Journal of Machine Learning for Biomedical Imaging 2024:004 vol. 2, pp. 657–685 Submitted 08/2023
Special issue: Generative Models Published 02/2024
Guest editors: Mert Sabuncu, Sotirios A. Tsaftaris

CoNeS: Conditional neural fields with shift modulation for multi-sequence MRI translation

Yunjie Chen https://orcid.org/0000-0001-9478-6953 y.chen@lumc.nl
Department of Radiology, Leiden University Medical Center, Leiden, the Netherlands

Marius Staring https://orcid.org/0000-0003-2885-5812 m.staring@lumc.nl
Department of Radiology, Leiden University Medical Center, Leiden, the Netherlands

Olaf M. Neve https://orcid.org/0000-0002-5104-8448 o.m.neve@lumc.nl
Department of Otorhinolaryngology and Head & Neck Surgery, Leiden University Medical Center, Leiden, the Netherlands

Stephan R. Romeijn https://orcid.org/0000-0002-4634-447X s.r.romeijn@lumc.nl
Department of Radiology, Leiden University Medical Center, Leiden, the Netherlands

Erik F. Hensen https://orcid.org/0000-0002-4393-7421 e.f.hensen@lumc.nl
Department of Otorhinolaryngology and Head & Neck Surgery, Leiden University Medical Center, Leiden, the Netherlands

Berit M. Verbist https://orcid.org/0000-0002-1010-2583 b.m.verbist@lumc.nl
Department of Radiology, Leiden University Medical Center, Leiden, the Netherlands

Jelmer M. Wolterink https://orcid.org/0000-0001-5505-475X j.m.wolterink@utwente.nl
Department of Applied Mathematics, Technical Medical Center, University of Twente, Enschede, the Netherlands

Qian Tao https://orcid.org/0000-0001-7480-0703 q.tao@tudelft.nl
Department of Imaging Physics, Delft University of Technology, Delft, the Netherlands

Abstract

Multi-sequence magnetic resonance imaging (MRI) has found wide applications in both modern clinical studies and deep learning research. However, in clinical practice, it frequently occurs that one or more of the MRI sequences are missing due to different image acquisition protocols or contrast agent contraindications of patients, limiting the utilization of deep learning models trained on multi-sequence data. One promising approach is to leverage generative models to synthesize the missing sequences, which can serve as a surrogate acquisition. State-of-the-art methods tackling this problem are based on convolutional neural networks (CNN) which usually suffer from spectral biases, resulting in poor reconstruction of high-frequency fine details. In this paper, we propose Conditional Neural fields with Shift modulation (CoNeS), a model that takes voxel coordinates as input and learns a representation of the target images for multi-sequence MRI translation. The proposed model uses a multi-layer perceptron (MLP) instead of a CNN as the decoder for pixel-to-pixel mapping. Hence, each target image is represented as a neural field that is conditioned on the source image via shift modulation with a learned latent code. Experiments on BraTS 2018 and an in-house clinical dataset of vestibular schwannoma patients showed that the proposed method outperformed state-of-the-art methods for multi-sequence MRI translation both visually and quantitatively. Moreover, we conducted spectral analysis, showing that CoNeS was able to overcome the spectral bias issue common in conventional CNN models. To further evaluate the usage of synthesized images in clinical downstream tasks, we tested a segmentation network using the synthesized images at inference. The results showed that CoNeS improved the segmentation performance when some MRI sequences were missing and outperformed other synthesis models. We concluded that neural

https://doi.org/10.59275/j.melba.2024-d61g

fields are a promising technique for multi-sequence MRI translation. Our code is available at `https://github.com/cyjdswx/CoNeS.git`.

Keywords: Neural fields, Magnetic Resonance Imaging, generative models, image-to-image translation, segmentation

1. Introduction

Multi-sequence magnetic resonance imaging (MRI) plays a key role in radiology and medical image computing. One advantage of MRI is the availability of various pulse sequences, such as T1-weighted MRI (T1), T2-weighted MRI (T2), T1-weighted with contrast (T1ce), and T2-fluid-attenuated inversion recovery MRI (FLAIR), which can provide complementary information to clinicians (Cherubini et al., 2016). The importance of the availability of multi-sequence MRI was also indicated by recent deep learning research (Cercignani and Bouyagoub, 2018), which shows that the more sequences were used for segmentation, the better results could be obtained. However, due to clinical restrictions on the use of contrast agents and the diversity in imaging protocols in different medical centers, it is difficult and time-consuming to always obtain exactly the same MRI sequences for training and inference, which may damage the generalization and performance of deep learning segmentation models.

One way to tackle this problem is to generate missing sequences from existing images based on the information learned from a set of paired images, known as image-to-image translation. Like in other computer vision tasks, convolutional neural networks (CNNs) with an encoder and decoder architecture are normally used for this specific task (Sevetlidis et al., 2016; Joyce et al., 2017; Wei et al., 2019). Despite the significant improvement over traditional non-deep-learning methods, these methods still suffer from the limitation of a pixel-wise loss function, such as the L1 or MSE loss, which tends to result in blurry results with undesirable loss of details in image structures (Isola et al., 2017; Dalmaz et al., 2022). To overcome this limitation, generative adversarial networks (GANs) were introduced for image-to-image translation and rapidly became a training protocol benchmark for medical image translation (Li et al., 2019; Nie et al., 2018; Armanious et al., 2020). GANs improve translation results both visually and quantitatively owing to the adversarial learning loss, which penalizes the images that are correctly classified by the discriminator.

However, research showed that generative models that use a CNN as a backbone network consisting of ReLU activation functions and transposed or up-convolutional layers usually suffer from spectral biases (Rahaman et al., 2019; Durall et al., 2020). Therefore, these generative models fit low-frequency signals first and may again fail to capture details in image structures during training. Transformers, which instead use multi-head self-attention blocks and multi-layer perceptrons (MLPs) have gained tremendous attention in computer vision research (Liu et al., 2023b; Jiang et al., 2021). Due to the absence of convolutional layers, transformers show great potential for preserving fine details and long-range dependencies and have recently been applied to medical image translation (Liu et al., 2023a; Dalmaz et al., 2022). However, despite the numerous efforts made by these studies, such as hybrid architectures and image patch-based processing, the training of transformers is still considered heavy and data-demanding (Dosovitskiy et al., 2021; Esser et al., 2021). The inherently high computational complexity of the transformer block and expensive mem-

ory cost of low-level tasks, such as denoising and super-resolution, further complicate the application of transformers in medical image translation (Chen et al., 2021a).

To address these limitations, we propose image-to-image translation using neural fields (Xie et al., 2022). In contrast to CNN or transformer-based methods, a neural field represents the target images on a continuous domain using a coordinate-based network, which can be conditioned on the information extracted from the source images. We previously proposed an image-to-image translation approach based on neural fields (Chen et al., 2023). Here, we substantially extend this model by proposing **Co**nditional **Ne**ural fields with **S**hift modulation (CoNeS). In contrast to traditional deep learning computer vision techniques, CoNeS parameterizes the target images as neural fields that can be queried on a grid to provide pixel-wise predictions. Specifically, we use an MLP as the decoder to map the voxel coordinates to the intensities on the target images. To capture instance-specific information, we condition the neural fields on the latent codes extracted from the source images. By applying shift modulation, the neural fields can be further varied across the coordinates to enhance their ability to preserve high-frequency signals.

Although plenty of work has shown great progress in medical image translation, most previous works have been evaluated based on image similarity metrics and only a few papers have evaluated the benefits of using synthesized images for downstream analysis. Amirrajab et al. (2023) fine-tuned a segmentation model with synthesized cardiac images to improve the performance of different modalities; Skandarani et al. (2020) introduced a variational autoencoder (VAE) based network for image translation based data augmentation to improve the generalization capabilities of a segmentation model. In practice, however, it would be more straightforward and beneficial to use the synthesized images directly without fine-tuning or training a new network. In this study, we perform downstream experiments using a pre-trained segmentation model to further evaluate different image translation models.

The main contributions of our work are:

- We developed a novel generative adversarial network for medical image translation based on conditional neural fields. In the proposed model, we build neural fields on the coordinates across the image to fit the target image. To improve the performance and the stability of the model, we introduce shift modulation, which conditions the neural fields on the output of a hypernetwork.

- We evaluated the proposed model by synthesizing various MRI sequences on two paired multi-sequence brain MRI datasets. The results show that the proposed model outperforms state-of-the-art methods both visually and quantitatively. We additionally performed spectral analysis, which indicates that our method is not affected by spectral biases in the way that traditional CNN-based generative models are.

- We compare different medical image translation models in downstream tasks by testing a segmentation model with the synthesized images. Our experiments indicate that by applying image translation, we can improve segmentation performance for incomplete MRI acquisition and our synthesized images outperform the state-of-the-art methods.

2. Related work

Missing MRI sequences Several studies have dealt with the missing MRI sequences problem in medical image analysis (Azad et al., 2022a). One early idea was to translate all available sequences into a shared latent space for downstream analysis. Following this idea, Havaei et al. (2016) developed the Hetero-Modal Image Segmentation (HeMIS) method, where sequence-specific convolutional layers are applied to each image sequence for establishing a common representation which enables robust segmentation when some images are missing. Later, Hu et al. (2020) and Azad et al. (2022b) introduced knowledge distillation to improve the segmentation performance in the same situation. In such a model, a network using all modalities as input (teacher network) and another network using a subset of them (student network) are optimized synchronously. During training, information from all modalities is distilled from the teacher to the student network to improve the performance of the student network. Recently, Liu et al. (2023b) developed a transformer-based model for Alzheimer's classification that can handle missing data scenarios. All these models managed to build a robust model for the situation that only a part of the modalities are available. However, since the missing MRIs are not explicitly constructed, it is still difficult for medical doctors to interpret and make decisions with these methods in clinical practice.

Image-to-image translation Image-to-image translation, on the contrary, focuses on synthesizing missing images from existing ones based on prior knowledge learned from the dataset. By predicting the missing images, clinicians can offer comprehensive diagnoses and also find an explanation of the results in downstream analysis. Recent progress in generative modeling, such as generative adversarial networks (GANs), variational auto-encoders (VAE), and diffusion models, has shown extraordinary capabilities in image generation (Isola et al., 2017; Kawar et al., 2022). In the domain of medical image translation, Dar et al. (2019) proposed pGAN based on a conditional GAN combined with a pixel-wise loss and a perceptual loss. Sharma and Hamarneh (2019) proposed a multi-modal GAN (MM-GAN) that extends the idea by using multi-modality imputation for arbitrary input and output modalities. Recently, Yurt et al. (2021) proposed mustGAN that enhanced the synthesis performance by aggregating multiple translation streams. Inspired by the recent progress of the transformer model, Dalmaz et al. (2022) proposed ResViT based on a hybrid architecture that consists of convolutional operators and transformer blocks. Although promising, most studies focus on the image quality of the output images, and only a few have extended their work to the use of synthesized images in downstream tasks (Iglesias et al., 2013; Van Tulder and de Bruijne, 2015; Amirrajab et al., 2023).

Neural fields Neural fields, also known as implicit neural representations (INRs) or coordinate-based networks, are increasingly popular in computer vision and medical image analysis (Xie et al., 2022; Molaei et al., 2023). The core idea behind neural fields is that neural networks are not used to learn an operator between signals, as in CNNs or vision transformers, but to *represent* a complex signal on a continuous spatial or spatiotemporal domain. Neural fields can be used to solve a wide range of problems, including 3D scene reconstruction and generative shape modeling. Park et al. (2019) proposed DeepSDF which learns a continuous signed distance function to represent 3D surfaces. One distinguished benefit of using neural fields is the capability to handle data with variable resolution because

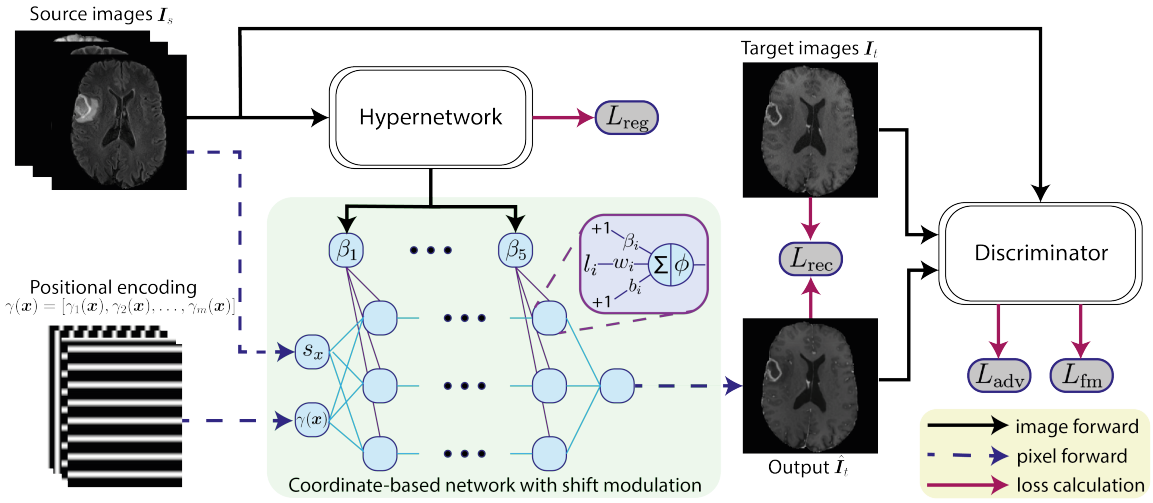


Figure 1: The overall architecture of CoNeS. The generator in the proposed models consists of a hypernetwork and a coordinate-based network. We condition the coordinate-based network on a varying latent code, which is generated by the hypernetwork, across coordinates via shift modulation. The conditional discriminator, which takes both the source images and real/fake images as input, further improves the performance of the generator. The proposed model is optimized using a reconstruction loss L_{rec}, an adversarial loss L_{adv}, a feature matching loss L_{fm} and latent code regularization L_{reg}.

of the absence of up-sampling architectures. Inspired by this, Chen et al. (2021b) proposed a Local Implicit Image Function (LIIF) for image super-resolution, which also shows the potential of handling image generation. Similarly in the field of medical imaging, McGinnis et al. (2023) performed multi-contrast MRI super-resolution via neural fields without any high-resolution training data. Wolterink et al. (2022) proposed to use INRs to represent a transformation function for deformable image registration. Amiranashvili et al. (2022) proposed to reconstruct anatomical shapes from sparse measurements via neural fields. Recently, Shaham et al. (2021) developed Spatially-Adaptive Pixelwise Networks (ASAP-Net), which is most relevant to our work, to speed up image-to-image translation by locally conditioned MLPs. Different from prior work, the neural fields in CoNeS are conditioned on a latent code varying across the coordinates through shift modulation, inspired by Dupont et al. (2022). Specifically, CoNeS consists of a global MLP shared by the whole image and a varying latent code, which determines pixel-wise affine transformations to modulate the neural fields.

3. Methods

3.1 Model overview

To formulate the problem, let $\boldsymbol{I}_t = \{I_t^i\}_{i=0}^{N_t}$ be the set of missing MRI sequences and $\boldsymbol{I}_s = \{I_s^i\}_{i=0}^{N_s}$ be the set of available MRI sequences, where N_t is the number of target sequences and N_s the number of source sequences. We assume that all images from an instance are

co-registered so that there is no extra deformation between the images. As a result, our problem is identical to learning a mapping function $\Phi : \boldsymbol{I}_s \rightarrow \boldsymbol{I}_t$ using the training dataset, which can be applied to all patients in the test dataset and generate the corresponding missing image $\boldsymbol{I}_t$. Similar to traditional GAN models, the proposed model consists of a generator that performs the mapping and a discriminator that aims to tell the real target image and the synthesized one apart. As introduced in pix2pix (Isola et al., 2017), we apply a conditional discriminator that takes both the source and predicted image as its input. The overall architecture of our approach is shown in Fig. 1. In the following section, we introduce how to use a coordinate-based network to model conditional neural fields for image-to-image translation.

3.2 Coordinate-based network

In a typical neural field algorithm, a quantity defined over spacetime, such as an RGB intensity or a signed distance function, is represented as a neural network that maps coordinates to the quantity. Specifically in our problem, we train an MLP that takes coordinates as input and outputs intensities of the target MRI sequences. Given a normalized d-dimensional coordinate $\boldsymbol{x} \in \mathbb{R}^d$, where each component lies in [-1,1], we use $\boldsymbol{t}_x = \{t^i_{\boldsymbol{x}}\}$ and $\boldsymbol{s}_x = \{s^i_{\boldsymbol{x}}\}$ to denote the intensities at position $\boldsymbol{x}$, where t^i_x refers to the intensity value in I^i_t and s^i_x refers to the intensity value in I^i_s, respectively. Hence, the function Φ can be formulated as a pixel-wise mapping that generates intensities over a d-dimensional space:

$$\boldsymbol{t}_{\boldsymbol{x}} = \Phi(\boldsymbol{x}; \boldsymbol{z}), \tag{1}$$

where $\boldsymbol{z}$ is a latent code that contains instance-specific information. During inference, the target images $\hat{\boldsymbol{I}}_t = \{\hat{I}^i_t\}$ are obtained by intensity prediction via sampling from the entire grid via Φ.

A network directly operating on the Cartesian coordinates tends to fit the low-frequency signals first and, as a result, fails to reconstruct the high-frequency image details (Mildenhall et al., 2021; Rahaman et al., 2019). One popular approach to overcome this problem is to map the Cartesian coordinates to a higher dimensional space via positional encoding $\gamma : \mathbb{R}^d \rightarrow \mathbb{R}^m$. In the proposed model, we use sinusoidal functions to perform positional encoding as follows (Zhong et al., 2020):

$$\gamma(\boldsymbol{x}) = [\gamma_1(\boldsymbol{x}), \gamma_2(\boldsymbol{x}), \ldots, \gamma_m(\boldsymbol{x})], \tag{2}$$

$$\gamma_{2i}(\boldsymbol{x}) = \sin(2^{i-1}\pi\boldsymbol{x}), \tag{3}$$

$$\gamma_{2i+1}(\boldsymbol{x}) = \cos(2^{i-1}\pi\boldsymbol{x}), \tag{4}$$

where m is a frequency parameter. Positional encoding can also be seen as Fourier feature mapping of the Cartesian coordinates. By using positional encoding as the input of the MLP, we enable the network to fit the neural field containing high-frequency variation.

3.3 Conditional neural fields

To let the neural field adapt to different input images, we condition it on a set of latent codes $\boldsymbol{z}$, which contain instance-specific information. In the proposed model, we introduce

a hypernetwork H that generates the latent code from the source images: $\boldsymbol{z} = H(\boldsymbol{I}_s)$. By extracting $\boldsymbol{z}$, we can then vary and adapt the neural fields to different instances. Below, we explain how we obtain the latent code $\boldsymbol{z}$ and how the proposed method parameterizes the neural fields with the conditioning via $\boldsymbol{z}$.

3.3.1 Hypernetwork

A hypernetwork refers to an extra neural network that generates parameters for the main network (Ha et al., 2017). The main network behaves like a typical neural network, while the hypernetwork encodes information from the inputs and transfers the information to the main network via the generated parameters. For clarity, we use $\boldsymbol{z}_i = [\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i]$ to denote the latent code used by the i-th layer of the MLP, where $\boldsymbol{\alpha}_i$ are the weights and $\boldsymbol{\beta}_i$ are the biases, both generated by H. Hence, for each layer of the MLP, we have:

$$\boldsymbol{l}_{i+1} = \phi(\boldsymbol{\alpha}_i \boldsymbol{l}_i + \boldsymbol{\beta}_i), \tag{5}$$

where $\boldsymbol{l}_i$ is the input feature of the i-th layer, and ϕ is the activation function. Both $\boldsymbol{\beta}_i$ and $\boldsymbol{l}_i$ are column vectors of size $n_{i+1} \times 1$ and $\boldsymbol{\alpha}_i$ is a matrix of size $n_{i+1} \times n_i$, where n_i is the number of neurons of the i-th layer. Inspired by ASAP-Net (Shaham et al., 2021), we vary the neural field of each pixel by varying the latent code across the coordinates, which can be denoted as $\boldsymbol{z}_i(\boldsymbol{x}) = [\boldsymbol{\alpha}_i(\boldsymbol{x}), \boldsymbol{\beta}_i(\boldsymbol{x})]$, to improve the representation capability. We use $H_{\boldsymbol{x}}$ to represent the latent code mapping for each pixel, and thus, Φ can be denoted as:

$$t_{\boldsymbol{x}} = \Phi(\gamma(\boldsymbol{x}); \boldsymbol{z}(\boldsymbol{x})) = \Phi(\gamma(\boldsymbol{x}); H_{\boldsymbol{x}}(\boldsymbol{I}_s)), \tag{6}$$

and each layer of the MLP can be denoted as:

$$\boldsymbol{l}_{i+1}(\boldsymbol{x}) = \phi(\boldsymbol{\alpha}_i(\boldsymbol{x}) \boldsymbol{l}_i(\boldsymbol{x}) + \boldsymbol{\beta}_i(\boldsymbol{x})), \tag{7}$$

where $\boldsymbol{l}_i(\boldsymbol{x})$ refers to the i-th input feature at position $\boldsymbol{x}$. Different from ASAP-Net, we adapt the bottom-up pathway from a feature pyramid network (Lin et al., 2017) as the hypernetwork H, which outputs the latent code $\boldsymbol{z}(\boldsymbol{x})$ for each pixel with a feasible memory cost (detailed in Section 4.2).

3.3.2 Shift modulation

By conditioning neural fields on varying latent codes across the coordinates, we can improve the representation capability of the network and better model the structure details (Xie et al., 2022; Peng et al., 2020). However, the number of parameters also increases with the spatial expansion of the neural fields, which may induce high computational costs and damage the performance due to over-fitting. This problem may become worse with larger input images. To compact the model while maintaining spatially varying neural fields, we propose to condition the neural network through feature-wise linear modulation (FiLM) (Perez et al., 2018). Instead of generating all parameters of the MLP per pixel, an affine transformation (scale and shift) is applied to every neuron of a single, global MLP. Thus, each layer of the one MLP can be denoted as:

$$\boldsymbol{l}_{i+1}(\boldsymbol{x}) = \overline{\boldsymbol{\alpha}_i}(\boldsymbol{x}) \phi(\boldsymbol{w}_i \boldsymbol{l}_i + \boldsymbol{b}_i) + \boldsymbol{\beta}_i(\boldsymbol{x}), \tag{8}$$

where the weights and biases of the MLP are now replaced by trainable parameters $\boldsymbol{w}_i$ and $\boldsymbol{b}_i$ that are shared by all coordinates, and $\overline{\boldsymbol{\alpha}_i}(\boldsymbol{x})$ is an $n_{i+1} \times n_{i+1}$ matrix that performs scaling to the neurons of the i-th layer by left matrix multiplication. Thus, we can obtain a modified neural field for each coordinate with fewer parameters. Furthermore, research shows that by using shifts only, which is so-called shift modulation, we can achieve comparable results with half of the parameters (Dupont et al., 2022). In this case, $\overline{\boldsymbol{\alpha}_i}$ is an identity matrix and all latent codes are used as shift parameters: $\boldsymbol{z}_i(\boldsymbol{x}) = \boldsymbol{\beta}_i(\boldsymbol{x})$. In practice, we split the biases of the MLP into two parts: trainable biases $\boldsymbol{b}_i$ and biases $\boldsymbol{\beta}_i(\boldsymbol{x})$ generated by H:

$$\boldsymbol{l}_{i+1}(\boldsymbol{x}) = \phi(\boldsymbol{w}_i \boldsymbol{l}_i + \boldsymbol{b}_i + \boldsymbol{\beta}_i(\boldsymbol{x})). \tag{9}$$

The hypernetwork H is optimized together with the MLP during training. In the experimental section, we will show by using shift modulation our model can achieve better performance at reduced complexity.

3.3.3 Intensity concatenation

In addition to shift modulation, we also condition the neural fields on the source images directly. Different from the latent codes, the pixel intensities provide first-hand uncoded local information. We concatenate the image intensities from all the source images as an additional input of the MLP. The mapping function of the neural fields therefore becomes:

$$t_{\boldsymbol{x}} = \Phi(\gamma(\boldsymbol{x}), \boldsymbol{s}_{\boldsymbol{x}}; H_{\boldsymbol{x}}(\boldsymbol{I}_s)). \tag{10}$$

3.4 Loss function

Like the standard GAN model, the discriminator and the generator in the proposed model are optimized alternately. In each iteration, we train the discriminator using the hinge loss (Lim and Ye, 2017):

$$L_D = \mathbb{E}_{\boldsymbol{I}_t, \boldsymbol{I}_s}[\max(0, 1 - D(\boldsymbol{I}_t, \boldsymbol{I}_s))] + \mathbb{E}_{\boldsymbol{I}_s}[\max(0, 1 + D(\hat{\boldsymbol{I}}_t, \boldsymbol{I}_s))], \tag{11}$$

where D is the discriminator and $\mathbb{E}$ is the expectation over the whole dataset.

The generator is trained by a loss function L that contains a reconstruction loss, an adversarial loss, a feature matching loss, and latent code regularization.

Reconstruction loss To ensure the synthesized images are as close to the real images as possible, we apply a reconstruction loss that maximizes the similarity between ground truth $\boldsymbol{I}_t$ and output images $\hat{\boldsymbol{I}}_t$, which are obtained by intensity prediction via sampling from the entire grid. We use the ℓ_1 loss function as suggested in (Isola et al., 2017):

$$L_{\text{rec}} = \mathbb{E}_{\boldsymbol{I}_t, \boldsymbol{I}_s}[\|\hat{\boldsymbol{I}}_t - \boldsymbol{I}_t\|_1]. \tag{12}$$

Adversarial loss Adversarial loss is applied to enforce that the generated images are good enough to fool the discriminator. Like the discriminator loss, we use the hinge function, which is defined as:

$$L_{\text{adv}} = -\mathbb{E}_{\boldsymbol{I}_s}[\log D(\hat{\boldsymbol{I}}_t, \boldsymbol{I}_s)]. \tag{13}$$

Feature matching loss To stabilize the training, we apply a feature matching loss introduced by Wang et al. (2018). Specifically, we feed both the real and generated images to the discriminator and extract the intermediate features from each forward pass. The two groups of the intermediate features are matched using the ℓ_1 loss function. Hence, the feature matching loss is defined as:

$$L_{\text{fm}} = \mathbb{E}_{\boldsymbol{I}_t, \boldsymbol{I}_s} \sum_{i=1}^{T} \frac{1}{N_i} [\|D_i(\boldsymbol{I}_t, \boldsymbol{I}_s) - D_i(\hat{\boldsymbol{I}}_t, \boldsymbol{I}_s)\|_1]. \tag{14}$$

where D_i denotes the i-th feature layer of the discriminator and N_i denotes the number of elements in each layer. T is the total number of layers of the discriminator.

Latent code regularization Last, we apply the ℓ_2 norm to $\boldsymbol{z}$ as a latent code regularization to stabilize the training:

$$L_{\text{reg}} = \mathbb{E}_{\boldsymbol{I}_s} \|H_{\boldsymbol{x}}(\boldsymbol{I}_s)\|_2. \tag{15}$$

Overall loss The overall loss function then becomes

$$L = \lambda_{\text{rec}} L_{\text{rec}} + \lambda_{\text{adv}} L_{\text{adv}} + \lambda_{\text{fm}} L_{\text{fm}} + \lambda_{\text{reg}} L_{\text{reg}}, \tag{16}$$

where λ_{rec}, λ_{adv}, λ_{fm}, and λ_{reg} are the weights of the loss functions.

4. Experiments and results

4.1 Dataset

To evaluate the proposed translation model, we conducted experiments on two datasets: (1) BraTS 2018 (Menze et al., 2014) and (2) an in-house Vestibular Schwannoma MRI (VS) dataset (Neve et al., 2022).

BraTS 2018 BraTS 2018 is a multi-sequence brain MRI dataset for tumor segmentation. The dataset consists of 285 patients for training and 66 patients for validation. Each patient has four co-registered MRI sequences: T1 (1-6 mm slice thickness), T1ce (1-6 mm slice thickness), T2 (2-6 mm slice thickness) and FLAIR (2-6 mm slice thickness). The tumor mask that includes the non-enhanced tumor, the enhanced tumor, and the edema was delineated by experts from multiple centers as a segmentation ground truth. All the scans in BraTS 2018 are resampled to 1 mm isotropic resolution.

Vestibular schwannoma MRI dataset The VS dataset is MRI scans of patients with vestibular schwannoma, which is a benign tumor arising from the neurilemma of the vestibular nerve. 191 patients were collected from 37 different hospitals using 12 different MRI scanners. In our study, 147 patients are selected for training and the remaining 44 patients are the validation set. All patients have a gadolinium-enhanced T1-weighted MRI (shortened to T1ce) and a high-resolution T2 (shortened to T2). The spatial resolution of the T1ce ranges from $0.27 \times 0.27 \times 0.9$ to $1.0 \times 1.0 \times 5.0$ mm, and the spatial resolution of T2 scans ranges from $0.23 \times 0.23 \times 0.5$ to $0.7 \times 0.7 \times 1.8$ mm. The intra- and extrameatal tumor was manually delineated by four radiologists. Different from BraTS 2018, there are only two sequences available in the VS dataset and the high resolution of the T2 offers better visibility of the lesion but may also degrade the image quality, which makes the image translation more challenging on this dataset.

4.2 Experimental setup

Network architecture The hypernetwork H of the proposed model is adapted from feature pyramid network (Lin et al., 2017). H consists of four convolutional modules containing [2, 4, 23, 3] residual blocks in sequence. Each convolutional module is followed by a 3×3 convolutional smoothing layer and up-sampling layer, computing a feature map that is downscaled by a factor of 2. We take the output of the last module, which has the same size as the input resolution, as the latent code $\boldsymbol{z}$. Adapted from (Shaham et al., 2021), the MLP in the proposed model contains five 64-channel layers. The Leaky ReLU function with a negative slope of 0.2 is applied as the activation function after all intermediate layers. The output layer is followed by a Tanh function which can constrain the range of the intensities to [-1, 1]. The discriminator is a 2D convolutional neural network that takes both the source image and prediction as input, both as a whole. The network contains five 4×4 convolutional blocks followed by a Leaky ReLU function except for the last layer. The strides and number of filters of the blocks are [2, 2, 2, 1, 1] and [64, 128, 256, 512, 1] respectively. Like pix2pix (Isola et al., 2017), the discriminator down-samples the inputs by 8 and penalizes structures at the scale of patches.

Pre-processing Registration was applied to the VS dataset before training. We considered the T1ce as the fixed image and performed rigid registration with the T2, for which we used Elastix software (Klein et al., 2009). All images from both datasets were then normalized to the range of [-1,1] and the background was cropped based on the bounding box of the foreground before training to reduce the image size. Both sequences in the VS dataset were resampled to 0.29×0.29 mm in-plane resolution, which is the median value of the T1ce domain. During training, random cropping was conducted on the images, with a cropping size of 160×128 for BraTS 2018 and a cropping size of 320×320 for the VS dataset, respectively.

Implementation details All experiments were conducted using Python 3.10 and PyTorch 1.12.1 on a mixed computation server equipped with Nvidia Quadro RTX 6000 and Nvidia Tesla V100 GPUs. The models were trained by the Adam optimizer using the Two Time-scale Update Rule (TTUR) training scheme, in which the generator and discriminator have a different initial learning rate (Heusel et al., 2017). We found that an initial learning rate of 1×10^{-4} for the generator and an initial learning rate of 4×10^{-4} for the discriminator worked best for our experiments. The learning rates were further decayed using a linear learning rate scheduler. Adapted from the choices of hyperparameters in Shaham et al. (2021), we set the frequency parameter $m = 6$ for positional encoding. We set $\lambda_{\text{adv}} = 1.0$ and $\lambda_{\text{rec}} = 100.0$, which gives us the best balance between sharp results and fewer artifacts as suggested in Isola et al. (2017). Both L_{fm} and L_{reg} help to stabilize the training, while large λ_{reg} and λ_{fm} may lead to poor reconstruction performance. We set $\lambda_{\text{fm}} = \lambda_{\text{reg}} = 10.0$, which ensures a stable training while maintaining reconstruction performance (Wang et al., 2018; Shaham et al., 2021). Lastly, we focus on 2D image translation in this paper and hence use 2D coordinates ($d = 2$).

Benchmark overview We compared our model with the following state-of-the-art methods: (1) pix2pix: pix2pix is a GAN-based image translation model, which consists of a UNet-based generator and a patch-based discriminator (Isola et al., 2017); (2) pGAN: pGAN is

Table 1: Quantitative comparison of different image translation models on BraTS 2018. The mean value and standard deviation of PSNR and SSIM are reported. The highest values per column are indicated in boldface; The † after each metric of the benchmarks indicates a significant difference ($p < .05$) compared to the proposed method.

model	**T1ce translation**		**T1 translation**		**T2 translation**		**FLAIR translation**	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
pix2pix	30.1 ±2.65†	0.941 ±0.014†	27.0 ±3.69	0.945 ±0.013†	28.0 ±2.63†	0.926 ±0.062†	27.6 ±3.03†	0.910 ±0.097†
pGAN	30.7 ±3.18†	0.943 ±0.015†	**27.5 ±3.72**	0.945 ±0.015†	29.2 ±2.80†	0.943 ±0.020†	28.5 ±3.26†	0.916 ±0.095†
ResViT	29.2 ±2.37†	0.935 ±0.014†	25.0 ±2.60†	0.918 ±0.014†	26.6 ±2.30†	0.923 ±0.020†	24.7 ±2.08†	0.876 ±0.092†
ASAP-Net	30.8 ±2.97†	0.948 ±0.017†	27.3 ±3.79	0.948 ±0.015†	28.6 ±2.74†	0.940 ±0.019†	28.4 ±3.10†	0.916 ±0.098†
CoNeS (proposed)	**31.2 ±3.11**	**0.951 ±0.017**	27.3 ±4.03	**0.953 ±0.014**	**29.6 ±3.03**	**0.950 ±0.021**	**29.1 ±2.99**	**0.926 ±0.097**

Table 2: Quantitative comparison of different image translation models after cropping on BraTS 2018. The mean value and standard deviation of PSNR and SSIM are reported. The highest values per column are indicated in boldface; The † after each metric of the benchmarks indicates a significant difference ($p < .05$) compared to the proposed method.

model	**T1ce translation**		**T1 translation**		**T2 translation**		**FLAIR translation**	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
pix2pix	19.9 ±3.30†	0.610 ±0.092†	15.7 ±4.41	0.636 ±0.089†	19.9 ±2.85†	0.658 ±0.082†	18.1 ±3.31†	0.585 ±0.081†
pGAN	20.6 ±3.73†	0.646 ±0.096†	**16.1 ±4.61**	0.663 ±0.099	20.8 ±3.23†	0.721 ±0.093†	18.9 ±3.70†	0.636 ±0.086†
ResViT	20.2 ±3.56†	0.612 ±0.098†	15.1 ±4.32†	0.599 ±0.090†	19.5 ±3.50†	0.672 ±0.093†	17.0 ±3.26†	0.545 ±0.111†
ASAP-Net	20.4 ±3.67†	0.634 ±0.115†	15.7 ±4.48	0.626 ±0.103†	20.3 ±3.05†	0.669 ±0.089†	18.5 ±3.60†	0.593 ±0.086†
CoNeS (proposed)	**20.9 ±3.66**	**0.667 ±0.099**	15.8 ±4.44	**0.666** ±0.094	**21.5 ±3.35**	**0.739 ±0.095**	**19.6 ±3.49**	**0.663 ±0.084**

a GAN-based image translation model using ResNet which follows an encoder-bottleneck-decoder architecture as backbone (Dar et al., 2019). Perceptual loss is introduced to improve the results; (3) ResViT: ResViT is an image translation model that combines pGAN with a transformer-based information bottleneck; (4) ASAP-Net: ASAP-Net is a neural field-based image translation model (Shaham et al., 2021). Different from the proposed model, ASAP-Net parameterizes patch-wise neural fields, which are conditioned through a UNet-shape hypernetwork without a shared MLP. For all implementations, we used the official GitHub repositories provided by the authors. We used the ℓ_1 loss as a reconstruction loss for all the benchmark methods. We used the least square loss function (Mao et al., 2017) as an adversarial loss for pix2pix, pGAN, and ResViT. Like the proposed method, we used the hinge loss function (Lim and Ye, 2017) as an adversarial loss for ASAP-net. All the

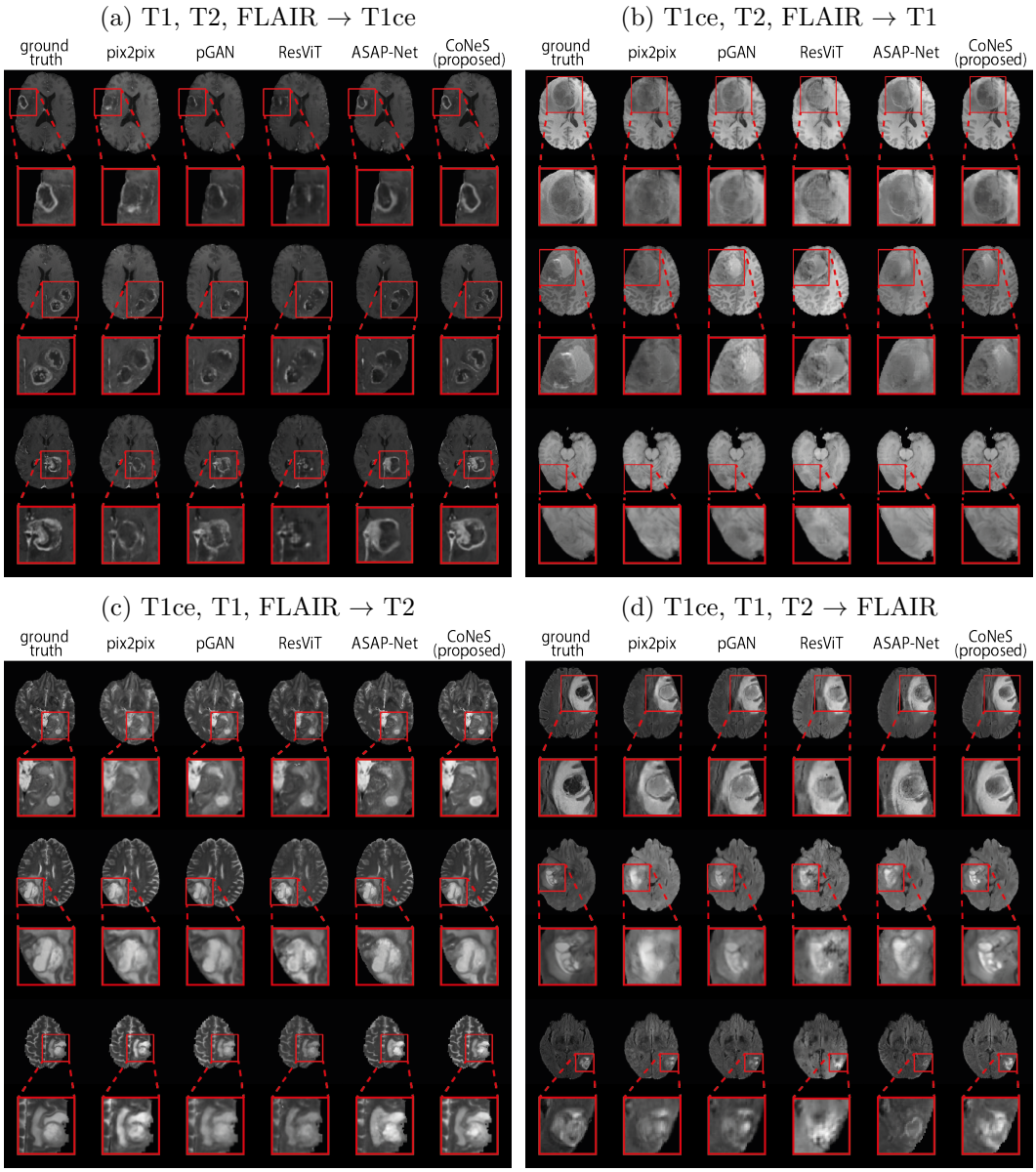


Figure 2: Comparison results of different image translation models on BraTS 2018: (a) T1, T2, FLAIR → T1ce; (b) T1ce, T2, FLAIR → T1; (c) T1ce, T1, FLAIR → T2; (d) T1ce, T1, T2 → FLAIR. For each translation experiment, three examples are selected for display. Each column shows the ground truth and translation results of the different models. Zoomed-in results indicated in red rectangles are shown below the whole images.

benchmark methods were trained using hyperparameters that were optimized by the original authors on the same dataset (BraTS). We trained ResViT with the pre-trained network as suggested in (Dalmaz et al., 2022), while all other models were trained from scratch.

Table 3: Quantitative comparison of different image translation models on VS dataset. The mean value and standard deviation of PSNR and SSIM are reported. The highest values per column are indicated in boldface; All metrics of the benchmarks in this table show significant differences ($p < .05$) compared to the proposed method.

model	**T1ce translation**		**T2 translation**	
	PSNR	SSIM	PSNR	SSIM
pix2pix	21.1 ± 1.39	0.602 ± 0.068	21.4 ± 1.78	0.506 ± 0.121
pGAN	21.6 ± 1.55	0.635 ± 0.077	22.2 ± 2.04	$\mathbf{0.575 \pm 0.131}$
ResViT	21.0 ± 1.58	0.575 ± 0.090	21.5 ± 1.80	0.489 ± 0.110
ASAP-Net	20.4 ± 1.24	0.552 ± 0.061	20.9 ± 1.97	0.500 ± 0.117
CoNeS (proposed)	$\mathbf{21.9 \pm 1.69}$	$\mathbf{0.638 \pm 0.077}$	$\mathbf{22.6 \pm 2.03}$	0.560 ± 0.126

Table 4: Quantitative comparison of different image translation models after cropping on VS dataset. The mean value and standard deviation of PSNR and SSIM are reported. The highest values per column are indicated in boldface; The † after each metric of the benchmarks indicates a significant difference ($p < .05$) compared to the proposed method.

model	**T1ce translation**		**T2 translation**	
	PSNR	SSIM	PSNR	SSIM
pix2pix	14.8 ± 3.28	$0.415 \pm 0.122^{\dagger}$	$16.6 \pm 1.72^{\dagger}$	$0.321 \pm 0.084^{\dagger}$
pGAN	$14.2 \pm 3.31^{\dagger}$	$0.417 \pm 0.133^{\dagger}$	$16.8 \pm 1.98^{\dagger}$	0.372 ± 0.134
ResViT	14.5 ± 2.75	$0.400 \pm 0.106^{\dagger}$	$16.7 \pm 1.66^{\dagger}$	$0.342 \pm 0.099^{\dagger}$
ASAP-Net	$13.0 \pm 2.93^{\dagger}$	$0.340 \pm 0.132^{\dagger}$	$15.3 \pm 1.64^{\dagger}$	$0.300 \pm 0.101^{\dagger}$
CoNeS (proposed)	$\mathbf{15.0 \pm 3.17}$	$\mathbf{0.451 \pm 0.118}$	$\mathbf{17.3 \pm 1.58}$	$\mathbf{0.379 \pm 0.101}$

4.3 Multi-sequence MRI translation

We first examined the quality of the images generated from the proposed model. Theoretically, CoNeS can be applied to any number of missing or present sequences by adapting input and output channels to N_s and N_t. For simplicity, we assumed one sequence was missing for all the patients during inference ($N_t = 1$), and thus, we trained models that generate one MRI sequence from the other sequences in the dataset for evaluation. Specifically, four image translation experiments were performed on BraTS 2018: (1) T1, T2, FLAIR $\rightarrow$ T1ce (shortened to T1ce translation); (2) T1ce, T2, FLAIR $\rightarrow$ T1 (shortened to T1 translation); (3) T1ce, T1, FLAIR $\rightarrow$ T2 (shortened to T2 translation); and (4) T1ce, T1, T2 $\rightarrow$ FLAIR (shortened to FLAIR translation). We used two different metrics for quantitative analysis in our study: peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM). Both the synthesized images and real images were normalized to [0,1] before evaluation. Wilcoxon signed-rank test between each benchmark and the proposed model was performed on all image translation experiments.

The quantitative results are listed in Table 1. As shown in the table, the proposed model performs significantly better ($p < .05$) than other state-of-the-art methods in most

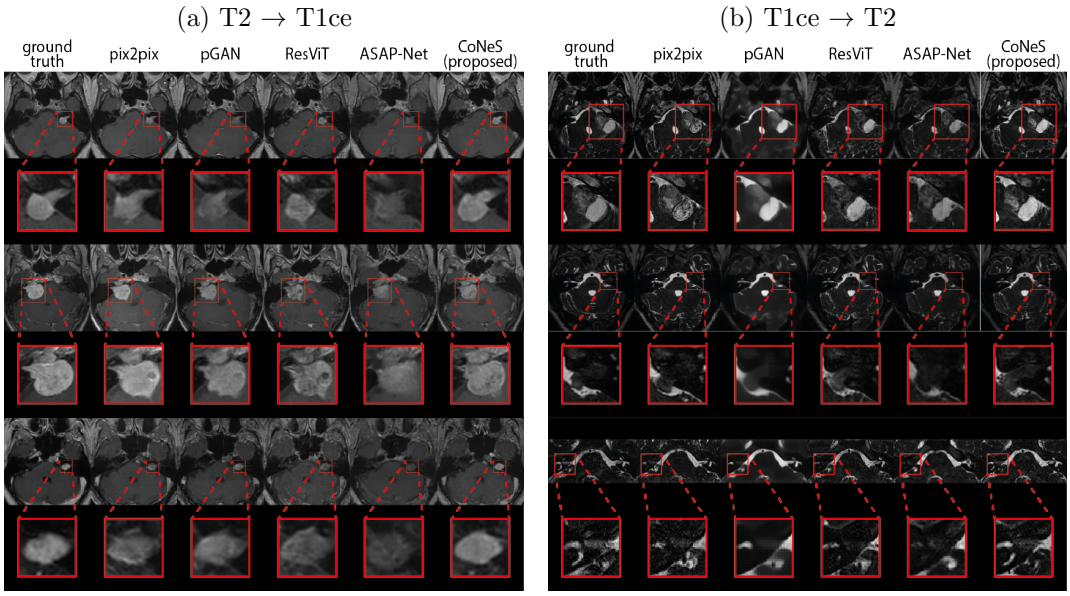


Figure 3: Comparison results of different image translation models on the VS dataset: (a) T2 → T1ce; (b) T1ce → T2. For each translation experiment, three examples are selected for display. Each column shows the ground truth and translation results of the different models. Zoomed-in results indicated in red rectangles are shown below the images.

metrics, except that pGAN obtains higher PSNR on T1 translation. Using T1ce translation on BraTS 2018 as an example, the PSNR and SSIM of the proposed model on BraTS 2018 are 31.2 dB and 0.951, which outperforms pix2pix by 1.1 dB PSNR and 1.0% SSIM, pGAN by 0.5 dB PSNR and 0.8% SSIM, ResViT by 2.0 dB PSNR and 1.6% SSIM, and ASAP-Net by 0.4 dB PSNR and 0.3% SSIM. Translation examples are shown in Fig. 2 in which we can see that the proposed model can recover more detailed structures, such as the contrast-enhanced tumor in T1ce, which is clinically highly relevant.

Both the PSNR and SSIM show global similarity, while the quality of the region around the tumor is more clinically interesting. To further evaluate the proposed model, we cropped the images using the bounding box of the tumor region and then evaluated the similarity of these sub-images using the aforementioned metrics. The bounding box was generated from the segmentation results of nnUNet (Isensee et al., 2021) for the reason that the segmentation ground truths of BraTS 2018 validation set are not available. The results are listed in Table 2. As we can see, the proposed model also performs significantly better ($p < .05$) in most tasks within this sub-region, which is consistent with our observation from zoomed-in results in Fig. 2. We observed that the performance of the proposed model decreased after cropping due to the lack of background. Again using T1ce translation as an example, the PSNR and SSIM of the proposed model are 20.9 dB and 0.667, which outperforms pix2pix by 1.0 dB PSNR and 5.7% SSIM, pGAN by 0.3 dB and 2.1% SSIM, ResViT by 0.7 dB and 5.5% SSIM, and ASAP-Net by 0.5 dB and 3.3% SSIM.

Next, we performed two image translation experiments on the VS dataset: (1) T2 → T1ce (shortened to T1ce translation) and (2) T1ce → T2 (shortened to T2 translation). We

again evaluated the entire image as well as the cropped region around the tumor, similar to BraTS 2018. Both quantitative results are listed in Table 3 and Table 4. All models struggle with the VS dataset and show decreased performance compared to BraTS 2018, and CoNeS still performs significantly better ($p < .05$) in most of the metrics. Taking T1ce translation as an example, CoNeS obtains a PSNR of 21.9 dB and a SSIM score of 0.638, which outperforms pix2pix by 0.8 dB PSNR and 3.6% SSIM, pGAN by 0.3 dB PSNR and 0.3% SSIM, ResViT by 0.9 dB PSNR and 6.3% SSIM, and ASAP-Net by 1.5 dB PSNR and 8.6% SSIM. Qualitatively, we can observe improved synthesized images using the proposed model as shown in Fig. 3. It is worth pointing out that although pGAN obtained better SSIM scores (0.575) in T2 translation, the visualization suggests that our results contain more informative details, while pGAN's results are blurry.

4.4 Spectral analysis

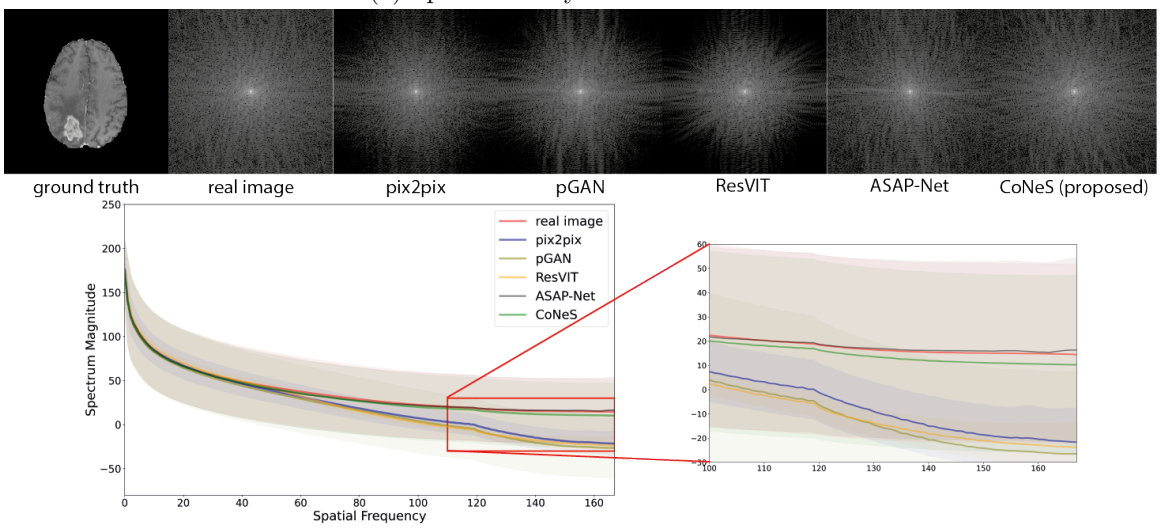
Research has shown that CNN-based generative models with up-sampling layers usually struggle with reproducing the spectral distribution correctly (Durall et al., 2020; Anokhin et al., 2021). On the contrary, coordinate-based networks like CoNeS build a direct pixel-to-pixel mapping without any up-sampling layer. In this section, we further evaluated the synthesized images in the frequency domain to demonstrate the improvement we obtained by performing spectral analysis on the T1ce translation model of both datasets. Specifically, we applied a 2D Fourier transform to all synthesized results as well as the real images, and then calculated a 1D representation of the 2D spectrum using Azimuthal integration (Durall et al., 2020). Azimuthal integration is defined as an integration over the radial frequencies:

$$\mathrm{AI}(\omega_k) = \int_0^{2\pi} \|\mathcal{F}(\omega_k \cos\theta, \omega_k \sin\theta)\omega_k\|_2 d\theta, \tag{17}$$

for $k = 0, \ldots, M/2 - 1$, and where $\mathcal{F}(m, n)$ is the Fourier Transform of a 2D image, θ is the radian, ω_k is the spatial frequency and M is the length of a square image.

A log transformation was performed to the 2D spectrum for better visualization, and we calculated the average 1D representation over the dataset to avoid biased sampling. As shown in Fig. 4, both ASAP-Net and CoNeS, which are coordinate-based networks, can reproduce the spectrum over the entire frequency range on BraTS 2018. Specifically, all the spectral curves are very close in the low-frequency range (spatial frequency < 50) which enables the generative models to reconstruct the general structure of the images. However, the spectral curves of GAN-based models dramatically drop in the high-frequency range (spatial frequency > 75), while the curves of ASAP-Net and CoNeS remain close to the real distribution. This shows that neural fields are able to overcome the spectral biases issue of convolutional neural networks. On the VS dataset, all the models yield higher spectrum magnitudes in the high-frequency range compared to the real images, which suggests that these translation models might add high-frequency noise to the synthesized images. Consistent with the similarity measurement results, ASAP-Net is not robust enough to reproduce the spectrum on the VS dataset and may induce more artifacts. On the contrary, CoNeS still outputs images whose spectrum is closest to the real images among all the translation models. The results indicate that by using neural fields conditioned via shift modulation, CoNeS is able to keep the representation capability and reproduce the spectrum distribution.

(a) Spectral analysis on BraTS 2018



(b) Spectral analysis on the VS dataset

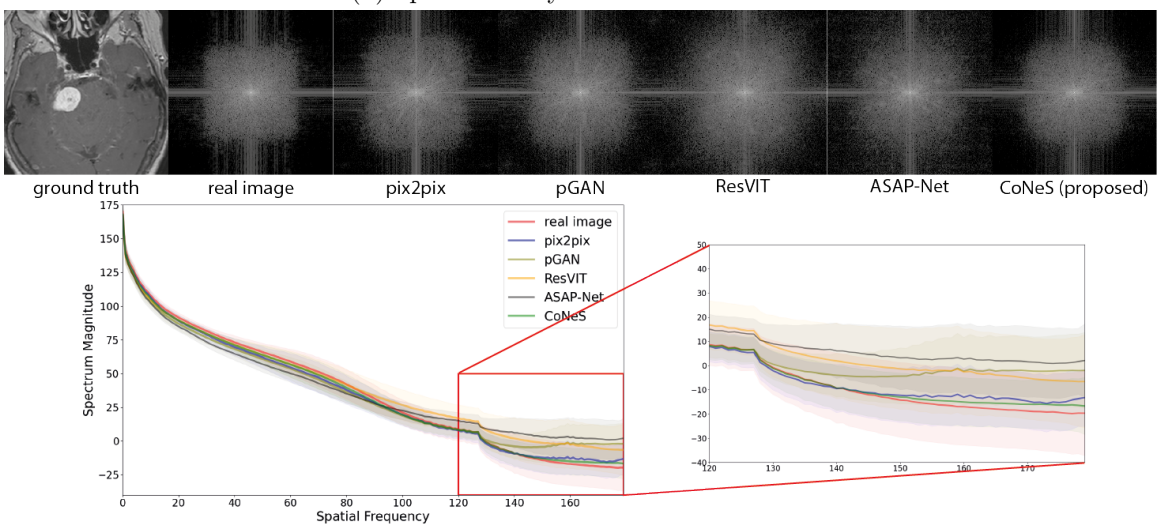


Figure 4: Spectral analysis of different image translation models. (a) and (b) show the analysis results on BraTS 2018 and the VS dataset, respectively. For each analysis, the Fourier transform of different synthesized images and the real image are shown in the top row. The bottom row shows the spectral distribution, in which the high-frequency range is zoomed in by the red rectangle.

4.5 Synthesized images for tumor segmentation

To further examine the impact of synthesized images in downstream analysis, we performed tumor segmentation using the synthesized images at inference. To do this, we first adopted the architecture from nnUnet (Isensee et al., 2021) and trained a segmentation network that uses all the sequences in the dataset as input. Note that all the images were normalized to

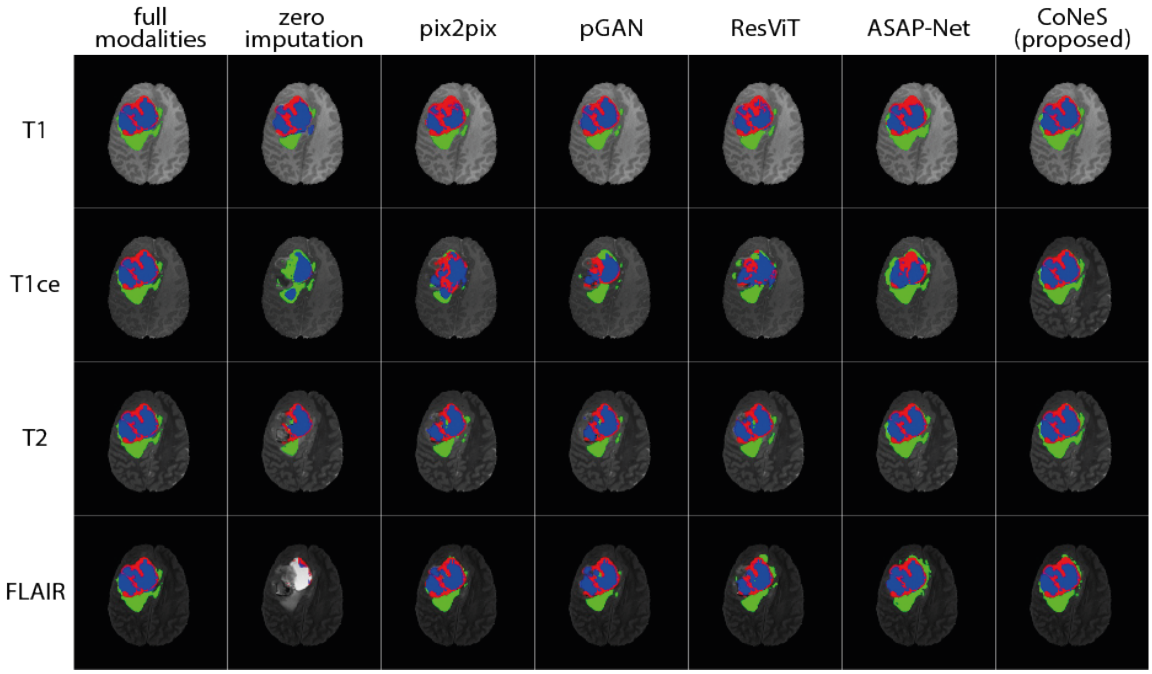
Table 5: Results of using different images for segmentation inference on BraTS 2018. The real sequences used are indicated by ✓, the missing ones by ✗, and the ones replaced by synthesized images by ∘. The mean of Dice scores and 95% HD (mm) of the enhanced tumor (ET), the whole tumor (WT), and the tumor core (TC) are reported. The highest values per column are indicated in boldface; The † after each metric of the benchmarks indicates a significant difference ($p < .05$) compared to inference using synthesized images from CoNeS.

input sequences				method	**Dice**			**95% HD (mm)**		
T1ce	T1	T2	FLAIR		ET	WT	TC	ET	WT	TC
✓	✓	✓	✓	\	0.770	0.888	0.822	4.49	6.27	8.92
✗	✓	✓	✓	zero imputation	0.068†	0.845†	0.362†	27.9†	8.80†	18.1†
∘	✓	✓	✓	pix2pix	0.191†	0.850†	0.537†	15.0†	8.10†	15.0†
∘	✓	✓	✓	pGAN	0.317†	0.858†	0.598†	14.9†	8.01†	14.0†
∘	✓	✓	✓	ResViT	0.223†	0.858†	0.555†	15.0†	7.87†	14.1†
∘	✓	✓	✓	ASAP-Net	0.332†	0.866†	0.597†	13.3†	**6.95**	13.2†
∘	✓	✓	✓	CoNeS	**0.386**	**0.870**	**0.662**	**13.1**	7.23	**13.0**
✓	✗	✓	✓	zero imputation	0.717†	0.865†	0.753†	6.53†	7.86†	11.3†
✓	∘	✓	✓	pix2pix	0.747	0.869†	0.780†	5.07†	7.70†	9.84†
✓	∘	✓	✓	pGAN	0.747†	0.868†	0.779†	5.60†	7.61†	10.2†
✓	∘	✓	✓	ResViT	0.751†	0.869†	0.784†	**4.62**	7.39†	9.75†
✓	∘	✓	✓	ASAP-Net	0.753†	0.881	0.806	5.43†	**6.73**	9.39
✓	∘	✓	✓	CoNeS	**0.764**	**0.885**	**0.808**	5.30	7.05	**8.94**
✓	✓	✗	✓	zero imputation	0.748†	0.835†	0.752†	5.64†	8.67†	11.6†
✓	✓	∘	✓	pix2pix	0.761	0.862†	0.784†	3.90†	7.53†	9.82†
✓	✓	∘	✓	pGAN	0.767	0.872†	0.797†	3.83†	7.34†	9.27
✓	✓	∘	✓	ResViT	0.759	0.855†	0.788†	4.19†	8.18†	9.74
✓	✓	∘	✓	ASAP-Net	0.764	0.880†	0.817	3.84†	6.50†	9.05
✓	✓	∘	✓	CoNeS	**0.778**	**0.886**	**0.829**	**3.15**	**6.01**	**8.34**
✓	✓	✓	✗	zero imputation	0.679†	0.403†	0.690†	27.8†	30.3†	23.2†
✓	✓	✓	∘	pix2pix	0.760	0.805†	0.771†	**3.74**	9.75†	11.1†
✓	✓	✓	∘	pGAN	0.766	0.833†	0.777†	4.60	8.59†	10.3†
✓	✓	✓	∘	ResViT	0.783	0.768†	0.752†	5.16†	11.7†	11.5†
✓	✓	✓	∘	ASAP-Net	**0.785**	0.823†	0.808†	3.80†	9.36†	**9.36**†
✓	✓	✓	∘	CoNeS	0.768	**0.853**	**0.809**	4.30	**7.56**	9.38

a range of [-1,1] during training to make the input channels consistent with the synthesized images. During inference, we tested the segmentation model with synthesized images and compared the results with the performance of the model when filling the missing channel with zeros, called zero imputation, in our experiments. For simplicity, we again assumed one specific sequence was missing and replaced this sequence while keeping the rest unchanged. Similar to the image translation experiments, we compared the segmentation performance using synthesized images generated from the proposed model to the other images via the Wilcoxon signed-rank test.

The tests were performed for each MRI sequence (T1ce, T1, T2, and FLAIR) on BraTS 2018. The performance was evaluated using three specific categories: 1) enhanced tumor (ET); 2) tumor core (TC, non-enhanced tumor, and edema); and 3) the whole tumor

(a) Segmentation example on BraTS 2018



(b) Segmentation example on the VS dataset

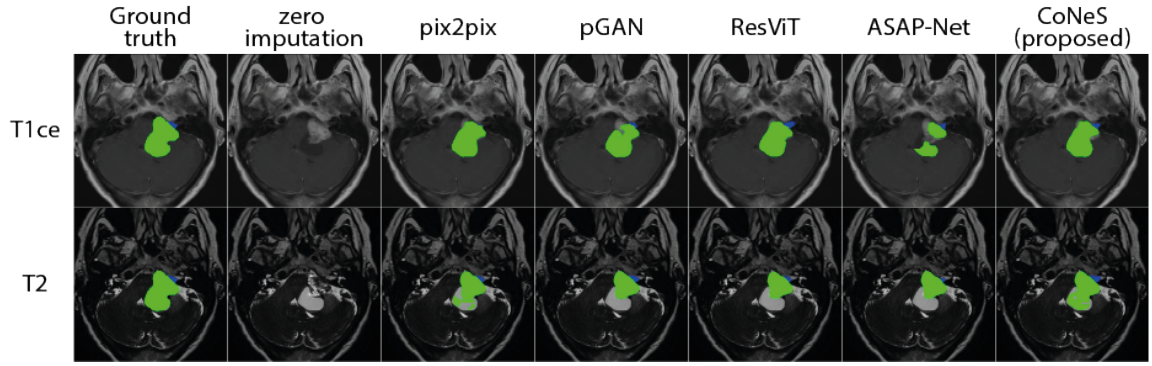


Figure 5: The results of segmentation experiments: (a) A segmentation example on BraTS 2018 and (b) an example on the VS dataset. The rows show the segmentation results with different MRI sequences replaced. The columns show ground truth (for BraTS 2018, segmentation results with full sequences) and segmentation results using different synthesized images.

(WT, enhanced tumor, non-enhanced tumor, and edema). Dice score and 95% Hausdorff distance (95% HD) of all the three categories are reported for quantitative evaluation in Table 5. We can see that the presence of sequences dramatically influences the performance of the segmentation model. For instance, when the T1ce is missing, the Dice score of the enhanced tumor is 0.068 because the enhanced information is only visible in the T1ce. As expected, most of the metrics show that inference with synthesized images performs

worse than inference with full sequences. However, we also noticed that when the real T2 or FLAIR were replaced with synthesized ones, we obtained a lower mean 95% HD. This occurs due to the influence of certain outliers. For example, sometimes the model can identify the enhanced tumor at the wrong position using real images, leading to a large 95% HD, while the other inferences using synthesized images completely miss the tumor. When we removed three outliers, the mean of 95% HD of the enhanced tumor became 2.99 mm, which is better than the others.

The best results among all the inferences using synthesized images (including zero imputation) for each sequence were highlighted in Table 5. The results indicate that using synthesized images for inference can significantly improve the segmentation performance and that the synthesized images of our model yield the best segmentation performance with a significant difference ($p < .05$) among all the translation models. Using the inferences without T1ce as examples, the Dice scores of the proposed model are 0.386, 0.870, and 0.662 in the enhanced tumor, the whole tumor, and the tumor core respectively. In comparison, the proposed model outperforms zero imputation by 31.8%, 2.5%, and 30.0%, pix2pix by 19.5%, 2.0%, and 12.5%, pGAN by 6.9%, 1.2%, and 6.4%, ResViT by 16.3%, 1.2%, and 10.7%, and ASAP-Net by 5.4%, 0.4%, and 6.5%. The 95% HDs of the proposed model are 13.1 mm, 7.23 mm, and 13.0 mm in the enhanced tumor, the whole tumor, and the tumor core respectively. In comparison, the proposed model outperforms zero imputation by 14.8 mm, 1.57 mm, and 5.1 mm, pix2pix by 1.9 mm, 0.87 mm, and 2.0 mm, pGAN by 1.8 mm, 0.78 mm, 1.0 mm, ResViT by 1.9 mm, 0.64 mm, 1.1 mm. Although ASAP-Net obtained a higher 95% HD (6.95 mm) in the whole tumor, we did not observe significant differences between it and the proposed model. Some example segmentation results are presented in Fig. 5. It is worth noting that the synthesized T1 of CoNeS performs better in segmentation than the ones from pGAN, although we got higher PSNR for pGAN in the former experiment.

We also performed the same segmentation experiments on the VS dataset. We evaluated the performance using three specific categories: 1) intrameatal tumor; 2) extrameatal tumor); and 3) the whole tumor (including intra- and extrameatal tumor). Dice score and 95% HD of all three categories are reported in Table 6. Similarly to BraTS 2018, all the synthesized images compensate for the performance loss due to the drop of sequences, and the proposed model performs significantly better ($p < .05$) than the other models. For instance, the synthesized T1ce generated by the proposed model obtained Dice scores of 0.567, 0.714, and 0.749 in the intrameatal tumor, the extrameatal tumor, and the whole tumor respectively. In comparison, the proposed model outperforms zero imputation by 56.6%, 68.4%, and 72.1%, pix2pix by 9.6%, 2.8%, and 3.6%, pGAN by 12.6%, 3.8%, and 8.8%, ResViT by 2.7%, 0.1%, and 3.0%, and ASAP-Net by 25.9%, 22.2%, and 24.1%. The 95% HDs of the proposed model are 2.33 mm, 3.54 mm, and 4.05 mm in the intrameatal tumor, the extrameatal tumor, and the whole tumor respectively. These results outperform zero imputation by 5.71 mm, 25.37 mm, and 30.05, pix2pix by 0.21 mm, 2.13 mm, and 2.45 mm, pGAN by 0.15 mm, 3.03 mm, and 3.82 mm, ASAP-Net by 0.77 mm, 3.72 mm, and 8.35 mm. We observed that ResViT obtained lower 95% HD (3.32 mm) in the extrameatal tumor, however, the proposed model still performs better than ResViT in most of the experiments. Example segmentation results are displayed in Fig. 5.

Table 6: Results of using different images for segmentation inference on the VS dataset. The real sequences used are indicated by ✓, the missing ones by ✗, and the ones replaced by synthesized images by ∘. The mean of Dice scores and 95% HD (mm) of the intrameatal tumor (IT), the extrameatal tumor (ET), and the whole tumor (WT) are reported. The highest values per column are indicated in boldface; The † after each metric of the benchmarks indicates significant differences ($p < .05$) compared to inference using synthesized images from CoNeS.

input sequences		method	**Dice**			**95% HD (mm)**		
T1ce	T2		IT	ET	WT	IT	ET	WT
✓	✓	\	0.761	0.853	0.896	1.34	1.71	1.45
✗	✓	zero imputation	0.001†	0.030†	0.028†	8.04†	28.91†	34.1†
∘	✓	pix2pix	0.471†	0.686†	0.713†	2.54	5.67†	6.50
∘	✓	pGAN	0.441†	0.676†	0.661†	2.48†	6.57†	7.87†
∘	✓	ResViT	0.540†	0.713†	0.719	2.36	**3.32**†	5.73
∘	✓	ASAP-Net	0.308†	0.492†	0.508†	3.10†	7.26†	12.4†
∘	✓	CoNeS	**0.567**	**0.714**	**0.749**	**2.33**	3.54	**4.05**
✓	✗	zero imputation	0.184†	0.397†	0.400†	4.06†	18.0†	22.2†
✓	∘	pix2pix	0.713†	0.856†	0.874†	1.54	2.19†	2.09
✓	∘	pGAN	0.701†	0.839†	0.844†	1.82†	2.60†	2.58†
✓	∘	ResViT	0.716†	0.831†	0.862	1.61	2.48†	2.32
✓	∘	ASAP-Net	0.677†	0.834†	0.854†	1.95†	2.63†	2.45†
✓	∘	CoNeS	**0.746**	**0.858**	**0.878**	**1.40**	**2.09**	**1.96**

4.6 Ablation study

Ablation studies were performed to verify the benefits of individual components in the proposed method. For simplicity, we trained a baseline CoNeS model that translates T1ce from T2 on BraTS 2018. We first examined the added value of the source image as input to the MLP by removing the intensity value from the input channel. In this case, the neural fields are conditioned on the latent code only. Next, we compared shift modulation against a full hypernetwork where all the parameters of the MLP are generated. Last, we trained the proposed method without the adversarial loss to show the contribution of the discriminator in our model. Wilcoxon signed-rank tests were performed between the baseline model and ablated models. Quantitative and qualitative results are shown in Table 7 and Fig. 6. We noticed that although the model without adversarial loss achieves marginally better SSIM and PSNR, the results, especially the tumor region, are visually blurry, which shows that the adversarial loss helps the model to reconstruct more details and outputs more realistic images. Apart from this, the proposed model obtained the best results among the ablated models that include the adversarial loss and showed significant differences ($p < .01$) in SSIM. These results also show that shift modulation helps to reduce the parameters from 14.5k to 0.26k, which is the number of neurons in the MLP, without loss of representation capability. Moreover, although the instance-specific information is already encoded in the latent code, conditioning the network on intensity directly can still add extra information and improve performance.

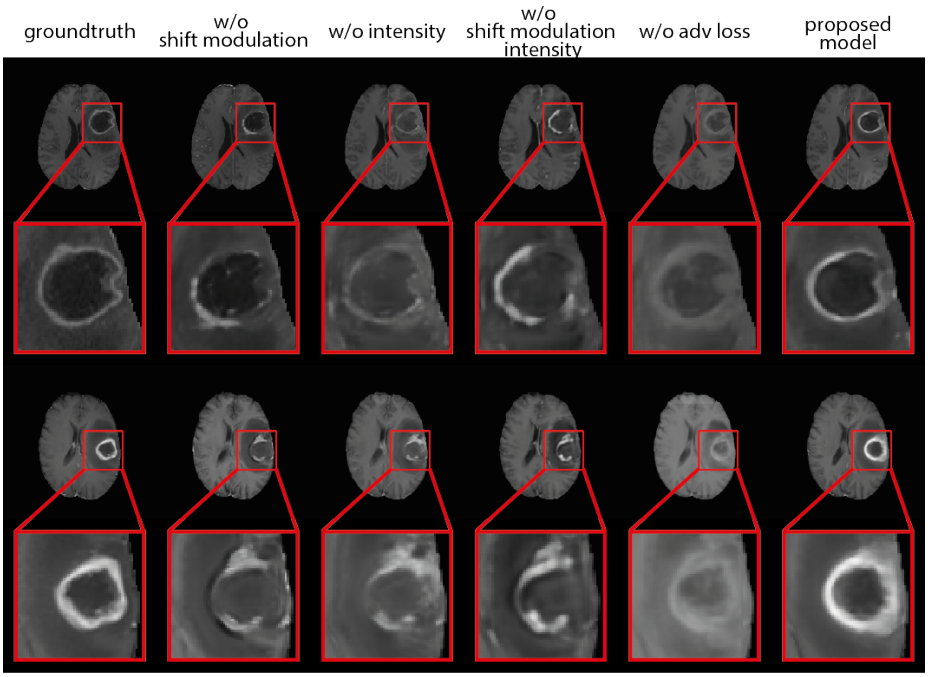


Figure 6: Example results of the ablated models. Zoomed-in results indicated with red rectangles are shown below the full images.

Table 7: Quantitative comparison of ablated models on BraTS 2018. The mean value and standard deviation of PSNR and SSIM are reported. The highest values per column are indicated in boldface; The † after each metric indicates a significant difference ($p < .01$) compared to the proposed model (the bottom row).

shift modulation	intensity	adversarial loss	#param generated	PSNR	SSIM
no	no	yes	14.5k	$29.6 \pm 2.13^{\dagger}$	$0.933 \pm 0.013^{\dagger}$
no	yes	yes	14.5k	29.9 ± 2.32	$0.938 \pm 0.013^{\dagger}$
yes	no	yes	0.26k	30.0 ± 2.13	$0.938 \pm 0.013^{\dagger}$
yes	yes	no	0.26k	$\mathbf{30.2 \pm 2.33}$	$\mathbf{0.943 \pm 0.013}^{\dagger}$
yes	yes	yes	0.26k	30.0 ± 2.22	0.941 ± 0.014

We next demonstrated the stability of the models by comparing the loss curves of the ablated models. Both the adversarial loss L_{adv} and the total loss L are shown in Fig. 7. We observe that both losses of the models using the full hypernetwork fluctuated substantially and L_{adv} increased midway through training. On the contrary, both loss curves of the models using shift modulation remained stable throughout the learning. The experiments suggested that by reducing the number of parameters generated, shift modulation is able to improve the stability of the image translation model.

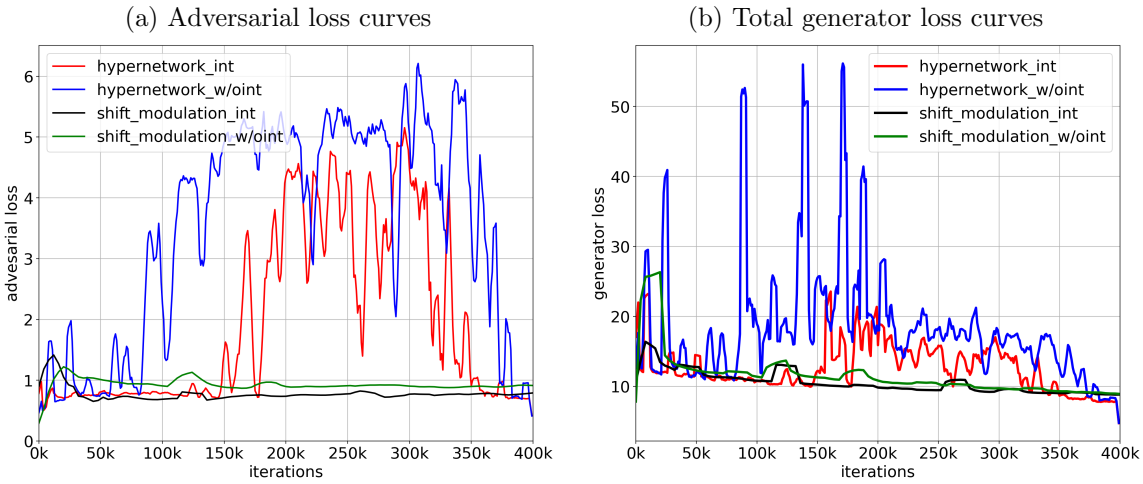


Figure 7: Training loss curves of the ablated models. (a) adversarial loss (b) total generator loss including reconstruction loss, adversarial loss, feature matching loss, and latent code regularization. The models using shift modulation show more stable training loss against the models using a full hypernetwork.

5. Discussion and conclusion

In this work, we proposed CoNeS, a novel conditional neural fields-based model for MRI translation. We modeled the image translation problem using neural fields, which are conditioned on the source images, and learned latent codes through a coordinate-based network. The proposed model adapts the predicted neural fields by varying the latent codes across coordinates to ensure better local representations. We then introduced a shift modulation strategy for the conditioning to reduce the model complexity and stabilize the training. We compared the proposed model with state-of-the-art image translation models and our experiments showed that CoNeS performs better in the entire image scope as well as the tumor region, which is clinically relevant. Through visualization results, we also showed that the proposed method can reproduce more structural details, while the other methods' results are visually more blurry. The transformer-based model (ResViT) performed on par with the other methods in our experiments, where on natural images they have been reported to outperform those (Kim et al., 2022; Shibasaki et al., 2022). Our datasets, however, are considerably smaller than what is used in the domain of natural images, while transformer-based models are considered data-demanding.

We performed a spectral analysis to demonstrate improvements in image translation when using neural fields. As expected, all the CNN-based models and ResViT, which is a hybrid transformer model containing transposed layers during decoding, were unable to reproduce high-frequency signals due to their spectral bias (Rahaman et al., 2019; Durall et al., 2020). In contrast, the proposed model was able to preserve the high-frequency information and reconstruct the spectrum in the entire frequency range on both datasets. We also observed that ASAP-Net, a neural field-based benchmark, did not show consistent performance across the two datasets and could not reproduce the spectral distribution on the

VS dataset either. These results are consistent with prior studies demonstrating that the full hypernetwork, in which all the parameters of the main network are generated, is sensitive to its initialization and difficult to optimize (Chang et al., 2019). The ablation studies further indicated that compared to a full hypernetwork, the conditioning via shift modulation can make the training of neural fields more stable and maintain the representation capability. Furthermore, the results also showed that by introducing the adversarial loss, the predicted images are more realistic and contain more textural detail, although the quantitative metrics (SSIM and PSNR) are slightly lower than the model without the adversarial loss. The reason may be that SSIM and PSNR are not able to measure the benefits of the adversarial loss, which is in line with the conclusion in previous research (Liu et al., 2023a; Dalmaz et al., 2022).

To evaluate the value of synthesized MRI in downstream analysis, we performed tumor segmentation experiments. We first demonstrated that dropping sequences during inference of a segmentation model can significantly damage the performance, which shows the complementary importance of multiple MRI sequences in segmentation. We next tested the segmentation model using different synthesized images and compared the results with the inference using incomplete input images. The experiments demonstrated that image translation models can significantly improve segmentation accuracy by replacing the missing input channel with synthesized images. Furthermore, the images generated by our proposed CoNeS model performed best among the state-of-the-art methods in most of the experiments, which is consistent with the visual improvement observed in the translation experiments. Nevertheless, we found that synthesized images cannot fully replace real images, and a baseline model trained on all real images performed best.

One limitation of our work is that in the clinic, the availability of MRI sequences may vary from patient to patient (Li et al., 2023). The proposed model, however, cannot handle arbitrarily missing sequences, unless separate models are trained for each case. Further work would be adapting the proposed model to random incomplete MRI scans by incorporating techniques like learning disentangled representations (Shen et al., 2020) or latent representation fusion (Chartsias et al., 2017). Moreover, the choice of the positional encoding frequency m may bias the network to fit the signal of a certain bandwidth (Wang et al., 2021). To ease the optimization and improve the generalization, it may be worthwhile to integrate periodic activation functions (Sitzmann et al., 2020) in our design instead of positional encoding for better representation capability.

In summary, we presented a neural fields-based model that synthesizes missing MRI from other sequences with excellent performance, which can be further integrated into the downstream analysis. All experiments showed improved performance compared to state-of-the-art translation models, while the spectrum analysis and ablation studies demonstrated the strengths of the proposed model over traditional CNN and neural fields models. Neural fields hold great promise in MRI translation to solve the missing MRI sequence problem in the clinic.

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding the treatment of animals or human subjects.

Conflicts of Interest

We declare we do not have conflicts of interest.

References

Tamaz Amiranashvili, David Lüdke, Hongwei Bran Li, Bjoern Menze, and Stefan Zachow. Learning shape reconstruction from sparse measurements with neural implicit functions. In *International Conference on Medical Imaging with Deep Learning*, pages 22–34. PMLR, 2022.

Sina Amirrajab, Yasmina Al Khalil, Cristian Lorenz, Jürgen Weese, Josien Pluim, and Marcel Breeuwer. Pathology synthesis of 3D-Consistent cardiac MR images using 2D VAEs and GANs. *Machine Learning for Biomedical Imaging*, 2:288–311, 2023.

Ivan Anokhin, Kirill Demochkin, Taras Khakhulin, Gleb Sterkin, Victor Lempitsky, and Denis Korzhenkov. Image generators with conditionally-independent pixel synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14278–14287, 2021.

Karim Armanious, Chenming Jiang, Marc Fischer, Thomas Küstner, Tobias Hepp, Konstantin Nikolaou, Sergios Gatidis, and Bin Yang. MedGAN: Medical image translation using GANs. *Computerized Medical Imaging and Graphics*, 79:101684, 2020.

Reza Azad, Nika Khosravi, Mohammad Dehghanmanshadi, Julien Cohen-Adad, and Dorit Merhof. Medical image segmentation on MRI images with missing modalities: a review. *arXiv preprint arxiv:2203.06217*, 2022a.

Reza Azad, Nika Khosravi, and Dorit Merhof. SMU-Net: Style matching U-Net for brain tumor segmentation with missing modalities. In *International Conference on Medical Imaging with Deep Learning*, pages 48–62, 2022b.

Mara Cercignani and Samira Bouyagoub. Brain microstructure by multi-modal MRI: Is the whole greater than the sum of its parts? *NeuroImage*, 182:117–127, 2018.

Oscar Chang, Lampros Flokas, and Hod Lipson. Principled weight initialization for hypernetworks. In *International Conference on Learning Representations*, 2019.

Agisilaos Chartsias, Thomas Joyce, Mario Valerio Giuffrida, and Sotirios A Tsaftaris. Multimodal MR synthesis via modality-invariant latent representation. *IEEE Transactions on Medical Imaging*, 37(3):803–814, 2017.

Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12299–12310, 2021a.

Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8628–8638, 2021b.

Yunjie Chen, Marius Staring, Jelmer M. Wolterink, and Qian Tao. Local implicit neural representations for multi-sequence MRI translation. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2023. .

Andrea Cherubini, Maria Eugenia Caligiuri, Patrice Péran, Umberto Sabatini, Carlo Cosentino, and Francesco Amato. Importance of multimodal MRI in characterizing brain tissue and its potential application for individual age prediction. *IEEE Journal of Biomedical and Health Informatics*, 20(5):1232–1239, 2016.

Onat Dalmaz, Mahmut Yurt, and Tolga Çukur. ResViT: Residual vision transformers for multimodal medical image synthesis. *IEEE Transactions on Medical Imaging*, 41(10): 2598–2614, 2022.

Salman UH Dar, Mahmut Yurt, Levent Karacan, Aykut Erdem, Erkut Erdem, and Tolga Cukur. Image synthesis in multi-contrast MRI with conditional generative adversarial networks. *IEEE Transactions on Medical Imaging*, 38(10):2375–2388, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Emilien Dupont, Hyunjik Kim, SM Ali Eslami, Danilo Jimenez Rezende, and Dan Rosenbaum. From data to functa: Your data point is a function and you can treat it like one. In *International Conference on Machine Learning*, pages 5694–5725, 2022.

Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7890–7899, 2020.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021.

David Ha, Andrew M. Dai, and Quoc V. Le. HyperNetworks. In *International Conference on Learning Representations*, 2017.

Mohammad Havaei, Nicolas Guizard, Nicolas Chapados, and Yoshua Bengio. HeMIS: Hetero-modal image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 469–477, 2016.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, volume 30, 2017.

Minhao Hu, Matthis Maillard, Ya Zhang, Tommaso Ciceri, Giammarco La Barbera, Isabelle Bloch, and Pietro Gori. Knowledge distillation from multi-modal to mono-modal segmentation networks. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 772–781, 2020.

Juan Eugenio Iglesias, Ender Konukoglu, Darko Zikic, Ben Glocker, Koen Van Leemput, and Bruce Fischl. Is synthesizing MRI contrast useful for inter-modality analysis? In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 631–638, 2013.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017.

Yifan Jiang, Shiyu Chang, and Zhangyang Wang. TransGAN: Two pure transformers can make one strong GAN, and that can scale up. In *Advances in Neural Information Processing Systems*, volume 34, pages 14745–14758, 2021.

Thomas Joyce, Agisilaos Chartsias, and Sotirios A Tsaftaris. Robust multi-modal MR image synthesis. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 347–355, 2017.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems*, volume 35, pages 23593–23606, 2022.

Soohyun Kim, Jongbeom Baek, Jihye Park, Gyeongnyeon Kim, and Seungryong Kim. InstaFormer: Instance-aware image-to-image translation with transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18321–18331, June 2022.

Stefan Klein, Marius Staring, Keelin Murphy, Max A Viergever, and Josien PW Pluim. elastix: a toolbox for intensity-based medical image registration. *IEEE Transactions on Medical Imaging*, 29(1):196–205, 2009.

Hongwei Li, Johannes C Paetzold, Anjany Sekuboyina, Florian Kofler, Jianguo Zhang, Jan S Kirschke, Benedikt Wiestler, and Bjoern Menze. DiamondGAN: Unified multi-modal generative adversarial networks for MRI sequences synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 795–803, 2019.

Hongwei Bran Li, Gian Marco Conte, Syed Muhammad Anwar, Florian Kofler, Koen van Leemput, Marie Piraud, Ivan Ezhov, Felix Meissen, Maruf Adewole, Anastasia Janas, et al. The brain tumor segmentation (BraTS) challenge 2023: Brain MR image synthesis for tumor segmentation (BraSyn). *arXiv preprint arXiv:2305.09011*, 2023.

Jae Hyun Lim and Jong Chul Ye. Geometric GAN. *arXiv preprint arXiv:1705.02894*, 2017.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.

Jiang Liu, Srivathsa Pasumarthi, Ben Duffy, Enhao Gong, Keshav Datta, and Greg Zaharchuk. One model to synthesize them all: Multi-contrast multi-scale transformer for missing data imputation. *IEEE Transactions on Medical Imaging*, 2023a.

Linfeng Liu, Siyu Liu, Lu Zhang, Xuan Vinh To, Fatima Nasrallah, Shekhar S Chandra, Alzheimer's Disease Neuroimaging Initiative, et al. Cascaded multi-modal mixing transformers for Alzheimer's disease classification with incomplete data. *NeuroImage*, page 120267, 2023b.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

Julian McGinnis, Suprosanna Shit, Hongwei Bran Li, Vasiliki Sideri-Lampretsa, Robert Graf, Maik Dannecker, Jiazhen Pan, Nil Stolt-Ansó, Mark Mühlau, Jan S Kirschke, et al. Single-subject multi-contrast MRI super-resolution via implicit neural representations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 173–183. Springer, 2023.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2014.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Amirali Molaei, Amirhossein Aminimehr, Armin Tavakoli, Amirhossein Kazerouni, Bobby Azad, Reza Azad, and Dorit Merhof. Implicit neural representation in medical imaging: a comparative survey. *arXiv preprint arXiv:2307.16142*, 2023.

Olaf M Neve, Yunjie Chen, Qian Tao, Stephan R Romeijn, Nick P de Boer, Willem Grootjans, Mark C Kruit, Boudewijn PF Lelieveldt, Jeroen C Jansen, Erik F Hensen, et al. Fully automated 3D vestibular schwannoma segmentation with and without gadolinium-based contrast material: A multicenter, multivendor study. *Radiology: Artificial Intelligence*, 4(4):e210300, 2022.

Dong Nie, Roger Trullo, Jun Lian, Li Wang, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen. Medical image synthesis with deep convolutional adversarial networks. *IEEE Transactions on Biomedical Engineering*, 65(12):2720–2730, 2018.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019.

Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *European Conference on Computer Vision*, pages 523–540, 2020.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pages 5301–5310, 2019.

Vasileios Sevetlidis, Mario Valerio Giuffrida, and Sotirios A Tsaftaris. Whole image synthesis using a deep encoder-decoder network. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 127–137, 2016.

Tamar Rott Shaham, Michaël Gharbi, Richard Zhang, Eli Shechtman, and Tomer Michaeli. Spatially-adaptive pixelwise networks for fast image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14882–14891, June 2021.

Anmol Sharma and Ghassan Hamarneh. Missing MRI pulse sequence synthesis using multimodal generative adversarial network. *IEEE Transactions on Medical Imaging*, 39(4): 1170–1183, 2019.

Liyue Shen, Wentao Zhu, Xiaosong Wang, Lei Xing, John M Pauly, Baris Turkbey, Stephanie Anne Harmon, Thomas Hogue Sanford, Sherif Mehralivand, Peter L Choyke, et al. Multi-domain image completion for random missing input data. *IEEE Transactions on Medical Imaging*, 40(4):1113–1122, 2020.

Kei Shibasaki, Shota Fukuzaki, and Masaaki Ikehara. 4K real time image to image translation network with transformers. *IEEE Access*, 10:73057–73067, 2022.

Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Advances in neural information processing systems*, volume 33, pages 7462–7473, 2020.

Youssef Skandarani, Nathan Painchaud, Pierre-Marc Jodoin, and Alain Lalande. On the effectiveness of GAN generated cardiac MRIs for segmentation. In *Medical Imaging with Deep Learning*, 2020.

Gijs Van Tulder and Marleen de Bruijne. Why does synthesized data improve multi-sequence classification? In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 531–538, 2015.

Peng-Shuai Wang, Yang Liu, Yu-Qi Yang, and Xin Tong. Spline positional encoding for learning 3D implicit signed distance fields. In *International Joint Conference on Artificial Intelligence*, 2021.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018.

Wen Wei, Emilie Poirion, Benedetta Bodini, Stanley Durrleman, Olivier Colliot, Bruno Stankoff, and Nicholas Ayache. Fluid-attenuated inversion recovery MRI synthesis from multisequence MRI using three-dimensional fully convolutional networks for multiple sclerosis. *Journal of Medical Imaging*, 6(1):014005, 2019.

Jelmer M Wolterink, Jesse C Zwienenberg, and Christoph Brune. Implicit neural representations for deformable image registration. In *International Conference on Medical Imaging with Deep Learning*, pages 1349–1359, 2022.

Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. In *Computer Graphics Forum*, volume 41, pages 641–676, 2022.

Mahmut Yurt, Salman UH Dar, Aykut Erdem, Erkut Erdem, Kader K Oguz, and Tolga Çukur. mustGAN: multi-stream generative adversarial networks for MR image synthesis. *Medical image analysis*, 70:101944, 2021.

Ellen D. Zhong, Tristan Bepler, Joseph H. Davis, and Bonnie Berger. Reconstructing continuous distributions of 3D protein structure from cryo-EM images. In *International Conference on Learning Representations*, 2020.