High Fidelity Medical Image-to-Image Translation

by

Clinton Wang

B.S., Yale University (2015)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2020

Author .....
Department of Electrical Engineering and Computer Science
May 13, 2020

Certified by .....
Polina Golland
Henry Ellis Warren (1894) Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by .....
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

High Fidelity Medical Image-to-Image Translation

by

Clinton Wang

Submitted to the Department of Electrical Engineering and Computer Science
on May 13, 2020, in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science and Engineering

Abstract

Despite much recent progress in image-to-image translation, it remains challenging to apply such techniques to medical images. We develop a novel parameterization of conditional generative adversarial networks that achieves high image fidelity when trained to transform magnetic resonance images (MRIs) conditioned on a patient's age and disease severity. The spatial-intensity transform generative adversarial network (SIT-GAN) constrains the generator to a smooth spatial transform composed with sparse intensity changes. This technique improves image quality and robustness to artifacts, and generalizes to different scanners. Our model achieves state of the art predictions of longitudinal brain MRIs without supervised training on paired scans. We also demonstrate SIT-GAN on a large clinical image dataset of stroke patients, where it captures associations between ventricle expansion and aging, as well as between white matter hyperintensities and stroke severity. Additionally, SIT-GAN provides a disentangled view of anatomical and textural changes with each transformation, making it easier to interpret the model's predictions in terms of physiological phenomena. As conditional generative models become increasingly versatile tools for data exploration, visualization and forecasting, such techniques for improving robustness are critical for their translation to clinical settings.

Thesis Supervisor: Polina Golland
Title: Henry Ellis Warren (1894) Professor of Electrical Engineering and Computer Science

Acknowledgments

I would like to thank my advisor Polina Golland for her strong support, leadership and guidance, which has made me a better researcher and communicator. She has provided me with many insights and opportunities resulting in a wonderful research experience. Thanks to all the other members of the Golland group: Razvan Marinescu, Daniel Moyer, Danielle Pace, Ray Liao, Maz Abulnaga, Nalini Singh, Peiqi Wang, and Bernhard Egger, for being supportive and helpful labmates and friends. They fostered an open, friendly and dynamic lab environment with many opportunities to learn and improve as a researcher. Finally, deepest thanks to my family for their unwavering support and unconditional love. Their encouragement and hard work have made it possible for me to have so many wonderful opportunities and experiences in my life.

Contents

Chapter 1

Introduction

Many common tasks in computer vision and medical image analysis require mapping images in one distribution to images in another distribution. In medical contexts, models that can change an input image along a set of controlled attributes (e.g., imaging modality or patient phenotype) are useful for a wide range of tasks including data augmentation [4], super-resolution [32], MR-to-CT translation [42], and prediction of disease trajectories [33]. With the development of conditional generative adversarial networks (cGANs), it became possible to tackle such image-to-image translation tasks with a single approach.

cGANs achieve state of the art results in applications as diverse as sketch to photo conversion [25], image colorization [46], image inpainting [44], and style transfer [22]. However, medical image-to-image translation remains a challenging problem, and cGANs have seen less success in this domain. Their application has been largely restricted to large datasets of high-quality research scans. When the target distribution is underrepresented in the training data or the data consists of lower quality clinical scans, such models may introduce artifacts as we illustrate in this work.

We seek to address this shortcoming by exploring a novel parameterization for generators based on spatial and intensity transforms. These transforms are commonly used in other areas of medical image analysis including data augmentation and image registration [4, 48]. In many medical applications, transformations between images can be well represented by a smooth deformation and a sparse intensity difference

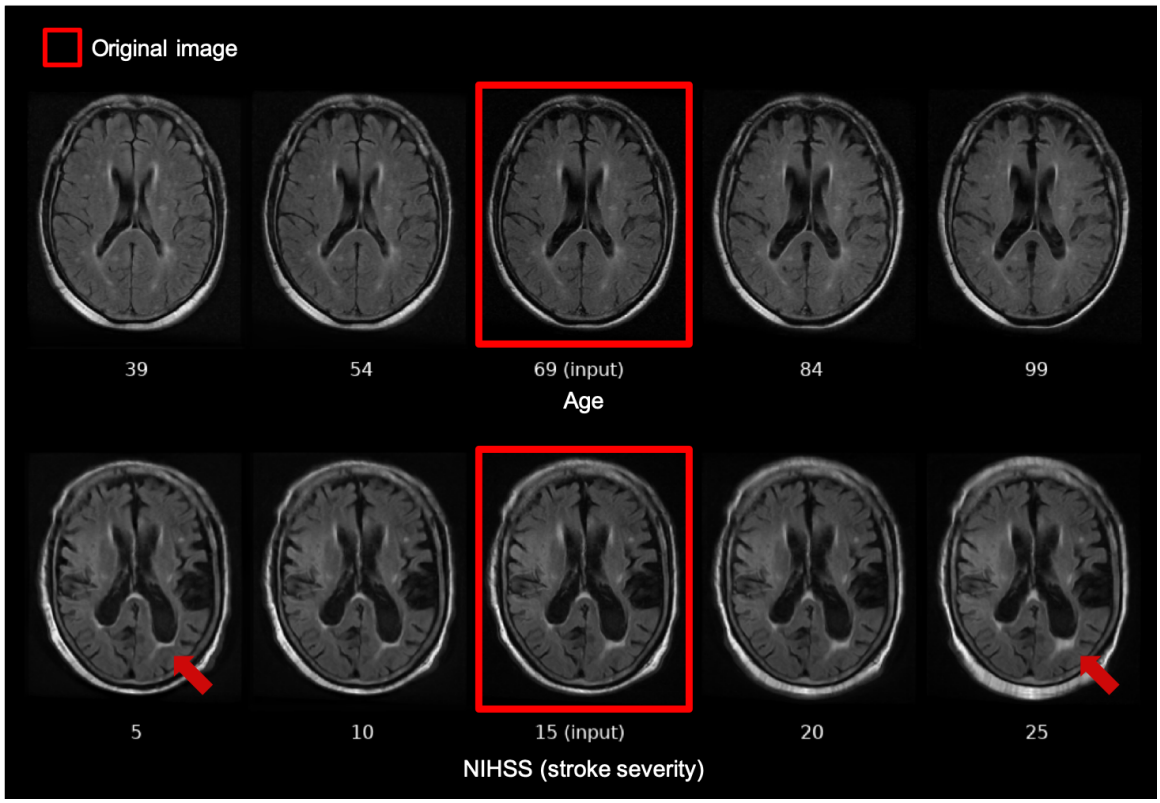

Figure 1-1: Synthetic fluid-attenuated inversion recovery (FLAIR) MRIs of acute ischemic stroke patients generated by the spatial-intensity transform generative adversarial network (SIT-GAN). The images in the red boxes were transformed into their neighboring images by conditioning on changes in age (top) and stroke severity (bottom). Our model replicates known physiological phenomena: age correlates with increasing ventricular volume and widening of the sulci, and stroke severity correlates with increasing volume of white matter hyperintensities around the ventricles (the bright spot near the red arrow).

transform, suggesting that a network that is parameterized by such transformations will be less likely to introduce spurious distortions. Indeed, even GANs that produce perceptually convincing outputs have been found to introduce subtle artifacts into their images [26, 47]. Thus, finding appropriate priors to constrain the transforms of image-to-image translation models is particularly important for their application to medical problems, where it may be more difficult to spot artifacts and where such artifacts could influence radiological findings.

In order for our model to learn to translate images conditioned on a particular attribute, it is sufficient for the training dataset to contain patient scans for various values of that attribute, even when those images belong to different subjects. For example, conditioning on age does not require supervised learning with longitudinal imaging from the same patient. This enables our model to be applied to a wide range of imaging datasets and conditioning attributes.

The main contributions of this thesis are as follows:

- We introduce SIT-GAN, a novel parameterization of conditional generative adversarial networks that leverages spatial-intensity transforms to improve image fidelity and robustness to artifacts in medical image-to-image translation tasks.

- We achieve state of the art performance on prediction of aging trajectories in T1-weighted brain magnetic resonance images (MRIs) without supervised training on paired scans.

- After training on clinical images of stroke patients, SIT-GAN is able to highlight the expansion of the ventricles associated with aging, as well as the growth in white matter hyperintensities associated with stroke severity. The model additionally provides a disentangled view of morphological and intensity changes associated with each transformation.

Roadmap Chapter 2 presents a brief overview of central concepts in deep learning, focusing on its applications in computer vision. We describe generative adversarial networks and their extension to image-to-image translation in unpaired and multi-

domain settings. We also outline previous uses of the spatial-intensity transform for medical image registration. Chapter 3 introduces SIT-GAN, specifying its network architecture, loss functions, and the use of spatial-intensity transforms to parameterize the generator. Chapter 4 reports two experiments comparing SIT-GAN with previous models to demonstrate that spatial-intensity transforms are an effective prior for medical image-image translation tasks. The first experiment, involving clinical quality scans of stroke patients from multiple sites, shows our method's robustness to low quality scans and its ability to generalize to unseen scanners. The second experiment, involving longitudinal scans, demonstrates our method's ability to predict the trajectories of subjects' brain scans. Finally, we provide concluding remarks in Chapter 5 and discuss possible applications and directions for future work.

Chapter 2

Background

In this chapter, we describe previous work that is relevant to understanding our model and its context. We offer a brief review of deep learning and convolutional neural networks, highlighting some modeling choices that appear in our network. We outline the evolution of image-to-image translation models in computer vision, focusing on techniques that we also use in our model. Finally, we discuss the use of spatial and intensity transforms in medical image analysis, which motivates our development of the spatial-intensity transform generative adversarial network (SIT-GAN).

2.1 Deep Learning

Deep learning is a powerful class of data-driven techniques for learning functions on high-dimensional data, with diverse applications in processing and generating images, videos, music, natural language, and many other types of data. In deep learning, a function of interest is parameterized as a neural network: the composition of a series of parameterized linear transforms (called layers) alternating with non-linear transforms (called activations). The parameters of all layers are optimized simultaneously via stochastic gradient descent (SGD) to match a set of input-output pairs that demonstrate the desired behavior of the function. Neural networks often have tens to hundreds of layers, giving rise to the name of deep learning.

Many networks have millions to hundreds of millions of parameters, in stark con-

trast to previous machine learning methods in which the number of model parameters was often kept relatively low, and almost certainly less than the number of datapoints to train on. But many studies have found that the over-parameterization of neural networks is key to their success: in the limit of infinite parameters, a neural network is capable of approximating any continuous function to arbitrary precision [28]; neural networks with a large number of parameters have improved theoretical guarantees on their convergence to the global optimum of the training data [1, 29].

This capacity for neural networks to train millions of parameters simultaneously has enabled their application in high dimensional domains where previous techniques required feature engineering: the manual selection of a small number of statistics that are ("mostly") sufficient with respect to a particular modeling task. Such a set of sufficient statistics was challenging to find on real-world tasks, and neural networks have surpassed these traditional modeling techniques on many computer vision, natural language processing, and audio processing tasks [16].

Almost all neural networks implement some form of normalization on either the intermediate values or parameters of the network, which can help training by making the optimization landscape smoother [38], making gradient magnitudes more consistent throughout the network [37], or providing smoothness guarantees about the network function throughout training [31]. Batch normalization [24], the most frequently used form of normalization, rescales the outputs of a specified layer to have the same (learned) mean and standard deviation within each minibatch.

2.1.1 Convolutional Neural Networks

The most common type of neural network for image processing is the convolutional neural network (CNN), which constrains the weight matrices of each layer in the network to correspond to convolutions with learned patches. Without any constraints on the weight matrix (in which case the layer is called fully connected), there are too many parameters to even store in memory. For example, just a single network layer between two RGB images of size 256×256 would require 38.7 billion parameters. Beyond this practical limitation, convolutions are a natural choice for image processing

applications, where features of interest are often shift-invariant[1]: if a noisy image is shifted, the output of a denoising algorithm should appear shifted by the same amount (modulo boundary effects); if objects are shifted in a scene, their bounding boxes and segmentations are translated by the same amount. It is well known that every linear shift-invariant system can be represented as a convolution with some impulse response. In neural networks, the impulse response is constrained to have small width (typically no larger than 5×5 in size), in analogy to the human visual system where each neuron receives stimuli from only a neighborhood of neurons at the previous level.

CNNs often feature several other types of layers. Strided convolutions or pooling layers are used to reduce the spatial resolution of their inputs, which may make it easier to learn global image features in a manner reminiscent of image pyramids. To increase spatial resolution, interpolation or deconvolution layers are often used. Other important innovations in the design of CNNs include dropout [39], which randomly zero out layer outputs, residual blocks [19], which parameterize layer outputs as a residual that is added to the layer input, and skip connections [21], where a signal passes through multiple layers before being concatenated with itself.

2.2 Image-to-Image Translation

Image-to-image translation, a term popularized in [25], refers to a broad category of tasks that learn a mapping from images in one distribution to images in another distribution, where the distributions and desired properties of the function depend on the underlying task. This includes a broad range of tasks including style transfer, photo enhancement, image super-resolution, synthesizing images from segmentations, object transfiguration, image colorization, image denoising, and image reconstruction. With the development of neural networks, these disparate tasks were able to be unified into a single problem: given pairs of example images from both domains, teach a convolu-

[1]Sometimes the property described here is called shift covariance, and shift invariance is instead used to refer to outputs that are independent of translations of the input.

tional neural network to map the input images to the output images. Although simple CNN architectures tended to produce blurry images, the development of U-Nets [34] and generative adversarial networks (GANs) [17] led to dramatic improvements in performance.

U-Net Prior to the U-Net, the typical architecture for building image-to-image mappings was an encoder-decoder. The encoder is a series of convolutional layers alternating with pooling or strided convolutional layers that reduce the spatial dimensions of the feature space, while the decoder is a series of convolutional layers alternating with deconvolutional or upsampling layers that increase the spatial dimensions of the feature space until they match the target image resolution (which is usually the same as the input image resolution, but may differ for tasks like image super-resolution). The U-Net incorporates skip connections between layers of the encoder and decoder at the same spatial resolution, which allows the network to propagate information about local details in the input image that are otherwise discarded by the encoder as it progressively downsamples its features. With only a pixel-wise loss in image space however, the U-Net outputs still suffer from blurry or unrealistic outputs, which would be addressed with the development of GANs.

Generative adversarial networks GANs were a major advancement for image synthesis, yielding the first models that could produce photorealistic images from random noise. A GAN consists of a generator network that outputs images, and a discriminator network that is trained to distinguish between the outputs of the generator and real images from the distribution of interest. The generator is simultaneously trained to output images that prevent the discriminator from making this distinction. When trained successfully, this adversarial approach can help the generator produce realistic outputs, as a good discriminator will be able to correctly classify images with artifacts as synthetic. In addition, GANs have a unique Nash equilibrium at which the generator will sample exactly from the target distribution. A generator that oversamples particular regions of the target distribution would allow the discriminator

to classify these regions as synthetic. In practice however, mode collapse remains a notoriously difficult problem to overcome, although many works have developed techniques to combat this phenomenon [18, 30]. Although we described GANs in the context of image synthesis, the input to the generator can be an arbitrary distribution, making it suitable for image-to-image translation tasks. When adversarial training is incorporated into the loss function of the U-Net we described earlier, its outputs become photorealistic [25]. The resulting model is sometimes referred to as pix2pix or a conditional GAN.

Cycle consistency loss In many image-to-image translation tasks, it may be difficult to find paired data from the source and target image distributions, but much easier to obtain independent samples from the distributions. Without paired data, there is no ground truth for the output image given an input image from our dataset. This problem was circumvented by CycleGAN [50], which pioneered the cycle consistency loss. In CycleGAN, two GANs are trained, with one generator mapping the source domain to the target domain and the other learning the reverse mapping. The cycle consistency loss minimizes the pixel-wise change in samples that pass through both generators.

Multi-domain translation Pix2pix and CycleGAN models were constrained to translation tasks between two distinct domains, such as day to night, horses to zebras, or sketches to photos. These models are unable to condition on continuous or categorical variables, e.g., changing a face to match an age, or modifying the size of objects in a photo. StarGAN [6] addresses this issue by modifying pix2pix in several ways. A single generator learns mappings between all domains, by receiving a set of conditional attribute values associated with the target image distribution. The discriminator is modified to produce a second output: the predicted values of the conditional attributes for input images. This predictor is trained on images from the training dataset, and the generator has an additional loss term that encourages it to generate images that the predictor matches to the target attribute values. To

allow StarGAN to learn from unpaired data, it adopts CycleGAN's cycle consistency loss, which is now computed by passing images through the same generator twice. The first pass can be conditioned on any set of target attribute values, while the second pass conditions on the ground truth attribute values associated with the image. ModularGAN [49] tackles the multi-domain translation problem with a different approach: instead of a single network that learns all mappings, it encodes input images and learns how to map latent vectors between different domains. However, this approach is difficult to adapt to continuous conditional variables. Later works such as StarGAN v2 [7] designed improved architectures for style transfer applications and image-to-image translation between categorical domains, although the original StarGAN remains the most suitable architecture for domains represented by multiple continuous, partially observed variables. With some tuning of the architecture and hyperparameters, StarGAN serves as the baseline (unconstrained model) for our experiments as it represents the state of the art in continuous-domain image-to-image translation.

2.3 Spatial and Intensity Transforms

Spatial deformation models have been a staple of medical image registration, which seeks to align the anatomical structures of medical images in order to establish a common coordinate system for downstream tasks such as segmentation and voxel-based analysis. For nonrigid image registration, the spatial transform is usually parameterized as a smooth deformation field, as most anatomical variation does not involve large local changes in shape. In these settings, the transform can be optimized independently for each input image [2, 35], or generated by a neural network trained for this task [3, 27]. Some works will also constrain the transform to be diffeomorphic [9], which guarantees that the topology of the original image is preserved.

In general, neural networks outperform optimization methods for estimating such deformation fields. FlowNet [13] and its successors [23] significantly advanced the state of the art on optical flow estimation: the task of estimating the displacement

of pixels between two frames of a video. The parameterization of GANs with spatial transforms has also found success in video-to-video translation [40]. To our knowledge, their application to image-to-image translation has not yet been explored.

Spatial transforms have been coupled with intensity transforms to improve medical image registration when there is variation in both anatomy and texture. Active Appearance Models [8] build statistical models of shape and intensity that can be used to register images with different tissue intensities. In more recent works, neural networks were used to learn spatial and intensity transforms mapping a fixed atlas image to any given input image [48]. Further work introduces atlases that are conditioned on a particular attribute and constructed using spatial transforms [11]. Spatial-intensity transforms have also been applied as a learned data augmentation technique for semi-supervised segmentation [4].

In the next chapter, we introduce SIT-GAN. Its design draws on many of the deep learning techniques we described for image-to-image translation: U-Nets, adversarial training, cycle consistency loss, and multi-domain translation. By additionally constraining our model to generate images using spatial and intensity transforms, we are able to improve its robustness and image fidelity on radiographic data.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 3

Spatial-Intensity Transform Generative Adversarial Network

In this chapter, we describe SIT-GAN, our novel parameterization of medical image-to-image translation models. First, we outline the overall components and training scheme of our model, which allow us to learn from unpaired data containing multiple partially observed conditional attributes. We then parameterize the model generator as a smooth deformation and sparse intensity difference transform. Lastly, we provide details about the architecture and implementation of our network.

3.1 Unpaired Image-to-Image Translation with Partially Observed Attributes

Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ of images $x_i \in \mathcal{X} : \Omega \to \mathbb{R}$ and conditional attributes $y_i \in \mathcal{Y}$ (e.g., age and stroke severity), we would like to train a generator to transform images such that their conditional attributes are shifted by a specified amount. Our network consists of a generator $G : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$, discriminator $D : \mathcal{X} \to \mathbb{R}$ (logits), and regressor $R : \mathcal{X} \to \mathcal{Y}$. Here we consider continuous vector attributes $y_i = (y_{i,1}, ..., y_{i,m})$ that may have missing values. Categorical attributes can be included by adding a classifier to the network.

Generator The generator G transforms a given input image such that the transformed image appears to take on different attribute values from the input image, but preserves aspects of the input image that are unrelated to the conditional attributes, such as non-pathological anatomy. Define $z_i = (x_i, y_i)$, $z_j = (x_j, y_j)$, $\Delta y = y_j - y_i$. During training, the generator is updated using the following loss terms:

$$\ell_{adv} = -D(G(x_i, \Delta y)) \qquad \text{Wasserstein adversarial loss} \tag{3.1}$$

$$\ell_{attr} = \frac{1}{m}\left\| \big(R(G(x_i, \Delta y)) - R(x_i)\big) - \Delta y \right\|_2^2 \qquad \text{relative attribute loss} \tag{3.2}$$

$$\ell_{cc} = \|G(G(x_i, \Delta y), -\Delta y) - x_i\|_1 \qquad \text{cycle consistency loss} \tag{3.3}$$

Parameterizing G in terms of Δy enables evaluation of the cycle consistency loss even when images have missing attributes [43]. To compute Δy in such cases, we introduce the convention that $y_{j,k} - y_{i,k} = 0$ if the kth attribute is missing. Putting the terms together, the total generator loss is:

$$\mathcal{L}_G = \mathbb{E}_{z_i, z_j}\big[\ell_{adv} + \lambda_{attr}\ell_{attr} + \lambda_{cc}\ell_{cc}\big] \tag{3.4}$$

where λ_{attr} and λ_{cc} are empirically determined weights.

Discriminator We simultaneously train the discriminator D with the Wasserstein GAN losses and gradient penalty [18].

$$\mathcal{L}_D = \mathbb{E}_{z_i, z_j}\big[D(G(x_i, \Delta y))\big] - \mathbb{E}_{z_i}\big[D(x_i)\big] - \mathbb{E}_{\hat{x}}\big[\lambda_{GP}(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2\big] \tag{3.5}$$

where $\hat{x}$ is obtained by interpolating real and translated images as described in [18], and λ_{GP} is a weight.

Regressor The regressor R is trained to predict the attributes of real images, using a mean squared error loss.

$$\mathcal{L}_R = \mathbb{E}_{z_i}\big[\frac{1}{m}\|R(x_i) - y_i\|_2^2\big] \tag{3.6}$$

We share layers between the discriminator and regressor, so a single optimizer is assigned to both subnetworks and updated using $\mathcal{L}_D + \lambda_R \mathcal{L}_R$.

3.2 Spatial-Intensity Transform Generator

To constrain the generator to spatial-intensity transforms, we define its outputs as the deformation field $F : \Omega \to \mathbb{R}^d$ for image dimensionality d, with corresponding transform $T_F : \mathcal{X} \to \mathcal{X}$, and the intensity difference map $\Delta x : \Omega \to \mathbb{R}$.

Rather than requiring the generator to produce the target image, it outputs F and Δx, then transforms the input image as $T_F(x_{in} + \Delta x)$. In addition, we add regularization terms to the generator's loss function that encourage the deformation field to be smooth and the intensity difference map to be sparse. Specifically, we used the discrete total variation norm [5] to regularize the deformation field and the L1-norm to regularize the intensity change:

$$\|F\|_{TV} = \frac{1}{|\Omega|}\sum_{\omega\in\Omega}\|\nabla F(\omega)\|_2 \tag{3.7}$$

$$\|\Delta x\|_1 = \frac{1}{|\Omega|}\sum_{\omega\in\Omega}|\Delta x(\omega)| \tag{3.8}$$

where $\|\nabla F(\omega)\|_2$ is approximated using finite differences. The total generator loss now becomes:

$$\mathcal{L}_G = \mathbb{E}_{z_i,z_j}\left[\ell_{adv} + \lambda_{attr}\ell_{attr} + \lambda_{cc}\ell_{cc} + \lambda_F\|F\|_{TV} + \lambda_{\Delta x}\|\Delta x\|_1\right] \tag{3.9}$$

for empirically determined weights λ_F and $\lambda_{\Delta x}$.

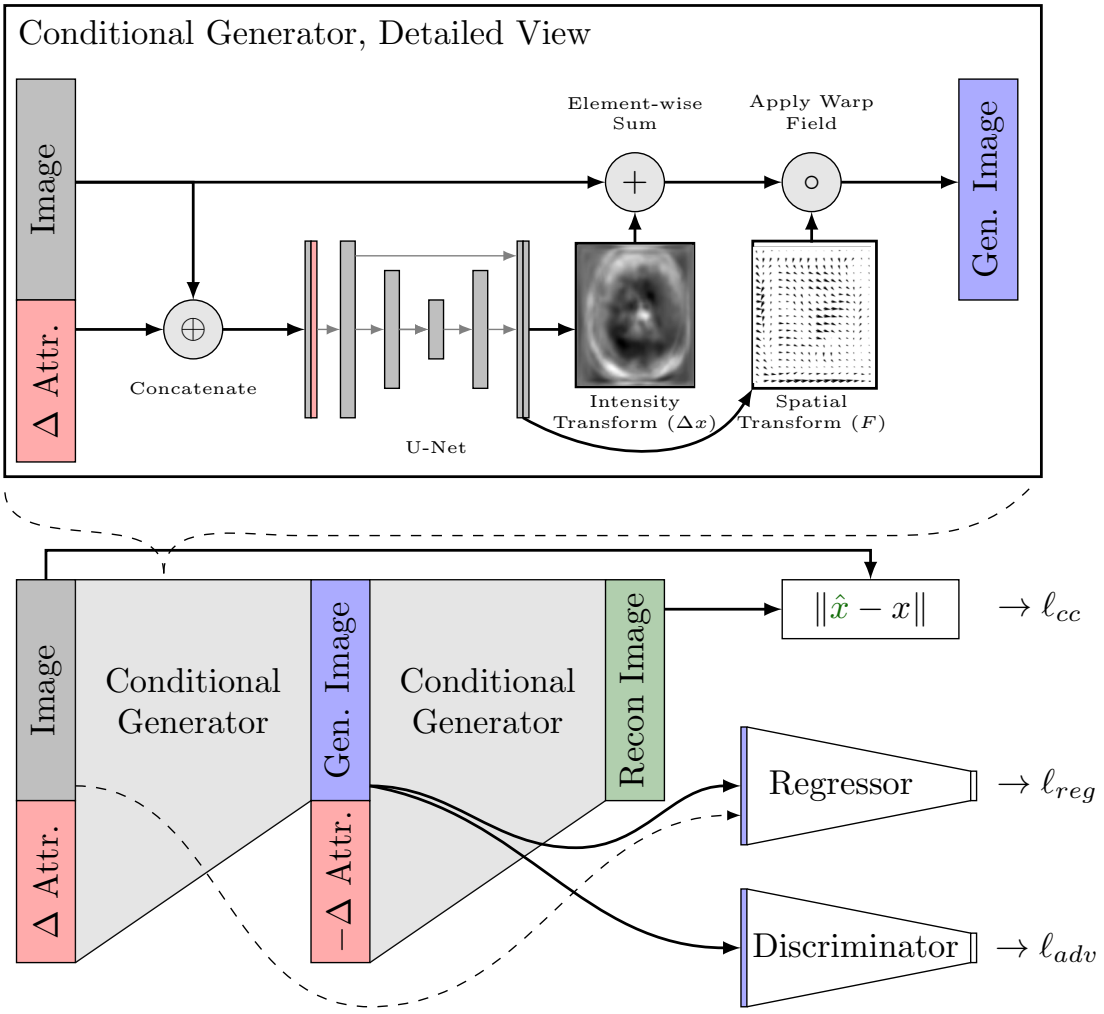

Figure 3-1: The generator takes in an image and the desired change in each attribute. In SIT-GAN, the generated image is obtained by applying a intensity difference map and deformation field to the input image. The parameters of the generator are updated from three loss terms: a cyclic consistency loss that discourages unnecessary changes to the input image, an attribute loss that encourages the generated image to match the desired attribute values, and an adversarial loss that penalizes unrealistic outputs.

3.3 Network Architecture and Implementation

SIT-GAN's generator was implemented as a 2D U-Net that takes in Δy by replicating each dimension spatially and concatenating channel-wise to x_i. It has 4 spatial resolutions, with 200 channels and 6 residual blocks at the lowest resolution. The discriminator and regressor share 5 down-sampling blocks, then split into fully connected layers of the appropriate dimension (1 output for the discriminator, m outputs for the regressor).

Batch normalization is used for all convolutional layers. Down-sampling blocks in the U-Net use convolutional layers alternating with max blur pooling [45]. Up-sampling blocks in the U-Net use bilinear upsampling between convolutional layers. The generator uses ReLU activations and the discriminator/regressor uses leaky ReLU activations. We use He initialization for the weights of all convolutional layers and set all biases to zero.

The subnetworks were trained with Adam optimizers, with one step in G's optimizer for every two steps in D/R's optimizer. D/R were trained for 50K iterations with a learning rate of 1.2×10^{-5}, and G was trained for 25K iterations with a learning rate of 1.5×10^{-4}. Both optimizers used a minibatch size of 4, and moving average parameter $\beta_1 = 0.86$. We used the following loss weights: $\lambda_R = 18$, $\lambda_{attr} = 3.5$, $\lambda_{cc} = 2.1$, $\lambda_F = 16$, $\lambda_{\Delta x} = 49$, and $\lambda_{GP} = 1.1$.

The following chapter will present experimental evaluation of the proposed SIT-GAN model. We evaluate the network on a dataset of clinical MRIs from acute ischemic stroke patients, where we measure the image fidelity and proximity of synthesized images to the target domain. We compare its performance to networks that do not transform the input image or use alternative transforms. We then test SIT-GAN's ability to forecast future scans using a dataset of longitudinal research scans from patients with various degrees of cognitive decline.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 4

Experiments

We conducted experiments on two cohorts: a set of clinical quality MRIs from patients with acute ischemic stroke obtained from the MRI-GENetics Interface Exploration (MRI-GENIE) study [15], and a set of research scans obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The stroke cohort tests our model's performance on lower quality scans and its ability to generalize to different clinical sites, while the longitudinal data of ADNI allows us to assess our model's ability to predict a person's aging trajectory.

4.1 Image-to-Image Translation of Stroke MRIs

4.1.1 Data

We used axial brain fluid-attenuated inversion recovery (FLAIR) MRIs obtained within 48 hours of symptom onset from the MRI-GENIE study. After excluding repeat scans as well as scans with extreme artifacts, we had 1821 scans from across 12 clinical sites. 418 images acquired from the largest site were used for 5-fold cross validation. Our models were then tested on the 1403 scans from all other clinical sites. Age was available for all patients, and stroke severity (measured on a scale from 0-36 called NIHSS) was available for 746 patients.

MRIs were preprocessed with resampling to isotropic 1mm resolution, N4 bias

Table 4.1: Parameterizations of the generator output.

Parameterization	***G* Outputs**	**Generated Image**	**Regularizers**
Unconstrained	x_{out}	x_{out}	N/A
Difference Transform	Δx	$x_{in} + \Delta x$	$\|\Delta x\|_1$
Optical Flow	F	$T_F(x_{in})$	$\|F\|_{TV}$
Weighted Flow	F, w	$w \odot T_F(x_{in}) + (\mathbf{1} - w) \odot x_{in}$	$\|F\|_{TV}$
SIT-GAN	$F, \Delta x$	$T_F(x_{in} + \Delta x)$	$\|F\|_{TV}, \|\Delta x\|_1$

field correction, ANTS registration to a FLAIR atlas, normalization of the white matter intensity, and cropping to 224 × 192. Native resolution varies, but is typically around 1mm × 1mm × 6mm. The thick slices introduce significant partial volume effects. The 15 middle axial slices of each subject were used, and all slices from the same subject were grouped into the same validation fold. We scaled age and stroke severity so that the empirical distribution of each attribute within the training data has a mean of 0 and a standard deviation of 1. The images were also augmented using horizontal flips and random affine transformations.

4.1.2 Baseline Methods

We compare SIT-GAN to several baseline methods, including a network whose generator does not transform the image, as well as several networks whose generators use alternate transformations of the input image. The different parameterizations are summarized in Table 4.1.

In the unconstrained network, the generator follows the standard practice of directly synthesizing a new image [6, 25, 50]. In practice, the skip connections of the U-Net and the cycle consistency loss tend to produce output images that are similar to the input images.

In the difference transform network, the generator is constrained to a sparse intensity difference transform of the input image. While it has the same expressiveness as the unconstrained network, it uses explicit regularization to penalize output images that differ significantly from their inputs, making it a suitable parameterization for capturing image-to-image translations that only involve small regions of the image.

The optical flow network is constrained to smooth deformations of the input image.

It assumes that every point in the output image originates from some nearby point in the input image [13], allowing it to capture morphological variation, but not intensity changes within anatomical structures.

The weighted flow network outputs a weighted sum of the input image and a smooth deformation of it, with pixel-wise weights output by the generator. It is the type of model used to synthesize successive frames in video-to-video translation models [40].

We trained two variants of the unconstrained model: one that has identical hyperparameters to SIT-GAN and the other networks, and one in which we tuned the number of layers, types of layers, loss term weights, and type of optimizer to make it as competitive with SIT-GAN as possible. In the tuned model, the discriminator and regressor were trained with a learning rate of 8.6×10^{-5}, and the generator was trained with a learning rate of 1.1×10^{-4}. The U-Net had 3 spatial resolutions with 96 channels and 3 residual blocks at the lowest resolution. Strided convolutions were used for downsampling. The discriminator/regressor had 6 downsampling blocks using max blur pooling. The tuned network used loss weights of $\lambda_R = 21$, $\lambda_{attr} = 1$, $\lambda_{cc} = 4$, and $\lambda_{GP} = 8$. The Adam optimizer had moving average parameter $\beta_1 = 0.46$.

4.1.3 Evaluation

To quantify the quality of model outputs in the absence of paired data, we computed the Fréchet Inception Distance (FID) [20] between the distribution of generated images and the distribution of validation or test images. We also used Precision and Recall for Distributions (PRD) [36] to compute the precision ($F_{1/8}$) and recall (F_8) of our generator. A high $F_{1/8}$ suggests that most modes of the generated distribution belong to the true distribution, whereas a high F_8 suggests that most modes of the true distribution belong to the generated distribution. Modes are estimated by finding clusters of images in Inception v3 embedding space. Note that because our goal is not to find a bijection between image distributions, these distributional metrics should not be interpreted as key measures of performance, but rather as indicators about whether a network may suffer from mode collapse or other issues.

Table 4.2: Performance metrics for translation of FLAIR MRIs conditioned on age and stroke severity (NIHSS), averaged over 5 runs. FID = Fréchet Inception Distance, P/R = Precision ($F_{1/8}$) and Recall (F_8) as defined in [36].

Model Type	**FID**	**P/R**	**Age MSE**	**NIHSS MSE**
Cross-validation				
Unconstrained	152.1	0.01/0.01	1.51	2.18
Unconstrained (tuned)	61.4	0.07/0.21	**0.51**	1.12
Difference Transform	57.2	**0.38**/**0.59**	1.37	1.14
Optical Flow	59.5	0.30/0.52	0.71	**1.09**
Weighted Flow	60.6	0.23/0.46	0.85	1.31
SIT-GAN	**38.6**	0.35/**0.59**	0.85	1.16
Test				
Unconstrained	180.5	0.07/0.02	1.11	1.21
Unconstrained (tuned)	51.0	0.41/0.21	**0.99**	**1.01**
Difference Transform	68.4	0.53/0.68	1.25	1.12
Optical Flow	28.4	**0.62**/**0.69**	1.16	1.11
Weighted Flow	35.0	0.56/0.59	1.32	1.14
SIT-GAN	**27.6**	0.53/0.66	1.28	1.12

We also evaluated the effectiveness of each model in transforming the target attribute by measuring the performance of an Inception v3 regressor on our generated images. This regressor was pre-trained on ImageNet [12] and fine-tuned to predict age and stroke severity from FLAIR MRIs. We emphasize that this regressor is different from the regressor used during training of the GAN, as the generator may have learned to exploit pecularities in the particular regressor it is trained alongside. By using a separately trained regressor with a different architecture, we expect that any gains that the generator accrued in this manner can be mitigated. We measure the mean squared error (MSE) of age and stroke severity (NIHSS) respectively, normalized to the empirical standard deviation of the attribute. The MSE of the Inception regressor on held out subjects in the cross-validation set is 0.24 on age and 0.70 on NIHSS, while it is 0.34 on age and 0.62 on NIHSS in the test set.

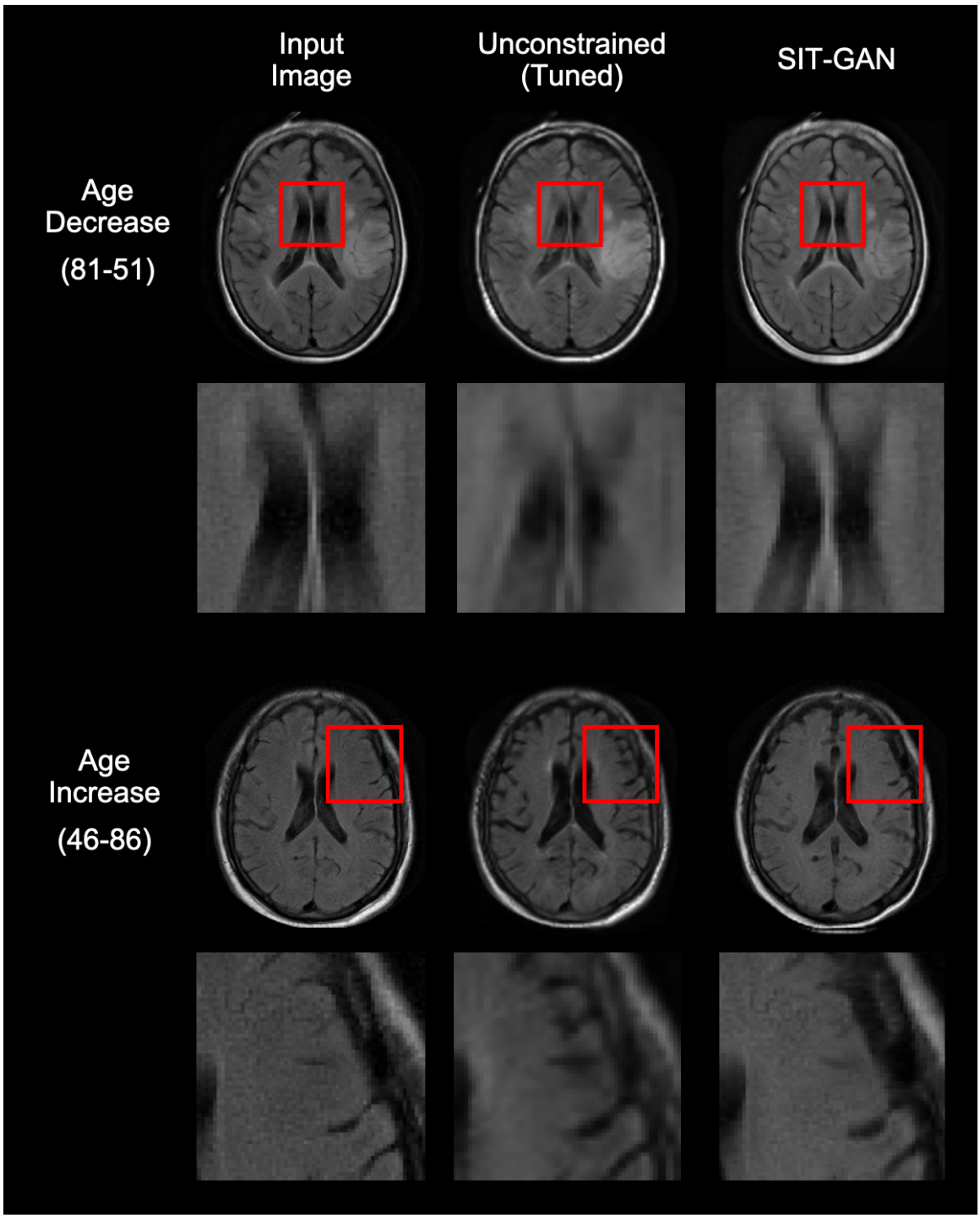

Figure 4-1: Comparison of stroke MRIs translated to a different age using the unconstrained model and our model. While both models change the ventricle shape appropriately, the unconstrained model blurs the ventricles (top rows) and excessively darkens the gray matter (bottom rows).

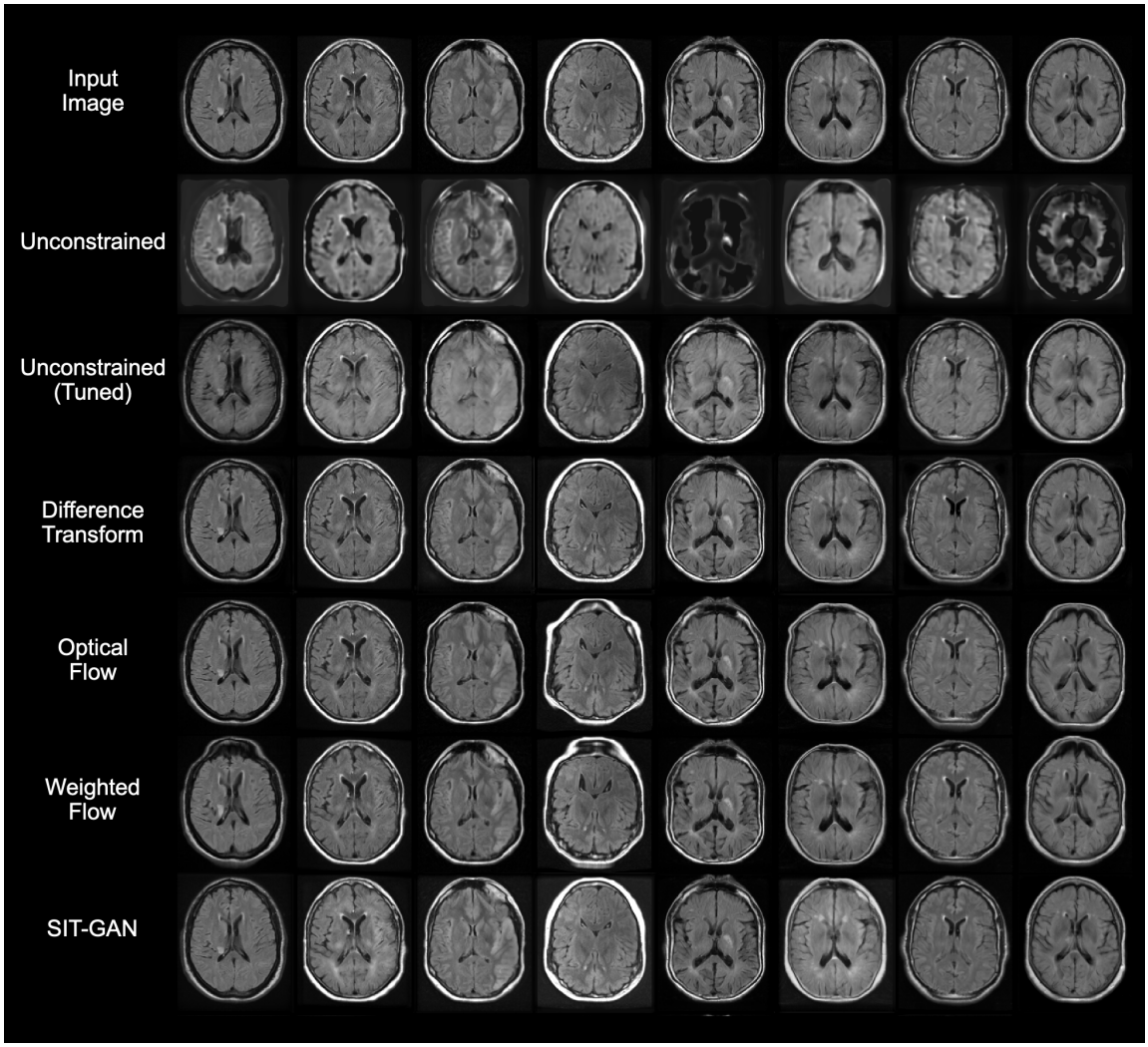

Figure 4-2: Uncurated examples of outputs from all models on the MRI-GENIE test set. Target attribute values are sampled from $\mathcal{N}(0, 4)$ to show the artifacts induced by larger transformations. (By comparison, target attribute values are distributed with mean 0 and standard deviation 1 during training.)

4.1.4 Results

Even after tuning, the unconstrained model suffers from high FID and low precision/recall as shown in Table 4.2. Figure 4-1 illustrates its tendency to introduce artifacts in translated images such as dark streaking of the gray matter with increasing age, and partial volume-like filling of the ventricles with decreasing age. With large translations, it often changes the intensity of large areas of the brain, as Figure 4-2 shows. However, the tuned unconstrained model performs well at target domain transfer for both age and NIHSS. Morphological changes of the ventricles and sulci are highly visible in its outputs.

The difference transform model dramatically improves precision and recall, but at a significant cost to its ability to match the target domain. Both of these effects can be explained by the close similarity of its outputs to the input images, as observed in Figure 4-2.

We see a similar pattern with both optical flow and weighted flow models. They perform well on distributional metrics but underperform the unconstrained model on target domain transfer, although the gap here is smaller than in the case of the difference transform model. Qualitatively, they induce noticeable morphological changes in the ventricles, although they also cause undesired distortions to the rest of the image.

SIT-GAN attains the best image fidelity, and performs similarly to the optical and weighted flow models in matching output images to their target domain. Often it is overly conservative in transforming input images. But when it succeeds, it is able to capture the expansion of the ventricles correlated with aging as well as the increase in white matter hyperintensities associated with stroke severity (see Figure 1-1), while producing much less severe artifacts than the unconstrained model as seen in Figure 4-1.

These results suggest that an unconstrained generator sacrifices image quality to capture more variation in the conditional attribute, while the constrained generators tend to be overly conservative in their transformations of input images.

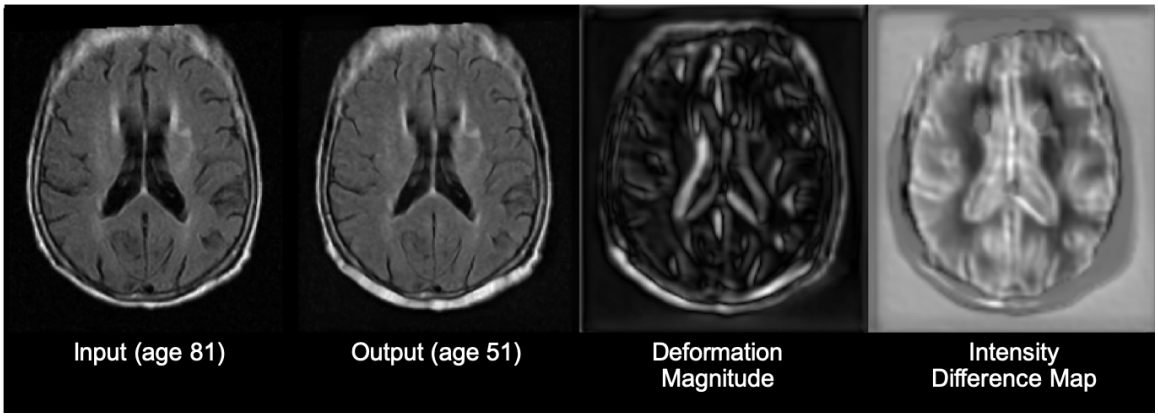

Figure 4-3: The magnitude of the deformation field and intensity difference map of SIT-GAN for an example transformation. The shrinkage of the ventricles and sulci are well captured by the deformation field, while tissue appearance changes are reflected in the difference map.

4.1.5 Disentangled Visualization

The deformation field and intensity difference map used in SIT-GAN to transform each input image can also be visualized separately. Figure 4-3 shows that the deformation highlights changes in morphology associated with age, while the intensity difference map show subtle changes in apparent tissue intensity that are not immediately apparent from the generated image. These effects, which would be inseparable with other parameterizations, are able to be visualized separately with our model. This may be valuable for detecting artifacts in generated images, as well as for finding and visualizing true correlations.

4.2 Predicting Aging Trajectories

4.2.1 Data

We performed image-to-image translation on longitudinal T1-weighted MRIs from ADNI conditioned on age and baseline diagnosis. The diagnostic categories were control, mild cognitive impairment, or Alzheimer's disease, encoded as -1, 0 and 1 respectively. The training set consisted of 3228 scans drawn from 77 subjects with unpaired data (i.e., a single timepoint scan) as well as 609 subjects with multiple timepoints (5.2 scans on average, separated by 0.79 years on average). The test set consisted of 749 scans from 149 subjects with multiple timepoints (4.7 scans on average, separated by 0.81 years on average).

Each scan was preprocessed with resampling to 1mm isotropic voxels, affine spatial normalization using FreeSurfer [14], and cropping to 224 $\times$ 192 slices [10]. The 15 middle axial slices of each subject were used. We scaled age so that its empirical distribution within the training data has a mean of 0 and a standard deviation of 1. During training, the images were augmented using horizontal flips and random affine transformations.

4.2.2 Evaluation

The architecture and hyperparameters of all models were kept identical to those used on the stroke dataset, so no separate validation set was needed. For every subject, we randomly select up to 5 pairs of timepoints. For each pair, we take the most central slice of the scan at the earlier timepoint x_1 and have our trained model predict the later timepoint image x_2. We compare the output image $\hat{x}_2$ to the actual scan obtained at the second timepoint, using the root mean square error (RMSE) of pixel intensities $\frac{1}{\sqrt{N}}\|\hat{x}_2 - x_2\|_2$ as well as their structural dissimilarity (DSSIM) [41] which compares images based on their patch statistics. The DSSIM of identical images is 0, and the DSSIM of images in which every patch is uncorrelated is 0.5.

Because RMSE and DSSIM do not distinguish between errors from target domain

Table 4.3: Performance metrics for longitudinal MRI prediction in ADNI, averaged over 5 runs. RMSE = pixel-wise root mean square error, DSSIM = structural dissimilarity.

Model Type	**RMSE**	**DSSIM**	**Age MSE**
Unconstrained	0.067	0.211	0.96
Unconstrained (tuned)	0.038	0.113	**0.52**
Difference Transform	**0.030**	0.098	0.83
Optical Flow	0.036	0.123	0.62
Weighted Flow	0.034	**0.094**	0.63
SIT-GAN	0.033	0.097	0.81

mismatch or from artifacts, we also evaluate whether the generated images match the target age. We use an Inception v3 regressor to evaluate the MSE with respect to age, similarly to the case of stroke scans. The regressor MSE on held out subjects in ADNI is 0.48.

4.2.3 Results

After tuning, the unconstrained model achieves the lowest regressor error as shown in Table 4.3, and some cases it best captures changes in ventricle morphology (e.g., the 4th column of Figure 4-5). However, it is somewhat inconsistent and often introduces significant artifacts, usually in the form of global intensity changes. As a result, its pixel-wise errors and structural dissimilarity from ground truth scans are relatively high.

The outputs of the difference transform network are more similar to ground truth scans, although generally they appear very similar to the input image, resulting in a poor match to the target domain. Additionally, the output images contain bright spot artifacts.

Although the optical flow does not improve on the unconstrained network, the weighted flow network achieves superior similarity to the ground truth, at a small cost to regressor accuracy. Because aging trajectories are dominated by morphological changes, the weighted flow network is able to perform relatively well at target domain transfer. Still, under large translations, it occasionally induces significant distortions,

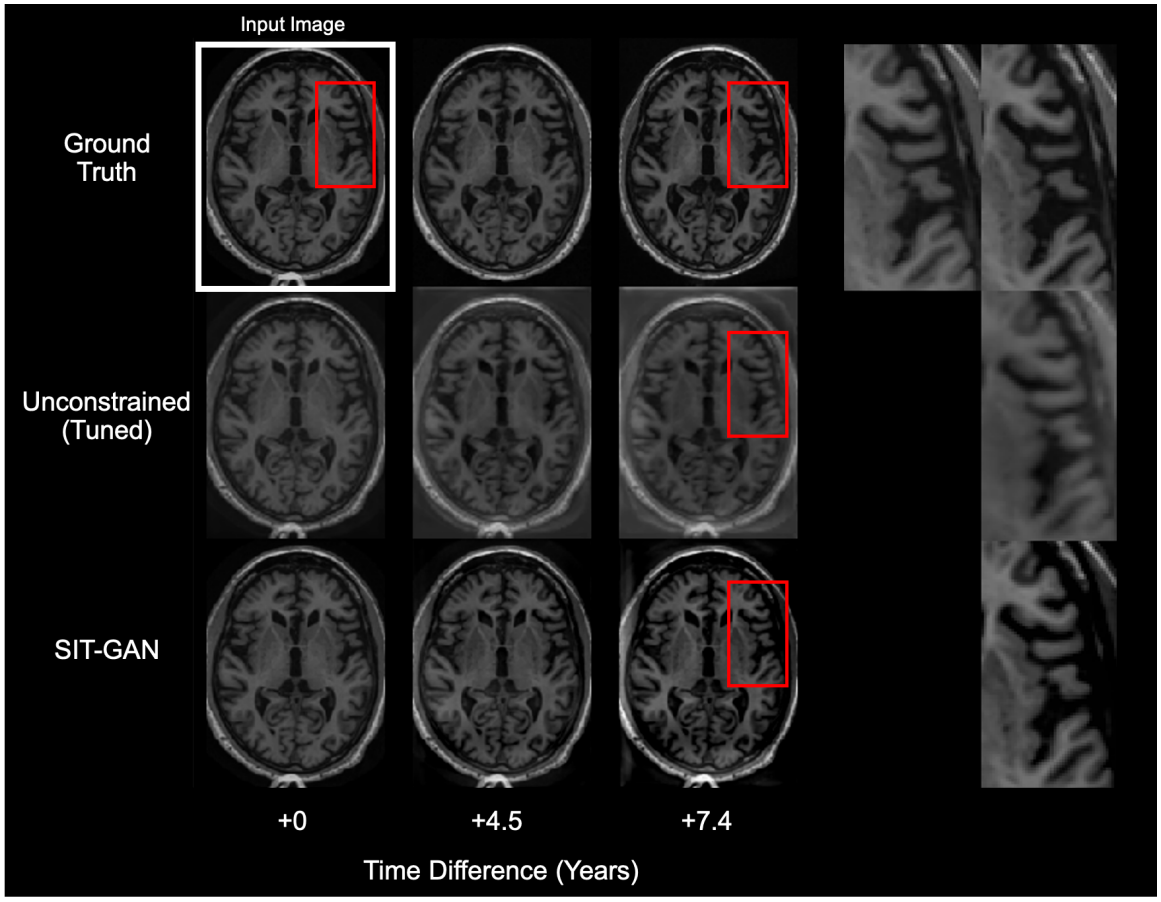

Figure 4-4: True and predicted longitudinal MRIs from the unconstrained model and SIT-GAN.

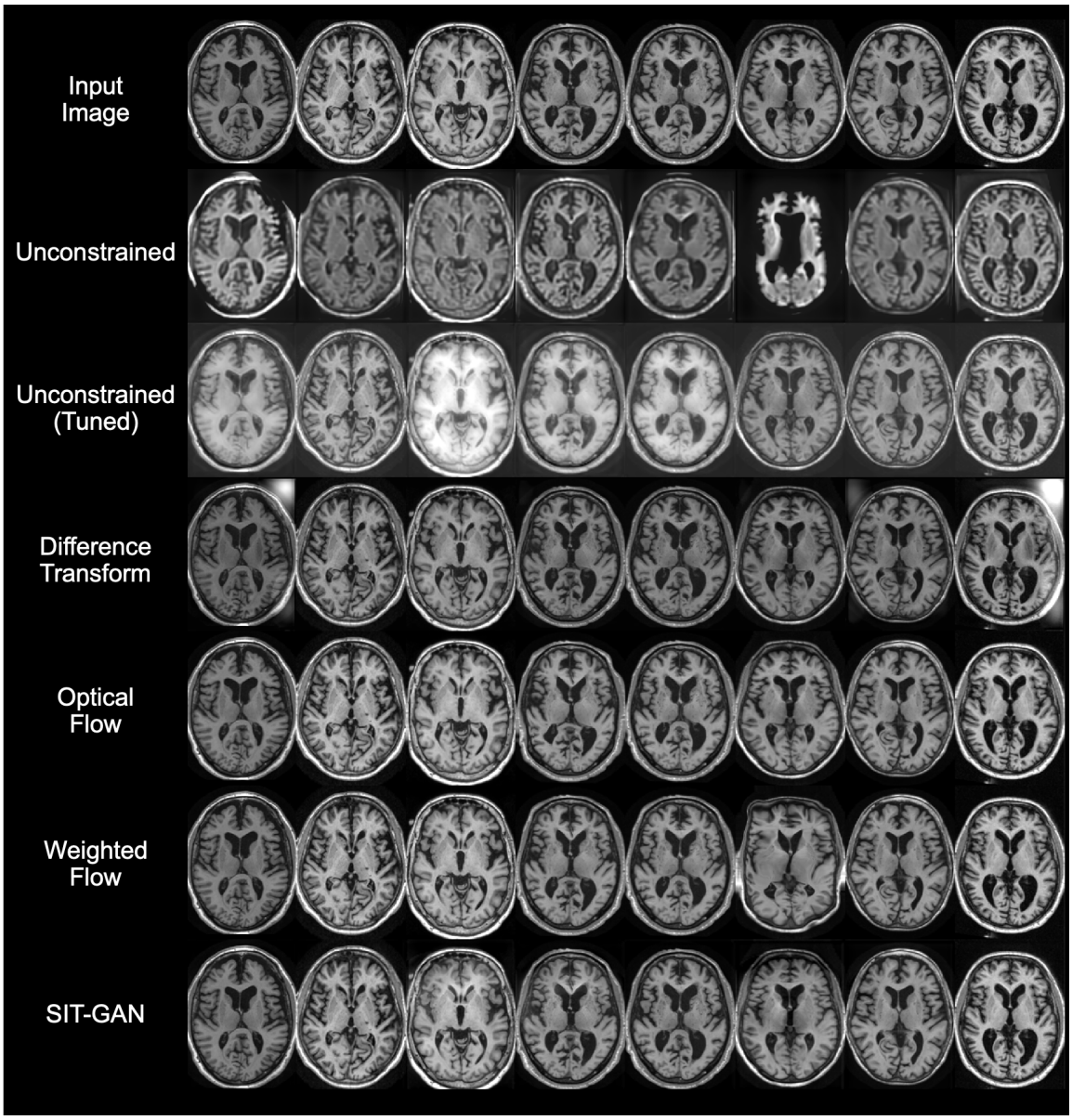

Figure 4-5: Uncurated examples of outputs from all models on the ADNI dataset. Target attribute values sampled from $\mathcal{N}(0, 4)$.

as seen in the 6th column of Figure 4-5.

SIT-GAN performs similarly to the difference transform model, but avoids introducing severe artifacts even under large translations. In general it is overly conservative, but it is capable of matching longitudinal scans quite closely. In Figure 4-4, the unconstrained model simulates increasing age by darkening the ventricles and white matter relative to the background, whereas SIT-GAN properly widens ventricles and sulci as reflected in the ground truth.

We note that most of our models outperform the DSSIM score of the current state of the art model (0.19 ± 0.08, [33]) on longitudinal MRI prediction in ADNI. SIT-GAN's DSSIM of 0.097 ± 0.020 is almost half of that, although our results are not directly comparable as our data processing pipelines differed.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 5

Conclusion

We presented the spatial-intensity transform generative adversarial network, a novel parameterization of GANs for medical image-to-image translation that improves image fidelity and robustness to artifacts. We demonstrated our model on a challenging dataset of clinical quality MRIs of stroke patients from multiple clinical sites, where it outperformed an unconstrained model on distributional metrics at the cost of more conservative transformations. Our model can visualize the correlation between age and ventricle expansion, as well as between the volume of white matter hyperintensities and stroke severity. The network additionally provides a disentangled view of changes in anatomical shape and tissue appearance through the deformation field and intensity difference image respectively. Without further hyperparameter tuning, SIT-GAN can achieve the state of the art on predicting longitudinal T1-weighted brain MRIs from unpaired data.

In each dataset, we saw that SIT-GAN and other types of constrained models could achieve superior image fidelity by sacrificing their ability to match the target domain. This consistent trade-off suggests that SIT-GAN should be applied in scenarios where robustness is particularly important, for example where the output images are directly used for data visualization, exploration or forecasting, but that an unconstrained model may be more powerful for applications such as data augmentation. Our work leaves open questions about how to navigate or circumvent this trade-off. For example, there may be opportunities to compensate the target domain

transfer by carefully relaxing the constraint (reducing the regularization weights) over the course of training, or by incorporating priors over anatomical structures.

We demonstrated our technique on two modalities of brain MRIs, yet SIT-GAN may offer even more potential for analyzing other organs and disease processes without a standard coordinate frame. For example, it is challenging to visualize the radiographic progression of lesions, abscesses, aneurysms, and many other pathologies on either a patient-specific or population-wide basis, despite a large amount of existing longitudinal data. In many cases, the progression of these pathologies is dominated by a combination of localized morphological and textural changes, making SIT-GAN a suitable model for learning to visualize such disease progression. Indeed, training high-quality unconstrained models on unaligned images of faces or people remains a difficult problem in computer vision, and so the robustness offered by SIT-GAN may be particularly important in applications to pathologies with variable locations.

As suggested by our stroke experiment, SIT-GAN may also be valuable in visualizing morphological and textural variation of organs or radiological findings conditioned on patient phenotype. It can provide a visual representation of counterfactuals and known correlations on a patient-specific basis, and because spatial transforms propagate segmentations from the original image to the synthetic image, one could also characterize and compare changes across different anatomical structures. These capabilities may even be helpful for generating hypotheses for clinical research, although it remains unclear how to best integrate such a model into a clinical research pipeline in order to identify promising directions.

Another key question of ongoing study in image-to-image translation is how to quantify and visualize uncertainty in model outputs, as well as identify distinct modes in the output distribution. This remains a difficult task for conditional GANs, and is particularly relevant in many clinical applications such as prediction of patient-specific as well as population-wide disease trajectories. Spatial-intensity transform constraints may be useful in this context as a way to constrain the search space, since probabilistic formulations of image-to-image translation may be less likely to penalize unrealistic outputs over the course of training.

The development of robust conditional GANs is particularly crucial in the context of the unpredictable ways that such models can induce artifacts, as well as the need for reliable and reproducible methods in clinical research and practice. Image-to-image translation is increasingly important for medical image analysis and clinical research as large, multi-site and longitudinal imaging datasets become available for a wider range of diseases and modalities. With their long history of success in medical image registration, it seems likely that spatial-intensity transforms will continue to play a key role as a prior for models in medical image-to-image translation.

THIS PAGE INTENTIONALLY LEFT BLANK

Bibliography

[1] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. *CoRR*, abs/1811.03962, 2018.

[2] Ruzena Bajcsy and Stane Kovačič. Multiresolution elastic matching. *Computer Vision, Graphics, and Image Processing*, 46(1):1 – 21, 1989.

[3] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. An unsupervised learning model for deformable medical image registration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9252–9260, 2018.

[4] Krishna Chaitanya, Neerav Karani, Christian F. Baumgartner, Olivio Donati, Anton S. Becker, and Ender Konukoglu. Semi-supervised and task-driven data augmentation. *CoRR*, abs/1902.05396, 2019.

[5] A. Chambolle, M. Novaga, D. Cremers, and T. Pock. An introduction to total variation for image analysis. In *in Theoretical Foundations and Numerical Methods for Sparse Recovery, De Gruyter*, 2010.

[6] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *CoRR*, abs/1711.09020, 2017.

[7] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains, 2019.

[8] T. F. Cootes, C. Beeston, G. J. Edwards, and C. J. Taylor. A unified framework for atlas matching using active appearance models. In Attila Kuba, Martin Šáamal, and Andrew Todd-Pokropek, editors, *Information Processing in Medical Imaging*, pages 322–333, Berlin, Heidelberg, 1999. Springer Berlin Heidelberg.

[9] Adrian V Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu. Unsupervised learning for fast probabilistic diffeomorphic registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 729–738. Springer, 2018.

[10] Adrian V Dalca, John Guttag, and Mert R Sabuncu. Anatomical priors in convolutional networks for unsupervised biomedical segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9290–9299, 2018.

[11] Adrian V. Dalca, Marianne Rakic, John Guttag, and Mert R. Sabuncu. Learning conditional deformable templates with convolutional networks, 2019.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[13] Philipp Fischer, Alexey Dosovitskiy, Eddy Ilg, Philip Häusser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. *CoRR*, abs/1504.06852, 2015.

[14] Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

[15] Anne Katrin Giese, Markus D. Schirmer, Kathleen L. Donahue, Lisa Cloonan, Robert Irie, Stefan Winzeck, Mark J.R.J. Bouts, Elissa C. McIntosh, Steven J. Mocking, Adrian V. Dalca, Ramesh Sridharan, Huichun Xu, Petrea Frid, Eva Giralt-Steinhauer, Lukas Holmegaard, Jaume Roquer, Johan Wasselius, John W. Cole, Patrick F. McArdle, Joseph P. Broderick, Jordi Jimenez-Conde, Christina Jern, Brett M. Kissela, Dawn O. Kleindorfer, Robin Lemmens, Arne Lindgren,

James F Meschia, Tatjana Rundek, Ralph L. Sacco, Reinhold Schmidt, Pankaj Sharma, Agnieszka Slowik, Vincent Thijs, Daniel Woo, Bradford B. Worrall, Steven J. Kittner, Braxton D. Mitchell, Jonathan Rosand, Polina Golland, Ona Wu, and Natalia S. Rost. Design and rationale for examining neuroimaging genetics in ischemic stroke: The mri-genie study. *Neurology: Genetics*, 3(5), 10 2017.

[16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[18] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein gans. *CoRR*, abs/1704.00028, 2017.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium. *CoRR*, abs/1706.08500, 2017.

[21] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016.

[22] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017.

[23] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. *CoRR*, abs/1612.01925, 2016.

[24] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015.

[25] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004, 2016.

[26] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. *arXiv preprint arXiv:1912.04958*, 2019.

[27] Julian Krebs, Tommaso Mansi, Hervé Delingette, Li Zhang, Florin C. Ghesu, Shun Miao, Andreas K. Maier, Nicholas Ayache, Rui Liao, and Ali Kamen. Robust non-rigid registration through agent-based action learning. In Maxime Descoteaux, Lena Maier-Hein, Alfred Franz, Pierre Jannin, D. Louis Collins, and Simon Duchesne, editors, *Medical Image Computing and Computer Assisted Intervention MICCAI 2017*, pages 344–352, Cham, 2017. Springer International Publishing.

[28] Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861 – 867, 1993.

[29] Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. *CoRR*, abs/1808.01204, 2018.

[30] Lars M. Mescheder. On the convergence properties of GAN training. *CoRR*, abs/1801.04406, 2018.

[31] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *CoRR*, abs/1802.05957, 2018.

[32] Tran Minh Quan, Thanh Nguyen-Duc, and Won-Ki Jeong. Compressed sensing mri reconstruction using a generative adversarial network with a cyclic loss. *IEEE transactions on medical imaging*, 37(6):1488–1497, 2018.

[33] Daniele Ravi, Daniel C. Alexander, and Neil P. Oxtoby. Degenerative adversarial neuroimage nets: Generating images that mimic disease progression. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 164–172, Cham, 2019. Springer International Publishing.

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[35] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes. Nonrigid registration using free-form deformations: application to breast mr images. *IEEE Transactions on Medical Imaging*, 18(8):712–721, 1999.

[36] Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall, 2018.

[37] Tim Salimans and Diederik P. Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *CoRR*, abs/1602.07868, 2016.

[38] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization?, 2018.

[39] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

[40] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *CoRR*, abs/1808.06601, 2018.

[41] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[42] Jelmer M Wolterink, Anna M Dinkla, Mark HF Savenije, Peter R Seevinck, Cornelis AT van den Berg, and Ivana Išgum. Deep mr to ct synthesis using unpaired data. In *International workshop on simulation and synthesis in medical imaging*, pages 14–23. Springer, 2017.

[43] Po-Wei Wu, Yu-Jing Lin, Che-Han Chang, Edward Y. Chang, and Shih-Wei Liao. Relgan: Multi-domain image-to-image translation via relative attributes, 2019.

[44] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018.

[45] Richard Zhang. Making convolutional networks shift-invariant again. *CoRR*, abs/1904.11486, 2019.

[46] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.

[47] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in GAN fake images. *CoRR*, abs/1907.06515, 2019.

[48] Amy Zhao, Guha Balakrishnan, Frédo Durand, John V. Guttag, and Adrian V. Dalca. Data augmentation using learned transforms for one-shot medical image segmentation. *CoRR*, abs/1902.09383, 2019.

[49] Bo Zhao, Bo Chang, Zequn Jie, and Leonid Sigal. Modular generative adversarial networks. *CoRR*, abs/1804.03343, 2018.

[50] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593, 2017.