ResViT: Residual Vision Transformers for Multimodal Medical Image Synthesis

Onat Dalmaz, *Graduate Student Member, IEEE*, Mahmut Yurt, and Tolga Çukur, *Senior Member, IEEE*

***Abstract*— Generative adversarial models with convolutional neural network (CNN) backbones have recently been established as state-of-the-art in numerous medical image synthesis tasks. However, CNNs are designed to perform local processing with compact filters, and this inductive bias compromises learning of contextual features. Here, we propose a novel generative adversarial approach for medical image synthesis, ResViT, that leverages the contextual sensitivity of vision transformers along with the precision of convolution operators and realism of adversarial learning. ResViT's generator employs a central bottleneck comprising novel aggregated residual transformer (ART) blocks that synergistically combine residual convolutional and transformer modules. Residual connections in ART blocks promote diversity in captured representations, while a channel compression module distills task-relevant information. A weight sharing strategy is introduced among ART blocks to mitigate computational burden. A unified implementation is introduced to avoid the need to rebuild separate synthesis models for varying source-target modality configurations. Comprehensive demonstrations are performed for synthesizing missing sequences in multi-contrast MRI, and CT images from MRI. Our results indicate superiority of ResViT against competing CNN- and transformer-based methods in terms of qualitative observations and quantitative metrics.**

***Index Terms*— Medical image synthesis, transformer, residual, vision, adversarial, generative, unified.**

I. Introduction

MEDICAL imaging plays a pivotal role in modern healthcare by enabling in vivo examination of pathology in the human body. In many clinical scenarios, multi-modal protocols are desirable that involve a diverse collection of images from multiple scanners (e.g., CT, MRI) [1], or multiple acquisitions from a single scanner (multi-contrast MRI) [2]. Complementary information about tissue morphology, in turn, empower physicians to diagnose with higher accuracy and confidence. Unfortunately, numerous factors including uncooperative patients and excessive scan times prohibit ubiquitous multi-modal imaging [3], [4]. As a result, there has been ever-growing interest in synthesizing unacquired images in multi-modal protocols from the subset of available images, bypassing costs associated with additional scans [5], [6].

Medical image synthesis aims to predict target-modality images for a subject given source-modality images acquired under a limited scan budget [7]. This is an ill-posed inverse problem since medical images are high dimensional, target-modality data are absent during inference, and there exist nonlinear differences in tissue contrast across modalities [8]–[13]. Unsurprisingly, recent adoption of deep learning methods for solving this difficult problem has enabled major performance leaps [14]–[21]. In learning-based synthesis, network models effectively capture a prior on the joint distribution of source-target images [22]–[24]. Earlier studies using CNNs for this purpose reported significant improvements over traditional approaches [22], [23], [25]–[28]. Generative adversarial networks (GANs) were later introduced that leverage an adversarial loss to increase capture of detailed tissue structure [24], [29]–[35]. Further improvements were attained by leveraging enhanced architectural designs [36]–[39], and learning strategies [40]–[42]. Despite their prowess, prior learning-based synthesis models are fundamentally based on convolutional architectures that use compact filters to extract local image features [43], [44]. Exploiting correlations among small neighborhoods of image pixels, this inductive bias reduces the number of model parameters to facilitate learning. However, it also limits expressiveness for contextual features that reflect long-range spatial dependencies [45], [46].

Medical images contain contextual relationships across both healthy and pathological tissues. For instance, bone in the skull or CSF in the ventricles broadly distribute over spatially contiguous or segregated brain regions, resulting in dependencies among distant voxels. While pathological tissues have less regular anatomical priors, their spatial distribution (e.g., location, quantity, shape) can still show disease-specific patterns [47]. For instance, multiple diffuse brain lesions are present in multiple sclerosis (MS) and Alzheimer's (AD); commonly located near periventricular and juxtacortical regions in MS, and near hippocampus, entorhinal cortex and isocortex in AD [48]. Meanwhile, few lesions manifest as spatially-contiguous clumps in cancer; with lesions typically located near the cerebrum and cerebellum in gliomas, and near the skull in meningiomas [48]. Thus, the distribution of pathology

Manuscript received 28 February 2022; revised 5 April 2022; accepted 12 April 2022. Date of publication 18 April 2022; date of current version 30 September 2022. The work of Onat Dalmaz was supported in part by the Scientific and Technological Research Council of Turkey BIDEB Scholarship. The work of Tolga Cukur was supported in part by the Turkish Academy of Sciences GEBIP 2015 Fellowship and in part by the Science Academy BAGEP 2017 Fellowship. *(Corresponding author: Tolga Çukur.)*

Onat Dalmaz and Mahmut Yurt are with the National Magnetic Resonance Research Center (UMRAM), Department of Electrical and Electronics Engineering, Bilkent University, 06800 Ankara, Turkey (e-mail: onat@ee.bilkent.edu.tr; mahmut@ee.bilkent.edu.tr).

Tolga Çukur is with the Neuroscience Program, Sabuncu Brain Research Center, and the National Magnetic Resonance Research Center (UMRAM), Department of Electrical and Electronics Engineering, Bilkent University, 06800 Ankara, Turkey (e-mail: cukur@ee.bilkent.edu.tr).

Digital Object Identifier 10.1109/TMI.2022.3167808

also involves context regarding the position and structure of lesions with respect to healthy tissue. In principle, synthesis performance can be enhanced by priors that capture these relationships. Vision transformers are highly promising for this goal since attention operators that learn contextual features can improve sensitivity for long-range interactions [49], and focus on critical image regions for improved generalization to atypical anatomy such as lesions [50]. However, adopting vanilla transformers in tasks with pixel-level outputs is challenging due to computational burden and limited localization [51]. Recent studies instead consider hybrid architectures or computation-efficient attention operators to adopt transformers in medical imaging tasks [52]–[57].

Here, we propose a novel deep learning model for medical image synthesis, ResViT, that translates between multi-modal imaging data. ResViT combines the sensitivity of vision transformers to global context, the localization power of CNNs, and the realism of adversarial learning. ResViT's generator follows an encoder-decoder architecture with a central bottleneck to distill task-critical information. The encoder and decoder contain CNN blocks to leverage local precision of convolution operators [58]. The bottleneck comprises novel aggregated residual transformer (ART) blocks to synergistically preserve local and global context, with a weight-sharing strategy to minimize model complexity. To improve practical utility, a unified ResViT implementation is introduced that consolidates models for numerous source-target configurations. Demonstrations are performed for synthesizing missing sequences in multi-contrast MRI, and CT from MRI. Comprehensive experiments on imaging datasets from healthy subjects and patients clearly indicate the superiority of the proposed method against competing methods. Code to implement the ResViT model is publicly available at https://github.com/icon-lab/ResViT.

Contributions

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

II. Related Work

The immense success of deep learning in inverse problems has motivated its rapid adoption in medical imaging [59], [60]. Medical image synthesis is a particularly ill-posed problem since target images are predicted without any target-modality data [32]. Earlier studies in this domain have proposed local networks based on patch-level processing [16], [61], [62]. While local networks offer benefits over traditional approaches, they can show limited sensitivity to broader context across images [22]. Later studies adopted deep CNNs for image-level processing with increasing availability of large imaging databases. CNN-based synthesis has been successfully demonstrated in various applications including synthesis across MR scanners [32], [63]–[65], multi-contrast MR synthesis [22], [23], [25]–[28], and CT synthesis [66]–[69]. Despite significant improvements they enable, CNNs trained with pixel-wise loss terms tend to suffer from undesirable loss of detailed structure [24], [43], [44].

To improve capture of structural details, GANs [29] were proposed to learn the distribution of target modalities conditioned on source modalities [70]. Adversarial losses empower GANs to capture an improved prior for recovery of high-spatial-resolution information [24], [43], [44]. In recent years, GAN-based methods were demonstrated to offer state-of-the-art performance in numerous synthesis tasks, including data augmentation as well as multi-modal synthesis [24], [34], [71], [72]. Important applications of GAN models include CT to PET [73], [74], MR to CT [75]–[77], unpaired cross-modality [78]–[81], 3T-to-7T [82], [83], and multi-contrast MRI synthesis [24], [30]–[42].

While GAN models have arguably emerged as a gold standard in recent years, they are not without limitation. In particular, GANs are based on purely convolutional operators known to suffer from poor across-subject generalization to atypical anatomy and sub-optimal learning of long-range spatial dependencies [45], [46]. Recent studies have incorporated spatial or channel attention mechanisms to modulate CNN-derived feature maps [37], [50], [84]–[88]. Such modulation motivates the network to give greater focus to regions that may suffer from greater errors [50], [85]. While attention maps might be distributed across image regions, multiplicative gating of local CNN features offers limited expressiveness in modeling of global context [51], [89], [90].

To incorporate contextual representations, transformer-based methods have received recent interest in imaging tasks such as segmentation [51], [89], [91], reconstruction [52]–[54], and synthesis [55]–[57]. Among relevant methods are Transformer GAN that suppresses noise in low-dose PET images [52], TransCT that suppresses noise in low-dose CT images [53], and SLATER that recovers MR images from undersampled k-space acquisitions [54]. While these methods reconstruct images for single-modality data, ResViT translates imaging data across separate modalities. Furthermore, Transformer GAN is an adversarial model with convolutional encoder-decoder and a bottleneck that contains a transformer without external residual connections. TransCT is a non-adversarial model where CNN blocks first learn textural components of low-frequency (LF) and high-frequency (HF) image parts; and a transformer without external residual connections then combines encoded HF and textural LF maps. In comparison, ResViT is an adversarial model that employs a hybrid architecture in its bottleneck comprising a cascade of residual transformer and residual CNN modules. Unlike SLATER based on an unconditional model that maps latent variables to images via cross-attention transformers, ResViT is a conditional model based on self-attention transformers.

Few recent studies have independently introduced transformer-based methods for medical image synthesis. VTGAN generates retinal angiograms from fundus

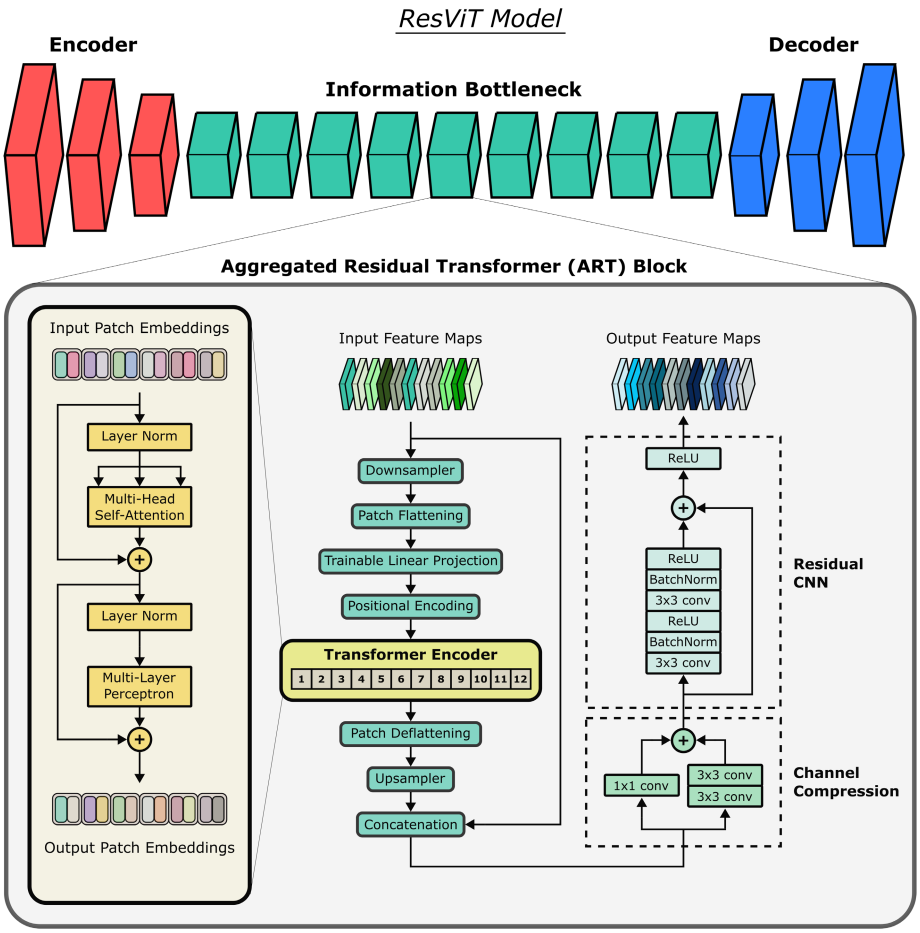

Fig. 1. The generator in ResViT follows an encoder-decoder architecture bridged with a central information bottleneck to distill task-specific information. The encoder and decoder comprise convolutional layers to maintain local precision and inductive bias in learned structural representations. Meanwhile, the information bottleneck comprises a stack of novel aggregated residual transformer (ART) blocks. ART blocks learn contextual representations via vision transformers, and synergistically fuse CNN-based local and transformer-based global representations.

photographs [55] and GANBERT performs MR-to-PET synthesis [56], whereas ResViT performs multi-contrast MRI and MR-to-CT synthesis. Both VTGAN and GANBERT use entirely convolutional generators and only include transformers in their discriminators. In contrast, ResViT incorporates transformers in its generator to explicitly leverage long-range context. The closest study to our work is PTNet that performs one-to-one translation between T_1- and T_2-weighted images in infant MRI [57]. However, PTNet is a non-adversarial model without a discriminator, and it follows a convolution-free architecture. In contrast, ResViT is an adversarial model with a hybrid CNN-transformer architecture to achieve high localization and contextual sensitivity along with a high degree of realism in synthesized images. Furthermore, a broader set of tasks are considered for ResViT including one-to-one and many-to-one translation.

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

III. Theory and Methods

A. Residual Vision Transformers

Here we propose a novel adversarial method for medical image synthesis named residual vision transformers, ResViT, that can unify various source-target modality configurations into a single model for improved practicality. ResViT leverages a hybrid architecture of deep convolutional operators and transformer blocks to simultaneously learn high-resolution structural and global contextual features (Fig. 1). The generator subnetwork follows an encoder - information bottleneck - decoder pathway, and the discriminator subnetwork is composed of convolutional operators. The generator's bottleneck contains a stack of novel aggregated residual transformer (ART) blocks. Each ART block is organized as the cascade of a transformer module that extracts hidden contextual features, and a CNN module that extracts hidden local features of input feature maps. Importantly, external skip connections are inserted around both modules to create multiple paths of information flow through the block. These paths propagate multiple sets of features to the output: (a) Input features from the previous network layer passing through skip connections of transformer and CNN modules; (b) Contextual features computed by the transformer module passing through the skip

connection of the CNN module; (c) Local features computed by the CNN module based on input features reaching through the skip connection of the transformer module; (d) Hybrid local-contextual features computed by the transformer-CNN cascade. Therefore, the main motivation for use of residual transformer and residual CNN modules in ART blocks is to learn an aggregated representation that synergistically combines lower-level input features along with their contextual, local, and hybrid local-contextual features.

The central segment of ResViT containing ART blocks acts as an information bottleneck for spatial and feature dimensions of medical image representations. On the one hand, the central segment processes feature maps that have been spatially downsampled by the encoder. This increases the relative emphasis on mid- to high-level spatial information over lower-level information [58]. On the other hand, ART blocks contain channel-compression (CC) modules that process concatenated feature maps from the previous ART block and the transformer module. CC modules downsample the concatenated maps in the feature dimension to distill a task-relevant set of convolutional and contextual features.

Given the computational efficiency of convolutional layers, CNNs pervasively process feature maps at high spatial resolution to improve sensitivity for local features [58]. In contrast, vision transformers include computationally exhaustive self-attention layers, so they typically process feature maps at relatively lower resolution [49]. To ensure that both the residual CNNs and transformers in ART blocks receive input feature maps at their expected resolutions, we incorporated down and upsampling blocks respectively at the input and output of transformer modules. This design ensures compatibility between the resolutions of feature maps extracted from CNN and transformer modules. In the remainder of this section, we explain the detailed composition of each architectural component, and we describe the loss functions to train ResViT.

1) Encoder: The first component of ResViT is a deep encoder network that contains a series of convolutional layers to capture a hierarchy of localized features of source images. Note that ResViT can serve as a unified synthesis model, so its encoder receives as input the full set of modalities within the imaging protocol, both source and target modalities (Fig. 2). Source modalities are input via an identity mapping, whereas unavailable target modalities are masked out:

$$X_i^G = a_i \cdot m_i \tag{1}$$

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

$$a_i = \begin{cases} 1 & \text{if } m_i \text{ is a source modality} \\ 0 & \text{if } m_i \text{ is a target modality} \end{cases} \tag{2}$$

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

2) Information Bottleneck: Next, ResViT employs a residual bottleneck to distill task-relevant information in the encoded features. Note that convolution operators have greater power in capturing localized features, whereas attention operators are more sensitive to context-driven features. To simultaneously maintain localization power and contextual sensitivity, we introduce ART blocks that aggregate the information from residual convolutional and transformer branches (Fig. 1). Receiving as input the jth layer feature maps $f_j \in \mathbb{R}^{N_C, H, W}$, an ART block first processes the feature maps via a vision transformer. Due to computational constraints, the transformer expects feature maps at smaller resolutions compared to convolutional layers. Thus, the spatial dimensions (H, W) of $f_j \in \mathbb{R}^{N_C, H, W}$ are lowered by a downsampling block (***DS***):

$$f_j' \in \mathbb{R}^{N_C', H', W'} = \boldsymbol{DS}(f_j) \tag{3}$$

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

$$z_0 = [f_j^1 P_E; f_j^2 P_E; \ldots; f_j^{N_P} P_E] + P_E^{pos} \tag{4}$$

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

Next, the transformer encoder processes patch embeddings via a cascade of L layers of multi-head self-attention (***MSA***) [92] and multi-layer perceptrons (***MLP***) [93]. The output of the lth layer in the transformer encoder is given as:

$$z_l' = \boldsymbol{MSA}(\boldsymbol{LN}(z_{l-1})) + z_{l-1} \tag{5}$$

$$z_l = \boldsymbol{MLP}(\boldsymbol{LN}(z_l')) + z_l' \tag{6}$$

MSA layers in Eq. 5 employ S separate self-attention heads:

$$\boldsymbol{MSA}(z) = [\boldsymbol{SA}_1(z); \boldsymbol{SA}_2(z); \ldots; \boldsymbol{SA}_S(z)] U_{msa} \tag{7}$$

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

$$A_{a,b} = softmax(q_a k_b^T / {N_D}^{0.5}) \tag{8}$$

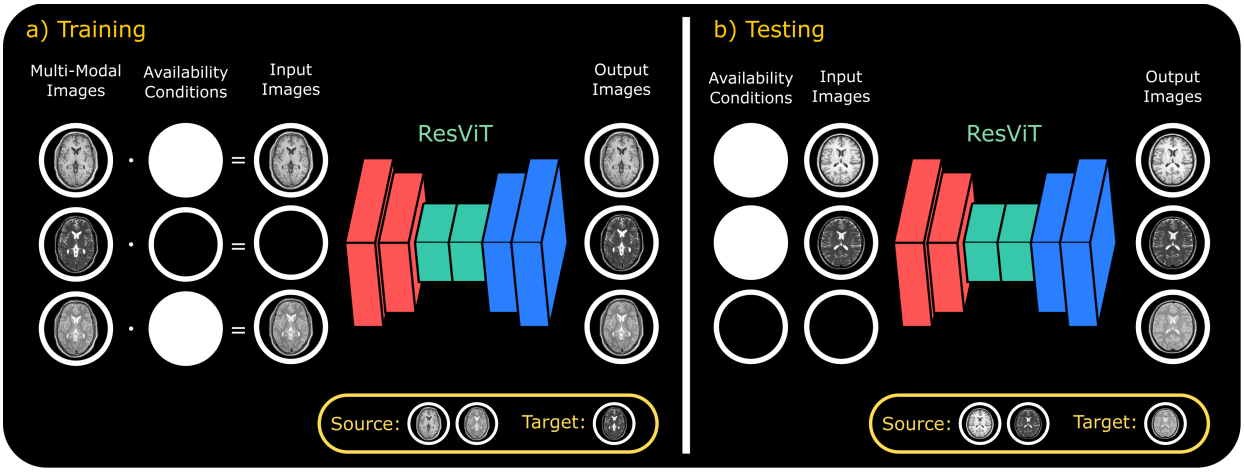

Fig. 2. ResViT is a conditional image synthesis model that can unify various source-target modality configurations into a single model for improved practicality. a) During training, ResViT takes as input the entire set of images within the multi-modal protocol, including both source and target modalities. For model consolidation across multiple synthesis tasks, various configurations of source-target modalities are expressed in terms of availability conditions in ResViT. b) During inference, the specific source-target configuration is determined via the availability conditions in each given test subject.

Note that q, k, v are respectively obtained as learnable projections T_q, T_k, T_v of z.

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

$$g_j \in \mathbb{R}^{N_C,H,W} = \boldsymbol{US}(g'_j) \tag{9}$$

where $g_j \in \mathbb{R}^{N_C,H,W}$ are upsampled feature maps output by the transformer module. Channel-wise concatenation is performed to fuse global context learned via the transformer with localized features captured via convolutional operators. To distill learned structural and contextual representations, the channels of the concatenated feature maps are then compressed via a channel compression ($\boldsymbol{CC}$) module:

$$h_j \in \mathbb{R}^{N_C,H,W} = \boldsymbol{CC}(concat(f_j, g_j)) \tag{10}$$

where h_j are compressed feature maps. $\boldsymbol{CC}$ uses two parallel convolutional branches of varying kernel size. Finally, the feature maps are processed via a residual CNN ($\boldsymbol{ResCNN}$) [58]:

$$f_{j+1} \in \mathbb{R}^{N_C,H,W} = \boldsymbol{ResCNN}(h_j) \tag{11}$$

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

3) Decoder: The last component of the generator is a deep decoder based on transposed convolutional layers. Because ResViT can serve as a unified model, its decoder can synthesize all contrasts within the multi-modal protocol regardless of the specific source-target configuration (Fig. 2). The decoder receives as input the feature maps f_A distilled by the bottleneck and produces multi-modality images $\hat{Y}_i^G \in \hat{Y}^G$ in separate channels, where A is the total number of ART blocks, and $\hat{Y}_i^G$ denotes the ith synthesized modality.

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

5) Discriminator: The discriminator in ResViT is based on a conditional PatchGAN architecture [43]. The discriminator performs patch-level differentiation between acquired and synthetic images. This implementation increases sensitivity to localized details related to high-spatial-frequency information. As ResViT can serve as a unified model by generating all modalities in the multi-modal protocol including sources, an availability-guided selective discriminator is employed:

$$X_i^D(source) = X_i^G = a_i \cdot m_i \tag{12}$$

$$X_i^D(syn\,target) = (1 - a_i) \cdot Y_i^G \tag{13}$$

$$X_i^D(acq\,target) = (1 - a_i) \cdot m_i \tag{14}$$

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

$$X^D(synthetic) = concat(X_i^D(source), X_i^D(syn\,target)) \tag{15}$$

$$X^D(acquired) = concat(X_i^D(source), X_i^D(acq\,target)) \tag{16}$$

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

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

$$L_{pix} = \sum_{i=1}^{I}(1 - a_i)\mathrm{E}[||(X^G)_i - m_i||_1] \quad (17)$$

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

$$L_{rec} = \sum_{i=1}^{I} a_i\mathrm{E}[||G(X^G)_i - m_i||_1] \quad (18)$$

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

$$L_{adv} = -\mathrm{E}[D(X^D(acquired)^2] - \mathrm{E}[(D(X^D(synthetic)) - 1)^2] \quad (19)$$

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

$$L_{ResViT} = \lambda_{pix}L_{pix} + \lambda_{rec}L_{rec} + \lambda_{adv}L_{adv} \quad (20)$$

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

B. Datasets

We demonstrated the proposed ResViT model on two multi-contrast brain MRI datasets (IXI: https://brain-development.org/ixi-dataset/, BRATS [94]) and a multi-modal pelvic MRI-CT dataset [95].

1) IXI Dataset: T_1-weighted, T_2-weighted, and PD-weighted brain MR images from 53 healthy subjects were analyzed. 25 subjects were reserved for training, 10 were reserved for validation, and 18 were reserved for testing. From each subject, 100 axial cross-sections containing brain tissues were selected. Acquisition parameters were as follows. T_1-weighted images: TE = 4.603ms, TR = 9.813ms, spatial resolution = $0.94 \times 0.94 \times 1.2\mathrm{mm}^3$. T_2-weighted images: TE = 100ms, TR = 8178.34ms, spatial resolution = $0.94 \times 0.94 \times 1.2\mathrm{mm}^3$. PD-weighted images: TE = 8ms, TR = 8178.34ms, spatial resolution = $0.94 \times 0.94 \times 1.2\mathrm{mm}^3$. The multi-contrast images in this dataset were unregistered. Hence, T_2- and PD-weighted images were spatially registered onto T_1-weighted images prior to modelling. Registration was performed via an affine transformation in FSL [96] based on mutual information.

2) BRATS Dataset: T_1-weighted, T_2-weighted, post-contrast T_2-weighted, and T_2 Fluid Attenuation Inversion Recovery (FLAIR) brain MR images from 55 subjects were analyzed. 25 subjects were reserved for training, 10 were reserved for validation, and 20 were reserved for testing. From each subject, 100 axial cross-sections containing brain tissues were selected. Please note that the BRATS dataset contains images collected under various clinical protocols and scanners at multiple institutions. As publicly shared, multi-contrast images are co-registered to the same anatomical template, interpolated to $1 \times 1 \times 1\mathrm{mm}^3$ resolution and skull-stripped.

3) MRI-CT Dataset: T_2-weighted MR and CT images of the male pelvis from 15 subjects were used. 9 subjects were reserved for training, 2 were reserved for validation, and 4 were reserved for testing. From each subject, 90 axial cross-sections were analysed. Acquisition parameters were as follows. T_2-weighted images: Group 1, TE = 97ms, TR = 6000-6600ms, spatial resolution = $0.875 \times 0.875 \times 2.5\mathrm{mm}^3$. Group 2, TE = 91-102ms, TR = 12000-16000ms, spatial resolution = $0.875\text{-}1.1 \times 0.875\text{-}1.1 \times 2.5\mathrm{mm}^3$. CT images: Group 1, spatial resolution = $0.98 \times 0.98 \times 3\mathrm{mm}^3$, Kernel = B30f. Group 2: spatial resolution = $0.1 \times 0.1 \times 2\mathrm{mm}^3$, Kernel = FC17. This dataset contains images collected under various protocols and scanners for each modality. As publicly shared, multi-modal images are co-registered onto T_2-weighted MR scans.

C. Competing Methods

We demonstrated the proposed ResViT model against several state-of-the-art image synthesis methods. The baseline methods included convolutional models (task-specific models: pGAN [24], pix2pix [43], medSynth [32]; unified models: MM-GAN [41], $\mathrm{pGAN}_{\mathrm{uni}}$), attention-augmented convolutional models (A-UNet [50], SAGAN [85]), and transformer models (task-specific: TransUNet [51], PTNet [57]; unified: $\mathrm{TransUNet}_{\mathrm{uni}}$). Hyperparameters of each competing method were optimized via identical cross-validation procedures.

1) Convolutional Models:

pGAN A convolutional GAN model with ResNet backbone was considered [24]. pGAN comprises CNN-based generator and discriminator networks. Its generator follows an encoder-bottleneck-decoder pathway, where the encoder and decoder are identical to those in ResViT. The bottleneck contains a cascade of residual CNN blocks.

pix2pix A convolutional GAN model with U-Net backbone was considered [43]. pix2pix has a CNN-based generator with an encoder-decoder structure tied with skip connections.

medSynth A convolutional GAN model with residual U-Net backbone was considered as provided at https://github.com/ginobilinie/medSynthesisV1 [32]. The generator of medSynth contains a long-skip connection from the first to the last layer.

MM-GAN A unified synthesis model based on a convolutional GAN was considered [41]. MM-GAN comprises CNN-based generator and discriminator networks, where the generator is based on U-Net. MM-GAN trains a single network under various source-target modality configurations. The original MM-GAN architecture was directly adopted, except for curriculum learning to ensure standard sample selection for all competing methods. The unification strategy in MM-GAN matches the unification strategy in ResViT.

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

2) Attention-Augmented Convolutional Models:

Attention U-Net (A-UNet) A CNN-based U-Net architecture with additive attention gates was considered [50]. Here we adopted the original A-UNet model as the generator of a conditional GAN model, where the discriminator was identical to that in ResViT.

Self-Attention GAN (SAGAN) A CNN-based GAN model with self-attention modules incorporated into the generator was considered [85]. Here we adapted the original SAGAN model designed for unconditional mapping by inserting the self-attention modules into the pGAN model as described in [97]. For fair comparison, the number and position of attention modules in SAGAN were matched to those of transformer modules in ResViT.

3) Transformer Models:

TransUNet A recent hybrid CNN-transformer architecture was considered [51]. Here, we adopted the original TransUNet model as the generator of a conditional GAN architecture with an identical discriminator to ResViT. We further replaced the segmentation head with a convolutional layer for synthesis.

PTNet A recent convolution-free transformer architecture was considered [57]. Here we adopted the original PTNet model as the generator of a conditional GAN architecture with an identical discriminator to ResViT.

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

D. Architectural Details

The encoder in the ResViT model contained three convolutional layers of kernel size 7, 3, 3 respectively. The feature map in the encoder output was of size $\mathbb{R}^{256,64,64}$, and this dimensionality was retained across the information bottleneck. The decoder contained three convolutional layers of kernel size 3, 3, 7 respectively. The information bottleneck contained nine ART blocks. The downsampling blocks preceding transformers contained two convolutional layers with stride 2 and kernel size 3. The upsampling blocks succeeding transformers contained two transposed convolutional layers with stride 2 and kernel size 3. Down and upsampling factors were set to $M = 4$. Channel compression lowered the number of channels from 512 to 256. The transformer encoder was adopted by extracting the transformer component of the ImageNet-pretrained model R50+ViT-B/16 (https://github.com/google-research/vision_transformer). The transformer encoder expected an input map of 16×16 spatial resolution. Patch flattening was performed with size $P = 1$ yielding a sequence length of 256 [49]. Note that transformer modules contain substantially higher number of parameters compared to convolutional modules. Thus, retaining a transformer in each ART block results in significant model complexity, inducing computational burden and suboptimal learning. To alleviate these issues, transformer modules in ART blocks utilized tied weights, and they were only retained in a subset of ART blocks while remaining blocks reduced to residual CNNs.

The configuration of transformer modules, i.e. their total number and position, was selected via cross-validation experiments. Due to the extensive number of potential configurations, a pre-selection process was implemented. Accordingly, performance for a transformer module inserted in a single ART block (A_1, A_2, ..., A_9) was measured, and the top half of positions was pre-selected. Composite configurations with multiple transformer modules were then formed based on the pre-selected blocks ($A_1 - A_5$, $A_1 - A_6 - A_9$ etc.). We observed that retaining more than 2 modules elevated complexity without any performance benefits. Validation performance for the best performing configurations ($A_1 - A_5$, $A_1 - A_6$, $A_1 - A_9$, $A_5 - A_9$, $A_4 - A_9$, $A_5 - A_9$, $A_1 - A_6 - A_9$) are listed in Table I for three representative tasks (T_1, $T_2 \rightarrow$ PD in IXI, T_1, $T_2 \rightarrow$ FLAIR in BRATS, and MRI $\rightarrow$ CT in MRI-CT). Consistently across tasks, the ($A_1 - A_6$) configuration yielded near-optimal performance and so it was selected for all experiments thereafter.

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

TABLE I

VALIDATION PERFORMANCE OF CANDIDATE RESVIT CONFIGURATIONS IN REPRESENTATIVE SYNTHESIS TASKS. PERFORMANCE IS TAKEN AS PSNR (dB) BETWEEN SYNTHESIZED AND REFERENCE TARGET IMAGES. A_i DENOTES THE PRESENCE OF A TRANSFORMER MODULE IN THE iTH ART BLOCK

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

TABLE II

VALIDATION PERFORMANCE OF RESVIT MODELS WITH VARYING SIZES OF TRANSFORMER MODULES IN REPRESENTATIVE SYNTHESIS TASKS

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

TABLE III

AVERAGE INFERENCE TIMES (MSEC) PER CROSS-SECTION, MODEL COMPLEXITY (MILLIONS OF PARAMETERS), AND MEMORY LOAD (GIGABYTES) FOR COMPETING METHODS

	ResViT	pGAN	pix2pix	medSynth	A-UNet	SAGAN	TransUNet	PTNet
Inference Time (msec)	98	60	60	81	70	63	78	224
Model Complexity (M)	132.49	14.14	57.18	35.21	37.64	15.30	108.08	30.80
GPU VRAM Usage (GB)	4.55	2.32	2.63	2.88	2.83	2.71	3.97	10.77

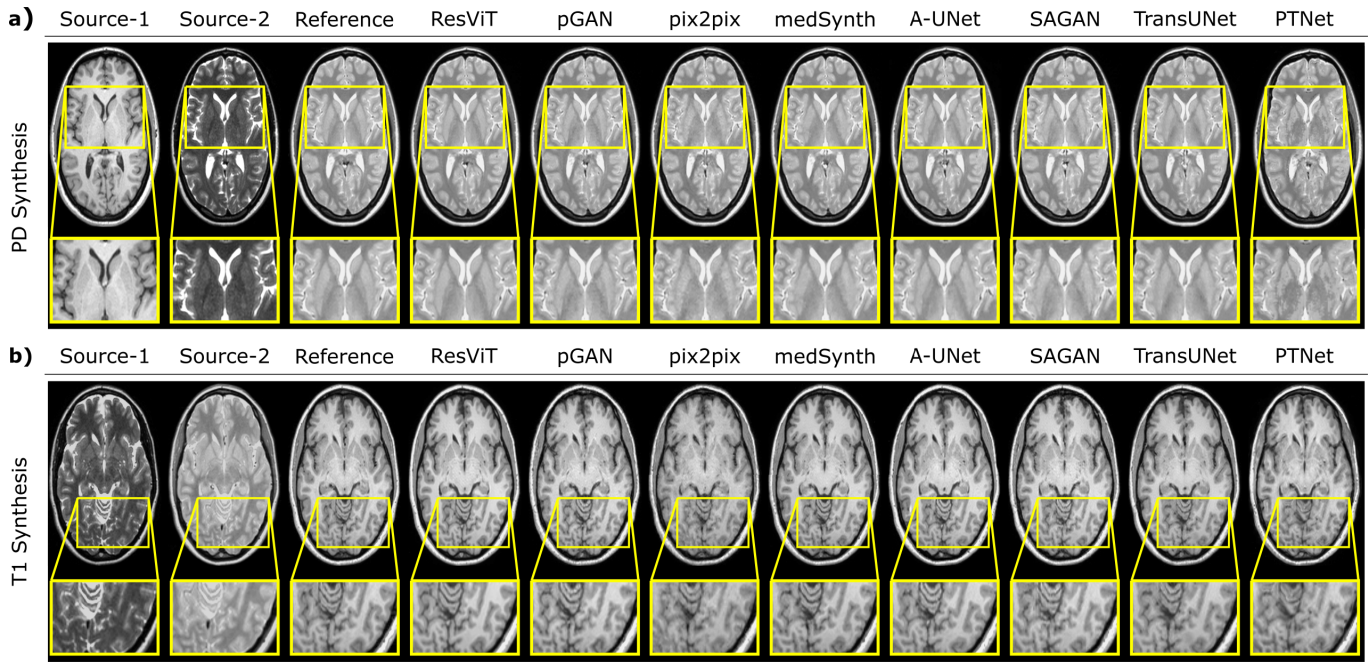

Fig. 3. ResViT was demonstrated on the IXI dataset for two representative many-to-one synthesis tasks: a) $T_1, T_2 \rightarrow PD$ b) $T_2, PD \rightarrow T_1$. Synthesized images from all competing methods are shown along with the source images and the reference target image. ResViT improves synthesis performance in regions that are depicted sub-optimally in competing methods. Overall, ResViT generates images with lower artifact and noise levels and sharper tissue depiction.

E. Modeling Procedures

For fair comparisons among competing methods, all models were implemented adversarially using the same PatchGAN discriminator and the loss function in Eq. 20. Task-specific models used adversarial and pixel-wise losses, whereas unified models used adversarial, pixel-wise, and reconstruction losses. Learning rate, number of epochs, and loss-term weighting were selected via cross-validation. Validation performance was measured as Peak Signal to Noise Ratio (PSNR) on three representative tasks (T_1, $T_2 \rightarrow$ PD in IXI, T_1, $T_2 \rightarrow$ FLAIR in BRATS, and MRI $\rightarrow$ CT in MRI-CT). We considered different learning rates in the set $\{10^{-5}, 10^{-4}, 2\text{x}10^{-4}, 5\text{x}10^{-4}, 10^{-3}\}$ and number of epochs in the set {5, 10, …, 200}. Eq. 20 contains only two degrees of freedom regarding the loss-term weights, and prior studies have reported models with higher weighting for pixel-wise over adversarial loss [24], [40]. Thus, we considered λ_{pix} in {20, 50, 100, 150} and $\lambda_{adv} = 1$. Note that $\lambda_{rec} = 0$ by definition in task-specific models with fixed configuration of source and target modalities, while $\lambda_{rec} = \lambda_{pix}$ was used in unified models as both loss terms measure the L_1-norm difference between reference and generated images for individual modalities. To minimize potential biases among competing methods, a common set of parameters that consistently yielded near-optimal performance were prescribed for all methods: 2×10^{-4} learning rate, 100 training epochs, $\lambda_{adv} = 1, \lambda_{pix} = 100$ for task-specific models, and $\lambda_{adv} = 1, \lambda_{rec} = 100, \lambda_{pix} = 100$ for unified models. All competing methods were trained via the Adam optimizer [98] with $\beta_1 = 0.5$, $\beta_2 = 0.999$. The learning rate was constant for the first 50 epochs and linearly decayed to 0 in the remaining epochs. Transformer modules in TransUNet and ResViT were initiated with ImageNet pre-trained versions for object classification [99]. ART blocks were initiated without transformer modules and then fine-tuned for 50 epochs following insertion of transformers at a higher learning rate of 10^{-3} as in [49]. Elevated learning rate during the second half of the training procedure was not adopted for other methods as it diminished performance. Modelling was performed via the PyTorch framework on Nvidia RTX A4000 GPUs. Inference times, model complexity, and memory load for all methods are listed in Table III. The hybrid ResViT and TransUNet models have comparable inference times and memory usage with pure convolutional architectures, while incurring notably higher model complexity due to the dense connections in self-attention and MLP layers. Although the convolution-free PTNet model uses an efficient attention operator to mitigate model complexity, it has significantly higher memory use and inference time compared to remaining models.

Synthesis quality was assessed via PSNR and Structural Similarity Index (SSIM) [100]. Metrics were calculated between ground truth and synthesized target images. Mean and standard deviations of metrics were reported across an independent test set, non-overlapping with training-validation sets. Significance of performance differences were evaluated with signed-rank tests ($p<0.05$). Tests were conducted on subject-average metrics, except MRI $\rightarrow$ CT where cross-sectional metrics were tested in each subject due to limited number of test subjects.

F. Experiments

1) Multi-Contrast MRI Synthesis: Experiments were conducted on the IXI and BRATS datasets to demonstrate synthesis performance in multi-modal MRI. In the IXI dataset, one-to-one tasks of $T_2 \rightarrow$ PD; PD $\rightarrow T_2$ and

TABLE IV
PERFORMANCE OF TASK-SPECIFIC SYNTHESIS MODELS IN MANY-TO-ONE (T_1, T_2 → PD, T_1, PD → T_2, AND T_2, PD → T_1) AND ONE-TO-ONE (T_2 → PDAND PD → T_2) TASKS IN THE IXI DATASET. PSNR (DB) AND SSIM ARE LISTED AS MEAN±STD ACROSS TEST SUBJECTS. BOLDFACE INDICATES THE TOP-PERFORMING MODEL FOR EACH TASK

	T_1, T_2 → PD		T_1, PD → T_2		T_2, PD → T_1		T_2 → PD		PD → T_2	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
ResViT	**33.92 ±1.44**	**0.977 ±0.004**	**35.71 ±1.20**	**0.977 ±0.005**	**29.58 ±1.37**	**0.952 ±0.011**	**32.90 ±1.20**	**0.972 ±0.005**	**34.24 ±1.09**	**0.972 ±0.005**
pGAN	32.91 ±0.94	0.966 ±0.005	33.95 ±1.06	0.965 ±0.006	28.71 ±1.08	0.941 ±0.013	32.20 ±1.00	0.963 ±0.005	33.05 ±0.95	0.963 ±0.007
pix2pix	32.25 ±1.24	0.974 ±0.006	33.62 ±1.31	0.973 ±0.009	28.35 ±1.24	0.949 ±0.016	30.72 ±1.28	0.956 ±0.007	30.74 ±1.63	0.950 ±0.012
medSynth	33.23 ±1.09	0.967 ±0.005	32.66 ±1.30	0.963 ±0.007	28.43 ±1.01	0.938 ±0.013	32.20 ±1.10	0.964 ±0.006	30.41 ±3.98	0.956 ±0.025
A-UNet	32.24 ±0.92	0.963 ±0.014	32.43 ±1.36	0.959 ±0.007	28.95 ±1.21	0.916 ±0.013	32.05 ±1.04	0.960 ±0.009	33.32 ±1.08	0.961 ±0.007
SAGAN	32.50 ±0.93	0.964 ±0.005	33.71 ±1.00	0.965 ±0.006	28.62 ±1.10	0.942 ±0.013	32.07 ±0.98	0.963 ±0.006	32.96 ±1.01	0.962 ±0.007
TransUNet	32.53 ±0.97	0.968 ±0.005	32.49 ±1.18	0.960 ±0.008	28.21 ±1.30	0.941 ±0.013	30.90 ±1.35	0.960 ±0.006	31.73 ±1.44	0.958 ±0.008
PTNet	30.92 ± 0.99	0.952 ±0.006	32.62 ±1.96	0.954 ±0.019	27.59 ±1.36	0.923 ±0.021	31.58 ±1.30	0.958 ±0.007	30.84 ±2.54	0.947 ±0.033

TABLE V
PERFORMANCE OF TASK-SPECIFIC SYNTHESIS MODELS IN MANY-TO-ONE TASKS (T_1, T_2 → FLAIR, T_1, FLAIR → T_2, AND T_2, FLAIR → T_1) AND ONE-TO-ONE TASKS (T_2→FLAIR AND FLAIR→T_2) ACROSS TEST SUBJECTS IN THE BRATS DATASET. BOLDFACE INDICATES THE TOP-PERFORMING MODEL FOR EACH TASK

	T_1, T_2 → FLAIR		T_1, FLAIR → T_2		T_2, FLAIR → T_1		T_2→FLAIR		FLAIR→T_2	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
ResViT	**25.84 ±1.13**	**0.886 ±0.014**	**26.90 ±1.20**	**0.938 ±0.011**	**26.20 ±1.31**	**0.924 ±0.009**	**24.97 ±1.07**	0.870 ±0.014	**25.78 ±0.92**	**0.908 ±0.015**
pGAN	24.89 ±1.10	0.867 ±0.015	26.51 ±1.13	0.922 ±0.012	25.72 ±1.54	0.918 ±0.011	24.01 ±1.15	0.864 ±0.014	25.09 ±1.52	0.894 ±0.015
pix2pix	24.31 ±1.21	0.862 ±0.015	26.12 ±1.53	0.920 ±0.012	25.80 ±1.72	0.918 ±0.011	23.15 ±1.93	0.869 ±0.016	24.52 ±0.88	0.883 ±0.014
medSynth	23.93 ±1.45	0.863 ±0.016	26.44 ±0.76	0.921 ±0.011	25.72 ±1.62	0.914 ±0.012	23.36 ±1.88	0.864 ±0.017	24.41 ±0.82	0.888 ±0.014
A-UNet	24.36 ±1.24	0.857 ±0.017	26.48 ±1.21	0.924 ±0.012	25.67 ±1.35	0.918 ±0.010	23.69 ±1.57	**0.873 ±0.015**	24.56 ±0.94	0.891 ±0.014
SAGAN	24.62 ±1.17	0.869 ±0.014	26.41 ±1.22	0.919 ±0.012	25.91 ±1.42	0.918 ±0.011	24.02 ±1.35	0.860 ±0.015	25.10 ±0.88	0.893 ±0.014
TransUNet	24.34 ±1.26	0.872 ±0.014	26.51 ±0.92	0.920 ±0.010	25.76 ±1.69	0.921 ±0.011	23.70 ±1.75	0.864 ±0.015	24.62 ±0.81	0.891 ±0.015
PTNet	23.78 ±1.24	0.851 ±0.031	25.09 ±1.23	0.905 ±0.016	22.19 ±1.88	0.920 ±0.014	23.01 ±0.85	0.851 ±0.014	24.78 ±0.88	0.894 ±0.015

many-to-one tasks of T_1, T_2 → PD; T_1, PD → T_2; T_2, PD → T_1 were considered. In the BRATS dataset, one-to-one tasks of T_2→FLAIR; FLAIR→T_2, many-to-one tasks of T_1, T_2 → FLAIR; T_1, FLAIR → T_2; T_2, FLAIR → T_1 were considered. In both datasets, task-specific ResViT models were compared against pGAN, pix2pix, medSynth, A-UNet, SAGAN, TransUNet, and PTNet. Meanwhile, unified ResViT models were demonstrated against pGAN_{uni}, MM-GAN, and $\text{TransUNet}_{\text{uni}}$.

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

3) Ablation Studies: Several lines of ablation experiments were conducted to demonstrate the value of the individual components of the ResViT model, including both architectural design elements and training strategies. Experiments were performed on three representative tasks: namely T_1, T_2 → PD in IXI, T_1, T_2 → FLAIR in BRATS, and MRI → CT. First, we assessed the performance contribution of the three main components in ResViT: transformer modules, convolutional modules and adversarial learning. Variant models were trained when transformer modules were ablated from ART blocks, when residual CNNs were ablated from transformer-retaining ART blocks, and when the adversarial loss term and the discriminator were ablated. In addition to PSNR and SSIM, we measured the Fréchet inception distance (FID) [101] between the synthesized and ground truth images to evaluate the importance of adversarial learning.

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

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

Next, we inspected the relative strength of contextual features in the distilled task-relevant representations in ART blocks. For a quantitative assessment, we compared the L_2-norm of the contextual feature map derived by the transformer module against that of the input feature map to the ART block relayed through the transformer's skip connection. Note that these two maps are distilled via the channel compression (CC) module following concatenation. Thus, we also compared the L_2-norm of the combination weights in the CC module for the contextual versus input features.

To interpret the information that self-attention mechanisms focus on during synthesis tasks, we computed and visualized the attention maps as captured by the transformer modules in ResViT. Attention maps were calculated based on the Attention Rollout technique, and a single average map was extracted for a given transformer module [102].

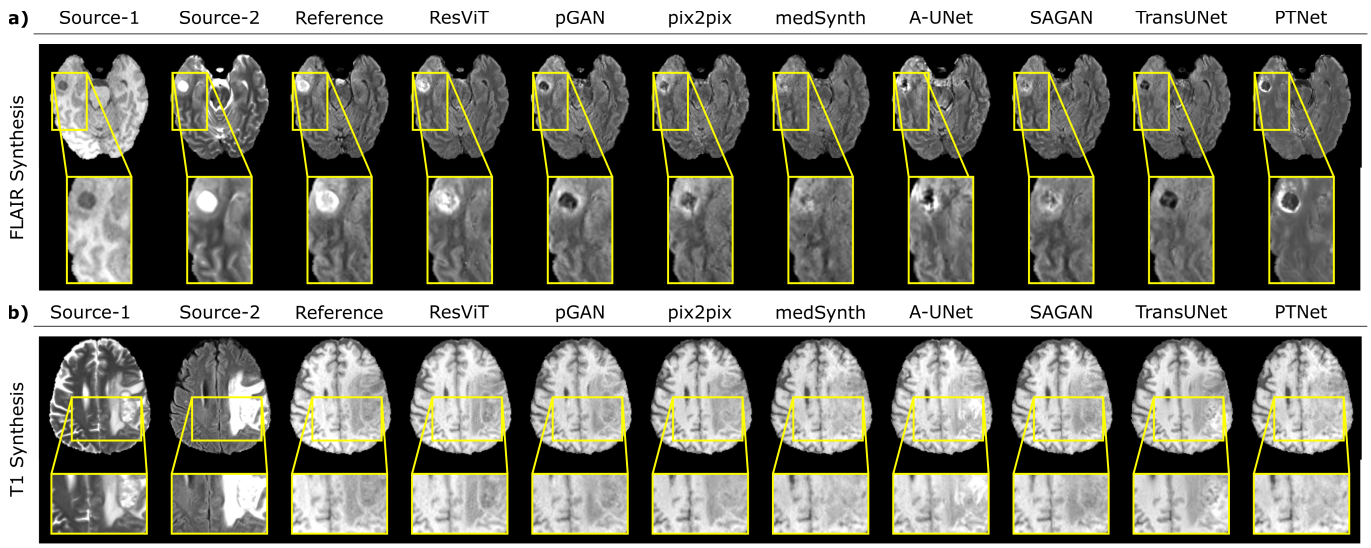

Fig. 4. ResViT was demonstrated on the BRATS dataset for two representative many-to-one synthesis tasks: a) T_1, T_2 → FLAIR, b) T_2, FLAIR → T_1. Synthesized images from all competing methods are shown along with the source images and the reference image. ResViT improves synthesis performance, especially in pathological regions (e.g., tumors, lesions) in comparison to competing methods. Overall, ResViT images have better-delineated tissue boundaries and lower artifact/noise levels.

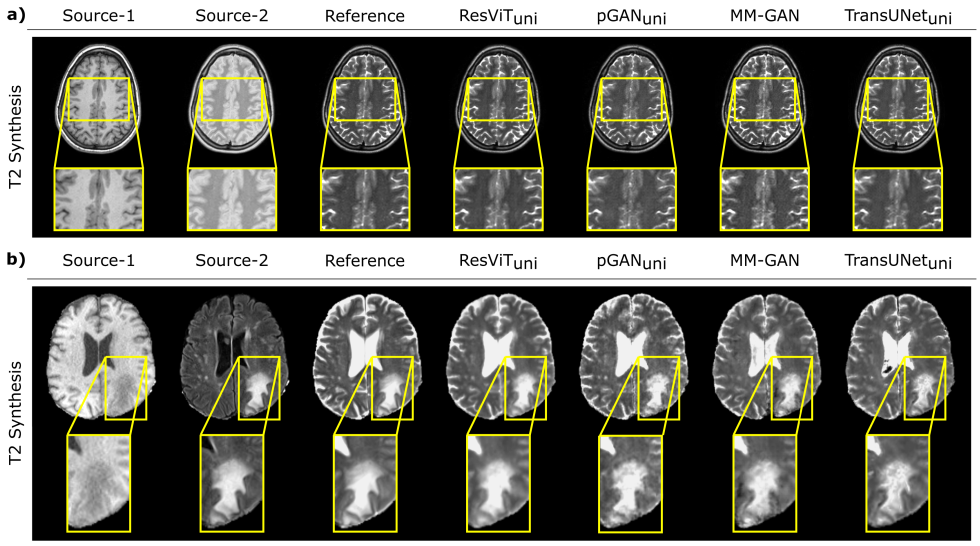

Fig. 5. ResViT$_{uni}$ was demonstrated against other unified models on brain MRI datasets for two representative tasks: a) T_1, PD → T_2 in IXI, b) T_1, FLAIR → T_2 in BRATS. Synthesized images from all competing methods are shown along with the source images and the reference target image. ResViT$_{uni}$ improves synthesis performance especially in pathological regions (tumors, lesions) in comparison to competing methods. Overall, ResViT$_{uni}$ generates images with lower artifact and noise levels and more accurate tissue depiction for tasks in both datasets.

IV. Results

A. Multi-Contrast MRI Synthesis

1) Task-Specific Synthesis Models: We demonstrated the performance of ResViT in learning task-specific synthesis models for multi-contrast MRI. ResViT was compared against convolutional models (pGAN, pix2pix, medSynth), attention-augmented CNNs (A-UNet, SAGAN), and recent transformer architectures (TransUNet, PTNet). First, brain images of healthy subjects in the IXI dataset were considered. PSNR and SSIM metrics are listed in Table IV for many-to-one and one-to-one tasks. ResViT achieves the highest performance in both many-to-one ($p<0.05$) and one-to-one tasks ($p<0.05$). On average, ResViT outperforms convolutional models by 1.71dB PSNR and 1.08% SSIM, attention-augmented models by 1.40dB PSNR and 1.45% SSIM, and transformer models by 2.33dB PSNR and 1.79% SSIM ($p<0.05$). Representative images for T_1, T_2 → PD and T_2, PD → T_1 are displayed in Fig. 3a,b. Compared to baselines, ResViT synthesizes target images with lower artifact levels and sharper tissue depiction.

We then demonstrated task-specific ResViT models on the BRATS dataset containing images of glioma patients. PSNR and SSIM metrics are listed in Table V for many-to-one and one-to-one tasks. ResViT again achieves the highest performance in many-to-one ($p<0.05$) and one-to-one tasks ($p<0.05$), except T_2→FLAIR where A-UNet has slightly higher SSIM. On average, ResViT outperforms convolutional models by 1.01dB PSNR and 1.41% SSIM, attention-augmented models by 0.84dB PSNR and 1.24% SSIM, and transformer models by 1.56dB PSNR and 1.63% SSIM ($p<0.05$). Note that the BRATS dataset contains pathology with large across-subject variability.

TABLE VI
PERFORMANCE OF UNIFIED SYNTHESIS MODELS IN MANY-TO-ONE TASKS (T_1, T_2 → PD, T_1, PD → T_2, AND T_2, PD → T_1) ACROSS TEST SUBJECTS IN THE IXI DATASET. BOLDFACE INDICATES THE TOP-PERFORMING MODEL FOR EACH TASK

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

TABLE VII
PERFORMANCE OF UNIFIED SYNTHESIS MODELS IN MANY-TO-ONE TASKS (T_1, T_2 → FLAIR, T_1, FLAIR → T_2, AND T_2, FLAIR → T_1) ACROSS TEST SUBJECTS IN THE BRATS DATASET. BOLDFACE INDICATES THE TOP-PERFORMING MODEL FOR EACH TASK

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

As expected, attention-augmented models show relative benefits against pure convolutional models, yet ResViT that explicitly models contextual relationships still outperforms all baselines. Representative target images for T_1, T_2 → FLAIR and T_2, FLAIR → T_1 are displayed in Fig. 3a,b, respectively. Compared to baselines, ResViT synthesizes target images with lower artifact levels and sharper tissue depiction. Importantly, ResViT reliably captures brain lesions in patients in contrast to competing methods with inaccurate depictions including TransUNet.

Superior depiction of pathology in ResViT signals the importance of ART blocks in simultaneously maintaining local precision and contextual consistency in medical image synthesis. In comparison, transformer-based TransUNet and PTNet yield relatively limited synthesis quality that might be attributed to several fundamental differences between the models. TransUNet uses only a transformer in its bottleneck while propagating shallow convolutional features via encoder-decoder skip connections, and its decoder increases spatial resolution via bilinear upsampling that might be ineffective in suppressing high-frequency artifacts [103]. In contrast, ResViT continues encoding and propagating convolutional features across the information bottleneck to create a deeper feature representation, and it employs transposed convolutions within upsampling modules to mitigate potential artifacts. PTNet is a convolution-free architecture that relies solely on self-attention operators that have limited localization ability [49]. Instead, ResViT is devised as a hybrid CNN-transformer architecture to improve sensitivity for both local and contextual features.

2) Unified Synthesis Models: Task-specific models are trained and tested to perform a single synthesis task to improve performance, but a separate model has to be built for each task. Next, we demonstrated ResViT in learning unified synthesis models for multi-contrast MRI. A unified ResViT ($ResViT_{uni}$) was compared against unified convolutional ($pGAN_{uni}$, MM-GAN) and transformer models ($TransUNet_{uni}$). Performance of unified models were evaluated at test time on many-to-one tasks in IXI (Table VI) and BRATS (Table VII). $ResViT_{uni}$ maintains the highest performance in many-to-one tasks in both IXI ($p<0.05$) and BRATS ($p<0.05$). In IXI, $ResViT_{uni}$ outperforms $pGAN_{uni}$ by 1.12dB PSNR and 0.70% SSIM, MM-GAN by 2.37dB PSNR and 1.80% SSIM, and $TransUNet_{uni}$ by 2.69dB PSNR and 1.67% SSIM ($p<0.05$). In BRATS, ResViT outperforms $pGAN_{uni}$ by 0.74dB PSNR and 0.93% SSIM, MM-GAN by 0.77dB PSNR and 0.90% SSIM, and $TransUNet_{uni}$ by 1.08dB PSNR and 1.43% SSIM ($p<0.05$). Representative target images are displayed in Fig. 5. ResViT synthesizes target images with lower artifacts and sharper depiction than baselines. These results suggest that a unified ResViT model can successfully consolidate models for varying source-target configurations.

TABLE VIII
PERFORMANCE FOR THE ACROSS-MODALITY SYNTHESIS TASK (T_2-WEIGHTED MRI → CT) ACROSS TEST SUBJECTS IN THE PELVIC MRI-CT DATASET. BOLDFACE INDICATES THE TOP-PERFORMING MODEL FOR EACH TASK

		ResViT	pGAN	pix2pix	medSynth	A-UNet	SAGAN	TransUNet	PTNet
MRI → CT	PSNR	**28.45 ±1.35**	26.80 ±0.90	26.53 ±0.45	26.36 ±0.63	27.80 ±0.63	27.61 ±1.02	27.76 ±1.03	26.11 ±0.93
	SSIM	**0.931 ±0.009**	0.905 ±0.008	0.898 ±0.004	0.894 ±0.009	0.913 ±0.004	0.910 ±0.006	0.914 ±0.009	0.900 ±0.015

B. Across-Modality Synthesis

We also demonstrated ResViT in across-modality synthesis. T_2-weighted MRI and CT images in the pelvic dataset were considered. ResViT was compared against pGAN, pix2pix, medSynth, A-UNet, SAGAN, TransUNet, and PTNet. PSNR and SSIM metrics are listed in Table VIII. ResViT yields the highest performance in each subject ($p<0.05$). On average, ResViT outperforms convolutional models by 1.89dB PSNR and 3.20% SSIM, attention-augmented models by 0.75dB PSNR and 1.95% SSIM, and transformer models by 1.52dB PSNR and 2.40% SSIM ($p<0.05$). Representative target images are displayed in Fig. 6. Compared to baselines, ResViT synthesizes target images with lower artifacts and more accurate tissue depiction. Differently from multi-contrast MRI, attention-augmented models and TransUNet offer more noticeable performance benefits over convolutional models. That said, ResViT still maintains further elevated performance, particularly near bone structures in CT images. This finding suggests that the relative importance of contextual representations is higher in MRI-CT synthesis. With the help of its residual transformer blocks, ResViT offers reliable performance with accurate tissue depiction in this task.

C. Ablation Studies

We performed a systematic set of experiments to demonstrate the added value of the main components and training

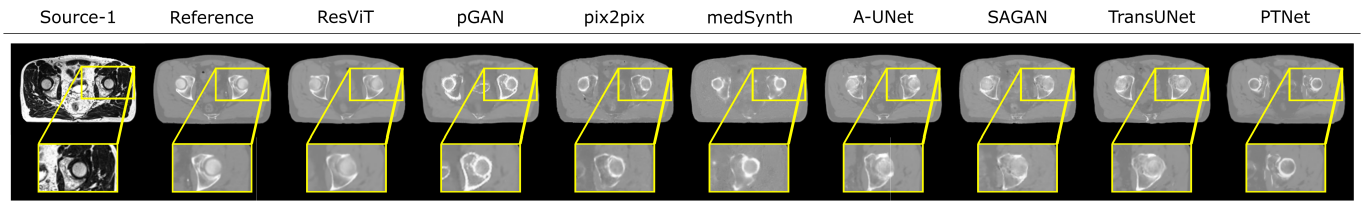

Fig. 6. ResViT was demonstrated on the pelvic MRI-CT dataset for the T_2-weighted MRI → CT task. Synthesized images from all competing methods are shown along with the source and reference images. ResViT enhances synthesis of relevant morphology in the CT domain as evidenced by the elevated accuracy near bone structures.

strategies used in ResViT. First, we compared ResViT against ablated variants where the convolutional modules in ART blocks, transformer modules in ART blocks, or the adversarial term in training loss were separately removed. Table IX lists performance metrics in the test set for three representative synthesis tasks. Consistently across tasks, ResViT yields optimal or near-optimal performance. ResViT achieves higher PSNR and SSIM in representative tasks compared to variants without transformer or convolutional modules ($p<0.05$). It also yields lower FID than these variants, except in MRI → CT where ablation of the convolutional module slightly decreases FID. Importantly, ResViT maintains notably lower FID compared to the variant without adversarial loss (albeit slightly lower SSIM in T_1, T_2 → FLAIR and PSNR, SSIM in MRI → CT). This is expected since FID is generally considered as a more suited metric to examine the perceptual benefits of adversarial learning than PSNR or SSIM that reflect heavier influence from relatively lower frequencies [101]. Representative synthesized images are also displayed in Fig. 7a. ResViT images more closely mimic the reference images, and show greater spatial acuity compared against the variant without adversarial loss. Taken together, these results indicate that adversarial learning enables ResViT to more closely capture the distributional properties of target-modality images.

Second, we compared ResViT against ablated variants where the weight tying procedure across transformer modules was neglected, or transformer modules in one of the two retaining ART blocks were removed. Table X lists performance metrics in the test set. ResViT yields higher performance than variants across representative tasks ($p<0.05$), except for the variant only retraining A_6 that yields similar SSIM in T_1, T_2 → PD. These results demonstrate the added value of the weight tying procedure and the transformer configuration in ResViT. We also compared ResViT against ablated variants where the pre-training of transformer modules or their delayed insertion during training were selectively neglected, as listed in Table XI. Our results indicate that ResViT outperforms all ablated variants ($p<0.05$), except for the variant without delayed insertion that yields similar SSIM in T_1, T_2 → PD.

Third, we examined the utility of the skip connections and down/upsampling blocks in the proposed architecture. We compared ResViT against variants built by removing the skip connection around the transformer module or around the CNN module in ART blocks. Table XII lists performance metrics in the test set. ResViT yields higher performance than all variants ($p<0.05$). Our results indicate that ResViT benefits substantially from residual learning in ART blocks. We also compared ResViT against variants built by replacing down/upsampling modules in ART blocks with unlearned maxpooling/bilinear interpolation modules, and by increasing encoder downsampling and decoder upsampling rates to remove down/upsampling modules in ART blocks entirely. ResViT outperforms all variants as listed in Table XII ($p<0.05$), except for MRI→CT where the variant with unlearned down/upsampling and ResViT yield similar

TABLE IX
TEST PERFORMANCE OF RESVIT AND VARIANTS ABLATED OF TRANSFORMER MODULES, CONVOLUTIONAL MODULES OR ADVERSARIAL LOSS. FID IS A SUMMARY METRIC ACROSS THE ENTIRE TEST SET. BOLDFACE INDICATES THE TOP-PERFORMING MODEL FOR EACH TASK

	T_1, T_2 → PD			T_1, T_2 → FLAIR			MRI → CT		
	PSNR	SSIM	FID	PSNR	SSIM	FID	PSNR	SSIM	FID
ResViT	**33.92 ±1.44**	**0.977 ±0.004**	**14.47**	**25.84 ±1.13**	0.886 ±0.014	**18.58**	28.45 ±1.35	0.931 ±0.009	60.28
w/o trans. modules	32.91 ±0.96	0.966 ±0.005	14.56	24.96 ±1.10	0.868 ±0.005	19.21	26.73 ±0.91	0.899 ±0.008	95.38
w/o conv. modules	33.49 ±1.34	0.971 ±0.005	14.84	25.11 ±1.02	0.874 ±0.014	20.30	28.19 ±1.15	0.922 ±0.009	**60.16**
w/o adv. loss	33.75 ±1.45	**0.977 ±0.005**	15.80	22.95 ±1.93	**0.891 ±0.015**	40.68	**28.58 ±1.13**	**0.932 ±0.007**	65.49

TABLE X
TEST PERFORMANCE OF RESVIT ($A_1 - A_6$) AND VARIANTS ABLATED OF WEIGHT TYING AND INDIVIDUAL TRANSFORMER MODULES. BOLDFACE INDICATES THE TOP-PERFORMING MODEL FOR EACH TASK

	T_1, T_2 → PD		T_1, T_2 → FLAIR		MRI → CT	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
$A_1 - A_6$	**33.92 ±1.44**	**0.977 ±0.004**	**25.84 ±1.13**	**0.886 ±0.014**	**28.45 ±1.35**	**0.931 ±0.009**
$A_1 - A_6$ (untied weights)	33.72 ±1.23	0.973 ±0.005	25.19 ±1.18	0.879 ± 0.014	28.16 ±1.11	0.923 ±0.007
A_1	33.51 ±1.15	0.971 ±0.005	24.98 ±1.60	0.883 ±0.015	28.06 ±1.31	0.921 ±0.008
A_6	33.78 ±1.34	**0.977 ±0.004**	25.25 ±1.20	0.880 ±0.014	27.95 ±1.22	0.921 ±0.008

TABLE XI
TEST PERFORMANCE OF RESVIT AND VARIANTS ABLATED OF PRE-TRAINING AND DELAYED INSERTION PROCEDURES FOR TRANSFORMERS. BOLDFACE INDICATES THE TOP-PERFORMING MODEL FOR EACH TASK

	T_1, T_2 → PD		T_1, T_2 → FLAIR		MRI → CT	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
ResViT	**33.92 ±1.44**	**0.977 ±0.004**	**25.84 ±1.13**	**0.886 ±0.014**	**28.45 ±1.35**	**0.931 ±0.009**
w/o pre-training	33.55 ±1.25	0.971 ±0.005	24.86 ±1.28	0.881 ± 0.016	27.94 ±1.25	0.912 ±0.009
w/o del. insertion	33.35 ±1.13	**0.977 ±0.004**	24.89 ±1.18	0.873 ±0.015	28.01 ±1.27	0.924 ±0.008
w/o pre-training or del. insertion	33.58 ±1.16	0.971 ±0.005	24.74 ±1.30	0.869 ±0.016	27.66 ±0.78	0.913 ±0.006

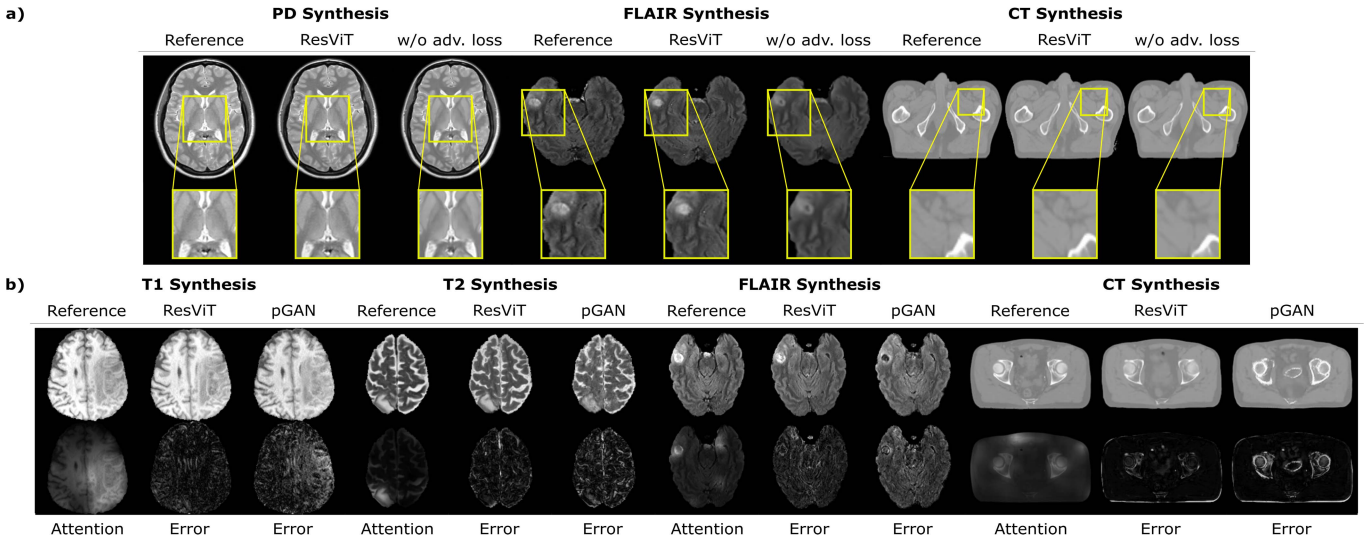

Fig. 7. a) ResViT was compared against a variant where the adversarial term was removed from the loss function. Representative results are shown for T_1, $T_2 \rightarrow$ PD in IXI, T_1, $T_2 \rightarrow$ FLAIR in BRATS, and MRI $\rightarrow$ CT in the pelvic dataset. Adversarial loss improves the acuity of synthesized images. b) Representative results from ResViT and pGAN are shown along with the reference images for T_2, FLAIR $\rightarrow$ T_1, T_1, FLAIR $\rightarrow$ T_2, and T_1, $T_2 \rightarrow$ FLAIR in BRATS; and MRI $\rightarrow$ CT in the pelvic dataset. Error maps between the synthetic and reference images for each method are displayed, along with the attention map for the first transformer module of ResViT. Here, the attention maps were overlaid onto the reference image for improved visualization. Attention maps focus on image regions where ResViT substantially reduces synthesis errors compared to pGAN.

TABLE XII

TEST PERFORMANCE OF RESVIT AND VARIANTS BUILT BY: REMOVING SKIP CONNECTIONS IN CONVOLUTIONAL MODULES, REMOVING SKIP CONNECTIONS IN TRANSFORMER MODULES, USING UNLEARNED DOWN/UPSAMPLING BLOCKS IN ART, REMOVING DOWN/UPSAMPLING BLOCKS IN ART VIA A HIGHER DEGREE OF DOWN/UPSAMPLING IN THE ENCODER/DECODER. BOLDFACE INDICATES THE TOP-PERFORMING MODEL FOR EACH TASK

	T_1, $T_2 \rightarrow$ PD		T_1, $T_2 \rightarrow$ FLAIR		MRI $\rightarrow$ CT	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
ResViT	**33.92 ±1.44**	**0.977 ±0.004**	**25.84 ±1.13**	**0.886 ±0.014**	**28.45 ±1.35**	**0.931 ±0.009**
w/o skip around conv. modules	28.24 ±1.27	0.942 ±0.009	25.02 ±0.98	0.864 ±0.016	26.94 ±0.73	0.906 ±0.007
w/o skip around trans. modules	31.53 ±1.26	0.962 ± 0.006	24.06 ±1.28	0.868 ±0.014	27.08 ±0.80	0.908 ±0.006
ART with unlearned down/upsampling	33.73 ±1.19	0.969 ±0.005	25.33 ±1.11	0.884 ±0.014	28.16 ±1.04	**0.931 ±0.007**
ART w/o down/upsampling	31.51 ±1.27	0.961 ±0.006	23.61 ±1.53	0.867 ±0.015	26.79 ±0.62	0.915 ±0.006

SSIM. These results demonstrate the benefits of the proposed down/upsampling scheme in ResViT.

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

Lastly, we wanted to visually interpret the benefits of the self-attention mechanisms in ResViT towards synthesis performance. Fig. 7b displays representative attention maps in ResViT. Synthetic images and error maps are also shown for ResViT as well as pGAN, which generally offered the closest performance to ResViT in our experiments. We find that the attention maps exhibit higher intensity in critical regions such as brain lesions in multi-contrast MRI and pelvic bone structure in MR-to-CT synthesis. Importantly, these regions of higher attentional focus are also the primary regions where the synthesis errors are substantially diminished with ResViT compared to pGAN. Taken together, these results suggest that the transformer-based ResViT model captures contextual relationships related to both healthy and pathological tissues to improve synthesis performance.

V. DISCUSSION

In this study, we proposed a novel adversarial model for image translation between separate modalities. Traditional GANs employ convolutional operators that have limited ability to capture long-range relationships among distant regions [46]. The proposed model aggregates convolutional and transformer branches within a residual bottleneck to preserve both local precision and contextual sensitivity. To our knowledge, this is the first adversarial model for medical image synthesis with a transformer-based generator. We further introduced a weight-sharing strategy among transformer modules to lower model complexity. Finally, a unification strategy was implemented to learn an aggregate model that copes with numerous source-target configurations without training separate models.

We demonstrated ResViT for missing modality synthesis in multi-contrast MRI and MRI-CT imaging. ResViT outperformed several state-of-the-art convolutional and transformer models in one-to-one and many-to-one tasks. We trained all models with an identical loss function to focus on architectural

TABLE XIII
FEATURE MAPS AND CORRESPONDING COMBINATION WEIGHTS FOR THE CHANNEL COMPRESSION (CC) MODULE WERE INSPECTED IN RESVIT. AVERAGED ACROSS THE TEST SET AND ART BLOCKS, L_2-NORM OF FEATURE MAPS FROM THE TRANSFORMER MODULE (g) AND FEATURE MAPS INPUT BY THE PREVIOUS ART BLOCK (f) ARE LISTED ALONG WITH COMBINATION WEIGHTS FOR g AND FOR f

	T_1, T_2 → PD	T_1, T_2 → FLAIR	MRI → CT
g	277.88	400.90	421.46
f	536.48	571.23	636.93
CC weights for g	96.5	112.04	72.88
CC weights for f	226.08	169.15	116.70

influences to synthesis performance. In unreported experiments, we also trained competing methods that were proposed with different loss functions using their original losses, including PTNet with mean-squared loss [57] and medSynth with mean-squared, adversarial and gradient-difference losses [32]. We observed that ResViT still maintains similar performance benefits over competing methods in these experiments. Yet, it remains important future work to conduct an in-depth assessment of optimal loss terms for ResViT, including gradient-difference and difficulty-aware losses for the generator [32], [104], [105], and edge-preservation and binary cross-entropy losses for the discriminator [105], [106].

Trained with image-average loss terms, CNNs have difficulty in coping with atypical anatomy that substantially varies across subjects [24], [43]. To improve generalization, recent studies have proposed self-attention mechanisms in GAN models over spatial or channel dimensions [50], [85]. Specifically, attention maps are used for multiplicative modulation of CNN-derived feature maps. This modulation encourages the network to focus on critical image regions with relatively limited task performance. While attention maps can be distributed across image regions, they mainly capture implicit contextual information via modification of local CNN features. Since feature representations are primarily extracted via convolutional filtering, the resulting model can still manifest limited expressiveness for global context. In contrast, the proposed architecture uses dedicated transformer blocks to explicitly model long-range spatial interactions in medical images.

Few recent studies have independently proposed transformer-based models for medical image synthesis [55]–[57]. In [56], a transformer is included in the discriminator of a traditional GAN for MR-to-PET synthesis. In [57], a UNet-inspired transformer architecture is proposed for infant MRI synthesis [57]. Differing from these efforts, our work makes the following contributions. (1) Compared to [56] that uses transformers to learn a prior for target PET images, we employ transformers in ResViT's generator to learn latent contextual representations of source images. (2) Unlike [57] that uses mean-squared error loss amenable to over-smoothing of target images [24], we leverage an adversarial loss to preserve realism. (3) While [57] uses a convolution-free transformer architecture, we instead propose a hybrid architecture that combines localization capabilities of CNNs with contextual sensitivity of transformers. (4) While [56] and [57] consider only task-specific, one-to-one synthesis models, here we uniquely introduce many-to-one synthesis models and a unified model that generalizes across multiple source-target configurations.

UNet-style models follow an encoder-decoder architecture with an hourglass structure [43]. Because spatial resolution is substantially lower in the midpoint of the hourglass (e.g. 16×16 maps), these models typically introduce skip connections between the encoder and decoder layers to facilitate preservation of low-level features. In contrast, ResViT is a ResNet-style model where encoded representations pass through a bottleneck of residual blocks before reaching the decoder [58], and encoder-decoder skip connections are omitted due to several reasons. First, ResViT maintains relatively high resolution at the output of its encoder (e.g. 64×64 maps), so its bottleneck represents relatively lower-level information. Second, each ART block is organized as a transformer-CNN cascade with skip connections around both modules, creating a residual path between the input and output of each block. This eventually bridges the encoder output to the decoder input, creating a native residual path in ResViT. Lastly, we observed during early stages of the study that a variant model that included encoder-decoder skip connections caused a minor performance drop, suggesting that these extra connections might reduce the effectiveness of the central information bottleneck.

Here, ResViT models were initialized with transformers pre-trained on 16×16 input feature maps. In turn, 256×256 images were 16-fold downsampled cumulatively across the encoder and transformer modules, and the transformer used a patch size of P=1 and sequence length of 256. Several strategies can be adopted to use ResViT at different image resolutions. In a first scenario, the downsampling rate and patch size can be preserved, while the sequence length is adjusted. For instance, a 512×512 image would be downsampled to a 32×32 feature map, resulting in a sequence of 1024 patches. While a transformer pre-trained on 32×32 maps would be ideal, vision transformers can reliably handle variable sequence lengths without retraining so the original transformer can still be used [49]. Note that longer sequences would incur a quadratic increase in processing and memory load in both cases [49]. In a second scenario, the original transformer with sequence length 256 can be maintained, while either the patch size or the downsampling rate is adjusted. For a 512×512 image, P=2 (2×2 patches) on a 32×32 map (16-fold downsampled) or P=1 on a 16×16 map (32-fold downsampled) could be used. Both options would achieve on par computational complexity to the original architecture, albeit the transformer would process feature maps at a relatively lower resolution compared to the resolution of the input image. It is unlikely that this would significantly affect ResViT's sensitivity to local features since the primary component of ART that captures local features is the residual CNN module whose resolution can be preserved. If the input image does not have a power-of-two size, the abovementioned strategies can be adopted after zero-padding to round up the resolution to the nearest power of two, or by implementing the encoder with non-integer downsampling rates [107]. Note that computer vision studies routinely fine-tune transformers at different image resolutions than encountered during pre-training without performance loss [49],

so ResViT might also demonstrate similar behavior. It remains important future work to investigate the comparative utility of the discussed resolution-adaptation strategies in medical image synthesis.

Several lines of development can help further improve ResViT's performance. Here, we considered synthesis tasks in which source and target modalities were registered prior to training, and they were paired across subjects. When registration accuracy is limited, a spatial registration block can be incorporated into the network. Furthermore, a cycle-consistency loss [44] can be incorporated in the optimization objective to allow the use of unregistered images. This latter strategy would also permit training of ResViT models on unpaired datasets [76], [77]. Data requirements for model training can be further alleviated by adopting semi-supervised strategies that allow mixing of paired and unpaired training data [75], or that would enable training of synthesis models directly from undersampled acquisitions [108]. Finally, ResViT might benefit from incorporation of multi-scale modules in the decoder to improve preservation of fine image details [106].

VI. Conclusion

Here we introduced a novel synthesis approach for multi-modal imaging based on a conditional deep adversarial network. In an information bottleneck, ResViT aggregates convolutional operators and vision transformers, thereby improving capture of contextual relations while maintaining localization power. A unified implementation was introduced that prevents the need to rebuild models for varying source-target configurations. ResViT achieves superior synthesis quality to state-of-the-art approaches in multi-contrast brain MRI and multi-modal pelvic MRI-CT datasets. Therefore, it holds promise as a powerful candidate for medical image synthesis.

References

[1] B. J. Pichler, M. S. Judenhofer, and C. Pfannenberg, *Multimodal Imaging Approaches: PET/CT and PET/MRI*. Berlin, Germany: Springer, 2008, pp. 109–132.

[2] B. Moraal *et al.*, "Multi-contrast, isotropic, single-slab 3D MR imaging in multiple sclerosis," *Eur. Radiol.*, vol. 18, no. 10, pp. 2311–2320, Oct. 2008.

[3] B. B. Thukral, "Problems and preferences in pediatric imaging," *Indian J. Radiol. Imag.*, vol. 25, no. 4, pp. 359–364, Oct. 2015.

[4] K. Krupa and M. Bekiesińska-Figatowska, "Artifacts in magnetic resonance imaging," *Polish J. Radiol.*, vol. 80, pp. 93–106, Feb. 2015.

[5] J. E. Iglesias, E. Konukoglu, D. Zikic, B. Glocker, K. Van Leemput, and B. Fischl, "Is synthesizing MRI contrast useful for inter-modality analysis?" in *Medical Image Computing and Computer-Assisted Intervention*. Cham, Switzerland: Springer, 2013, pp. 631–638.

[6] Y. Huo, Z. Xu, S. Bao, A. Assad, R. G. Abramson, and B. A. Landman, "Adversarial synthesis learning enables segmentation without target modality ground truth," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 1217–1220.

[7] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar, "Advances and challenges in super-resolution," *Int. J. Imag. Syst. Technol.*, vol. 14, no. 2, pp. 47–57, 2004.

[8] D. H. Ye, D. Zikic, B. Glocker, A. Criminisi, and E. Konukoglu, "Modality propagation: Coherent synthesis of subject-specific scans with data-driven regularization," in *Medical Image Computing and Computer-Assisted Intervention*. Cham, Switzerland: Springer, 2013, pp. 606–613.

[9] C. Catana *et al.*, "Toward implementing an MRI-based PET attenuation-correction method for neurologic studies on the MR-PET brain prototype," *J. Nucl. Med.*, vol. 51, no. 9, pp. 1431–1438, Sep. 2010.

[10] J. Lee, A. Carass, A. Jog, C. Zhao, and J. Prince, "Multi-atlas-based CT synthesis from conventional MRI with patch-based refinement for MRI-based radiotherapy planning," *Proc. SPIE*, vol. 10133, Feb. 2017, Art. no. 101331I.

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

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

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

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

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

[16] H. Van Nguyen, K. Zhou, and R. Vemulapalli, "Cross-domain synthesis of medical images using efficient location-sensitive deep network," in *Medical Image Computing and Computer-Assisted Intervention*. Cham, Switzerland: Springer, 2015, pp. 677–684.

[17] R. Vemulapalli, H. V. Nguyen, and S. K. Zhou, "Unsupervised cross-modal synthesis of subject-specific scans," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 630–638.

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

[19] D. C. Alexander, D. Zikic, J. Zhang, H. Zhang, and A. Criminisi, "Image quality transfer via random forest regression: Applications in diffusion MRI," in *Medical Image Computing and Computer-Assisted Intervention*. Cham, Switzerland: Springer, 2014, pp. 225–232.

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

[21] P. Coupe, J. V. Manjón, M. Chamberland, M. Descoteaux, and B. Hiba, "Collaborative patch-based super-resolution for diffusion-weighted images," *Neuroimage*, vol. 83, pp. 245–261, Dec. 2013.

[22] V. Sevetlidis, M. V. Giuffrida, and S. A. Tsaftaris, "Whole image synthesis using a deep encoder-decoder network," in *Simulation and Synthesis in Medical Imaging*. Cham, Switzerland: Springer, 2016, pp. 127–137.

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

[24] S. U. H. Dar, M. Yurt, L. Karacan, A. Erdem, E. Erdem, and T. Çukur, "Image synthesis in multi-contrast MRI with conditional generative adversarial networks," *IEEE Trans. Med. Imag.*, vol. 38, no. 10, pp. 2375–2388, Oct. 2019.

[25] C. Bowles *et al.*, "Pseudo-healthy image synthesis for white matter lesion segmentation," in *Simulation and Synthesis in Medical Imaging*. Cham, Switzerland: Springer, 2016, pp. 87–96.

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

[27] T. Joyce, A. Chartsias, and S. A. Tsaftaris, "Robust multi-modal MR image synthesis," in *Medical Image Computing and Computer Assisted Intervention*. Cham, Switzerland: Springer, 2017, pp. 347–355.

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

[29] I. Goodfellow *et al.*, "Generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 24, 2014, pp. 2672–2680.

[30] A. Beers *et al.*, "High-resolution medical image synthesis using progressively grown generative adversarial networks," 2018, *arXiv:1805.03144*.

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

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

[33] K. Armanious *et al.*, "MedGAN: Medical image translation using GANs," *Computerized Med. Imag. Graph.*, vol. 79, Jan. 2020, Art. no. 101684.

[34] D. Lee, J. Kim, W.-J. Moon, and J. C. Ye, "CollaGAN: Collaborative GAN for missing image data imputation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2487–2496.

[35] H. Li *et al.*, "DiamondGAN: Unified multi-modal generative adversarial networks for MRI sequences synthesis," in *Medical Image Computing and Computer Assisted Intervention*. Cham, Switzerland: Springer, 2019, pp. 795–803.

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

[37] H. Lan, A. Toga, and F. Sepehrband, "SC-GAN: 3D self-attention conditional GAN with spectral normalization for multi-modal neuroimaging synthesis," *bioRxiv*, 2020, Art. no. 2020.06.09.143297.

[38] M. Yurt, S. U. Dar, A. Erdem, E. Erdem, K. K. Oguz, and T. Çukur, "MustGAN: Multi-stream generative adversarial networks for MR image synthesis," *Med. Image Anal.*, vol. 70, May 2021, Art. no. 101944.

[39] H. Yang *et al.*, "Synthesizing multi-contrast MR images via novel 3D conditional variational auto-encoding GAN," *Mobile Netw. Appl.*, vol. 26, pp. 1–10, Oct. 2021.

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

[41] A. Sharma and G. Hamarneh, "Missing MRI pulse sequence synthesis using multi-modal generative adversarial network," *IEEE Trans. Med. Imag.*, vol. 39, no. 4, pp. 1170–1183, Apr. 2020.

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

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

[44] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.

[45] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[46] N. Kodali, J. Abernethy, J. Hays, and Z. Kira, "On convergence and stability of GANs," 2017, *arXiv:1705.07215*.

[47] A. Adam, A. Dixon, J. Gillard, C. Schaefer-Prokop, R. Grainger, and D. Allison, *Grainger & Allison's Diagnostic Radiology*. Amsterdam, The Netherlands: Elsevier, 2014.

[48] D. Ellison, *Neuropathology: A Reference Text of CNS Pathology*. Amsterdam, The Netherlands: Elsevier, 2012.

[49] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

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

[51] J. Chen *et al.*, "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[52] Y. Luo *et al.*, "3D transformer-GAN for high-quality PET reconstruction," in *Medical Image Computing and Computer Assisted Intervention*. Cham, Switzerland: Springer, 2021, pp. 276–285.

[53] Z. Zhang, L. Yu, X. Liang, W. Zhao, and L. Xing, "TransCT: Dual-path transformer for low dose computed tomography," in *Medical Image Computing and Computer Assisted Intervention*. Cham, Switzerland: Springer, 2021, pp. 55–64.

[54] Y. Korkmaz, S. U. Dar, M. Yurt, M. Özbey, and T. Çukur, "Unsupervised MRI reconstruction via zero-shot learned adversarial transformers," *IEEE Trans. Med. Imag.*, early access, Jan. 27, 2022, doi: 10.1109/TMI.2022.3147426.

[55] S. A. Kamran, K. F. Hossain, A. Tavakkoli, S. Lee Zuckerbrod, and S. A. Baker, "VTGAN: Semi-supervised retinal image synthesis and disease prediction using vision transformers," 2021, *arXiv:2104.06757*.

[56] H.-C. Shin *et al.*, "GANBERT: Generative adversarial networks with bidirectional encoder representations from transformers for MRI to PET synthesis," 2020, *arXiv:2008.04393*.

[57] X. Zhang *et al.*, "PTNet: A high-resolution infant MRI synthesizer based on transformer," 2021, *arXiv:2105.13993*.

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

[59] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Med. Image Anal.*, vol. 58, Dec. 2019, Art. no. 101552.

[60] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[61] R. Li *et al.*, "Deep learning based imaging data completion for improved brain disease diagnosis," in *Medical Image Computing and Computer-Assisted Intervention*. Cham, Switzerland: Springer, 2014, pp. 305–312.

[62] A. Torrado-Carvajal *et al.*, "Fast patch-based pseudo-CT synthesis from T1-weighted MR images for PET/MR attenuation correction in brain studies," *J. Nucl. Med.*, vol. 57, no. 1, pp. 136–143, Jan. 2016.

[63] K. Bahrami, F. Shi, I. Rekik, and D. Shen, "Convolutional neural network for reconstruction of 7T-like images from 3T MRI using appearance and anatomical features," in *Deep Learning and Data Labeling for Medical Applications*. Cham, Switzerland: Springer, 2016, pp. 39–47.

[64] K. Bahrami, F. Shi, X. Zong, H. W. Shin, H. An, and D. Shen, "Reconstruction of 7T-like images from 3T MRI," *IEEE Trans. Med. Imag.*, vol. 35, no. 9, pp. 2085–2097, Sep. 2016.

[65] Y. Zhang, P.-T. Yap, L. Qu, J.-Z. Cheng, and D. Shen, "Dual-domain convolutional neural networks for improving structural information in 3T MRI," *Magn. Reson. Imag.*, vol. 64, pp. 90–100, Dec. 2019.

[66] X. Han, "MR-based synthetic CT generation using a deep convolutional neural network method," *Med. Phys.*, vol. 44, no. 4, pp. 1408–1419, 2017.

[67] D. Nie, X. Cao, Y. Gao, L. Wang, and D. Shen, "Estimating CT image from MRI data using 3D fully convolutional networks," in *Deep Learning and Data Labeling for Medical Applications*. Cham, Switzerland: Springer, 2016, pp. 170–178.

[68] H. Arabi, G. Zeng, G. Zheng, and H. Zaidi, "Novel deep learning-based CT synthesis algorithm for MRI-guided PET attenuation correction in brain PET/MR imaging," in *Proc. IEEE Nucl. Sci. Symp. Med. Imag. Conf. Proc. (NSS/MIC)*, Nov. 2018, pp. 1–3.

[69] K. Klaser *et al.*, "Improved MR to CT synthesis for PET/MR attenuation correction using imitation learning," in *Simulation and Synthesis in Medical Imaging*. Cham, Switzerland: Springer, 2019, pp. 13–21.

[70] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[71] V. Sandfort, K. Yan, P. Pickhardt, and R. Summers, "Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks," *Sci. Rep.*, vol. 9, Nov. 2019, Art. no. 16884.

[72] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, Dec. 2018.

[73] A. Ben-Cohen *et al.*, "Cross-modality synthesis from CT to PET using FCN and GAN networks for improved automated lesion detection," *Eng. Appl. Artif. Intell.*, vol. 78, pp. 186–194, Feb. 2019.

[74] G. Santini *et al.*, "Unpaired PET/CT image synthesis of liver region using CycleGAN," in *Proc. 16th Int. Symp. Med. Inf. Process. Anal.*, Nov. 2020, pp. 247–257.

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

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

[77] J. Wolterink, A. M. Dinkla, M. Savenije, P. Seevinck, C. Berg, and I. Isgum, "Deep MR to CT synthesis using unpaired data," in *Simulation and Synthesis in Medical Imaging*. Cham, Switzerland: Springer, 2017, pp. 14–23.

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

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

[80] Y. Hiasa *et al.*, "Cross-modality image synthesis from unpaired data using CycleGAN: Effects of gradient consistency loss and training data size," in *Simulation and Synthesis in Medical Imaging*. Cham, Switzerland: Springer, 2018, pp. 31–41.

[81] A. Chartsias, T. Joyce, R. Dharmakumar, and S. A. Tsaftaris, "Adversarial image synthesis for unpaired multi-modal cardiac data," in *Simulation and Synthesis in Medical Imaging*. Cham, Switzerland: Springer, 2017, pp. 3–13.

[82] H. Do, P. Bourdon, D. Helbert, M. Naudin, and R. Guillevin, "7T MRI super-resolution with generative adversarial network," in *Proc. Int. Symp. Electron. Imag.*, 2021, pp. 106-1–106-7.

[83] L. Xiang, Y. Li, W. Lin, Q. Wang, and D. Shen, "Unpaired deep cross-modality synthesis with fast training," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2018, pp. 155–164.

[84] V. Kearney *et al.*, "Attention-aware discrimination for MR-to-CT image translation using cycle-consistent generative adversarial networks," *Radiol., Artif. Intell.*, vol. 2, no. 2, Mar. 2020, Art. no. e190027.

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

[86] J. Zhao *et al.*, "Tripartite-GAN: Synthesizing liver contrast-enhanced MRI to improve tumor detection," *Med. Image Anal.*, vol. 63, Jul. 2020, Art. no. 101667.

[87] Z. Yuan *et al.*, "SARA-GAN: Self-attention and relative average discriminator based generative adversarial networks for fast compressed sensing MRI reconstruction," *Frontiers Neuroinform.*, vol. 14, p. 58, Nov. 2020.

[88] M. Li, W. Hsu, X. Xie, J. Cong, and W. Gao, "SACNN: Self-attention convolutional neural network for low-dose CT denoising with self-supervised perceptual loss network," *IEEE Trans. Med. Imag.*, vol. 39, no. 7, pp. 2289–2301, Jul. 2020.

[89] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "CoTr: Efficiently bridging CNN and transformer for 3D medical image segmentation," 2021, *arXiv:2103.03024*.

[90] Y. Dai and Y. Gao, "TransMed: Transformers advance multi-modal medical image classification," 2021, *arXiv:2103.05940*.

[91] D. Karimi, S. Vasylechko, and A. Gholipour, "Convolution-free medical image segmentation using transformers," 2021, *arXiv:2102.13645*.

[92] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–11.

[93] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Upper Saddle River, NJ, USA: Prentice-Hall, 1994.

[94] B. H. Menze *et al.*, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans. Med. Imag.*, vol. 34, no. 10, pp. 1993–2024, Oct. 2014.

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

[96] M. Jenkinson and S. Smith, "A global optimisation method for robust affine registration of brain images," *Med. Image Anal.*, vol. 5, no. 2, pp. 143–156, Jun. 2001.

[97] H. Lan, A. W. Toga, and F. Sepehrband, "Three-dimensional self-attention conditional GAN with spectral normalization for multi-modal neuroimaging synthesis," *Magn. Reson. Med.*, vol. 86, no. 3, pp. 1718–1733, Sep. 2021.

[98] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–13.

[99] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor, "ImageNet-21K pretraining for the masses," 2021, *arXiv:2104.10972*.

[100] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[101] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6629–6640.

[102] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 4190–4197.

[103] R. Durall, M. Keuper, and J. Keuper, "Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7887–7896.

[104] B. Zhan *et al.*, "LR-cGAN: Latent representation based conditional generative adversarial network for multi-modality MRI synthesis," *Biomed. Signal Process. Control*, vol. 66, Apr. 2021, Art. no. 102457.

[105] D. Nie and D. Shen, "Adversarial confidence learning for medical image segmentation and synthesis," *Int. J. Comput. Vis.*, vol. 128, nos. 10–11, pp. 2494–2513, Nov. 2020.

[106] Y. Luo *et al.*, "Edge-preserving MRI image synthesis via adversarial network with iterative multi-scale fusion," *Neurocomputing*, vol. 452, pp. 63–77, Sep. 2021.

[107] L.-H. Chen, C. G. Bampis, Z. Li, C. Chen, and A. C. Bovik, "Convolutional block design for learned fractional downsampling," 2021, *arXiv:2105.09999*.

[108] M. Yurt, S. U. H. Dar, M. Özbey, B. Tınaz, K. K. Oğuz, and T. Çukur, "Semi-supervised learning of mutually accelerated MRI synthesis without fully-sampled ground truths," 2020, *arXiv:2011.14347*.