ResViT in maintaining contextual and local representations.

Trained with image-average loss terms, CNNs have difficulty in coping with atypical anatomy that substantially varies across subjects [73, 53]. To improve generalization, recent studies have proposed self-attention mechanisms in GAN models over spatial or channel dimensions [128, 127]. Specifically, attention maps are used for multiplicative modulation of CNN-derived feature maps. This modulation encourages the network to focus on critical image regions with relatively limited task performance. While attention maps can be distributed across image regions, they mainly capture implicit contextual information via modification of local CNN features. Since feature representations are primarily extracted via convolutional filtering, the resulting model can still manifest limited expressiveness for global context. In contrast, the proposed architecture uses dedicated transformer blocks to explicitly model long-range spatial interactions in medical

images.

Here, ResViT models were initialized with pre-trained transformers that expect a 16×16 feature map. In turn, 256×256 input images were 16-fold downsampled cumulatively across the encoder and transformer modules, and the transformer used a patch size of $P=1$ and sequence length of 256. Several strategies can be adopted to use ResViT at different input image resolutions. In a first scenario, the downsampling rate and patch size can be preserved, while the sequence length is adjusted or a transformer pretrained at a different resolution is used. For both cases, a 512×512 image would be downsampled to a 32×32 feature map, yielding a sequence length of 1024. While vision transformers can cope with variable sequence lengths, this would incur a quadratic increase in processing and memory load [98]. In a second scenario, the pre-trained transformer and sequence length can be maintained, while the patch size or the downsampling rate is adjusted. For a 512×512 image downsampled to a 32×32 feature map, $P=2$ (2×2 patches) or 32-fold downsampling would attain a sequence length of 256. Both options would achieve on par computational complexity to the original architecture, albeit the transformer would process feature maps at a lower resolution. It is unlikely that this would significantly affect ResViT’s sensitivity to local features since the primary component of ART that captures local features is the residual CNN module whose resolution is preserved. If the input image does not have a power-of-two size, the abovementioned strategies can be adopted after zero-padding to round up the resolution to the nearest power of two, or by implementing an encoder with non-integer downsampling rates [155]. Note that computer vision studies routinely fine-tune transformers at different image resolutions than encountered during pretraining without performance loss [98]. It remains important future work to investigate the performance of the discussed resolution-adaptation strategies in medical image synthesis.

Few recent studies have independently proposed transformer-based models for medical image synthesis tasks [139, 140, 138]. In [139], a transformer was included in the discriminator of a traditional GAN for MRI-to-PET synthesis. In [140], a UNet-inspired transformer architecture was proposed for infant MRI synthesis [140]. Differing from these efforts, our work makes the following contributions.

(1) Compared to [139] that uses transformers to learn a prior for target PET images, we employ transformers in ResViT’s generator to learn latent contextual representations of source images. (2) Unlike [140] that uses mean-squared error loss amenable to over-smoothing of target images [53], we leverage an adversarial loss to preserve realism. (3) [140] uses transformers across all segments of a UNet-based architecture. We instead propose a ResNet-inspired hybrid architecture that combines localization capabilities of CNNs with contextual sensitivity of transformers. (4) While [139] and [140] consider only task-specific, one-to-one synthesis models, here we uniquely introduce many-to-one synthesis models and a unified model that generalizes across multiple source-target configurations.

Several lines of development can help further improve ResViT’s performance. Here, we considered synthesis tasks in which source and target modalities were registered prior to training, and they were paired across subjects. When registration accuracy is limited, a spatial registration block can be incorporated into the network. Furthermore, a cycle-consistency loss [74] can be incorporated in the optimization objective to allow the use of unregistered images. This latter strategy would also permit training of ResViT models on unpaired datasets [119, 118]. Finally, data requirements for model training can be further alleviated by adopting semi-supervised strategies that allow mixing of paired and unpaired training data [117], or that would enable training of synthesis models directly from undersampled acquisitions [156, 157].

Chapter 4

Unsupervised Medical Image Translation with Adversarial Diffusion Models

The content of this chapter reflects the study described in the following publications:

- M. Özbey*, O. Dalmaz*, S. Dar, A. Bedel, S. Özturk, A. Güngör, and T. Çukur, “Unsupervised Medical Image Translation with Adversarial Diffusion Models,” in *IEEE Transactions on Medical Imaging*, June 2023. Available: doi: 10.1109/TMI.2023.3290149. *: equal contribution
- M. Özbey, O. Dalmaz, A. Bedel, S. Dar, Ş. Özturk, A. Güngör, and T. Çukur, “Adversarial Diffusion Models for Unsupervised Medical Image Synthesis,” *NeurIPS Medical Imaging Meets*, Dec. 2022. (Virtual Conference)
- O. Dalmaz, M. Özbey, A. Bedel, S. Dar, Ş. Özturk, A. Güngör, and T. Çukur, “Cycle-Consistent Adversarial Diffusion For Unsupervised Medical Image Translation,” in *IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, Virtual Conference, Apr. 2023. (Virtual Conference)

- M. Özbey, O. Dalmaz, A. Bedel, S. Dar, Ş. Öztürk, A. Güngör, and T. Çukur, “Adversarial Diffusion Probabilistic Models for Unpaired MRI Contrast Translation,” in *31st annual meeting of International Society for Magnetic Resonance Imaging (ISMRM)*, Toronto, June 2023. Conference)

4.1 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 [15]. Unfortunately, broad adoption of multi-modal protocols is fraught with challenges due to economic and labor costs [83, 94, 89, 50]. Medical image translation is a powerful solution that involves synthesis of a missing target modality under guidance from an acquired source modality [48, 92, 88, 87]. This recovery is an ill-conditioned problem given nonlinear variations in tissue signals across modalities [85, 93, 86]. At this juncture, learning-based methods are offering performance leaps by incorporating nonlinear data-driven priors to improve problem conditioning [90, 91, 49, 109].

Learning-based image translation involves network models trained to capture a prior on the conditional distribution of target given source images [45, 46, 47]. In recent years, generative adversarial network (GAN) models have been broadly adopted for translation tasks, given their exceptional realism in image synthesis [158, 54, 55, 56, 57, 58]. 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 [63, 64, 65, 59, 60]. Based on this adversarial mechanism, state-of-the-art results have been reported with GANs in numerous translation tasks including synthesis across MR scanners [55], multi-contrast MR synthesis [158, 57, 63, 61], and cross-modal synthesis [121, 117, 159].

While powerful, GAN models indirectly characterize the distribution of the target modality through a generator-discriminator interplay without evaluating

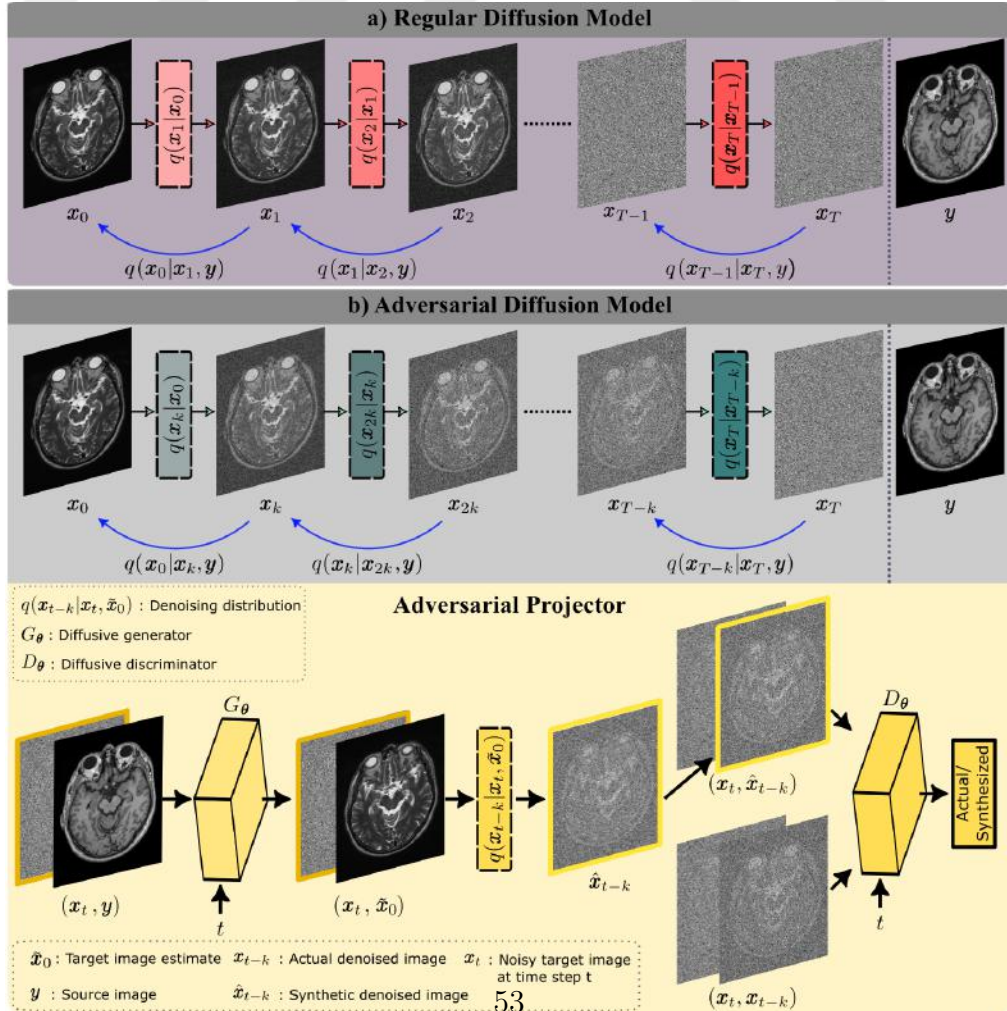
likelihood [73]. Such implicit characterization is potentially amenable to learning biases, including premature convergence and mode collapse. Moreover, GAN models commonly employ a rapid one-shot sampling process without intermediate steps, inherently limiting the reliability of the mapping performed by the network. In turn, these issues can limit the quality and diversity of synthesized images [75]. 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 [160, 75]. 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 [160].

Here, we propose a novel adversarial diffusion model for medical image synthesis, SynDiff, to perform efficient and high-fidelity modality translation (Fig. 4.1). 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. 4.2). 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.

Contributions

- We introduce the first adversarial diffusion model in the 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

Figure 4.1: **a)** Regular diffusion models gradually transform between actual image samples for the target modality ($\mathbf{x}_0$) and isotropic Gaussian noise ($\mathbf{x}_T$) 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 $q(\mathbf{x}_{t+1}|\mathbf{x}_t)$. Each reverse step (left arrows) suppresses the added noise to create a denoised sample. For image translation, a source modality ($\mathbf{y}$) can also be provided as conditioning input to the reverse steps resulting in a reverse transition probability $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{y})$ assumed to be Gaussian, and operationalized via a neural network $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{y})$ that estimates its mean. **b)** The proposed adversarial diffusion model performs fast transformation between $\mathbf{x}_0$ and $\mathbf{x}_T$ 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 $q(\mathbf{x}_{t-k}|\mathbf{x}_t, \mathbf{y})$. To improve accuracy, reverse diffusion steps are operationalized via a novel adversarial projector that uses a generator G_θ and a discriminator D_θ . G_θ first produces an estimate of the target image $\tilde{\mathbf{x}}_0$ given $\mathbf{x}_t$ and $\mathbf{y}$, and a denoised image sample $\hat{\mathbf{x}}_{t-k}$ is then synthesized from the denoising distribution $q(\mathbf{x}_{t-k}|\mathbf{x}_t, \tilde{\mathbf{x}}_0)$. Meanwhile, D_θ distinguishes between actual ($\mathbf{x}_{t-k}$) and synthetic samples ($\hat{\mathbf{x}}_{t-k}$) for the denoised image.



modalities.

- We propose a novel source-conditional adversarial projector to capture reverse transition probabilities over large step sizes for efficient image sampling.

4.2 Related Work

To translate medical images, conditional GANs perform one-shot source-to-target mapping via a generator trained using an adversarial loss [55]. Adversarial loss terms are known to improve sensitivity to high-frequency details in tissue structure over canonical pixel-wise losses [158]. 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 [56, 64, 58, 63, 59, 65]. For improved flexibility, other studies proposed cycle-consistency or mutual information losses to enable unsupervised learning from unpaired data [123, 161, 122, 121, 158, 162, 118, 120]. In general, enhanced spatial acuity and realism have been reported in target images synthesized with GANs when compared to vanilla convolutional models [158]. That said, several problems can arise in GAN models, including lower mapping reliability for the one-shot sampling process [75], premature convergence of the discriminator before the generator is properly trained [60], and poor representational diversity due to mode collapse [73]. 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 [160, 75]. 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 [163, 164, 165, 166, 167], unconditional image generation [168], and anomaly detection [169, 170]. 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 [85]. 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 [87]. 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 [55]. Yet, multi-modal imaging of a large cohort is challenging due to economic and time costs [158]. 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 [160]. 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 [171]. 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 [172]. 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 [173]. 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 4.1: Description of variables related to images, diffusion processes, networks and probability distributions. Throughout the manuscript, vectorial quantities are annotated in bold font.

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

4.3 Theory

4.3.1 Denoising Diffusion Models

Regular diffusion models map between pure noise samples and actual images through a gradual process over T time steps (Fig. 4.1a). In the forward direction, a small amount of Gaussian noise is added repeatedly onto an input image $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ to obtain a sample $\mathbf{x}_T$ from an isotropic Gaussian distribution for sufficiently large T . Forward diffusion forms a Markov chain where the mapping from $\mathbf{x}_{t-1}$ to $\mathbf{x}_t$ and the respective forward transition probability are:

$$\mathbf{x}_t = \sqrt{1 - \beta_t} \mathbf{x}_{t-1} + \sqrt{\beta_t} \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad (4.1)$$

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}) \quad (4.2)$$

where β_t is noise variance, $\boldsymbol{\epsilon}$ is added noise, $\mathcal{N}$ is a Gaussian distribution, $\mathbf{I}$ is an identity covariance matrix. Reverse diffusion also forms a Markov chain from $\mathbf{x}_T$ onto $\mathbf{x}_0$, albeit each step aims to gradually denoise the samples. Under large T and small β_t , the reverse transition probability between $\mathbf{x}_{t-1}$ and $\mathbf{x}_t$ can be approximated as a Gaussian distribution [174, 175]:

$$q(\mathbf{x}_{t-1} | \mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}(\mathbf{x}_t, t), \boldsymbol{\Sigma}(\mathbf{x}_t, t)) \quad (4.3)$$

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:

$$L_{vb} = \mathbb{E}_{q(\mathbf{x}_{0:T})} \left[\log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T} | \mathbf{x}_0)} \right] \leq \mathbb{E}_{q(\mathbf{x}_0)} [\log p_{\boldsymbol{\theta}}(\mathbf{x}_0)] \quad (4.4)$$

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, $\mathbf{x}_{0:T}$ denote the collection of image samples between time steps 0 and T , and $\mathbf{x}_{1:T} | \mathbf{x}_0$ denote image samples between time steps 1 and T conditioned on the sample at time step 0.

The bound can be decomposed as:

$$L_{vb} = \log p_{\boldsymbol{\theta}}(\mathbf{x}_0|\mathbf{x}_1) - \sum_{t=1}^T KL(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) || p_{\boldsymbol{\theta}}(\mathbf{x}_{t-1}|\mathbf{x}_t)) \quad (4.5)$$

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

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) = \frac{1}{\sqrt{\psi_t}} \left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1 - \bar{\psi}_t}} \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) \right) \quad (4.6)$$

where $\psi_t = 1 - \beta_t$ and $\bar{\psi}_t = \prod_{r=[0,1,\dots,t]} \psi_r$. In Eq. 4.6, $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ can be derived if the network is used to estimate the added noise $\boldsymbol{\epsilon}$ by minimizing the following loss [176]:

$$L_{err} = \mathbb{E}_{t, \mathbf{x}_0, \boldsymbol{\epsilon}} \left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}, t)\|_2^2 \right] \quad (4.7)$$

where t , $\mathbf{x}_0$ and $\boldsymbol{\epsilon}$ are sampled from the discrete uniform distribution $\mathcal{U}(0, T)$, $q(\mathbf{x}_0)$ and $\mathcal{N}(\mathbf{0}, \mathbf{I})$, respectively. During inference, reverse diffusion steps are performed starting from a random sample $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. For each step $t \in T \dots 1$, $\boldsymbol{\mu}$ is derived using Eq. 4.6 based on the network estimate $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$, and $\mathbf{x}_{t-1}$ is sampled based on Eq. 4.3.

4.3.2 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. 4.1b). It also employs a non-diffusive module for estimating source images paired with corresponding target images, so as to enable unsupervised learning (Fig. 4.2). The adversarial diffusion process that forms the basis of the diffusive module, the network architecture, and the learning procedures for SynDiff are detailed below.

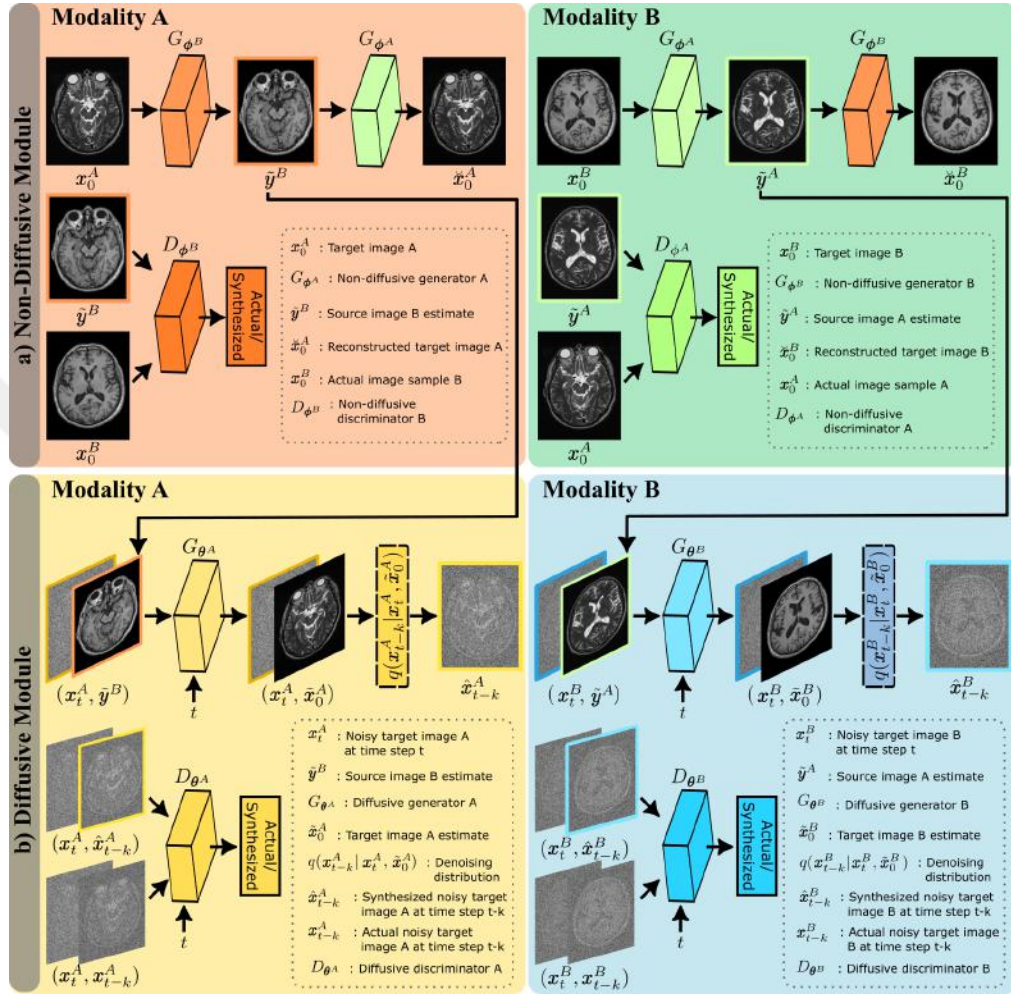


Figure 4.2: For unsupervised learning, SynDiff leverages a cycle-consistent architecture that bilaterally translates between two modalities (A , B). For synthesizing a target image $\hat{x}_0^A$ of modality A , the diffusive module in Fig. 4.1b requires guidance from a source image y^B of modality B 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 $\tilde{y}^B$ from x_0^A . Similarly, for synthesizing a target image $\hat{x}_0^B$ of modality B with the diffusive module, the non-diffusive module first estimates a paired source image $\tilde{y}^A$ from x_0^B . **a)** To do this, the non-diffusive module comprises two generator-discriminator pairs ($G_{\phi^{A,B}}$, $D_{\phi^{A,B}}$) that generate initial translation estimates for $x_0^A \rightarrow \tilde{y}^B$ (orange) and $x_0^B \rightarrow \tilde{y}^A$ (green). **b)** These initial translation estimates $\tilde{y}^{A,B}$ 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 ($G_{\theta^{A,B}}$, $D_{\theta^{A,B}}$) to generate denoised image estimates for $(x_t^A, \tilde{y}^B, t) \rightarrow \hat{x}_{t-k}^A$ (yellow) and $(x_t^B, \tilde{y}^A, t) \rightarrow \hat{x}_{t-k}^B$ (blue).

4.3.2.1 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. 4.3, but this limits efficiency in image generation. Here, we instead propose fast diffusion with the following forward steps:

$$\mathbf{x}_t = \sqrt{1 - \gamma_t} \mathbf{x}_{t-k} + \sqrt{\gamma_t} \boldsymbol{\epsilon} \quad (4.8)$$

$$q(\mathbf{x}_t | \mathbf{x}_{t-k}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \gamma_t} \mathbf{x}_{t-k}, \gamma_t \mathbf{I}) \quad (4.9)$$

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

$$\gamma_t = 1 - e^{\bar{\beta}_{\min} \frac{k}{T} - (\bar{\beta}_{\max} - \bar{\beta}_{\min}) \frac{2tk - k^2}{2T^2}} \quad (4.10)$$

$\bar{\beta}_{\min}$ and $\bar{\beta}_{\max}$ control the progression of noise variance in an exponential schedule [177].

Guidance from a source image ($\mathbf{y}$) 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 $q(\mathbf{x}_{t-k} | \mathbf{x}_t, \mathbf{y})$ and the normality assumption used to compute Eq. 4.4 breaks down [160]. Here we introduce a novel source-conditional adversarial projector to capture the complex transition probability $q(\mathbf{x}_{t-k} | \mathbf{x}_t, \mathbf{y})$ 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 $q(\mathbf{x}_{t-k} | \mathbf{x}_t)$ [171]. In SynDiff, a conditional generator $G_\theta(\mathbf{x}_t, \mathbf{y}, t)$ performs gradual denoising in each reverse step to synthesize $\hat{\mathbf{x}}_{t-k} \sim p_\theta(\mathbf{x}_{t-k} | \mathbf{x}_t, \mathbf{y})$. G_θ receives the image pair $(\mathbf{x}_t, \mathbf{y})$ as a two-channel input, and it extracts intermediate feature maps $\mathbf{f}_i$ where $i \in [1, \dots, N]$ is the subblock index in an encoder-decoder structure [177]. A learnable temporal embedding $\mathbf{m}$ is computed given t , and this embedding is added as a channel-specific bias term onto the feature maps in each subblock [177]: $\mathbf{f}'_i = \mathbf{f}_i + \mathbf{m}$. Meanwhile, a discriminator $D_\theta(\{\hat{\mathbf{x}}_{t-k} \text{ or } \mathbf{x}_{t-k}\}, \mathbf{x}_t, t)$ distinguishes samples drawn from estimated versus true denoising distributions ($p_\theta(\mathbf{x}_{t-k} | \mathbf{x}_t, \mathbf{y})$ vs. $q(\mathbf{x}_{t-k} | \mathbf{x}_t, \mathbf{y})$). D_θ receives either $(\mathbf{x}_t, \hat{\mathbf{x}}_{t-k})$ or $(\mathbf{x}_t, \mathbf{x}_{t-k})$ as a two-channel input. The temporal embedding $\mathbf{m}$ is also added as a bias term onto the feature maps across D_θ . A non-saturating

adversarial loss is adopted for G_{θ} [178]:

$$L_{G_{\theta}} = \mathbb{E}_{t,q(\mathbf{x}_t|\mathbf{x}_0,\mathbf{y}),p_{\theta}(\mathbf{x}_{t-k}|\mathbf{x}_t,\mathbf{y})}[-\log(D_{\theta}(\hat{\mathbf{x}}_{t-k}))] \quad (4.11)$$

where $t \sim \mathcal{U}(\{0, k, \dots, T\})$, and the discriminator arguments are abbreviated for brevity. D_{θ} also adopts a non-saturating adversarial loss with gradient penalty [179]:

$$\begin{aligned} L_{D_{\theta}} = \mathbb{E}_{t,q(\mathbf{x}_t|\mathbf{x}_0,\mathbf{y})} & \left[\mathbb{E}_{q(\mathbf{x}_{t-k}|\mathbf{x}_t,\mathbf{y})} [-\log(D_{\theta}(\mathbf{x}_{t-k}))] \right. \\ & + \mathbb{E}_{p_{\theta}(\mathbf{x}_{t-k}|\mathbf{x}_t,\mathbf{y})} [-\log(1 - D_{\theta}(\hat{\mathbf{x}}_{t-k}))] \\ & \left. + \eta \mathbb{E}_{q(\mathbf{x}_{t-k}|\mathbf{x}_t,\mathbf{y})} \left\| \nabla_{\mathbf{x}_{t-k}} D_{\theta}(\mathbf{x}_{t-k}) \right\|_2^2 \right] \end{aligned} \quad (4.12)$$

where η is the weight for the gradient penalty.

Evaluation of Eqs. 4.11-4.12 require sampling from $q(\mathbf{x}_{t-k}|\mathbf{x}_t, \mathbf{y})$ that is unknown. Yet, $\mathbf{x}_0$ and $\mathbf{y}$ are non-linearly related images of the same anatomy and $\mathbf{x}_t$ is conditionally independent of $\mathbf{y}$ given $\mathbf{x}_0$. Thus, if the non-linear mapping from a particular $\mathbf{y}$ onto $\mathbf{x}_0$ is injective (i.e., one-to-one), the reverse transition probability can be expressed as $q(\mathbf{x}_{t-k}|\mathbf{x}_t, \mathbf{x}_0, \mathbf{y}) = q(\mathbf{x}_{t-k}|\mathbf{x}_t, \mathbf{x}_0)$ [160]. Bayes' rule can then be used to express the denoising distribution in terms of forward transition probabilities:

$$q(\mathbf{x}_{t-k}|\mathbf{x}_t, \mathbf{x}_0) = q(\mathbf{x}_t|\mathbf{x}_{t-k}, \mathbf{x}_0) \frac{q(\mathbf{x}_{t-k}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)} \quad (4.13)$$

Using Eq. 4.8, it can then be shown that $q(\mathbf{x}_{t-k}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-k}; \bar{\boldsymbol{\mu}}(\mathbf{x}_t, \mathbf{x}_0), \bar{\gamma}\mathbf{I})$ with the following parameters:

$$\bar{\boldsymbol{\mu}} = \frac{\sqrt{\bar{\alpha}_{t-k}}\gamma_t}{1 - \bar{\alpha}_t} \mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-k})}{1 - \bar{\alpha}_t} \mathbf{x}_t, \bar{\gamma} = \frac{1 - \bar{\alpha}_{t-k}}{1 - \bar{\alpha}_t} \gamma_t \quad (4.14)$$

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

Eqs. 4.11-4.12 also require sampling from the network-parameterized denoising distribution $p_{\theta}(\mathbf{x}_{t-k}|\mathbf{x}_t, \mathbf{y})$. A trivial albeit deterministic sample would be the generator output, i.e. $\hat{\mathbf{x}}_{t-k} \sim \delta(\mathbf{x}_{t-k} - G_{\theta}(\mathbf{x}_t, \mathbf{y}, t))$. To maintain stochasticity, we instead operationalize the generator distribution as follows:

$$p_{\theta}(\mathbf{x}_{t-k}|\mathbf{x}_t, \mathbf{y}) := q(\mathbf{x}_{t-k}|\mathbf{x}_t, \tilde{\mathbf{x}}_0 = G_{\theta}(\mathbf{x}_t, \mathbf{y}, t)) \quad (4.15)$$

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

4.3.2.2 Network Architecture

To synthesize a target-modality image, the reverse diffusion steps parametrized in Eq.4.15 require guidance from a source-modality image of the same anatomy. However, the training set might include only unpaired images $\mathbf{x}_0^A, \mathbf{x}_0^B$ 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 $\tilde{\mathbf{y}}^B$ of modality B is produced given a target image $\mathbf{x}_0^A$ of modality A ; and a source-image estimate $\tilde{\mathbf{y}}^A$ is produced given a target image $\mathbf{x}_0^B$. To do this, two generator-discriminator pairs (G_{ϕ^A}, D_{ϕ^A}) and (G_{ϕ^B}, D_{ϕ^B}) with parameters $\phi^{A,B}$ are employed [158]. The generators produce the estimates $\tilde{\mathbf{y}}^{A,B}$ as:

$$\begin{aligned}\tilde{\mathbf{y}}^B &= G_{\phi^B}(\mathbf{x}_0^A) \\ \tilde{\mathbf{y}}^A &= G_{\phi^A}(\mathbf{x}_0^B)\end{aligned}\tag{4.16}$$

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

$$L_{G_{\phi}} = \mathbb{E}_{p_{\phi}(\mathbf{y}|\mathbf{x}_0)}[-\log(D_{\phi}(\tilde{\mathbf{y}}))]\tag{4.17}$$

where $p_{\phi}(\mathbf{y}|\mathbf{x}_0)$ denotes the network parametrization of the conditional distribution of the source given the target image, and the conditioning input $\mathbf{x}_0$ 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:

$$\begin{aligned}L_{D_{\phi}} &= \mathbb{E}_{q(\mathbf{y}|\mathbf{x}_0)}[-\log(D_{\phi}(\mathbf{y}))] + \\ &\quad \mathbb{E}_{p_{\phi}(\mathbf{y}|\mathbf{x}_0)}[-\log(1 - D_{\phi}(\tilde{\mathbf{y}}))]\end{aligned}\tag{4.18}$$

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

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 $\hat{\mathbf{x}}^A$ is produced given $\tilde{\mathbf{y}}^B$; and a synthetic target image $\hat{\mathbf{x}}^B$ is produced given $\tilde{\mathbf{y}}^A$. To do this, two adversarial diffusion processes are used with respective generator-discriminator pairs $(G_{\theta^A}, D_{\theta^A})$ and $(G_{\theta^B}, D_{\theta^B})$ of parameters $\theta^{A,B}$. Starting with Gaussian noise images $\mathbf{x}_T^{A,B}$ 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. 4.3.2.1:

$$\begin{aligned}\tilde{\mathbf{x}}_0^A &= G_{\theta^A}(\mathbf{x}_t^A, \mathbf{y} = \tilde{\mathbf{y}}^B, t) \\ \tilde{\mathbf{x}}_0^B &= G_{\theta^B}(\mathbf{x}_t^B, \mathbf{y} = \tilde{\mathbf{y}}^A, t)\end{aligned}\tag{4.19}$$

Afterwards, the denoising distribution for each modality as described in Eq. 4.15 is used to synthesize target images:

$$\begin{aligned}\hat{\mathbf{x}}_{t-k}^A &\sim q(\mathbf{x}_{t-k}^A | \mathbf{x}_t^A, \tilde{\mathbf{x}}_0^A) \\ \hat{\mathbf{x}}_{t-k}^B &\sim q(\mathbf{x}_{t-k}^B | \mathbf{x}_t^B, \tilde{\mathbf{x}}_0^B)\end{aligned}\tag{4.20}$$

4.3.2.3 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 $\hat{\mathbf{x}}_0^{A,B}$. In the non-diffusive module, source-image estimates are projected to the target domain via the generators:

$$\begin{aligned}\check{\mathbf{x}}_0^A &= G_{\phi^A}(\tilde{\mathbf{y}}^B) \\ \check{\mathbf{x}}_0^B &= G_{\phi^B}(\tilde{\mathbf{y}}^A)\end{aligned}\tag{4.21}$$

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

$$L_{\text{cyc}} = \mathbb{E}_{t,q(\mathbf{x}_0^{A,B}),q(\mathbf{x}_t^{A,B}|\mathbf{x}_0^{A,B})} [\lambda_{1\phi}(\|\mathbf{x}_0^A - \check{\mathbf{x}}_0^A\|_1 + \|\mathbf{x}_0^B - \check{\mathbf{x}}_0^B\|_1) + \lambda_{1\theta}(\|\mathbf{x}_0^A - \hat{\mathbf{x}}_0^A\|_1 + \|\mathbf{x}_0^B - \hat{\mathbf{x}}_0^B\|_1)] \quad (4.22)$$

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 [158]. The diffusive and non-diffusive modules are trained jointly without any pretraining procedures. Accordingly, the overall generator loss is:

$$L_G^{\text{total}} = \lambda_{2\phi}(L_{G_\phi^A} + L_{G_\phi^B}) + \lambda_{2\theta}(L_{G_\theta^A} + L_{G_\theta^B}) + L_{\text{cyc}} \quad (4.23)$$

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_ϕ} is defined as in Eq. 4.17 and L_{G_θ} is defined as in Eq. 4.11. The overall discriminator loss is given as:

$$L_D^{\text{total}} = \lambda_{2\phi}(L_{D_\phi^A} + L_{D_\phi^B}) + \lambda_{2\theta}(L_{D_\theta^A} + L_{D_\theta^B}) \quad (4.24)$$

with L_{D_ϕ} defined as in Eq. 4.18 and L_{D_θ} defined as in Eq. 4.12.

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 $\rightarrow$ target), $G_{\theta^B}(\mathbf{x}_t^B, \mathbf{y}^A, t)$ is used where $\mathbf{x}_t^B$ is the target-image sample of modality B at time step t and $\mathbf{y}^A$ is the acquired source image of modality A provided as input. Inference starts at time step T with a Gaussian noise sample $\mathbf{x}_T^B$ drawn from $\mathcal{N}(\mathbf{0}, \mathbf{I})$, 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. 4.19-4.20 to attain $\hat{\mathbf{x}}_0^B$ at time step 0 as the synthetic target image.

4.4 Methods

4.4.1 Datasets

We demonstrated SynDiff on two multi-contrast brain MRI datasets (IXI¹, BRATS [143]), and a multi-modal pelvic MRI-CT dataset [146]. 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 [147]. 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.

4.4.1.1 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= $0.94 \times 0.94 \times 1.2\text{mm}^3$.

¹<https://brain-development.org/ixi-dataset/>

4.4.1.2 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.

4.4.1.3 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, $0.88 \times 0.88 \times 3\text{mm}^3$ resolution, or TE=4.77ms, TR=7.46ms, $1.10 \times 1.10 \times 2\text{mm}^3$ resolution were prescribed. For T_2 scans, TE=97ms, TR=6000-6600ms, $0.88 \times 0.88 \times 2.50\text{mm}^3$ resolution, or TE=91-102ms, TR = 12000-16000ms, $0.88 \times 1.10 \times 0.88 \times 2.50\text{mm}^3$ resolution were prescribed. For CT scans, $0.10 \times 0.10 \times 3\text{mm}^3$ resolution, Kernel=B30f, or $0.10 \times 0.10 \times 2\text{mm}^3$ resolution, Kernel=FC17 were prescribed. To implement synthesis tasks from accelerated MRI scans [180, 156], fully-sampled MRI data were retrospectively undersampled 4-fold in two dimensions to attain low-resolution images at a 16x acceleration rate [180].

4.4.2 Competing Methods

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

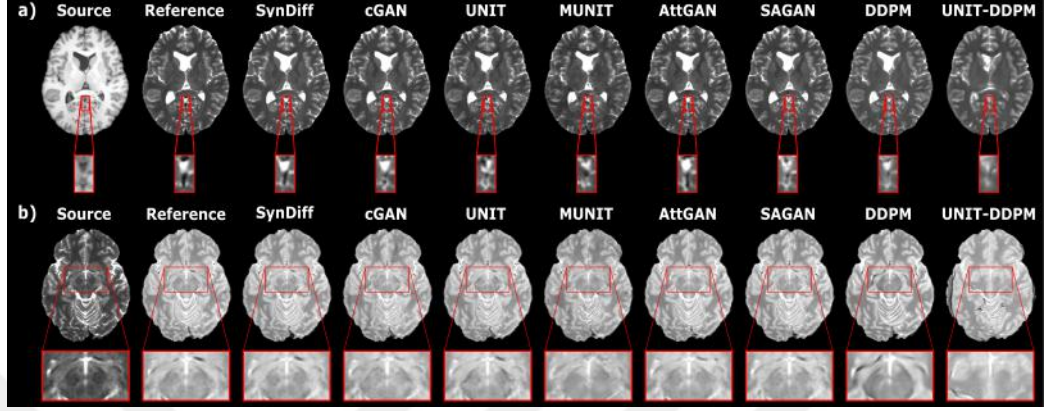


Figure 4.3: SynDiff was demonstrated on IXI for translation between MRI contrasts. Synthesized images from competing methods are displayed along with the source and the ground-truth target (reference) images for representative a) $T_1 \rightarrow T_2$, b) $T_2 \rightarrow PD$ tasks. Display windows of 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 4.2: Performance for multi-contrast MRI translation tasks in IXI. PSNR (dB) and SSIM (%) are listed as mean $\pm$ 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$\pm$1.40	94.77$\pm$1.26	30.32$\pm$1.46	94.28$\pm$1.32	30.09$\pm$1.36	94.99$\pm$1.17	30.85$\pm$1.56	94.03$\pm$1.12	33.64$\pm$0.86	96.58$\pm$0.36	35.47$\pm$1.15	96.98$\pm$0.36
cGAN	29.22 $\pm$ 1.20	93.46 $\pm$ 1.33	29.24 $\pm$ 1.26	93.01 $\pm$ 1.44	28.42 $\pm$ 1.03	93.38 $\pm$ 1.19	29.92 $\pm$ 1.45	93.19 $\pm$ 1.21	33.58 $\pm$ 0.75	96.46 $\pm$ 0.39	34.24 $\pm$ 1.00	96.09 $\pm$ 0.47
UNIT	29.17 $\pm$ 1.15	93.54 $\pm$ 1.34	28.34 $\pm$ 0.98	92.02 $\pm$ 1.45	28.10 $\pm$ 0.99	92.97 $\pm$ 1.21	29.29 $\pm$ 1.08	92.36 $\pm$ 1.24	32.57 $\pm$ 0.65	96.22 $\pm$ 0.38	34.74 $\pm$ 1.07	96.66 $\pm$ 0.39
MUNIT	26.35 $\pm$ 0.88	89.78 $\pm$ 1.78	26.61 $\pm$ 0.86	88.28 $\pm$ 1.90	25.99 $\pm$ 0.89	89.73 $\pm$ 1.70	27.59 $\pm$ 1.02	88.71 $\pm$ 1.60	29.17 $\pm$ 0.71	92.01 $\pm$ 1.01	29.80 $\pm$ 0.82	91.61 $\pm$ 1.00
AttGAN	29.27 $\pm$ 1.38	93.74 $\pm$ 1.36	28.37 $\pm$ 1.08	92.21 $\pm$ 1.45	28.02 $\pm$ 1.07	92.83 $\pm$ 1.19	29.65 $\pm$ 1.42	92.98 $\pm$ 1.23	32.15 $\pm$ 0.67	95.93 $\pm$ 0.44	35.11 $\pm$ 1.11	96.76 $\pm$ 0.40
SAGAN	28.85 $\pm$ 1.26	93.38 $\pm$ 1.40	29.01 $\pm$ 1.32	92.87 $\pm$ 1.43	27.93 $\pm$ 1.22	93.04 $\pm$ 1.29	29.58 $\pm$ 1.51	92.76 $\pm$ 1.25	32.44 $\pm$ 0.71	95.91 $\pm$ 0.46	34.75 $\pm$ 0.83	96.64 $\pm$ 0.38
DDPM	24.93 $\pm$ 0.69	89.49 $\pm$ 1.69	28.04 $\pm$ 1.03	91.14 $\pm$ 1.58	24.95 $\pm$ 0.74	89.08 $\pm$ 1.67	27.16 $\pm$ 0.95	90.45 $\pm$ 1.33	30.49 $\pm$ 0.84	94.74 $\pm$ 0.69	29.67 $\pm$ 0.71	93.18 $\pm$ 0.83
UNIT-DDPM	24.01 $\pm$ 0.72	86.59 $\pm$ 2.16	22.44 $\pm$ 1.26	81.64 $\pm$ 3.06	23.81 $\pm$ 0.97	86.62 $\pm$ 2.44	26.81 $\pm$ 1.35	88.57 $\pm$ 2.04	25.43 $\pm$ 0.49	88.08 $\pm$ 1.09	25.13 $\pm$ 1.42	84.47 $\pm$ 2.53

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.

4.4.2.1 SynDiff

In the non-diffusive module, generators used a ResNet backbone with three encoding, six residual, and three decoding blocks [181]; 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 [182]. 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 subblocks received a temporal embedding derived by projecting a 32-dimensional sinusoidal position encoding through a two-layer multi-layer perceptron (MLP) [177]. They also received 256-dimensional random latents from a three-layer MLP to modulate feature maps via adaptive normalization [183]. 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, $\mu=0.5$, $T=1000$, 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 $\bar{\beta}_{\min}=0.1$, $\bar{\beta}_{\max}=20$.

4.4.2.2 cGAN

A cycle-consistent GAN model was considered with architecture and loss functions adopted from [158]. 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, 2×10^{-4} learning rate linearly decayed to 0 in the last 50 epochs. Weights for cycle-consistency and adversarial losses were 100 and 1.

4.4.2.3 UNIT

An unsupervised GAN model that assumes a shared latent space between source-target modalities was considered, with architecture and loss functions adopted from [184]. UNIT comprised two discriminators and two translators with ResNet

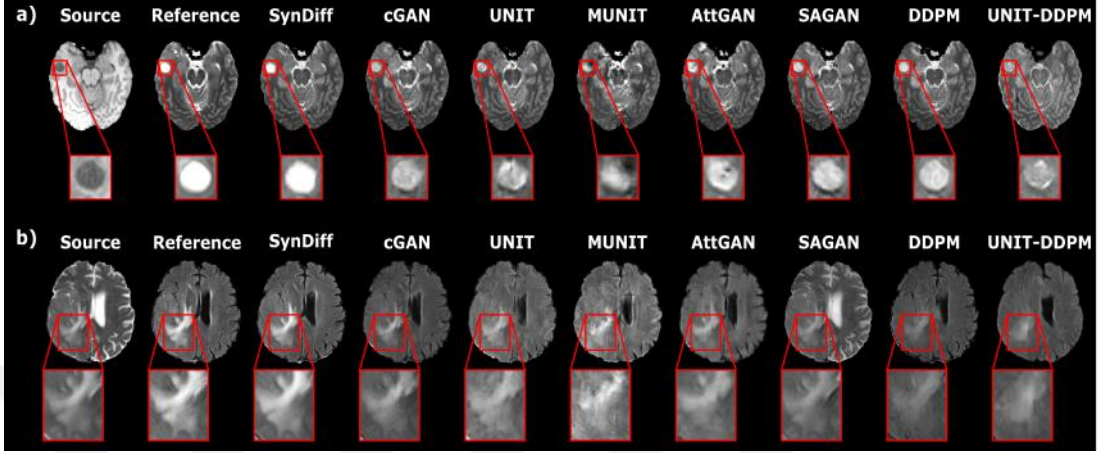


Figure 4.4: SynDiff was demonstrated on BRATS for translation between MRI contrasts. Synthesized images are displayed along with the source and the ground-truth target (reference) images for representative a) $T_1 \rightarrow T_2$, b) $T_2 \rightarrow \text{FLAIR}$ tasks. Display windows of a) $[0 \ 0.75]$, b) $[0 \ 0.80]$ are used. SynDiff lowers noise/artifact levels and more accurately depicts detailed structure compared to baselines.

Table 4.3: Performance for multi-contrast MRI translation tasks in BRATS. PSNR (dB) and SSIM (%) listed as mean $\pm$ std across the test set.

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

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

4.4.2.4 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 [185]. 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.

4.4.2.5 AttGAN

A cycle-consistent GAN model with attentional generators [128] was adopted for unsupervised translation. AttGAN comprised two convolutional attention UNet generators and two patch discriminators [128]. Cross-validated hyperparameters were 100 epochs, 2×10^{-4} learning rate linearly decayed to 0 in the last 50 epochs. Weights for cycle-consistency and adversarial losses were 100 and 1.

4.4.2.6 SAGAN

A cycle-consistent GAN model with self-attention generators [127] 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 [127]. Cross-validated hyperparameters were 100 epochs, 2×10^{-4} learning rate linearly decayed to 0 in the last 50 epochs. Weights for cycle-consistency and adversarial losses were 100 and 1.

4.4.2.7 DDPM

A recent diffusion model with improved sampling efficiency was considered, with architecture and loss functions adopted from [186]. 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 [186]. Weight for cycle-consistency loss was 1.

4.4.2.8 UNIT-DDPM

A recent diffusion model allowing unsupervised training was considered, with architecture and loss functions adopted from [172]. 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 [186]. Weight for cycle-consistency loss was 1, and the release time was 1 as in [172].

4.4.3 Modeling Procedures

All models were implemented in Python using the PyTorch framework. Models were trained using Adam optimizer with $\beta_1=0.5$, $\beta_2=0.9$. 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 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

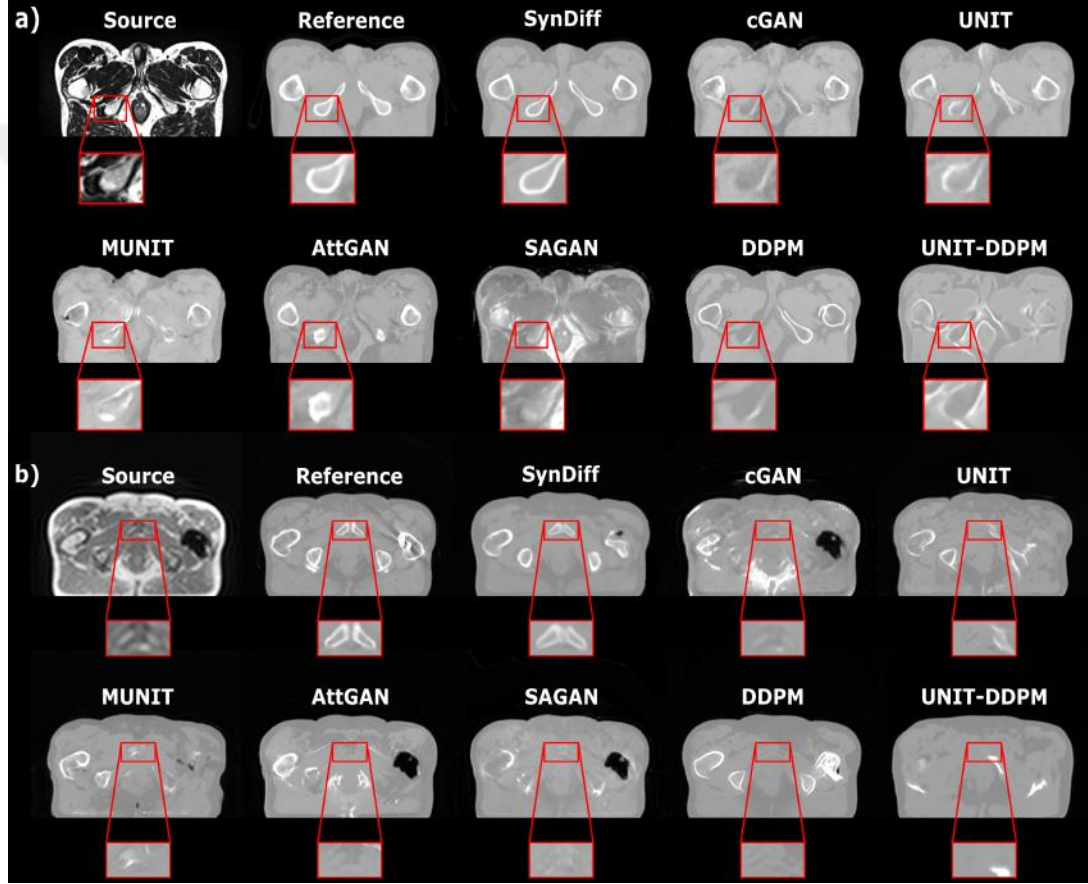


Figure 4.5: SynDiff was demonstrated on the pelvic dataset for multi-modal MRI-CT translation. Synthesized images are displayed along with the source and the ground-truth target (reference) images for representative a) $T_2 \rightarrow CT$, b) accelerated $T_1 \rightarrow CT$ tasks. Display windows of 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.

Table 4.4: Performance for multi-modal MRI-CT translation tasks in the pelvic dataset. PSNR (dB) and SSIM (%) listed as mean $\pm$ std across the test set. ‘acc.’ stands for accelerated.

	T ₂ →CT		T ₁ →CT		acc. T ₂ →CT		acc. T ₁ →CT	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SynDiff	26.86	87.94	25.16	86.02	26.71	87.32	25.47	85.00
	± 0.51	± 2.53	± 1.53	± 2.05	± 0.63	± 2.84	± 1.09	± 2.10
cGAN	25.07	84.91	24.11	77.81	21.24	69.62	20.35	64.73
	± 0.17	± 1.84	± 1.00	± 1.84	± 0.51	± 0.85	± 0.32	± 1.47
UNIT	26.10	86.40	25.04	82.62	25.20	84.83	24.92	81.44
	± 0.49	± 2.71	± 0.39	± 1.52	± 0.37	± 1.43	± 0.39	± 1.13
MUNIT	22.90	77.42	24.76	79.81	23.44	77.88	24.42	79.64
	± 1.05	± 2.17	± 0.62	± 1.20	± 0.77	± 2.04	± 0.34	± 1.05
AttGAN	23.81	74.35	24.76	82.48	23.91	76.47	21.34	67.24
	± 0.18	± 0.84	± 1.06	± 2.49	± 0.29	± 0.66	± 0.51	± 1.52
SAGAN	21.03	67.77	23.89	77.05	19.61	61.92	23.28	70.02
	± 0.33	± 0.86	± 1.02	± 2.87	± 0.78	± 0.32	± 0.96	± 2.85
DDPM	24.66	83.24	24.92	82.63	24.35	83.25	24.62	83.04
	± 0.19	± 2.62	± 0.81	± 3.64	± 0.47	± 1.70	± 0.59	± 2.40
UNIT-DDPM	21.49	80.23	20.26	76.79	21.89	77.69	21.45	77.10
	± 0.72	± 2.69	± 1.17	± 1.37	± 0.77	± 3.06	± 0.23	± 2.83

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 ($p < 0.05$).

4.5 Results

4.5.1 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,

Table 4.5: 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

experiments were performed on brain images from healthy subjects in IXI. Table 4.2 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. 4.3. 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 glioma patients in BRATS. Table 4.3 lists performance metrics for $T_2 \rightarrow T_1$, $T_1 \rightarrow T_2$, $FLAIR \rightarrow T_1$, $T_1 \rightarrow FLAIR$, $FLAIR \rightarrow T_2$, and $T_2 \rightarrow FLAIR$ tasks. SynDiff again achieves the highest synthesis performance in all tasks ($p < 0.05$), except for cGAN that yields similar PSNR in $T_1 \rightarrow FLAIR$, and performs similarly in $FLAIR \rightarrow T_1$. 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. 4.4. 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.

4.5.2 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 4.4 lists performance metrics for $T_2 \rightarrow CT$, $T_1 \rightarrow CT$, accelerated $T_2 \rightarrow CT$, and accelerated $T_1 \rightarrow CT$ 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. 4.5. 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. 4.5b 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 [169]. In turn, the trained models synthesize CT images with regular tissue appearance as opposed to artifacts near metal.

4.5.3 Model Complexity

A practical concern for medical image translation is the computational complexity of the applied models. Table 4.5 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.

4.5.4 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.

4.5.5 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

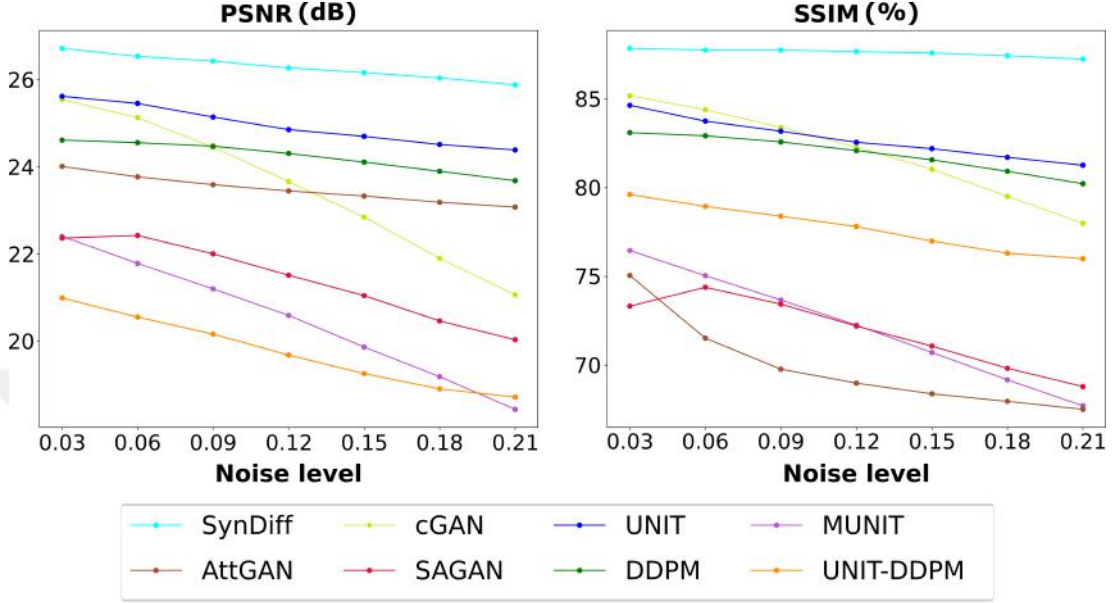


Figure 4.6: Performance of competing methods as a function of added noise level on source-modality images. Results shown for the representative $T_2 \rightarrow CT$ task in terms of PSNR (left), SSIM (right).

Table 4.6: 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

multi-modal $T_2 \rightarrow CT$ task. Zero-mean bivariate Gaussian white noise was added onto each cross-section at std. values ranging in $[0.03 \ 0.21]$ relative to the mean pixel intensity [167]. Fig. 4.6 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.

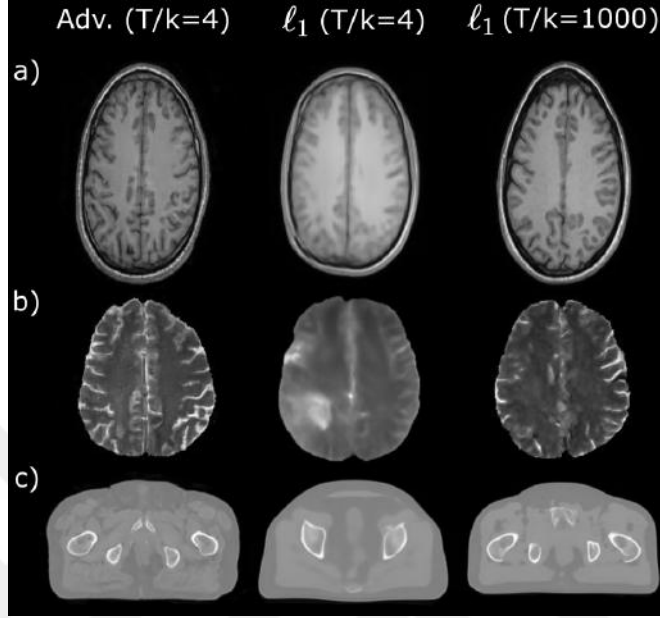


Figure 4.7: The adversarial projector in SynDiff with $T/k=4$ steps was compared against a variant model using an ℓ_1 -loss based projector with $T/k=4$ and $T/k=1000$. Image samples are shown for the unconditional synthesis tasks: a) T_1 in IXI, b) T_2 in BRATS and c) CT in pelvic datasets. Display windows of a) $[0 \ 0.90]$, b) $[0 \ 0.80]$ for MRI images, and c) $[-1000 \ 1300]$ HU for CT images are used.

4.5.6 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

Table 4.7: Performance of variant models ablated of adversarial loss, cycle-consistency loss and the diffusive module. PSNR and SSIM listed as mean $\pm$ std across the test set.

	PD $\rightarrow$ T ₁		T ₁ $\rightarrow$ T ₂		T ₂ $\rightarrow$ 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

in Fig. 4.7, and FID scores are listed in Table 4.6. 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.

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 4.7, 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 4.8, 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

Table 4.8: 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 $\pm$ std across the test set.

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

remaining loss-term 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 4.9, there

Table 4.9: Performance of variant models as mean $\pm$ 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 $\rightarrow$ T ₁		T ₁ $\rightarrow$ T ₂		T ₂ $\rightarrow$ CT	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SynDiff	30.09	94.99	27.10	92.35	26.86	87.94
	$\pm$ 1.36	$\pm$ 1.17	$\pm$ 1.26	$\pm$ 1.27	$\pm$ 0.51	$\pm$ 2.53
Pretrained-frozen	29.19	94.22	27.23	92.50	26.77	87.65
	$\pm$ 1.28	$\pm$ 1.23	$\pm$ 1.30	$\pm$ 1.31	$\pm$ 0.82	$\pm$ 2.81
Pretrained-trained	29.24	94.24	27.44	92.75	26.97	88.40
	$\pm$ 1.17	$\pm$ 1.21	$\pm$ 1.23	$\pm$ 1.24	$\pm$ 0.7	$\pm$ 2.74

Table 4.10: Performance of variant models as mean $\pm$ 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 $\rightarrow$ T ₁		T ₁ $\rightarrow$ T ₂		T ₂ $\rightarrow$ CT	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
$n_{ND} = 5$	19.88	70.31	25.26	89.99	22.59	74.59
	$\pm$ 0.60	$\pm$ 2.70	$\pm$ 1.08	$\pm$ 1.13	$\pm$ 0.39	$\pm$ 2.44
$n_{ND} = 10$	27.68	92.12	26.04	91.12	24.29	81.21
	$\pm$ 0.70	$\pm$ 1.17	$\pm$ 1.13	$\pm$ 1.26	$\pm$ 0.19	$\pm$ 2.71
$n_{ND} = 25$	29.51	94.45	26.47	91.41	26.03	86.14
	$\pm$ 1.14	$\pm$ 1.12	$\pm$ 1.33	$\pm$ 1.48	$\pm$ 0.35	$\pm$ 2.52
$n_{ND} = 50$	30.09	94.99	27.10	92.35	26.86	87.94
	$\pm$ 1.36	$\pm$ 1.17	$\pm$ 1.26	$\pm$ 1.27	$\pm$ 0.51	$\pm$ 2.53

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 $\rightarrow$ T₁ task, albeit it generally increases their performance in the relatively hard T₁ $\rightarrow$ T₂ and T₂ $\rightarrow$ 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 $PD \rightarrow T_1$, 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 4.10 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.

4.6 Discussion

4.6.1 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 [160, 186]. 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 [56, 158]. 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 [158, 187]. 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.

4.6.2 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 [179]. 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 [179].

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 [55, 63, 188].

4.6.3 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 [158, 189]. Performance improvements might also be viable by expanding the size of training datasets based on a collection of undersampled source- and target-modality acquisitions [156], or a combination of paired and unpaired source-target modality data [117].

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 [190, 159, 135]. 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 [191].

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 [192] or running the diffusion process in a compact latent space [168]. 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 [171].

4.6.4 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 [158]. 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 [57]. In such cases, translation performance can be improved by incorporating multiple source modalities that capture more diverse tissue information [64, 58, 65].

Here we considered native MRI contrasts for source-to-target syntheses. Yet, the proposed framework can be adopted to perform translation tasks that involve other MRI pulse sequences [193, 194], e.g. diffusion MRI [195, 196]. That said, given adequate training sets, SynDiff can learn to translate between other imaging modalities, e.g. MPI [197, 198, 199, 200, 201, 202, 203], or functional MRI [204,

205, 206, 207, 208, 209, 210, 211, 212, 213, 214, 215, 212, 213, 216].

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 [217, 218]. 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 [219]. 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. Lastly, the fast conditional diffusion process in SynDiff might also offer performance benefits over GANs in other applications such as denoising and super-resolution[167, 220, 221].

Chapter 5

Conclusion

In this thesis, first, a novel synthesis approach for multi-modal medical image translation based on an adversarial model with a transformer backbone was proposed. In an information bottleneck, the proposed residual vision transformer model ResViT aggregates convolutional operators and vision transformers, thereby improving capture of contextual relations while maintaining localization power. A unified implementation was introduced that mitigates the need to rebuild models for varying source-target configurations. ResViT achieves superior synthesis quality to approaches in multi-contrast brain MRI and multi-modal pelvic MRI-CT datasets.

Next, a novel adversarial diffusion model for medical image translation between source and target modalities was introduced. SynDiff leverages a fast diffusion process to efficiently synthesize target images, and a conditional adversarial projector for accurate reverse diffusion sampling. Unsupervised learning is achieved via a cycle-consistent architecture that embodies coupled diffusion processes between the two modalities. A comparative study of four stochastic image translation models is also presented in this thesis: SynDiff, MUNIT, DDPM and UNIT-DDPM. We have also analyzed the variability of the target images synthesized by each method for a given source image. Our results show that all methods produce high-quality target images with low variability across samples,

except for UNIT-DDPM which has slightly higher variability. This suggests that the random noise variables used by the stochastic methods do not induce significant diversity in the output images. Therefore, we conclude that stochastic image translation models are effective and consistent in generating realistic target images for a given source image, but they may not capture the full range of possible variations in the target domain for conditional tasks. 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.

Several directions of future work can be explored to further improve the performance of the models presented in this thesis. First, the benefits of combining the transformer and diffusion architectures can be investigated to create a unified model that can leverage the advantages of both approaches. For example, a transformer-based generator can be used in SynDiff to enhance the contextual sensitivity and generalization ability of the model. Alternatively, a diffusive generator can be used in ResViT to improve fidelity and diversity in the output images. Second, federated learning techniques can be adopted to train the models on decentralized and distributed data sources, which can mitigate privacy concerns and increase data availability [44]. For example, partial model aggregation or personalization blocks can be used to cope with data heterogeneity and domain shifts across different sites [43]. Third, the computational efficiency of the models can be optimized by employing acceleration strategies, such as latent space diffusion or intermediate image initialization, to reduce the inference time and training time of the models [168].

Bibliography

- [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*, pp. 631–638, 2013.
- [2] B. J. Pichler, M. S. Judenhofer, and C. Pfannenberger, *Multimodal Imaging Approaches: PET/CT and PET/MRI*, pp. 109–132. Springer, 2008.
- [3] B. Moraal, S. Roosendaal, P. Pouwels, H. Vrenken, R. Schijndel, D. Meier, C. Guttman, J. Geurts, and F. Barkhof, “Multi-contrast, isotropic, single-slab 3d MR imaging in multiple sclerosis,” vol. 18, pp. 2311–2320, 2008.
- [4] T. Çukur, N. K. Bangerter, and D. G. Nishimura, “Enhanced spectral shaping in steady-state free precession imaging,” *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 58, no. 6, pp. 1216–1223, 2007.
- [5] T. Çukur and D. G. Nishimura, “Fat-water separation with alternating repetition time balanced ssfp,” *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 60, no. 2, pp. 479–484, 2008.
- [6] T. Çukur and D. G. Nishimura, “Multiple repetition time balanced steady-state free precession imaging,” *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 62, no. 1, pp. 193–204, 2009.

- [7] T. Çukur, J. H. Lee, N. K. Bangerter, B. A. Hargreaves, and D. G. Nishimura, “Non-contrast-enhanced flow-independent peripheral mr angiography with balanced ssfp,” *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 61, no. 6, pp. 1533–1539, 2009.
- [8] T. Çukur, M. Yamada, W. R. Overall, P. Yang, and D. G. Nishimura, “Positive contrast with alternating repetition time ssfp (parts): A fast imaging technique for spio-labeled cells,” *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 63, no. 2, pp. 427–437, 2010.
- [9] K. Keskin, U. Yılmaz, and T. Çukur, “Constrained ellipse fitting for efficient parameter mapping with phase-cycled bssfp mri,” *IEEE Transactions on Medical Imaging*, vol. 41, no. 1, pp. 14–26, 2022.
- [10] E. Bıyık, K. Keskin, S. UH Dar, A. Koç, and T. Çukur, “Factorized sensitivity estimation for artifact suppression in phase-cycled bssfp mri,” *NMR in Biomedicine*, vol. 33, no. 4, p. e4228, 2020.
- [11] T. Çukur, M. Lustig, and D. G. Nishimura, “Multiple-profile homogeneous image combination: application to phase-cycled ssfp and multicoil imaging,” *Magnetic resonance in medicine*, vol. 60, no. 3, pp. 732–738, 2008.
- [12] T. Çukur, M. Lustig, E. U. Saritas, and D. G. Nishimura, “Signal compensation and compressed sensing for magnetization-prepared mr angiography,” *IEEE transactions on medical imaging*, vol. 30, no. 5, pp. 1017–1027, 2011.
- [13] T. Çukur, A. Shimakawa, H. Yu, B. A. Hargreaves, B. S. Hu, D. G. Nishimura, and J. H. Brittain, “Magnetization-prepared ideal bssfp: a flow-independent technique for noncontrast-enhanced peripheral angiography,” *Journal of Magnetic Resonance Imaging*, vol. 33, no. 4, pp. 931–939, 2011.
- [14] N. K. Bangerter, T. Cukur, B. A. Hargreaves, B. S. Hu, J. H. Brittain, D. Park, G. E. Gold, and D. G. Nishimura, “Three-dimensional fluid-suppressed t2-prep flow-independent peripheral angiography using balanced ssfp,” *Magnetic resonance imaging*, vol. 29, no. 8, pp. 1119–1124, 2011.

- [15] 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, p. 101331I, 2017.
- [16] M. Aksoy, C. Forman, M. Straka, T. Çukur, J. Hornegger, and R. Bammer, “Hybrid prospective and retrospective head motion correction to mitigate cross-calibration errors,” *Magnetic resonance in medicine*, vol. 67, no. 5, pp. 1237–1251, 2012.
- [17] B. Thukral, “Problems and preferences in pediatric imaging,” vol. 25, pp. 359–364, 2015.
- [18] K. Krupa and M. Bekiesińska-Figatowska, “Artifacts in magnetic resonance imaging,” vol. 80, pp. 93–106, 2015.
- [19] M. Lustig, D. Donoho, J. Santos, and J. Pauly, “Compressed sensing mri,” *Signal Processing Magazine, IEEE*, vol. 25, pp. 72 – 82, 04 2008.
- [20] M. Shahdloo, E. Ilıcak, M. Tofghi, E. U. Saritas, A. E. Çetin, and T. Çukur, “Projection onto epigraph sets for rapid self-tuning compressed sensing mri,” *IEEE transactions on medical imaging*, vol. 38, no. 7, pp. 1677–1689, 2018.
- [21] T. Çukur, J. M. Santos, J. M. Pauly, and D. G. Nishimura, “Variable-density parallel imaging with partially localized coil sensitivities,” *IEEE transactions on medical imaging*, vol. 29, no. 5, pp. 1173–1181, 2010.
- [22] B. Gözcü, R. K. Mahabadi, Y.-H. Li, E. Ilıcak, T. Cukur, J. Scarlett, and V. Cevher, “Learning-based compressive mri,” *IEEE transactions on medical imaging*, vol. 37, no. 6, pp. 1394–1406, 2018.
- [23] S. U. H. Dar, M. Özbey, A. B. Çatlı, and T. Çukur, “A transfer-learning approach for accelerated mri using deep neural networks,” *Magnetic resonance in medicine*, vol. 84, no. 2, pp. 663–685, 2020.
- [24] S. U. H. Dar, M. Yurt, M. Shahdloo, M. E. Ildız, B. Tınaz, and T. Çukur, “Prior-guided image reconstruction for accelerated multi-contrast MRI via

- generative adversarial networks,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 6, pp. 1072–1087, 2020.
- [25] T. Cukur, M. Lustig, and D. G. Nishimura, “Improving non-contrast-enhanced steady-state free precession angiography with compressed sensing,” *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 61, no. 5, pp. 1122–1131, 2009.
- [26] L. K. Senel, T. Kilic, A. Gungor, E. Kopanoglu, H. E. Guven, E. U. Saritas, A. Koc, and T. Cukur, “Statistically segregated k-space sampling for accelerating multiple-acquisition mri,” *IEEE transactions on medical imaging*, vol. 38, no. 7, pp. 1701–1714, 2019.
- [27] T. Çukur, “Accelerated phase-cycled ssfp imaging with compressed sensing,” *IEEE transactions on medical imaging*, vol. 34, no. 1, pp. 107–115, 2014.
- [28] T. Çukur and E. Ü. Saritaş, “Comparison of system-matrix-based and projection-based reconstructions for field free line magnetic particle imaging,” *International Journal on Magnetic Particle Imaging*, vol. 3, no. 1, pp. 1–8, 2017.
- [29] E. Ilıcak, S. Ozdemir, J. Zapp, L. R. Schad, and F. G. Zöllner, “Dynamic mode decomposition of dynamic mri for assessment of pulmonary ventilation and perfusion,” *Magnetic Resonance in Medicine*, 2023.
- [30] E. Kopanoglu, A. Güngör, T. Kilic, E. U. Saritas, K. K. Oguz, T. Çukur, and H. E. Güven, “Simultaneous use of individual and joint regularization terms in compressive sensing: Joint reconstruction of multi-channel multi-contrast mri acquisitions,” *NMR in Biomedicine*, vol. 33, no. 4, p. e4247, 2020.

- [31] T. Sanchez, B. Gözcü, R. B. van Heeswijk, A. Eftekhari, E. Ilıcak, T. Çukur, and V. Cevher, “Scalable learning-based sampling optimization for compressive dynamic mri,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8584–8588, IEEE, 2020.
- [32] T. Çukur, J. M. Santos, D. G. Nishimura, and J. M. Pauly, “Varying kernel-extent gridding reconstruction for undersampled variable-density spirals,” *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 59, no. 1, pp. 196–201, 2008.
- [33] S. U. H. Dar, M. Yurt, and T. Çukur, “A few-shot learning approach for accelerated mri via fusion of data-driven and subject-driven priors,” *arXiv preprint arXiv:2103.07790*, 2021.
- [34] A. Koç, B. Bartan, E. Gundogdu, T. Çukur, and H. M. Ozaktas, “Sparse representation of two-and three-dimensional images with fractional fourier, hartley, linear canonical, and haar wavelet transforms,” *Expert Systems with Applications*, vol. 77, pp. 247–255, 2017.
- [35] T. Çukur, J. M. Santos, J. M. Pauly, and D. G. Nishimura, “Variable-density parallel imaging with partially localized coil sensitivities,” *IEEE transactions on medical imaging*, vol. 29, no. 5, pp. 1173–1181, 2010.
- [36] A. Güngör, E. Kopanoğlu, T. Çukur, and H. E. Güven, “Fast recovery of compressed multi-contrast magnetic resonance images,” in *Medical Imaging 2017: Image Processing*, vol. 10133, pp. 503–508, SPIE, 2017.
- [37] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar, “Advances and challenges in super-resolution,” vol. 14, pp. 47–57, 2004.
- [38] O. Dalmaz, B. Saglam, K. Gönç, and T. Çukur, “edagan: Encoder-decoder attention generative adversarial networks for multi-contrast mr image synthesis,” in *2022 9th International Conference on Electrical and Electronics Engineering (ICEEE)*, pp. 320–324, IEEE, 2022.

- [39] O. Dalmaz, B. Sağlam, K. Gönç, S. U. Dar, and T. Çukur, “Bottleneck sharing generative adversarial networks for unified multi-contrast mr image synthesis,” in *2022 30th Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4, IEEE, 2022.
- [40] S. Y. Selçuk, O. Dalmaz, S. U. H. Dar, and T. Çukur, “Improving image synthesis quality in multi-contrast mri using transfer learning via autoencoders,” in *2022 30th Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4, IEEE, 2022.
- [41] O. Dalmaz, I. Aytekin, S. U. H. Dar, A. Erdem, E. Erdem, and T. Cukur, “Multi-contrast mri synthesis with channel-exchanging-network,” in *2022 30th Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4, IEEE, 2022.
- [42] M. U. Mirza, O. Dalmaz, and T. Çukur, “Skip connections for medical image synthesis with generative adversarial networks,” in *2022 30th Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4, IEEE, 2022.
- [43] O. Dalmaz, U. Mirza, G. Elmas, M. Özbey, S. U. Dar, E. Ceyani, S. Avestimehr, and T. Çukur, “One model to unite them all: Personalized federated learning of multi-contrast mri synthesis,” *arXiv preprint arXiv:2207.06509*, 2022.
- [44] O. Dalmaz, U. Mirza, G. Elmas, M. Özbey, S. U. Dar, and T. Çukur, “A specificity-preserving generative model for federated mri translation,” in *Distributed, Collaborative, and Federated Learning, and Affordable AI and Healthcare for Resource Diverse Global Health: Third MICCAI Workshop, DeCaF 2022, and Second MICCAI Workshop, FAIR 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18 and 22, 2022, Proceedings*, pp. 79–88, Springer, 2022.
- [45] C. Bowles, C. Qin, C. Ledig, R. Guerrero, R. Gunn, A. Hammers, E. Sakka, D. Dickie, M. Hernández, N. Royle, J. Wardlaw, H. Rhodius-Meester, B. Tijms, A. Lemstra, W. Flier, F. Barkhof, P. Scheltens, and D. Rueckert,

- “Pseudo-healthy image synthesis for white matter lesion segmentation,” in *Simul Synth Med Imaging*, pp. 87–96, 2016.
- [46] A. Chartsias, T. Joyce, M. V. Giuffrida, and S. A. Tsiftaris, “Multimodal MR synthesis via modality-invariant latent representation,” vol. 37, no. 3, pp. 803–814, 2018.
 - [47] W. Wei, E. Poirion, B. Bodini, S. Durrleman, O. Colliot, B. Stankoff, and N. Ayache, “Fluid-attenuated inversion recovery MRI synthesis from multisequence MRI using three-dimensional fully convolutional networks for multiple sclerosis,” vol. 6, no. 1, p. 014005, 2019.
 - [48] N. Cordier, H. Delingette, M. Le, and N. Ayache, “Extended modality propagation: Image synthesis of pathological cases,” vol. 35, pp. 2598–2608, 2016.
 - [49] V. Sevetlidis, M. V. Giuffrida, and S. A. Tsiftaris, “Whole image synthesis using a deep encoder-decoder network,” in *Simul Synth Med Imaging*, pp. 127–137, 2016.
 - [50] T. Joyce, A. Chartsias, and S. A. Tsiftaris, “Robust multi-modal MR image synthesis,” in *Med Image Comput Comput Assist Interv*, pp. 347–355, 2017.
 - [51] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” in *Adv Neural Inf Process Syst*, vol. 24, 2014.
 - [52] A. Beers, J. Brown, K. Chang, J. Campbell, S. Ostmo, M. Chiang, and J. Kalpathy-Cramer, “High-resolution medical image synthesis using progressively grown generative adversarial networks,” *arXiv:1805.03144*, 2018.
 - [53] S. U. Dar, M. Yurt, L. Karacan, A. Erdem, E. Erdem, and T. Çukur, “Image synthesis in multi-contrast MRI with conditional generative adversarial networks,” vol. 38, no. 10, pp. 2375–2388, 2019.
 - [54] 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.

- [55] D. Nie, R. Trullo, J. Lian, L. Wang, C. Petitjean, S. Ruan, and Q. Wang, “Medical image synthesis with deep convolutional adversarial networks,” vol. 65, no. 12, pp. 2720–2730, 2018.
- [56] K. Armanious, C. Jiang, M. Fischer, T. Küstner, T. Hepp, K. Nikolaou, S. Gatidis, and B. Yang, “MedGAN: Medical image translation using GANs,” vol. 79, p. 101684, 2019.
- [57] D. Lee, J. Kim, W.-J. Moon, and J. C. Ye, “CollaGAN: Collaborative GAN for missing image data imputation,” in *Comput Vis Pattern Recognit*, pp. 2487–2496, 2019.
- [58] H. Li, J. C. Paetzold, A. Sekuboyina, F. Kofler, J. Zhang, J. S. Kirschke, B. Wiestler, and B. Menze, “DiamondGAN: Unified multi-modal generative adversarial networks for MRI sequences synthesis,” in *Med. Image Comput Comput Assist Interv*, pp. 795–803, 2019.
- [59] T. Zhou, H. Fu, G. Chen, J. Shen, and L. Shao, “Hi-Net: Hybrid-fusion network for multi-modal MR image synthesis,” vol. 39, no. 9, pp. 2772–2781, 2020.
- [60] 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.
- [61] 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,” vol. 70, p. 101944, 2021.
- [62] H. Yang, X. Lu, S.-H. Wang, Z. Lu, J. Yao, Y. Jiang, and P. Qian, “Synthesizing multi-contrast MR images via novel 3D conditional variational auto-encoding GAN,” vol. 26, pp. 1–10, 2021.
- [63] 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,” vol. 38, no. 7, pp. 1750–1762, 2019.

- [64] A. Sharma and G. Hamarneh, “Missing MRI pulse sequence synthesis using multi-modal generative adversarial network,” vol. 39, pp. 1170–1183, 2020.
- [65] G. Wang, E. Gong, S. Banerjee, D. Martin, E. Tong, J. Choi, H. Chen, M. Wintermark, J. M. Pauly, and G. Zaharchuk, “Synthesize high-quality multi-contrast magnetic resonance imaging from multi-echo acquisition using multi-task deep generative model,” vol. 39, no. 10, pp. 3089–3099, 2020.
- [66] O. Dalmaz, U. Mirza, G. Elmas, M. Özbey, S. U. Dar, and T. Çukur, “A specificity-preserving generative model for federated mri translation,” in *Distributed, Collaborative, and Federated Learning, and Affordable AI and Healthcare for Resource Diverse Global Health: Third MICCAI Workshop, DeCaF 2022, and Second MICCAI Workshop, FAIR 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18 and 22, 2022, Proceedings*, pp. 79–88, Springer, 2022.
- [67] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *Comput Vis Pattern Recognit*, pp. 7794–7803, 2018.
- [68] N. Kodali, J. Hays, J. Abernethy, and Z. Kira, “On convergence and stability of GANs,” *arXiv:1705.07215*, 2017.
- [69] I. Aytekin, O. Dalmaz, H. Ankishan, E. U. Saritas, U. Bagci, T. Cukur, and H. Celik, “Detecting COVID-19 from respiratory sound recordings with transformers,” in *Medical Imaging 2022: Computer-Aided Diagnosis* (K. Drukker and K. M. Iftikharuddin, eds.), vol. 12033, p. 1203306, International Society for Optics and Photonics, SPIE, 2022.
- [70] I. Aytekin, O. Dalmaz, K. Gonc, H. Ankishan, E. U. Saritas, U. Bagci, H. Celik, and T. Cukur, “Covid-19 detection from respiratory sounds with hierarchical spectrogram transformers,” *arXiv preprint arXiv:2207.09529*, 2022.
- [71] Y. Korkmaz, M. Özbey, and T. Cukur, “Mri reconstruction with conditional adversarial transformers,” in *Machine Learning for Medical Image*

- Reconstruction: 5th International Workshop, MLMIR 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings*, pp. 62–71, Springer, 2022.
- [72] Y. Korkmaz, M. Yurt, S. U. H. Dar, M. Özbey, and T. Cukur, “Deep mri reconstruction with generative vision transformers,” in *Machine Learning for Medical Image Reconstruction: 4th International Workshop, MLMIR 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 4*, pp. 54–64, Springer, 2021.
 - [73] 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.
 - [74] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” *Int Conf Comput Vis*, pp. 2242–2251, 2017.
 - [75] P. Dhariwal and A. Nichol, “Diffusion models beat gans on image synthesis,” in *Adv Neural Inf Process Syst*, vol. 34, pp. 8780–8794, 2021.
 - [76] M. Özbey, S. U. Dar, H. A. Bedel, O. Dalmaz, Ş. Öztürk, A. Güngör, and T. Çukur, “Unsupervised medical image translation with adversarial diffusion models,” *arXiv preprint arXiv:2207.08208*, 2022.
 - [77] M. Yurt, M. Özbey, S. U. Dar, B. Tinaz, K. K. Oguz, and T. Çukur, “Progressively volumetrized deep generative models for data-efficient contextual learning of mr image recovery,” *Medical Image Analysis*, vol. 78, p. 102429, 2022.
 - [78] S. Turk, A. Demirkaya, M. Y. Turali, C. Hepdurgun, S. U. Dar, A. K. Karabulut, A. Azizova, M. Orman, I. Tamsel, U. Aydingoz, M. Argin, and T. Cukur, “Jointnet: A deep model for predicting active sacroiliitis from sacroiliac joint radiography,” 2023.
 - [79] F. T. Y. Vural, S. D. Newman, T. Çukur, and I. Ö. Ertugrul, “Machine learning methods for human brain imaging,” *Frontiers in Neuroinformatics*, vol. 17, 2023.

- [80] I. Sivgin, H. A. Bedel, Ş. Öztürk, and T. Çukur, “A plug-in graph neural network to boost temporal sensitivity in fmri analysis,” *arXiv preprint arXiv:2301.00439*, 2023.
- [81] Ş. Öztürk and T. Çukur, “Deep clustering via center-oriented margin free-triplet loss for skin lesion detection in highly imbalanced datasets,” *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 9, pp. 4679–4690, 2022.
- [82] Y. Huo, Z. Xu, S. Bao, A. Assad, R. G. Abramson, and B. A. Landman, “Adversarial synthesis learning enables segmentation without target modality ground truth,” in *Int. Symp. BioMed Imaging*, pp. 1217–1220, 2018.
- [83] 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*, pp. 606–613, 2013.
- [84] C. Catana, A. van der Kouwe, T. Benner, C. J. Michel, M. Hamm, M. Fenchel, B. Fischl, B. Rosen, M. Schmand, and A. G. Sorensen, “Toward implementing an MRI-based PET attenuation-correction method for neurologic studies on the MR-PET brain prototype,” vol. 51, no. 9, pp. 1431–1438, 2010.
- [85] S. Roy, A. Jog, A. Carass, and J. L. Prince, “Atlas based intensity transformation of brain MR images,” in *Multimodal Brain Image Anal.*, pp. 51–62, 2013.
- [86] 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.
- [87] Y. Huang, L. Shao, and A. F. Frangi, “Cross-modality image synthesis via weakly coupled and geometry co-regularized joint dictionary learning,” vol. 37, no. 3, pp. 815–827, 2018.

- [88] 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*, pp. 291–298, 2017.
- [89] A. Jog, A. Carass, S. Roy, D. L. Pham, and J. L. Prince, “Random forest regression for magnetic resonance image synthesis,” vol. 35, pp. 475–488, 2017.
- [90] 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*, pp. 677–684, 2015.
- [91] R. Vemulapalli, H. Van Nguyen, and S. K. Zhou, “Unsupervised cross-modal synthesis of subject-specific scans,” in *Int Conf Comput Vis*, pp. 630–638, 2015.
- [92] Y. Wu, W. Yang, L. Lu, Z. Lu, L. Zhong, M. Huang, Y. Feng, Q. Feng, and W. Chen, “Prediction of CT substitutes from MR images based on local diffeomorphic mapping for brain PET attenuation correction,” vol. 57, no. 10, pp. 1635–1641, 2016.
- [93] 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*, pp. 225–232, 2014.
- [94] T. Huynh, Y. Gao, J. Kang, L. Wang, P. Zhang, J. Lian, and D. Shen, “Estimating CT image from MRI data using structured random forest and auto-context model,” vol. 35, no. 1, pp. 174–183, 2016.
- [95] P. Coupe, J. V. Manjón, M. Chamberland, M. Descoteaux, and B. Hiba, “Collaborative patch-based super-resolution for diffusion-weighted images,” *NeuroImage*, vol. 83, pp. 245–261, 2013.
- [96] A. Adam, A. Dixon, J. Gillard, C. Schaefer-Prokop, R. Grainger, and D. Allison, *Grainger & Allison’s Diagnostic Radiology*. Elsevier, 2014.

- [97] D. Ellison, S. Love, L. Chimelli, B. Harding, J. Lowe, H. Vinters, S. Brandner, and W. Yong, *Neuropathology: A Reference Text of CNS Pathology*. Elsevier, 2012.
- [98] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv:2010.11929*, 2021.
- [99] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, “TransUNet: Transformers make strong encoders for medical image segmentation,” *arXiv:2102.04306*, 2021.
- [100] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Comput Vis Pattern Recognit*, pp. 770–778, 2016.
- [101] X. Yi, E. Walia, and P. Babyn, “Generative adversarial network in medical imaging: A review,” vol. 58, p. 101552, 2019.
- [102] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, “A survey on deep learning in medical image analysis,” vol. 42, pp. 60–88, 2017.
- [103] R. Li, W. Zhang, H.-I. Suk, L. Wang, J. Li, D. Shen, and S. Ji, “Deep learning based imaging data completion for improved brain disease diagnosis,” in *Med. Image Comput Comput Assist Interv*, pp. 305–312, 2014.
- [104] A. Torrado-Carvajal, J. L. Herraiz, E. Alcain, A. S. Montemayor, L. Garcia-Cañamaque, J. A. Hernandez-Tamames, Y. Rozenholc, and N. Malpica, “Fast patch-based pseudo-CT synthesis from T1-weighted MR images for PET/MR attenuation correction in brain studies,” vol. 57, no. 1, pp. 136–143, 2016.
- [105] K. Bahrami, F. Shi, I. Rekik, and D. Shen, “Convolutional neural network for reconstruction of 7T-like images from 3T MRI using appearance and anatomical features,” in *Deep Learn Data Label Med Appl*, pp. 39–47, 2016.

- [106] K. Bahrami, F. Shi, X. Zong, H. W. Shin, H. An, and D. Shen, “Reconstruction of 7T-like images from 3T MRI,” vol. 35, no. 9, pp. 2085–2097, 2016.
- [107] Y. Zhang, P.-T. Yap, L. Qu, J.-Z. Cheng, and D. Shen, “Dual-domain convolutional neural networks for improving structural information in 3T MRI,” vol. 64, pp. 90–100, 2019.
- [108] X. Han, “MR-based synthetic CT generation using a deep convolutional neural network method,” vol. 44, no. 4, pp. 1408–1419, 2017.
- [109] 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*, pp. 170–178, 2016.
- [110] H. Arabi, G. Zeng, G. Zheng, and H. Zaidi, “Novel deep learning-based CT synthesis algorithm for MRI-guided PET attenuation correction in brain PET/MR imaging,” in *Nucl. Sci. Symp.*, pp. 1–3, 2018.
- [111] K. Klaser, T. Varsavsky, P. Markiewicz, T. Vercauteren, D. Atkinson, K. Thielemans, B. Hutton, M. J. Cardoso, and S. Ourselin, “Improved MR to CT synthesis for PET/MR attenuation correction using imitation learning,” in *Simul Synth Med Imaging*, pp. 13–21, 2019.
- [112] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” *arXiv:1411.1784*, 2014.
- [113] V. Sandfort, K. Yan, P. Pickhardt, and R. Summers, “Data augmentation using generative adversarial networks (cycleGAN) to improve generalizability in CT segmentation tasks,” vol. 9, 11 2019.
- [114] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, “GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification,” *Neurocomput*, vol. 321, pp. 321–331, 2018.
- [115] A. Ben-Cohen, E. Klang, S. Raskin, S. Soffer, S. Ben-Haim, E. Konen, M. Amitai, and H. Greenspan, “Cross-modality synthesis from CT to PET

using FCN and GAN networks for improved automated lesion detection,” vol. 78, 2018.

- [116] G. Santini, C. Fourcade, N. Moreau, C. Rousseau, L. Ferrer, M. Lacombe, V. Fleury, M. Campone, P. Jezequel, and M. Rubeaux, “Unpaired PET/CT image synthesis of liver region using CycleGAN,” in *Int Symp Med Inf Process Anal*, vol. 11583, pp. 247 – 257, 2020.
- [117] C.-B. Jin, H. Kim, M. Liu, W. Jung, S. Joo, E. Park, Y. S. Ahn, I. H. Han, J. I. Lee, and X. Cui, “Deep CT to MR synthesis using paired and unpaired data,” *Sensors*, vol. 19, no. 10, p. 2361, 2019.
- [118] Y. Ge, D. Wei, Z. Xue, Q. Wang, X. Zhou, Y. Zhan, and S. Liao, “Unpaired MR to CT synthesis with explicit structural constrained adversarial learning,” in *Int. Symp. Biomed. Imaging*, pp. 1096–1099, 2019.
- [119] J. Wolterink, A. M. Dinkla, M. Savenije, P. Seevinck, C. Berg, and I. Isgum, “Deep MR to CT synthesis using unpaired data,” in *Simul Synth Med Imaging*, pp. 14–23, 2017.
- [120] X. Dong, T. Wang, Y. Lei, K. Higgins, T. Liu, W. Curran, H. Mao, J. Nye, and X. Yang, “Synthetic CT generation from non-attenuation corrected PET images for whole-body PET imaging,” vol. 64, no. 21, p. 215016, 2019.
- [121] H. Yang, J. Sun, A. Carass, C. Zhao, J. Lee, Z. Xu, and J. Prince, “Unpaired brain MR-to-CT synthesis using a structure-constrained cycleGAN,” *arXiv:1809.04536*, 2018.
- [122] Y. Hiasa, Y. Otake, M. Takao, T. Matsuoka, K. Takashima, A. Carass, J. Prince, N. Sugano, and Y. Sato, “Cross-modality image synthesis from unpaired data using cycleGAN: Effects of gradient consistency loss and training data size,” in *Simul Synth Med Imaging*, pp. 31–41, 2018.
- [123] A. Chartsias, T. Joyce, R. Dharmakumar, and S. A. Tsaftaris, “Adversarial image synthesis for unpaired multi-modal cardiac data,” in *Simul Synth Med Imaging*, pp. 3–13, 2017.

- [124] H. Do, P. Bourdon, D. Helbert, M. Naudin, and R. Guillemin, “7T MRI super-resolution with generative adversarial network,” in *Int. Symp. Electronic Imaging*, pp. 106–1 – 106–7, 2021.
- [125] L. Xiang, Y. Li, W. Lin, Q. Wang, and D. Shen, “Unpaired deep cross-modality synthesis with fast training,” in *Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support*, pp. 155–164, 2018.
- [126] V. Kearney, B. P. Ziemer, A. Perry, T. Wang, J. W. Chan, L. Ma, O. Morin, S. S. Yom, and T. D. Solberg, “Attention-aware discrimination for MR-to-CT image translation using cycle-consistent generative adversarial networks,” vol. 2, no. 2, p. e190027, 2020.
- [127] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, “Self-attention generative adversarial networks,” in *Int Conf Mach Learn*, vol. 97, pp. 7354–7363, 2019.
- [128] O. Oktay, J. Schlemper, L. L. Folgoc, M. J. Lee, M. Heinrich, K. Misawa, K. Mori, S. G. McDonagh, N. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, “Attention U-Net: Learning where to look for the pancreas,” *arXiv:1804.03999*, 2018.
- [129] J. Zhao, D. Li, Z. Kassam, J. Howey, J. Chong, B. Chen, and S. Li, “Tripartite-GAN: Synthesizing liver contrast-enhanced MRI to improve tumor detection,” vol. 63, p. 101667, 2020.
- [130] Z. Yuan, M. Jiang, Y. Wang, B. Wei, Y. Li, P. Wang, W. Menpes-Smith, Z. Niu, and G. Yang, “SARA-GAN: Self-attention and relative average discriminator based generative adversarial networks for fast compressed sensing MRI reconstruction,” vol. 14, p. 58, 2020.
- [131] M. Li, W. Hsu, X. Xie, J. Cong, and W. Gao, “SACNN: Self-attention convolutional neural network for low-dose CT denoising with self-supervised perceptual loss network,” vol. 39, no. 7, pp. 2289–2301, 2020.
- [132] Y. Xie, J. Zhang, C. Shen, and Y. Xia, “CoTr: Efficiently bridging CNN and transformer for 3D medical image segmentation,” *arXiv:2103.03024*, 2021.

- [133] Y. Dai and Y. Gao, “TransMed: Transformers advance multi-modal medical image classification,” *arXiv:2103.05940*, 2021.
- [134] D. Karimi, S. Vasylechko, and A. Gholipour, “Convolution-free medical image segmentation using transformers,” *arXiv:2102.13645*, 2021.
- [135] Y. Luo, Y. Wang, C. Zu, B. Zhan, X. Wu, J. Zhou, D. Shen, and L. Zhou, “3D Transformer-GAN for high-quality PET reconstruction,” in *Med Image Comput Comput Assist Interv*, pp. 276–285, 2021.
- [136] Z. Zhang, L. Yu, X. Liang, W. Zhao, and L. Xing, “TransCT: Dual-path transformer for low dose computed tomography,” in *Med. Image Comput Comput Assist Interv*, pp. 55–64, 2021.
- [137] Y. Korkmaz, S. U. H. Dar, M. Yurt, M. Özbey, and T. Çukur, “Unsupervised mri reconstruction via zero-shot learned adversarial transformers,” *IEEE Transactions on Medical Imaging*, vol. 41, no. 7, pp. 1747–1763, 2022.
- [138] S. A. Kamran, K. F. Hossain, A. Tavakkoli, S. L. Zuckerbrod, K. M. Sanders, and S. A. Baker, “VTGAN: Semi-supervised retinal image synthesis and disease prediction using vision transformers,” *arXiv:2104.06757*, 2021.
- [139] H.-C. Shin, A. Ihsani, S. Mandava, S. T. Sreenivas, C. Forster, J. Cha, and A. D. N. Initiative, “GANBERT: Generative adversarial networks with bidirectional encoder representations from transformers for MRI to PET synthesis,” *arXiv:2008.04393*, 2020.
- [140] X. Zhang, X. He, J. Guo, N. Ettehadi, N. Aw, D. Semanek, J. Posner, A. Laine, and Y. Wang, “PTNet: A high-resolution infant MRI synthesizer based on transformer,” *arXiv:2105.13993*, 2021.
- [141] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” *Adv Neural Inf Process Syst*, pp. 1–11, 2017.
- [142] S. Haykin, *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994.

- [143] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M.-A. Weber, T. Arbel, B. B. Avants, N. Ayache, P. Buendia, D. L. Collins, N. Cordier, J. J. Corso, A. Criminisi, T. Das, H. Delingette, Demiralp, C. R. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, K. M. Iftekharuddin, R. Jena, N. M. John, E. Konukoglu, D. Lashkari, J. A. Mariz, R. Meier, S. Pereira, D. Precup, S. J. Price, T. R. Raviv, S. M. S. Reza, M. Ryan, D. Sarikaya, L. Schwartz, H.-C. Shin, J. Shotton, C. A. Silva, N. Sousa, N. K. Subbanna, G. Szekely, T. J. Taylor, O. M. Thomas, N. J. Tustison, G. Unal, F. Vasseur, M. Wintermark, D. H. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes, and K. Van Leemput, “The multimodal brain tumor image segmentation benchmark (BRATS),” vol. 34, no. 10, pp. 1993–2024, 2015.
- [144] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, J. Freymann, K. Farahani, and C. Davatzikos, “Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features,” vol. 4, p. 170117, 2017.
- [145] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, and et al., “Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge,” *arXiv:1811.02629*, 2019.
- [146] T. Nyholm, S. Svensson, S. Andersson, J. Jonsson, M. Sohlén, C. Gustafsson, E. Kjellén, K. Söderström, P. Albertsson, L. Blomqvist, B. Zackrisson, L. E. Olsson, and A. Gunnlaugsson, “MR and CT data with multiobserver delineations of organs in the pelvic area—part of the gold atlas project,” vol. 45, no. 3, pp. 1295–1300, 2018.
- [147] M. Jenkinson and S. Smith, “A global optimisation method for robust affine registration of brain images,” vol. 5, pp. 143–156, 2001.
- [148] H. Lan, the Alzheimer Disease Neuroimaging Initiative, A. W. Toga, and F. Sepehrband, “Three-dimensional self-attention conditional GAN with

- spectral normalization for multimodal neuroimaging synthesis,” vol. 86, no. 3, pp. 1718–1733, 2021.
- [149] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Int Conf Learn Represent*, 2015.
 - [150] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor, “Imagenet-21k pretraining for the masses,” *arXiv:2104.10972*, 2021.
 - [151] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” vol. 13, pp. 600 – 612, 2004.
 - [152] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs trained by a two time-scale update rule converge to a local nash equilibrium,” in *Adv Neural Inf Process Syst*, p. 6629–6640, 2017.
 - [153] S. Abnar and W. Zuidema, “Quantifying attention flow in transformers,” in *Assoc. Comput. Linguist.*, pp. 4190–4197, 2020.
 - [154] R. Durall, M. Keuper, and J. Keuper, “Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions,” in *Comput Vis Pattern Recognit*, pp. 7887–7896, 2020.
 - [155] L.-H. Chen, C. G. Bampis, Z. Li, C. Chen, and A. C. Bovik, “Convolutional block design for learned fractional downsampling,” *arXiv:2105.09999*, 2021.
 - [156] M. Yurt, O. Dalmaz, S. U. H. Dar, M. Özbey, B. Tınaz, K. K. Oğuz, and T. Çukur, “Semi-supervised learning of MRI synthesis without fully-sampled ground truths,” *IEEE Trans Med Imaging*, vol. 41, no. 12, pp. 3895–3906, 2022.
 - [157] S. U. Dar, Ş. Öztürk, M. Özbey, and T. Çukur, “Learning deep mri reconstruction models from scratch in low-data regimes,” *arXiv preprint arXiv:2301.02613*, 2023.
 - [158] 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.

- [159] 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.
- [160] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” in *Adv Neural Inf Process Syst*, vol. 33, pp. 6840–6851, 2020.
- [161] 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), pp. 14–23, 2017.
- [162] 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), pp. 22–31, 2019.
- [163] 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, pp. 14938–14954, 2021.
- [164] H. Chung and J. C. Ye, “Score-based diffusion models for accelerated mri,” *Med Image Anal*, vol. 80, p. 102479, 2022.
- [165] Y. Song, L. Shen, L. Xing, and S. Ermon, “Solving inverse problems in medical imaging with score-based generative models,” 2022.
- [166] S. U. Dar, Şaban Öztürk, Y. Korkmaz, G. Elmas, M. Özbey, A. Güngör, and T. Çukur, “Adaptive diffusion priors for accelerated mri reconstruction,” *arXiv:2207.05876*, 2022.
- [167] H. Chung, E. S. Lee, and J. C. Ye, “Mr image denoising and super-resolution using regularized reverse diffusion,” *arXiv:2203.12621*, 2022.
- [168] W. H. L. Pinaya, P.-D. Tudosiu, J. Dafflon, P. F. da Costa, V. Fernandez, P. Nachev, S. Ourselin, and M. J. Cardoso, “Brain imaging generation with latent diffusion models,” *arXiv:2209.07162*, 2022.
- [169] J. Wolleb, F. Bieder, R. Sandkühler, and P. C. Cattin, “Diffusion models for medical anomaly detection,” vol. 13438, pp. 35–45, Springer, 2022.

- [170] W. H. L. Pinaya, M. S. Graham, R. Gray, P. F. Da Costa, P.-D. Tudosiu, P. Wright, Y. H. Mah, A. D. MacKinnon, J. T. Teo, R. Jager, D. Werring, G. Rees, P. Nachev, S. Ourselin, and M. J. Cardoso, “Fast Unsupervised Brain Anomaly Detection and Segmentation with Diffusion Models,” *arXiv:2206.03461*, 2022.
- [171] Z. Xiao, K. Kreis, and A. Vahdat, “Tackling the generative learning trilemma with denoising diffusion GANs,” in *Int Conf Learn Represent*, 2022.
- [172] H. Sasaki, C. G. Willcocks, and T. P. Breckon, “Unit-DDPM: Unpaired image translation with denoising diffusion probabilistic models,” *arXiv:2104.05358*, 2021.
- [173] X. Meng, Y. Gu, Y. Pan, N. Wang, P. Xue, M. Lu, X. He, Y. Zhan, and D. Shen, “A novel unified conditional score-based generative framework for multi-modal medical image completion,” *arXiv:2207.03430*, 2022.
- [174] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, “Deep unsupervised learning using nonequilibrium thermodynamics,” in *Int Conf Mach Learn*, pp. 2256–2265, 2015.
- [175] W. Feller, “On the theory of stochastic processes, with particular reference to applications,” in *Proc Berkeley Symp Math Stat Probab*, vol. 1, pp. 403–433, University of California Press, 1949.
- [176] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Adv Neural Inf Process Syst*, vol. 33, pp. 6840–6851, 2020.
- [177] 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.
- [178] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” *Adv Neural Inf Process Syst*, vol. 27, 2014.

- [179] L. Mescheder, A. Geiger, and S. Nowozin, “Which training methods for gans do actually converge?,” in *Int Conf Mach Learn*, pp. 3481–3490, 2018.
- [180] 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.
- [181] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *IEEE Conf Comput Vis Pattern Recognit*, pp. 770–778, 2016.
- [182] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Med Image Comput Comput Assist Inter*, pp. 234–241, Springer, 2015.
- [183] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” pp. 4401–4410, 2019.
- [184] M.-Y. Liu, T. Breuel, and J. Kautz, “Unsupervised image-to-image translation networks,” in *Adv Neural Inf Process Syst*, vol. 30, 2017.
- [185] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, “Multimodal unsupervised image-to-image translation,” in *European Conf Comput Vis*, pp. 172–189, 2018.
- [186] A. Q. Nichol and P. Dhariwal, “Improved denoising diffusion probabilistic models,” in *Int Conf Mach Learn*, pp. 8162–8171, 2021.
- [187] T. Xiang, M. Yurt, A. B. Syed, K. Setsompop, and A. Chaudhari, “DDM²: Self-Supervised Diffusion MRI Denoising with Generative Diffusion Models,” *arXiv:2302.03018*, 2023.
- [188] G. Elmas, S. U. Dar, Y. Korkmaz, E. Ceyani, B. Susam, M. Özbey, S. Avestimehr, and T. Çukur, “Federated Learning of Generative Image Priors for MRI Reconstruction,” *IEEE Trans Med Imaging*, pp. 1–1, 2022.
- [189] Q. Lyu and G. Wang, “Conversion between CT and MRI images using diffusion and score-matching models,” *arXiv:2209.12104*, 2022.

- [190] H. A. Bedel, I. Sivgin, O. Dalmaz, S. U. Dar, and T. Çukur, “Bolt: Fused window transformers for fmri time series analysis,” *Medical Image Analysis*, p. 102841, 2023.
- [191] Y. Korkmaz, S. U. H. Dar, M. Yurt, M. Özbey, and T. Çukur, “Unsupervised mri reconstruction via zero-shot learned adversarial transformers,” *IEEE Transactions on Medical Imaging*, vol. 41, no. 7, pp. 1747–1763, 2022.
- [192] 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*, pp. 12413–12422, 2022.
- [193] B. Quist, B. A. Hargreaves, T. Cukur, G. R. Morrell, G. E. Gold, and N. K. Bangerter, “Simultaneous fat suppression and band reduction with large-angle multiple-acquisition balanced steady-state free precession,” *Magnetic resonance in medicine*, vol. 67, no. 4, pp. 1004–1012, 2012.
- [194] K. T. Kwon, H. H. Wu, T. Shin, T. Çukur, M. Lustig, and D. G. Nishimura, “Three-dimensional magnetization-prepared imaging using a concentric cylinders trajectory,” *Magnetic resonance in medicine*, vol. 71, no. 5, pp. 1700–1710, 2014.
- [195] E. U. Saritas, D. Lee, T. Cukur, A. Shankaranarayanan, and D. G. Nishimura, “Hadamard slice encoding for reduced-fov diffusion-weighted imaging,” *Magnetic resonance in Medicine*, vol. 72, no. 5, pp. 1277–1290, 2014.
- [196] S. G. Kafali, T. Çukur, and E. U. Saritas, “Phase-correcting non-local means filtering for diffusion-weighted imaging of the spinal cord,” *Magnetic resonance in medicine*, vol. 80, no. 3, pp. 1020–1035, 2018.
- [197] A. Ozaslan, A. Alacaoglu, O. Demirel, T. Çukur, and E. Saritas, “Fully automated gridding reconstruction for non-cartesian x-space magnetic particle imaging,” *Physics in Medicine & Biology*, vol. 64, no. 16, p. 165018, 2019.

- [198] O. B. Demirel, T. Kilic, T. Çukur, and E. U. Saritas, “Anatomical measurements correlate with individual magnetostimulation thresholds for khz-range homogeneous magnetic fields,” *Medical physics*, vol. 47, no. 4, pp. 1836–1844, 2020.
- [199] S. Ilbey, C. B. Top, A. Güngör, T. Çukur, E. U. Saritas, and H. E. Güven, “Fast system calibration with coded calibration scenes for magnetic particle imaging,” *IEEE transactions on medical imaging*, vol. 38, no. 9, pp. 2070–2080, 2019.
- [200] A. Gungor, B. Aşkın, D. A. Soydan, E. U. Saritas, C. B. Top, and T. Çukur, “A denoiser scaling technique for plug-and-play mpi reconstruction,” *International Journal on Magnetic Particle Imaging IJMPI*, vol. 9, no. 1 Suppl 1, 2023.
- [201] A. Z. Alkilani, T. Çukur, and E. U. Saritas, “Fd-net: An unsupervised deep forward-distortion model for susceptibility artifact correction in epi,” 2023.
- [202] B. Askin, A. Güngör, D. Alptekin Soydan, E. U. Saritas, C. B. Top, and T. Cukur, “Pp-mpi: A deep plug-and-play prior for magnetic particle imaging reconstruction,” in *Machine Learning for Medical Image Reconstruction: 5th International Workshop, MLMIR 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings*, pp. 105–114, Springer, 2022.
- [203] X. Wu, P. Gao, P. Zhang, Y. Shang, B. He, L. Zhang, J. Jiang, H. Hui, and J. Tian, “Cross-domain knowledge transfer based parallel-cascaded multi-scale attention network for limited view reconstruction in projection magnetic particle imaging,” *Computers in Biology and Medicine*, vol. 158, p. 106809, 2023.
- [204] E. Çelik, S. U. H. Dar, Ö. Yılmaz, Ü. Keleş, and T. Çukur, “Spatially informed voxelwise modeling for naturalistic fmri experiments,” *NeuroImage*, vol. 186, pp. 741–757, 2019.

- [205] M. Shahdloo, E. Çelik, and T. Çukur, “Biased competition in semantic representation during natural visual search,” *Neuroimage*, vol. 216, p. 116383, 2020.
- [206] Ö. Yılmaz, E. Çelik, and T. Çukur, “Informed feature regularization in voxelwise modeling for naturalistic fmri experiments,” *European Journal of Neuroscience*, vol. 52, no. 5, pp. 3394–3410, 2020.
- [207] S. Tokgoz, D. Aydogdu, B. Ilhan, Y. Sahin, N. Bariseri, B. M. Ozturkler, and T. Çukur, “Musical mirror-symmetrical movement tasks: comparison of rhythm versus melody-playing,” *NeuroReport*, vol. 31, no. 7, pp. 523–529, 2020.
- [208] T. Çukur, A. G. Huth, S. Nishimoto, and J. L. Gallant, “Functional subdomains within human ffa,” *Journal of Neuroscience*, vol. 33, no. 42, pp. 16748–16766, 2013.
- [209] T. Çukur, A. G. Huth, S. Nishimoto, and J. L. Gallant, “Functional subdomains within scene-selective cortex: parahippocampal place area, retrosplenial complex, and occipital place area,” *Journal of Neuroscience*, vol. 36, no. 40, pp. 10257–10273, 2016.
- [210] T. Cukur, S. Nishimoto, A. G. Huth, and J. L. Gallant, “Attention during natural vision warps semantic representation across the human brain,” *Nature neuroscience*, vol. 16, no. 6, pp. 763–770, 2013.
- [211] T. Zhang, J. S. Gao, T. Çukur, and J. L. Gallant, “Voxel-based state space modeling recovers task-related cognitive states in naturalistic fmri experiments,” *Frontiers in neuroscience*, vol. 14, p. 1103, 2021.
- [212] L. K. Şenel, İ. Utlu, V. Yücesoy, A. Koç, and T. Çukur, “Generating semantic similarity atlas for natural languages,” in *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 795–799, IEEE, 2018.
- [213] L. K. Şenel, I. Utlu, V. Yücesoy, A. Koc, and T. Cukur, “Semantic structure and interpretability of word embeddings,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1769–1779, 2018.

- [214] I. Kiremitçi, Ö. Yilmaz, E. Çelik, M. Shahdloo, A. G. Huth, and T. Çukur, “Attentional modulation of hierarchical speech representations in a multitalker environment,” *bioRxiv*, pp. 2020–12, 2021.
- [215] E. Çelik, U. Keles, İ. Kiremitçi, J. L. Gallant, and T. Çukur, “Cortical networks of dynamic scene category representation in the human brain,” *Cortex*, vol. 143, pp. 127–147, 2021.
- [216] M. Shahdloo, E. Çelik, B. A. Urgan, J. L. Gallant, and T. Çukur, “Task-dependent warping of semantic representations during search for visual action categories,” *Journal of Neuroscience*, vol. 42, no. 35, pp. 6782–6799, 2022.
- [217] F. Wu and X. Zhuang, “Unsupervised domain adaptation with variational approximation for cardiac segmentation,” vol. 40, no. 12, pp. 3555–3567, 2021.
- [218] 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*, pp. 179–188, 2022.
- [219] 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.
- [220] 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.
- [221] Q. Yang, P. Yan, Y. Zhang, H. Yu, Y. Shi, X. Mou, M. K. Kalra, Y. Zhang, L. Sun, and G. Wang, “Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss,” vol. 37, no. 6, pp. 1348–1357, 2018.