# Longitudinal Prediction of Postnatal Brain Magnetic Resonance Images via a Metamorphic Generative Adversarial Network

Yunzhi Huang<sup>a,b</sup>, Sahar Ahmad<sup>b</sup>, Luyi Han<sup>c</sup>, Shuai Wang<sup>d</sup>, Zhengwang Wu<sup>b</sup>, Weili Lin<sup>b</sup>, Gang Li<sup>b</sup>, Li Wang<sup>b</sup>, Pew-Thian Yap<sup>b,\*</sup>

<sup>a</sup>School of Automation, Nanjing University of Information Science and Technology, Nanjing 210044, China

<sup>b</sup>Department of Radiology and Biomedical Research Imaging Center (BRIC), University of North Carolina, Chapel Hill, USA

<sup>c</sup>Department of Radiology and Nuclear Medicine, Radboud University Medical Center, Geert Grooteplein 10, 6525 GA, Nijmegen, The Netherlands <sup>d</sup>Department of Computer Science, Shandong University (Weihai), China

#### Abstract

Missing scans are inevitable in longitudinal studies due to either subject dropouts or failed scans. In this paper, we propose a deep learning framework to predict missing scans from acquired scans, catering to longitudinal infant studies. Prediction of infant brain MRI is challenging owing to the rapid contrast and structural changes particularly during the first year of life. We introduce a trustworthy metamorphic generative adversarial network (MGAN) for translating infant brain MRI from one time-point to another. MGAN has three key features: (i) Image translation leveraging spatial and frequency information for detail-preserving mapping; (ii) Quality-guided learning strategy that focuses attention on challenging regions. (iii) Multi-scale hybrid loss function that improves translation of tissue contrast and structural details. Experimental results indicate that MGAN outperforms existing GANs by accurately predicting both contrast and anatomical details.

Keywords: Infant brain MRI, Longitudinal prediction, Metamorphic GAN

Preprint submitted to Journal of Pattern Recognition

<sup>\*</sup>Corresponding author

Email address: ptyap@med.unc.edu (Pew-Thian Yap)

## 1. Introduction

Brain MRI is commonly used to investigate normative and aberrant brain evolution through infancy [1]. To precisely chart brain growth trajectories, *temporally dense* longitudinal datasets are often required but are difficult to acquire. Moreover, infant studies often involve incomplete longitudinal datasets, given the unique challenges associated with infant MRI acquisition. The missing data at different time points can be due to subject dropouts or failed scans owing to excessive motion, insufficient coverage, or imaging artifacts [2].

Longitudinal prediction of infant brain scans is challenging as brain MRI contrasts change rapidly through the first year of life. The brain volume doubles to about 65% of the adult brain by the end of the first year [3]. The gray matter (GM) follows a faster growth trajectory (108% - 149% increase) compared to white matter (WM; 11% increase) [4]. The rapid brain evolution is characterized by both structural and contrast variations [5, 6]. As shown in Figure 1, the WM appears to be darker than the GM during the neonatal phase as the brain is going through myelination, and by sixth month, WM and GM are almost indistinguishable due to the poor tissue contrast.

### 1.1. Related Work

The longitudinal prediction of infant brain MRI can be formulated as an image-to-image translation task — mapping images from a source time point to a target time point [7]. Several studies in the field of computer vision [7, 8, 9, 10] have shown that generative adversarial networks (GANs) [11] yield superior performance in translating images from a domain to another. In the field of medical image analysis, [12] introduced an auto-context GAN to progressively refine MRI-to-CT synthesis. In their follow-up study, [13] incorporated difficulty-aware attention mechanism to improve predictions in challenging regions. Similarly, [14] introduced self-attention to encourage the transformation of a foreground object while retaining the background. Medical image-to-image translation network (MedGAN) [15] uses a pre-trained classification network as



Figure 1: Appearance and structural changes at two time points during the first year of life. Wavelet decomposition for capturing structural details.

feature extractor to match textures and structural details of synthetic and target CT images. All the aforementioned methods for cross-modality synthesis focus on appearance changes and neglect morphological changes. The longitudinal prediction of infant MR brain images, however, requires dealing with fast-paced structural and appearance changes.

To promote structural consistency in cross-modality synthesis, several recent approaches incorporate segmentation similarity as a learning constraint [16, 17]. However, tissue segmentation of infant brain MRI is challenging due to the overlap of GM and WM intensity distributions (Fig. 1). Several approaches attempted to ensure structural consistency without relying on tissue maps. [18] employed gradient differences in a loss function to improve the prediction of boundaries. [19] incorporated gradient correlation differences in a structureconsistency loss to improve edge alignment in MRI-to-CT synthesis. Although successful, the gradient-based constraint introduces noise and fail to capture sufficient boundary information in images with low contrast. [20] incorporated a patch-based self-similarity loss by comparing each patch with all its neighbors in a pre-defined non-local region to ensure structural consistency. However, the search for corresponding non-local regions is computationally expensive.

#### 1.2. Contributions

In this paper, we employ CycleGAN [8], a cycle consistent generation framework, to simultaneously learn structural and appearance changes between two time points. Major contributions of our work are summarized below:

- (i) We propose a trustworthy adversarial learning metamorphosis framework that accounts for both the appearance and structural changes in infant brain MRI.
- (ii) We use a spatial-frequency transfer block equipped with wavelet decomposition to transform features from multiple frequency bands to learn the structural changes.
- (iii) We employ a quality guidance strategy to incorporate a quality-driven loss function to improve predictions in challenging regions.
- (iv) We devise a multi-scale hybrid loss function to improve the matching of both the textural details and the anatomical edges between the predicted image and the desired target image. The discriminator network is evoked at multiple resolutions via deep-supervision, thus allowing accurate prediction of anatomical structures through adversarial learning.

The rest of the paper is organized as follows: Section 2 details the proposed method. Section 3.2 describes the dataset used for evaluation and presents the experimental results. Section 4 provides additional discussion and concludes the paper.

## 2. Methods

In this work, we implement a framework for prediction of metamorphic changes using a GAN. Details of our method are described next.

## 2.1. Network Architecture

We propose a metamorphic GAN (MGAN) to predict the infant brain MR image scanned at time point  $t_b$  from a time point  $t_a$ . Without loss of generality, we assume that  $t_b > t_a$ . Our network architecture, shown in Fig. 2, is cycle-consistent and learns a reversible translation between the two timepoints. It consists of (i) a forward path for earlier-to-later time-point image prediction and (ii) a backward path for later-to-earlier time-point image prediction. The two generators  $G_a$  and  $G_b$  and their corresponding discriminators  $D_a$ and  $D_b$  follow an encoder-decoder architecture. Both the generators incorporate a spatial-frequency transfer (SFT) block to transform the appearance and structural features via multiple branches detailed in Fig. 3. The two discriminators to focus on challenging regions. We will describe the components of our network in the subsequent sections.

#### 2.1.1. Metamorphic Generator

The metamorphic generator (Fig. 3) takes a 3D patch of size  $64 \times 64 \times 64$  as input and predicts a 3D patch. The generator consists of an encoder, SFT block, and a decoder.

Encoder. The encoding path consists of two convolution blocks, each with a  $3 \times 3 \times 3$  convolution layer, followed by 3D instance normalization (IN) [21] and a rectified linear unit (ReLU) [22]. For downsampling, we use convolution with a stride of 2 instead of pooling to avoid potential information loss. We keep a 1-stride convolution in the first stage of the encoder to retain details, and use a 2-stride convolution in the second stage. The resulting numbers of feature maps in the two-stage encoder are 64 and 32.

Spatial-frequency transfer block. Longitudinal prediction requires translating both contrast and structure between two time points. We propose to embed a spatial-frequency transfer block in between enocder-decoder to extract the spatial and frequency domain information of feature maps. The SFT block is





Figure 2: Overview of the metamorphic GAN.

divided into two branches: (i) frequency transform branch, and (ii) spatial transform branch. The frequency transform branch is equipped with discrete wavelet transform (DWT) that takes into account the low frequency tissue contrast and high frequency structural details. The DWT layer decomposes the feature map into low frequency approximation and high frequency details along three dimensions, resulting in eight subvolumes: *LLL*, *LLH*, *LHL*, *LHH*, *HLL*, *HLH*, *HHL* and *HHH*. This decomposition allows more effective transfer of spatial-frequency details.



Figure 3: Network architecture of the metamorphic generator.

Given the *i*-th channel feature map  $f^i$  of size  $(s_x \times s_y \times s_z)$ , the decomposed feature map  $f^i_j$  at frequency band j is obtained by convolving  $f^i$  with wavelet filter  $w_j$ :

$$f_j^i = f^i \circledast w_j. \tag{1}$$

The wavelet filters for each frequency band are calculated by DWT decomposition and are preset in the convolution layer. Correspondingly, the feature maps are reconstructed in the decode path via inverse discrete wavelet transform (IDWT) layer. We show the representative feature maps from the DWT and IDWT layers in Fig. 4. The DWT layer is akin to pooling layer as the DWT decomposition halves the size of the input feature maps. The IDWT layer corresponds to the deconvolution operation with the fixed weights obtained via wavelet filters. There is also an intermediate *transfer* operation between the DWT and IDWT layer. This transfer operation is realized through 9 residual blocks [23]; the input of each block is processed by two  $3 \times 3 \times 3$  convolution layers with 64 channels followed by IN and ReLU for activation. A shortcut connection is added between the input and the output of every residual convolution block. Residual transfer learning simplifies feature generation and transfer from a source domain to a target domain.

The second branch in the SFT block — spatial transform branch — is integrated to compensate for the information truncated by the wavelets. It is



Figure 4: Feature maps obtained from the DWT and IDWT layers.

implemented using a convolution layer with kernel size  $3 \times 3 \times 3$  and stride of 2 to downsample the feature maps in spatial domain. These downsampled feature maps undergo transfer operation and are later upsampled by strided deconvolution layer with kernel size  $3 \times 3 \times 3$ .

Each branch in the SFT block is trained independently without weight sharing. The feature maps from both the frequency and spatial transform branches are concatenated using a  $3 \times 3 \times 3$  convolution layer and stride of 1, followed by IN and ReLU operation; capturing both the contrast and structural information for translating from the source domain to the target domain.

Decoder. Deep supervision [24] is leveraged in the decoding path to strengthen the gradient flow and encourage learning useful representations at multiple scales. The feature maps are upsampled by a 2-stride deconvolution layer and are then convolved with a  $3 \times 3 \times 3$  kernel to get the predicted output.

#### 2.1.2. Uncertainty Quantization

The uncertainty associated with the prediction stems from two aspects: *epistemic* uncertainty (model uncertainty) and *aleatoric* uncertainty (data uncertainty) [25, 26]. As shown in Fig. 3, two Monte-Carlo (MC) dropout layers are incorporated in our generator to estimate the epistemic uncertainty. MC dropout regularizes the network weights as *Bernoulli* distributions for variational Bayesian inference [27, 28]. Note, MC dropout is only enabled during inference. A set of predictions  $\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N\}$  are sampled from the distribution  $p(\hat{y}|I, \mathbf{w}_n)$  via N stochastic inferences using the metamorphic generator.

The epistemic uncertainty is estimated as the variance over the predictions:

$$\mathcal{U}_e = \sqrt{\frac{\sum_{n=1}^{N} (\hat{y}_n - \overline{y})^2}{N}},\tag{2}$$

where  $\hat{y}$  denotes to the prediction by feeding the generator G an input image I,  $\mathbf{w}_n$  represents the generator weights after the *n*-th dropout, N refers to the number of prediction instances, and  $\overline{y}$  is the mean of the predictions.

The aleatoric uncertainty is typically measured with the test-time augmentation technique [29, 30]. During inference, we perturb the input data with spatial transformations (flip and rotation) and random noise. Similar to the estimation of the epistemic uncertainty, we sample a set of predictions  $\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N\}$  from the distribution  $p(\hat{y}|I, S)$ ) and estimate the aleatoric uncertainty as the variance over the predictions:

$$\mathcal{U}_{a} = \sqrt{\frac{\sum_{n=1}^{N} (S^{-1}(\hat{y}(S(x+rn)) - \overline{y})^{2})}{N}},$$
(3)

where S represents the spatial transformation,  $S^{-1}$  corresponds to the inverse transformation, and rn corresponds to random noise.

## 2.1.3. Multi-scale discriminator

The discriminator in MGAN has a U-shaped architecture, as shown in Fig. 5, to locally distinguish the predicted images from real images. It takes as input a  $64 \times 64 \times 64$  image patch and outputs the quality probability map for the given 3D patch. The continuous probability map quantifies the quality of the predicted image patch. Inferior quality, associated with lower probability values, is commonly associated with complex structures, e.g., the cortical ribbon. Superior quality, associated with higher probability values, corresponds to flat regions with simple structures. In the encoding path of the discriminator, the input is downsampled three times; in the decoding path, the feature maps are upsampled three times. For downsampling/upsampling, we use a  $4 \times 4 \times 4$  convolution/deconvolution layer, followed by IN and ReLU activation. The numbers of feature channels are 64, 128, and 256 in the three stages of the discrimi-

nator. Deep supervision strategy [24] is incorporated in the decoding path to strengthen gradient back propagation.



Figure 5: Network architecture of the multi-scale discriminator.

## 2.2. Loss Functions

We incorporated supervised learning with multi-scale information via deep supervision strategy [24]. The loss function  $\mathcal{L}_{MGAN}$  is defined as:

$$\mathcal{L}_{\text{MGAN}} = \mathcal{L}_{s_1} + \mathcal{L}_{s_2} + \mathcal{L}_{s_3}, \tag{4}$$

where  $s_1$ ,  $s_2$ , and  $s_3$  refer to the three scales employed [31, 32]. For each scale, the objective function is composed of three loss functions to effectively learn the prediction task. The loss functions are described next.

## 2.2.1. Adversarial Loss

We propose to use the standard adversarial loss function, which aims to match the distribution of the predicted images with that of the real images. It is given by

$$\mathcal{L}(G_a, G_b, D_a, D_b) = \mathbb{E}_{I_{t_a}}[\log(D_a(I_{t_a})] \\ + \mathbb{E}_{I_{t_b}}[\log(1 - D_a(G_b(I_{t_b})))] \\ + \mathbb{E}_{I_{t_b}}[\log(D_b(I_{t_b})] \\ + \mathbb{E}_{I_{t_a}}[\log(1 - D_b(G_a(I_{t_a})))],$$
(5)

where  $I_{t_a}$  and  $I_{t_b}$  refer to the images at time-point  $t_a$  and  $t_b$ , respectively,  $G_a$ and  $G_b$  are the mapping functions, and  $D_a$  and  $D_b$  are the discriminators.

#### 2.2.2. Paired Loss

The generators  $G_a$  and  $G_b$  seek to minimize the difference between real and predicted images. We propose to enhance the performance of the generators by defining a paired loss function that constraints the difference at voxel-, feature-, and frequency-level. Our paired loss function  $\mathcal{L}_{\text{gen}}^{(\cdot)}$  consists of three loss terms: (i) quality-driven loss, (ii) texture loss, and (iii) frequency loss.

$$\mathcal{L}_{gen}^{G_a} = \mathcal{L}_{Q}^{G_a} + \mathcal{L}_{T}^{G_a} + \mathcal{L}_{F}^{G_a},$$

$$\mathcal{L}_{gen}^{G_b} = \mathcal{L}_{Q}^{G_b} + \mathcal{L}_{T}^{G_b} + \mathcal{L}_{F}^{G_b}.$$
(6)

Quality-driven loss. The low tissue contrast and the dramatic brain growth hinder translation of regions such as the convoluted cerebral cortex. Here, we present a quality-guided learning strategy to strengthen the transformation of the unfathomable regions. The discriminator outputs a quality map that defines the voxel-wise probabilities for each predicted image. The heterogeneous distribution of the probabilities in the quality map motivates us to treat voxels differently. Voxels with lower probabilities correspond to poor prediction and require more attention compared to those with higher probabilities. This enhances the image translation power of the generator at complex regions in the infant brain MRI. The quality-driven loss  $\mathcal{L}_{Q}^{(\cdot)}$  is defined as:

$$\mathcal{L}_{\mathbf{Q}}^{G_{a}}(G_{a};\theta^{G_{a}}) = \mathbb{E}_{I_{t_{a}},I_{t_{b}},Q^{D_{b}}}[\|I_{t_{b}} - G_{a}(I_{t_{a}})\|_{1} \odot (1 - Q^{D_{b}})^{\beta}],$$

$$\mathcal{L}_{\mathbf{Q}}^{G_{b}}(G_{b};\theta^{G_{b}}) = \mathbb{E}_{I_{t_{a}},I_{t_{b}},Q^{D_{a}}}[\|I_{t_{a}} - G_{b}(I_{t_{b}})\|_{1} \odot (1 - Q^{D_{a}})^{\beta}],$$
(7)

where  $Q^{(\cdot)}$  is the quality map,  $\theta^{(\cdot)}$  denotes the parameters of the network,  $\odot$  defines the element-wise multiplication and  $\beta$  represents the parameter that enables to focus on difficult-to-predict regions. If  $\beta$  is set to zero, then  $\mathcal{L}_{Q}^{(\cdot)}$  will be equivalent to  $\mathcal{L}_{1}$  norm; losing the ability to define adaptive weights based on quality map. In this study, we empirically set its value to 1.5.

*Texture loss.* This loss ensures that the predicted image has a texture similar to the target image, and it is defined as the mean square error (MSE) between the Gram matrix of the target and the predicted image [33, 34]:

$$\mathcal{L}_{T}^{G_{a}} = \|M(I_{t_{b}}) - M(G_{a}(I_{t_{a}}))\|_{2},$$
  
$$\mathcal{L}_{T}^{G_{b}} = \|M(I_{t_{a}}) - M(G_{b}(I_{t_{b}}))\|_{2}.$$
(8)

The gram matrix  $M(\cdot)$  is the inner product of the generated images.

Frequency loss. The frequency loss  $\mathcal{L}_{\mathrm{F}}^{(\cdot)}$  is incorporated via wavelet decomposition of the generators' outputs, which steers the effective prediction of the structural details.  $\mathcal{L}_{\mathrm{F}}^{(\cdot)}$  is defined as:

$$\mathcal{L}_{\rm F}^{G_a} = \sum_{k \in K} \| \text{DWT}(I_{t_b})^k - \text{DWT}(G_a(I_{t_a}))^k \|_1,$$

$$\mathcal{L}_{\rm F}^{G_b} = \sum_{k \in K} \| \text{DWT}(I_{t_a})^k - \text{DWT}(G_b(I_{t_b}))^k \|_1,$$
(9)

where  $K = \{LLL, LLH, LHL, HLL, LHH, HLH, HHL, HHH\}$ ; LLL corresponds to the approximation coefficients which encode the image contrast, and the remaining terms correspond to the detail coefficients, encoding the high frequency structural details. The wavelet coefficients are decomposed using bior1.3 [35], which is compactly supported by a biorthogonal spline wavelet [36].

## 2.2.3. Cycle Consistency Loss

The cycle consistency loss function  $\mathcal{L}_{cyc}$  ensures that the image prediction cycle brings the predicted image back to the original image, i.e.,  $G_b(G_a(I_{t_a})) \approx I_{t_a}$  and it is given by:

$$\mathcal{L}_{cyc}(G_a, G_b) = \mathbb{E}_{I_{t_a}}[\|I_{t_a} - G_b(G_a(I_{t_a}))\|_1], \\ + \mathbb{E}_{I_{t_b}}[\|I_{t_b} - G_a(G_b(I_{t_b}))\|_1].$$
(10)

This loss function constraints both the forward and backward image prediction cycles, causing  $G_a$  and  $G_b$  to be consistent with each other.

#### 3. Experimental Results

#### 3.1. Data Acquisition and Preprocessing

The dataset consists of longitudinal T1-weighted (T1w) and T2-weighted (T2w) MR images of healthy infant subjects enrolled in the Multi-visit Advanced Pediatric Brain Imaging (MAP) study. Informed written consent was obtained from the parents of all the participants and all study protocols were approved by the University of North Carolina at Chapel Hill Institutional Review Board. Each subject was scanned every three months in the first postnatal year. The imaging parameters for T1w MRI data were: TR = 1900 ms, TE = 4.38 ms, flip angle = 7°. All the images had 144 sagittal slices and 1 mm isotropic voxel resolution. The imaging parameters for T2w MR images were TR = 7380 ms, TE = 119 ms, flip angle = 150°, 64 sagittal slices, and  $1.25 \times 1.25 \times 1.95 \text{ mm}^3$  voxel size.

The dataset was preprocessed using our infant-dedicated preprocessing pipeline [37, 38]. Then, all the postnatal images of each subject were linearly aligned to their corresponding 12-months-old images and resampled to the size of  $256 \times 256 \times 256$  with  $1 \times 1 \times 1 \text{ mm}^3$  voxel resolution. We randomly split the MRI data from 30 healthy infants into 20 and 10 for training and testing, respectively. Five-fold cross-validation was performed to tune the hyper-parameters.

#### 3.2. Implementation Details

The proposed metamorphic GAN was implemented using TensorFlow library [39] on a single Nvidia TitanX (Pascal) GPU. Adam optimizer [40] was adopted with an initial learning rate of  $1 \times 10^{-4}$  and batch size of 1. Training, validation, and testing were performed separately for T1w and T2w images.

During training, we uniformly sampled 3D patches from each image encompassing the brain region with a dense stride of 10, providing sufficient samples for training. The generator was first trained with 5 epochs before the adversarial training. The adversarial training was stopped at 50 epochs.

During inference, the N = 20 inferences were performed for the estimation of epistemic and aleatoric uncertainty. The keep rate of the dropout layers was set to 0.8. Test-time data augmentation was carried out using a combination of random flip, rotation along each of the three axes, and random noise, which were modeled respectively with discrete Bernoulli distribution  $\mathcal{B}(0.5)$ , uniform distribution  $\mathcal{U}(0, 2\pi)$ , and normal distribution  $\mathcal{N}(0, 0.05)$ .

#### 3.3. Evaluation Criteria

We employed two commonly used metrics to evaluate the quality of the predicted images: (i) peak signal-to-noise ratio (PSNR), and (ii) structural similarity (SSIM) [41]. Higher PSNR and SSIM correspond to accurate image prediction.

## 3.4. Comparison with Existing Techniques

We compared MGAN with three widely used GANs: CycleGAN [16], Pix2Pix [7], and WGAN [42]. All the compared models were used to predict the 12-monthold brain MRI from the 2-week-old brain MRI. The prediction task is challenging due to the extent of changes between the two time points (Fig. 1). For fair comparison, we re-trained the GANs for optimal parameters.

The image prediction results shown for the compared models in Fig. 6 indicate that MGAN yields T1w and T2w image predictions that are closer to the ground truth with richer details than the other models. The error maps indicate that MGAN achieves the lowest error among all methods, especially around the ventricles and cerebral cortex. Summary statistics for PSNR and SSIM are reported in Table 1. MGAN achieves significant improvement (p < 0.05, paired *t*-test) for PSNR and SSIM over other methods.

### 3.5. Ablation Study

Here, we investigate the effectiveness of three components of MGAN — the SFT block, quality-guided learning, and the hybrid loss function. The influence of frequency transform on longitudinal prediction was verified using two variants of the metamorphic generator: (i) incorporating the SFT block equipped with both the frequency and spatial transform branches, and (ii) replacing the

Axial		x X				V V
Error Map						1 64 64 62 0
Sagittal						
Error Map						1 66 64 62 0
Coronal						
Error Map						1 84 64 62
	Source Image (2 weeks)	CycleGAN	Pixel2Pixel	WGAN	MGAN	Ground Truth (12 months)

(a) T1w image predictions



(b) T2w image predictions

Figure 6: Longitudinal image prediction with various GANs.

	T	lw	T	2w
Method	PSNR	SSIM $(\%)$	PSNR	SSIM $(\%)$
CycleGAN	$22.6{\pm}1.1$	74.2±2.8	21.8±1.0	75.3±2.2
Pix2Pix	$23.0{\pm}1.3$	$76.2 \pm 3.4$	$22.9{\pm}0.9$	$77.2 \pm 2.8$
WGAN	$24.1{\pm}1.2$	$79.4{\pm}2.4$	$23.4{\pm}0.9$	$79.5 {\pm} 2.0$
MGAN	$\textbf{26.4}{\pm}\textbf{0.9}$	$84.0{\pm}2.2$	$25.5{\pm}0.7$	$84.8{\pm}1.8$

Table 1: Summary statistics of PSNR and SSIM for different GANs.

SFT block with a conventional spatial transform branch. We also investigated the efficacy of quality-guided learning by conducting experiments with/without quality maps generated by the discriminators. The configurations are summarized as follows:

- Backbone: SFT with only spatial transform branch and without quality guidance.
- SFT-NCG: SFT without quality guidance.
- ST-CG: Conventional spatial transform and quality guidance.
- MGAN: SFT and quality guidance.

Table 2 indicates that MGAN achieves the highest PSNR and SSIM with a significant improvement (p < 0.05, paired *t*-test). SFT-NCG and ST-CG perform better than Backbone, validating that wavelet-based feature mapping and quality-guidance improve prediction accuracy.

Fig. 7 shows that Backbone and ST-CG predict the 12-month scan poorly due to the spatial complexity of the cortical ribbon. SFT-NCG generated unsatisfactory results at difficult-to-predict regions as indicated by the high values in the error map. MGAN yields the most accurate prediction, which matches the ground truth both in terms of tissue contrast and anatomical structure.

We investigated the contribution of the uncertainty-aware loss  $\mathcal{L}_{Q}$ , texture loss  $\mathcal{L}_{T}$ , and frequency loss  $\mathcal{L}_{F}$ . Table 3 indicates that including all loss



(a) T1w image predictions



Figure 7: Longitudinal image prediction results obtained with different MGAN configurations.

	T	1w	T	2w
Model	PSNR	SSIM $(\%)$	PSNR	SSIM $(\%)$
Backbone	24.9±0.5	$80.3 \pm 1.3$	24.2±0.4	81.1±1.2
SFT-NCG	$25.2{\pm}1.0$	$83.7 {\pm} 2.0$	$25.1 {\pm} 0.9$	$83.0{\pm}1.6$
ST-CG	$25.9{\pm}1.2$	$82.5 {\pm} 2.0$	$24.8{\pm}1.0$	$82.2 \pm 1.5$
MGAN	$26.4{\pm}0.9$	84.0±2.2	$25.5{\pm}0.7$	84.8±1.8

Table 2: Ablation study with different MGAN configurations.

terms (Eq. 6) yields the highest PSNR and SSIM. In contrast, using only the uncertainty-aware loss yields the lowest PSNR and SSIM. This implies that both the texture and frequency losses improve the predictive power of the generator.

Table 3: Ablation study with different combinations of losses.

			T1w		T	2w
$\mathcal{L}_{\mathrm{Q}}$	$\mathcal{L}_{\mathrm{T}}$	$\mathcal{L}_{\mathrm{F}}$	PSNR	SSIM $(\%)$	PSNR	SSIM $(\%)$
$\checkmark$			26.0±1.3	83.3±1.4	25.1±0.8	83.1±1.4
$\checkmark$	$\checkmark$		$26.2{\pm}1.0$	$83.5 {\pm} 2.0$	$25.3{\pm}0.9$	$83.7 {\pm} 1.9$
$\checkmark$		$\checkmark$	$26.1{\pm}1.0$	$83.7 {\pm} 1.8$	$25.2{\pm}0.8$	84.3±1.7
$\checkmark$	$\checkmark$	$\checkmark$	$26.4{\pm}0.9$	$84.0{\pm}2.2$	$25.5{\pm}0.7$	84.8±1.8

## 3.6. Longitudinal Prediction

We demonstrate the effectiveness MGAN in predicting a 12-month-old image from any earlier time-point, i.e., 2 weeks, 3 months, 6 months, and 9 months. Predictions from the forward and backward prediction paths are evaluated. The predicted images along with the error maps and uncertainty maps are shown in Fig. 8. The quantitative results are presented in Table 4. Despite the significant



(a) T1w image predictions





Figure 8: Longitudinal prediction results for different time points. (*Left*) The forward path predicts a 12-month-old image from images at earlier time points. (*Right*) The backward path predicts images of earlier time points from a 12-month-old image.

differences in appearance and structure, MGAN is able to predict the images with great resemblance to the ground-truth images in both tissue contrast and anatomical structure. This is validated by the high PSNR and SSIM values. The corresponding epistemic and aleatoric uncertainty maps of the predictions are also shown in Fig. 8. The epistemic and aleatoric uncertainty is positively



Figure 9: Quality visualization for the 0-to-12-month-old prediction.

correlated with prediction errors.

Table 4: Statistical summary of evaluation metrics for longitudinal prediction.

		T1w		T2w		
		PSNR	$\mathrm{SSIM}(\%)$	PSNR	$\mathrm{SSIM}(\%)$	
ion	$0m \rightarrow 12m$	$26.4{\pm}0.9$	84.0±2.2	$25.5 \pm 0.7$	84.8±1.8	
oredic	$3m\rightarrow 12m$	$26.1 \pm 1.3$	84.7±4.0	$25.7 {\pm} 0.9$	$83.8 \pm 2.2$	
forward p	$6m \rightarrow 12m$	$27.7 \pm 2.2$	$89.1 \pm 3.6$	$26.8{\pm}1.8$	87.5±2.8	
	$9m\rightarrow 12m$	$29.0{\pm}2.9$	89.9±4.6	$28.5 \pm 1.9$	88.3±2.8	
backward prediction	$12m \rightarrow 0m$	$27.1 \pm 0.9$	$86.7 {\pm} 0.2$	$26.5 \pm 1.1$	86.7±1.2	
	$12m \rightarrow 3m$	$26.9 \pm 1.7$	$86.9 \pm 3.1$	$26.4{\pm}1.2$	$86.4 \pm 2.2$	
	$12m \rightarrow 6m$	$27.8 \pm 1.7$	$89.7 \pm 2.7$	27.1±1.8	89.4±3.2	
	$12m \rightarrow 9m$	$28.4{\pm}2.5$	$90.5 \pm 3.1$	$27.6 \pm 2.2$	$90.1 \pm 3.4$	

## 4. Discussion

In this paper, we presented a metamorphic GAN that can be trained to predict infant brain MRI from one time point to another. Longitudinal prediction of infant brain MRI is challenging owing to rapid contrast and structural changes in the first year of life. To capture these changes, our MGAN incorporates a SFT block and integrates quality-guided learning via a hybrid loss function.

We compared our method with existing generative adversarial networks, such as CycleGAN, Pix2Pix, and WGAN. We found that these networks are effective in prediction structures at a global scale but are less effective in predicting fine-scale structural details, especially in the cortex (Fig. 6). In contrast, our prediction network capture spatially heterogeneous changes by employing both spatial and frequency transforms to generate feature maps. Particularly, DWTbased frequency transform decomposes the image into low and high frequency components to help the translation of image contrast and subtle details (Fig. 7).

The quality-guided learning strategy involves using an estimation map for characterizing voxel-wise prediction quality. Fig. 9 shows that regions with complex structures, e.g., the cerebral cortex, are associated with higher bias values. In contrast, regions with simple structure, e.g., lateral ventricles, are associated with lower bias values. As shown in Fig. 7, employing the qualitydriven  $\mathcal{L}_Q$  loss results in more accurate predictions at challenging regions with complex patterns. Additionally, the wavelet decomposition and gram matrix enhance the similarity between predictions and ground truths both in terms of content and style (Table 3).

## 5. Conclusion

We have proposed a trustworthy learning-based framework for longitudinal postnatal brain MRI prediction. The key feature our method is the utilization of wavelet transform to enable image prediction at multiple frequencies. We utilize quality guidance to strengthen the learning of prediction of challenging regions. We employ a hybrid loss function and a multi-scale discriminator to capture differences in global intensity, style, and structure. Experimental results demonstrate that our method achieves superior performance over several stateof-the-art image-to-image translation networks. Despite the effectiveness of our method, it is currently trained with paired data. In future, it can be extended to be trainable with unpaired data.

## Acknowledgments

This work was supported in part by United States National Institutes of Health (NIH) grants EB008374, EB006733, and AG053867. Y. Huang was supported by the China Scholarship Council and the National Natural Science Foundation of China under Grant 6210011424.

## References

- J. Dubois, M. Benders, A. Cachia, F. Lazeyras, R. Ha-Vinh Leuchter, S. V. Sizonenko, C. Borradori-Tolsa, J.-F. Mangin, P. S. Hüppi, Mapping the early cortical folding process in the preterm newborn brain, Cerebral Cortex 18 (6) (2008) 1444 – 1454.
- B. R. Howell, M. A. Styner, W. Gao, P.-T. Yap, L. Wang, K. Baluyot, E. Yacoub, G. Chen, T. Potts, A. Salzwedel, G. Li, J. H. Gilmore, J. Piven, J. K. Smith, D. Shen, K. Ugurbil, H. Zhu, W. Lin, J. T. Elison, The UNC/UMN baby connectome project (BCP): An overview of the study design and protocol development, NeuroImage 185 (2019) 891 905. doi:https://doi.org/10.1016/j.neuroimage.2018.03.049.
  URL http://www.sciencedirect.com/science/article/pii/S1053811918302593
- [3] J. H. Gilmore, R. C. Knickmeyer, W. Gao, Imaging structural and functional brain development in early childhood, Nature Reviews Neuroscience 19 (3) (2018) 123.

- [4] J. Matsuzawa, M. Matsui, T. Konishi, K. Noguchi, R. C. Gur, W. Bilker, T. Miyawaki, Age-related volumetric changes of brain gray and white matter in healthy infants and children, Cerebral Cortex 11 (4) (2001) 335 – 342. doi:https://doi.org/10.1093/cercor/11.4.335.
- [5] R. C. Knickmeyer, S. Gouttard, C. Kang, D. Evans, K. Wilber, J. K. Smith, R. M. Hamer, W. Lin, G. Gerig, J. H. Gilmore, A structural MRI study of human brain development from birth to 2 years, Journal of Neuroscience 28 (47) (2008) 12176 – 12182.
- [6] T. Paus, D. Collins, A. Evans, G. Leonard, B. Pike, A. Zijdenbos, Maturation of white matter in the human brain: a review of magnetic resonance studies, Brain research bulletin 54 (3) (2001) 255 – 266.
- [7] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with conditional adversarial networks, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 5967 – 5976.
- [8] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, 2017 IEEE International Conference on Computer Vision (ICCV) (2017) 2242 – 2251.
- [9] X. Huang, M.-Y. Liu, S. J. Belongie, J. Kautz, Multimodal unsupervised image-to-image translation, in: ECCV, 2018.
- [10] M.-Y. Liu, T. Breuel, J. Kautz, Unsupervised image-to-image translation networks, in: NIPS, 2017.
- [11] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, Y. Bengio, Generative adversarial networks, ArXiv abs/1406.2661.
- [12] D. Nie, R. Trullo, J. Lian, C. Petitjean, S. Ruan, Q. Wang, D. Shen, Medical image synthesis with context-aware generative adversarial networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2017, pp. 417 – 425.

- [13] D. Nie, D. Shen, Adversarial confidence learning for medical image segmentation and synthesis, International Journal of Computer Vision (2020) 1 – 20.
- [14] K. Lee, M.-K. Choi, H. Jung, DavinciGAN: Unpaired surgical instrument translation for data augmentation, in: MIDL, 2018.
- [15] K. Armanious, C. Yang, M. Fischer, T. Küstner, K. Nikolaou, S. Gatidis, B. Yang, MedGAN: Medical image translation using GANs, Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society 79 (2019) 101684.
- [16] Z. Zhang, L. Yang, Y. Zheng, Translating and segmenting multimodal medical volumes with cycle- and shape-consistency generative adversarial network, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018) 9242 – 9251.
- [17] A. Chartsias, T. Joyce, R. Dharmakumar, S. A. Tsaftaris, Adversarial image synthesis for unpaired multi-modal cardiac data, in: SASHIMI@MICCAI, 2017.
- [18] D. Nie, R. Trullo, J. Lian, L. Wang, C. Petitjean, S. Ruan, Q. Wang, D. Shen, Medical image synthesis with deep convolutional adversarial networks, IEEE Transactions on Biomedical Engineering 65 (12) (2018) 2720 – 2730.
- [19] Y. Hiasa, Y. Otake, M. Takao, T. Matsuoka, K. Takashima, A. Carass, J. L. Prince, N. Sugano, Y. Sato, Cross-modality image synthesis from unpaired data using CycleGAN, in: International workshop on simulation and synthesis in medical imaging, Springer, 2018, pp. 31 – 41.
- [20] H. Yang, J. Sun, A. Carass, C. Zhao, J. Lee, Z. Xu, J. L. Prince, Unpaired brain MR-to-CT synthesis using a structure-constrained CycleGAN, in: DLMIA/ML-CDS@MICCAI, 2018.

- [21] D. Ulyanov, A. Vedaldi, V. S. Lempitsky, Instance normalization: The missing ingredient for fast stylization, ArXiv abs/1607.08022.
- [22] X. Glorot, A. Bordes, Y. Bengio, Deep sparse rectifier neural networks, in: AISTATS, 2011.
- [23] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 770 – 778.
- [24] S. Xie, Z. Tu, Holistically-nested edge detection, International Journal of Computer Vision 125 (2015) 3–18.
- [25] A. Der Kiureghian, O. Ditlevsen, Aleatory or epistemic? does it matter?, Structural safety 31 (2) (2009) 105–112.
- [26] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, A. Khosravi, U. R. Acharya, V. Makarenkov, et al., A review of uncertainty quantification in deep learning: Techniques, applications and challenges, arXiv preprint arXiv:2011.06225.
- [27] A. Kendall, V. Badrinarayanan, R. Cipolla, Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding, arXiv preprint arXiv:1511.02680.
- [28] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: international conference on machine learning, PMLR, 2016, pp. 1050–1059.
- [29] M. S. Ayhan, P. Berens, Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks.
- [30] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks, Neurocomputing 338 (2019) 34–45.

- [31] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, B. Catanzaro, Highresolution image synthesis and semantic manipulation with conditional GANs, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8798 – 8807.
- [32] A. Karnewar, O. Wang, MSG-GAN: Multi-scale gradients for generative adversarial networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7799 – 7808.
- [33] L. Gatys, A. S. Ecker, M. Bethge, Texture synthesis using convolutional neural networks, in: Advances in neural information processing systems, 2015, pp. 262 – 270.
- [34] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, M.-H. Yang, Diversified texture synthesis with feed-forward networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3920 – 3928.
- [35] A. Cohen, Biorthogonal bases of compactly supported wavelets, 2006.
- [36] R. Szewczyk, K. Grabowski, M. Napieralska, W. Sankowski, M. Zubert, A. Napieralski, A reliable iris recognition algorithm based on reverse biorthogonal wavelet transform, Pattern Recognition Letters 33 (8) (2012) 1019 – 1026.
- [37] Y. Dai, F. Shi, L. Wang, G. Wu, D. Shen, iBEAT: a toolbox for infant brain magnetic resonance image processing, Neuroinformatics 11 (2) (2013) 211 - 225.
- [38] G. Li, J. Nie, L. Wang, F. Shi, W. Lin, J. H. Gilmore, D. Shen, Mapping region-specific longitudinal cortical surface expansion from birth to 2 years of age, Cerebral cortex 23 (11) (2013) 2724 – 2733.
- [39] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga,

S. Moore, D. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden,X. Zhang, TensorFlow: A system for large-scale machine learning.

- [40] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, CoRR abs/1412.6980.
- [41] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing 13 (2004) 600 – 612.
- [42] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein generative adversarial networks, in: ICML, 2017.